Journal of Machine Learning Research 1 (2015) xx-xx Submitted xx/15; Published xx/15

Importance weighting without importance weights: An efficient algorithm for combinatorial semi-bandits

Gergely Neu GERGELY.NEU@GMAIL.COM
SequeL team, INRIA Lille – Nord Europe

Gábor Bartók BARTOK@GOOGLE.COM
Department of Computer Science, ETH Zürich

Editor:

Abstract

We propose a sample-efficient alternative for importance weighting for situations where one only has sample access to the probability distribution that generates the observations. Our new method, called Recurrence Weighting (`RW`), is described and analyzed in the context of online combinatorial optimization under semi-bandit feedback, where a learner sequentially selects its actions from a combinatorial decision set so as to minimize its cumulative loss. In particular, we show that the well-known Follow-the-Perturbed-Leader (`FPL`) prediction method coupled with Recurrence Weighting yields the first computationally efficient reduction from offline to online optimization in this setting. We provide a thorough theoretical analysis for the resulting algorithm, showing that its performance is on par with previous, inefficient solutions. Our main contribution is showing that, despite the relatively large variance induced by the `RW` procedure, our performance guarantees hold with high probability rather than only in expectation. As a side result, we also improve the best known regret bounds for `FPL` in online combinatorial optimization with full feedback, closing the perceived performance gap between `FPL` and exponential weights in this setting.

Keywords: Online learning, Combinatorial optimization, Bandit problems, Semi-bandit feedback, Follow the Perturbed Leader

1. Introduction

Importance weighting is a crucially important tool used in many areas of machine learning, and specifically online learning with partial feedback. While most work assumes that importance weights are readily available or can be computed with little effort during runtime, this is often not the case in many practical settings, even when one has cheap sample access to the distribution generating the observations. Among other cases, such situations may arise when observations are generated by complex hierarchical sampling schemes, probabilistic programs, or, more generally, black-box generative models. In this paper, we propose a simple and efficient sampling scheme called *Recurrence Weighting* (`RW`) to compute reliable estimates of importance weights *using only sample access.*

Our main motivation is studying a specific online learning algorithm whose practical applicability in partial-feedback settings had long been hindered by the problem outlined above. Specifically, we consider the well-known *Follow-the-Perturbed-Leader* (`FPL`) prediction method that maintains implicit sampling distributions that usually cannot be expressed in closed form. In this paper, we endow `FPL` with our Recurrence Weighting scheme to construct the first known computationally efficient reduction from offline to online combinatorial optimization under an important partial-information scheme known as *semi-bandit feedback.* In the rest of this section, we describe our precise setting, present related work and outline our main results.

A preliminary version of this paper was published as Neu and Bartók (2013). Gábor Bartók is now with Google Zürich.

Parameters: set of decision vectors $\mathcal{S} \subseteq \{0,1\}^d$, number of rounds T;
For all $t = 1, 2, \dots, T$**, repeat**

1. The learner chooses a probability distribution $\boldsymbol{p}_t$ over $\mathcal{S}$.
2. The learner draws action $\boldsymbol{V}_t$ randomly according to $\boldsymbol{p}_t$.
3. The environment chooses loss vector $\boldsymbol{\ell}_t$.
4. The learner suffers loss $\boldsymbol{V}_t^\top \boldsymbol{\ell}_t$.
5. The learner observes some feedback based on $\boldsymbol{\ell}_t$ and $\boldsymbol{V}_t$.

Figure 1: The protocol of online combinatorial optimization.

1.1 Online combinatorial optimization

We consider a special case of online linear optimization known as online combinatorial optimization (see Figure 1). In every round $t = 1, 2, \dots, T$ of this sequential decision problem, the learner chooses an *action* $\boldsymbol{V}_t$ from the finite action set $\mathcal{S} \subseteq \{0,1\}^d$, where $\|\boldsymbol{v}\|_1 \le m$ holds for all $\boldsymbol{v} \in \mathcal{S}$. At the same time, the environment fixes a loss vector $\boldsymbol{\ell}_t \in [0,1]^d$ and the learner suffers loss $\boldsymbol{V}_t^\top \boldsymbol{\ell}_t$. The goal of the learner is to minimize the cumulative loss $\sum_{t=1}^T \boldsymbol{V}_t^\top \boldsymbol{\ell}_t$. As usual in the literature of online optimization (Cesa-Bianchi and Lugosi, 2006), we measure the performance of the learner in terms of the *regret* defined as

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

that is, the gap between the total loss of the learning algorithm and the best fixed decision in hindsight. In the current paper, we focus on the case of *non-oblivious* (or *adaptive*) environments, where we allow the loss vector $\boldsymbol{\ell}_t$ to depend on the previous decisions $\boldsymbol{V}_1, \dots, \boldsymbol{V}_{t-1}$ in an arbitrary fashion. Since it is well-known that no deterministic algorithm can achieve sublinear regret under such weak assumptions, we will consider learning algorithms that choose their decisions in a randomized way. For such learners, another performance measure that we will study is the *expected regret* defined as

$$\widehat{R}_T = \max_{\boldsymbol{v} \in \mathcal{S}} \sum_{t=1}^T \mathbb{E}\left[(\boldsymbol{V}_t - \boldsymbol{v})^\top \boldsymbol{\ell}_t\right] = \mathbb{E}\left[\sum_{t=1}^T \boldsymbol{V}_t^\top \boldsymbol{\ell}_t\right] - \min_{\boldsymbol{v} \in \mathcal{S}} \mathbb{E}\left[\sum_{t=1}^T \boldsymbol{v}^\top \boldsymbol{\ell}_t\right].$$

The framework described above is general enough to accommodate a number of interesting problem instances such as path planning, ranking and matching problems, finding minimum-weight spanning trees and cut sets. Accordingly, different versions of this general learning problem have drawn considerable attention in the past few years. These versions differ in the amount of information made available to the learner after each round t. In the simplest setting, called the *full-information* setting, it is assumed that the learner gets to observe the loss vector $\boldsymbol{\ell}_t$ regardless of the choice of $\boldsymbol{V}_t$. As this assumption does not hold for many practical applications, it is more interesting to study the problem under *partial-information* constraints, meaning that the learner only gets some limited feedback based on its own decision. In the current paper, we focus on a more realistic partial-information scheme known as *semi-bandit feedback* (Audibert, Bubeck, and Lugosi, 2014) where the learner only observes the components $\ell_{t,i}$ of the loss vector for which $V_{t,i} = 1$, that is, the losses associated with the components selected by the learner[1].

1 Here, $V_{t,i}$ and $\ell_{t,i}$ are the i^{th} components of the vectors $\boldsymbol{V}_t$ and $\boldsymbol{\ell}_t$, respectively.

1.2 Related work

The most well-known instance of our problem is the *multi-armed bandit* problem considered in the seminal paper of Auer, Cesa-Bianchi, Freund, and Schapire (2002): in each round of this problem, the learner has to select one of N *arms* and minimize regret against the best fixed arm while only observing the losses of the chosen arms. In our framework, this setting corresponds to setting $d = N$ and $m = 1$. Among other contributions concerning this problem, Auer et al. propose an algorithm called `Exp3` (Exploration and Exploitation using Exponential weights) based on constructing loss estimates $\widehat{\ell}_{t,i}$ for each component of the loss vector and playing arm i with probability proportional to $\exp(-\eta \sum_{s=1}^{t-1} \widehat{\ell}_{s,i})$ at time t, where $\eta > 0$ is a parameter of the algorithm, usually called the learning rate[2]. This algorithm is essentially a variant of the Exponentially Weighted Average (`EWA`) forecaster (a variant of weighted majority algorithm of Littlestone and Warmuth, 1994, and aggregating strategies of Vovk, 1990, also known as `Hedge` by Freund and Schapire, 1997). Besides proving that the *expected* regret of `Exp3` is $O\big(\sqrt{NT \log N}\big)$, Auer et al. also provide a general lower bound of $\Omega\big(\sqrt{NT}\big)$ on the regret of any learning algorithm on this particular problem. This lower bound was later matched by a variant of the Implicitly Normalized Forecaster (`INF`) of Audibert and Bubeck (2010) by using the same loss estimates in a more refined way. Audibert and Bubeck also show bounds of $O\big(\sqrt{NT/\log N} \log(N/\delta)\big)$ on the regret that hold with probability at least $1 - \delta$, uniformly for any $\delta > 0$.

The most popular example of online learning problems with actual combinatorial structure is the shortest path problem first considered by Takimoto and Warmuth (2003) in the full information scheme. The same problem was considered by György, Linder, Lugosi, and Ottucsák (2007), who proposed an algorithm that works with semi-bandit information. Since then, we have come a long way in understanding the "price of information" in online combinatorial optimization—see Audibert, Bubeck, and Lugosi (2014) for a complete overview of results concerning all of the information schemes considered in the current paper. The first algorithm directly targeting general online combinatorial optimization problems is due to Koolen, Warmuth, and Kivinen (2010): their method named `Component Hedge` guarantees an optimal regret of $O\big(m\sqrt{T \log(d/m)}\big)$ in the full information setting. In particular, this algorithm is an instance of a more general algorithm class known as Online Stochastic Mirror Descent (`OSMD`). Audibert, Bubeck, and Lugosi (2014) show that `OSMD`-based methods can also be used for proving expected regret bounds of $O\big(\sqrt{mdT}\big)$ for the semi-bandit setting, which is also shown to coincide with the minimax regret in this setting. For completeness, we note that the `EWA` forecaster is known to attain an expected regret of $O\big(m^{3/2}\sqrt{T \log(d/m)}\big)$ in the full information case and $O\big(m\sqrt{dT \log(d/m)}\big)$ in the semi-bandit case.

While the results outlined above might suggest that there is absolutely no work left to be done in the full information and semi-bandit schemes, we get a different picture if we restrict our attention to *computationally efficient* algorithms. First, note that methods based on exponential weighting of each decision vector can only be efficiently implemented for a handful of decision sets $\mathcal{S}$—see Koolen et al. (2010) and Cesa-Bianchi and Lugosi (2012) for some examples. Furthermore, as noted by Audibert et al. (2014), `OSMD`-type methods can be efficiently implemented by convex programming if the convex hull of the decision set can be described by a polynomial number of constraints. Details of such an efficient implementation are worked out by Suehiro, Hatano, Kijima, Takimoto, and Nagano (2012), whose algorithm runs in $O(d^6)$ time, which can still be prohibitive in practical applications. While Koolen et al. (2010) list some further examples where `OSMD` can be implemented efficiently, we conclude that there is no general efficient algorithm with near-optimal performance guarantees for learning in combinatorial semi-bandits.

The Follow-the-Perturbed-Leader (`FPL`) prediction method (first proposed by Hannan, 1957 and later rediscovered by Kalai and Vempala, 2005) offers a computationally efficient solution for the

2 In fact, Auer et al. mix the resulting distribution with a uniform distribution over the arms with probability ηN. However, this modification is not needed when one is concerned with the total expected regret, see, e.g., Bubeck and Cesa-Bianchi (2012, Section 3.1).

online combinatorial optimization problem given that the *static* combinatorial optimization problem $\min_{\boldsymbol{v}\in\mathcal{S}} \boldsymbol{v}^\mathsf{T}\boldsymbol{\ell}$ admits computationally efficient solutions for any $\boldsymbol{\ell} \in \mathbb{R}^d$. The idea underlying `FPL` is very simple: in every round t, the learner draws some random perturbations $\boldsymbol{Z}_t \in \mathbb{R}^d$ and selects the action that minimizes the perturbed total losses:

$$\boldsymbol{V}_t = \operatorname*{arg\,min}_{\boldsymbol{v}\in\mathcal{S}} \boldsymbol{v}^\mathsf{T} \left(\sum_{s=1}^{t-1} \boldsymbol{\ell}_s - \boldsymbol{Z}_t \right).$$

Despite its conceptual simplicity and computational efficiency, `FPL` have been relatively overlooked until very recently, due to two main reasons:

- The best known bound for `FPL` in the full information setting is $O\big(m\sqrt{dT}\big)$, which is worse than the bounds for both `EWA` and `OSMD` that scale only logarithmically with d.
- Considering bandit information, no efficient `FPL`-style algorithm is known to achieve a regret of $O\big(\sqrt{T}\big)$. On one hand, it is relatively straightforward to prove $O\big(T^{2/3}\big)$ bounds on the expected regret for an efficient `FPL`-variant (see, e.g., Awerbuch and Kleinberg, 2004 and McMahan and Blum, 2004). Poland (2005) proved bounds of $O\big(\sqrt{NT\log N}\big)$ in the N-armed bandit setting, however, the proposed algorithm requires $O\big(T^2\big)$ numerical operations per round.

The main obstacle for constructing a computationally efficient `FPL`-variant that works with partial information is precisely the lack of closed-form expressions for importance weights. In the current paper, we address the above two issues and show that an efficient `FPL`-based algorithm can achieve as good performance guarantees as `EWA` in online combinatorial optimization.

Our work contributes to a new wave of positive results concerning `FPL`. Besides the reservations towards `FPL` mentioned above, the reputation of `FPL` has been also suffering from the fact that the nature of regularization arising from perturbations is not as well-understood as the explicit regularization schemes underlying `OSMD` or `EWA`. Very recently, Abernethy et al. (2014) have shown that `FPL` implements a form of strongly convex regularization over the convex hull of the decision space. Furthermore, Rakhlin et al. (2012) showed that `FPL` run with a specific perturbation scheme can be regarded as a relaxation of the minimax algorithm. Another recently initiated line of work shows that intuitive *parameter-free* variants of `FPL` can achieve excellent performance in full-information settings (Devroye et al., 2013 and Van Erven et al., 2014).

1.3 Our results

In this paper, we propose a loss-estimation scheme called Recurrence Weighting to efficiently compute importance weights for the observed components of the loss vector. Building on this technique and the `FPL` principle, resulting in *an efficient algorithm for regret minimization under semi-bandit feedback*. Besides this contribution, our techniques also enable us to improve the best known regret bounds of `FPL` in the full information case. We prove the following results concerning variants of our algorithm:

- a bound of $O\big(m\sqrt{dT\log(d/m)}\big)$ on the expected regret under semi-bandit feedback (Theorem 1),
- a bound of $O\big(m\sqrt{dT\log(d/m)} + \sqrt{mdT}\log(1/\delta)\big)$ on the regret that holds with probability at least $1-\delta$, uniformly for all $\delta \in (0,1)$ under semi-bandit feedback (Theorem 2),
- a bound of $O\big(m^{3/2}\sqrt{T\log(d/m)}\big)$ on the expected regret under full information (Theorem 13).

We also show that both of our semi-bandit algorithms access the optimization oracle $O(dT)$ times over T rounds with high probability, increasing the running time only by a factor of d compared to the full-information variant. Notably, our results close the gaps between the performance bounds of `FPL` and `EWA` under both full information and semi-bandit feedback.

2. Recurrence Weighting

In this section, we introduce the main idea underlying Recurrence Weighting[3] in the specific context of N-armed bandits where $d = N$, $m = 1$ and the learner has access to the basis vectors $\{\boldsymbol{e}_i\}_{i=1}^d$ as its decision set $\mathcal{S}$. In this setting, components of the decision vector are referred to as *arms*. For ease of notation, define I_t as the unique arm such that $V_{t,I_t} = 1$ and $\mathcal{F}_{t-1}$ as the sigma-algebra induced by the learner's actions and observations up to the end of round $t-1$. Using this notation, we define $p_{t,i} = \mathbb{P}\left[I_t = i \middle| \mathcal{F}_{t-1}\right]$.

Most bandit algorithms rely on feeding some loss estimates to a sequential prediction algorithm. It is commonplace to consider *importance-weighted* loss estimates of the form

$$\widehat{\ell}^*_{t,i} = \frac{\mathbb{1}_{\{I_t = i\}}}{p_{t,i}} \ell_{t,i} \tag{2}$$

for all t, i such that $p_{t,i} > 0$. It is straightforward to show that $\widehat{\ell}^*_{t,i}$ is an unbiased estimate of the loss $\ell_{t,i}$ for all such t, i. Otherwise, when $p_{t,i} = 0$, we set $\widehat{\ell}^*_{t,i} = 0$, which gives $\mathbb{E}\left[\widehat{\ell}^*_{t,i}\middle| \mathcal{F}_{t-1}\right] = 0 \leq \ell_{t,i}$.

To our knowledge, all existing bandit algorithms operating in the non-stochastic setting utilize some version of the importance-weighted loss estimates described above. This is a very natural choice for algorithms that operate by first computing the probabilities $p_{t,i}$ and then sampling I_t from the resulting distributions. While many algorithms fall into this class (including the `Exp3` algorithm of Auer et al. (2002), the `Green` algorithm of Allenberg et al. (2006) and the `INF` algorithm of Audibert and Bubeck (2010), one can think of many other algorithms where the distribution $\boldsymbol{p}_t$ is specified implicitly and thus importance weights are not readily available. Arguably, `FPL` is the most important online prediction algorithm that operates with implicit distributions that are notoriously difficult to compute in closed form. To overcome this difficulty, we propose a different loss estimate that can be efficiently computed *even when* $\boldsymbol{p}_t$ *is not available for the learner.*

Our estimation procedure dubbed Recurrence Weighting (`RW`) is based on the simple observation that, even though p_{t,I_t} might not be computable in closed form, one can simply generate a geometric random variable with expectation $1/p_{t,I_t}$ by repeated sampling from $\boldsymbol{p}_t$. Specifically, we propose the following procedure to be executed in round t:

Recurrence Weighting for multi-armed bandits

1. The learner draws $I_t \sim \boldsymbol{p}_t$.
2. For $k = 1, 2, \ldots$
   (a) Draw $I'_t(k) \sim \boldsymbol{p}_t$.
   (b) If $I'_t(k) = I_t$, break.
3. Let $K_t = k$.

Observe that K_t generated this way is a geometrically distributed random variable given I_t and $\mathcal{F}_{t-1}$. Consequently, we have $\mathbb{E}\left[K_t \left| \mathcal{F}_{t-1}, I_t\right.\right] = 1/p_{t,I_t}$. We use this property to construct the estimates

$$\widehat{\ell}_{t,i} = K_t \mathbb{1}_{\{I_t = i\}} \ell_{t,i} \tag{3}$$

3 In the preliminary version of this paper(Neu and Bartók, 2013), the same method was called "Geometric Resampling". We chose to change the name to avoid association with the concept of resampling broadly used in statistics.

for all arms i. We can easily show that the above estimate is unbiased whenever $p_{t,i} > 0$:

$$\begin{aligned}\mathbb{E}\left[\widehat{\ell}_{t,i}\middle|\mathcal{F}_{t-1}\right] &= \sum_j p_{t,j}\mathbb{E}\left[\widehat{\ell}_{t,i}\middle|\mathcal{F}_{t-1}, I_t = j\right] \\ &= p_{t,i}\mathbb{E}\left[\ell_{t,i}K_t\left|\mathcal{F}_{t-1}, I_t = i\right.\right] \\ &= p_{t,i}\ell_{t,i}\mathbb{E}\left[K_t\left|\mathcal{F}_{t-1}, I_t = i\right.\right] \\ &= \ell_{t,i}.\end{aligned}$$

Notice that the above procedure produces $\widehat{\ell}_{t,i} = 0$ almost surely whenever $p_{t,i} = 0$, giving $\mathbb{E}\left[\widehat{\ell}_{t,i}\middle|\mathcal{F}_{t-1}\right] = 0$ for such t, i.

One practical concern with the above sampling procedure is that its worst-case running time is unbounded: while the expected number of necessary samples K_t is clearly N, the actual number of samples might be much larger. In the next section, we offer a remedy to this problem, as well as generalize the approach to work in the combinatorial semi-bandit case.

3. An efficient algorithm for combinatorial semi-bandits

In this section, we present our main result: an efficient reduction from offline to online combinatorial optimization under semi-bandit feedback. The most critical element in our technique is extending the Recurrence Weighting idea to the case of combinatorial action sets. For defining the procedure, let us assume that we are running a randomized algorithm mapping histories to probability distributions over the action set $\mathcal{S}$: letting $\mathcal{F}_{t-1}$ denote the sigma-algebra induced by the history of interaction between the learner and the environment, the algorithm picks action $\boldsymbol{v} \in \mathcal{S}$ with probability $p_t(\boldsymbol{v}) = \mathbb{P}\left[\boldsymbol{V}_t = \boldsymbol{v}\left|\mathcal{F}_{t-1}\right.\right]$. Also introducing $q_{t,i} = \mathbb{E}\left[V_{t,i}\left|\mathcal{F}_{t-1}\right.\right]$, we can define the counterpart of the standard importance-weighted loss estimates of Equation 2 as the vector $\widehat{\boldsymbol{\ell}}^*_t$ with components

$$\widehat{\ell}^*_{t,i} = \frac{V_{t,i}}{q_{t,i}}\ell_{t,i}. \tag{4}$$

Again, the problem with these estimates is that for many algorithms of practical interest, the importance weights $q_{t,i}$ cannot be computed in closed form. We now extend the Recurrence Weighting procedure defined in the previous section to estimate the importance weights in an efficient manner as follows:

Recurrence Weighting for combinatorial semi-bandits

1. The learner draws $\boldsymbol{V}_t \sim \boldsymbol{p}_t$.
2. For $k = 1, 2, \dots, M$, draw $\boldsymbol{V}'_t(k) \sim \boldsymbol{p}_t$.
3. For $i = 1, 2, \dots, d$,

$$K_{t,i} = \min\bigl(\bigl\{k : V'_{t,i}(k) = 1\bigr\} \cup \{M\}\bigr).$$

Based on these random variables, we construct our loss-estimate vector $\widehat{\boldsymbol{\ell}}_t \in \mathbb{R}^d$ with components

$$\widehat{\ell}_{t,i} = K_{t,i}V_{t,i}\ell_{t,i} \tag{5}$$

for all $i = 1, 2, \dots, d$. Since $V_{t,i}$ are nonzero only for coordinates for which $\ell_{t,i}$ is observed, these estimates are well-defined. It also follows that the sampling procedure can be terminated once for every i with $V_{t,i} = 1$, there is a copy $\boldsymbol{V}'_t(k)$ such that $V'_{t,i}(k) = 1$.

Now everything is ready to define our algorithm: `FPL+RW`, standing for Follow-the-Perturbed-Leader with Recurrence Weighted loss estimates. Defining $\widehat{\boldsymbol{L}}_t = \sum_{s=1}^t \widehat{\boldsymbol{\ell}}_s$, at time step t `FPL+RW`

```
function c = fpl_rw(eta,M,l)
    [d T] = size(l);
    c = zeros(1,T)./0;
    L_est = zeros(d,1);
    for t=1:T
        loss = l(t,:);
        % compute perturbed leader:
        Z = - log(rand(d,1));
        V = opt(eta*L_est - Z);
        % suffer loss:
        c(t) = V' * loss;
        % compute recurrence weights:
        K = rw(eta*L_est, V, M);
        % update loss estimates:
        L_est = L_est + K .* V .* loss;
    end
end
```

```
function K = rw(L,V,M)
    d   = size(L,1);
    K =  zeros(d,1);
    % initialize waiting list:
    waiting = V;
    for k=1:M
        % increment counter:
        K = K + waiting;
        % compute perturbed leader:
        Z = -log(rand(d,1));
        W = opt(L - Z);
        % update waiting list:
        waiting = waiting .* (1 - V.*W);
        if(~waiting)
            return
        end
    end
end
```

Figure 2: MATLAB code for `FPL+RW`, with `RW` shown as a subroutine. Both functions assume access to a linear optimization oracle `opt`. The code for `RW` uses a waiting list to compute the recurrence weights: `waiting(i)==1` if and only if `V(i)==1` and index `i` hasn't recurred yet.

draws the components of the perturbation vector $\boldsymbol{Z}_t$ independently from a standard exponential distribution and selects action[4]

$$\boldsymbol{V}_t = \operatorname*{arg\,min}_{\boldsymbol{v}\in\mathcal{S}} \boldsymbol{v}^\mathsf{T} \left(\eta\widehat{\boldsymbol{L}}_{t-1} - \boldsymbol{Z}_t\right), \tag{6}$$

where $\eta > 0$ is a parameter of the algorithm. As we mentioned earlier, the distribution $\boldsymbol{p}_t$, while implicitly specified by $\boldsymbol{Z}_t$ and the estimated cumulative losses $\widehat{\boldsymbol{L}}_{t-1}$, cannot usually be expressed in closed form for `FPL`[5]. However, sampling the actions $\boldsymbol{V}_t'(\cdot)$ can be carried out by drawing additional perturbation vectors $\boldsymbol{Z}_t'(\cdot)$ independently from the same distribution as $\boldsymbol{Z}_t$ and then solving a linear optimization task. We emphasize that the above additional actions are *never actually played by the algorithm*, but are only necessary for constructing the loss estimates. The power of `FPL+RW` is that, unlike other algorithms for combinatorial semi-bandits, its implementation only requires access to a linear optimization oracle over $\mathcal{S}$. We point the reader to Section 3.2 for a more detailed discussion of the running time of `FPL+RW`. A simple MATLAB code implementing `FPL+RW` is shown on Figure 2.

As we will show shortly, `FPL+RW` as defined above comes with strong performance guarantees that hold *in expectation*. One can think of several possible ways to robustify `FPL+RW` so that it provides bounds that hold with high probability. One possible way is to follow Auer et al. (2002) and define the loss-estimate vector $\widetilde{\boldsymbol{\ell}}_t^*$ with components

$$\widetilde{\ell}_{t,i}^* = \widehat{\ell}_{t,i} - \frac{\beta}{q_{t,i}}$$

for some $\beta > 0$. The obvious problem with this definition is that it requires perfect knowledge of the importance weights $q_{t,i}$ for all i—even for the ones for which $V_{t.i} = 0$. While it is possible to

4 By the definition of the perturbation distribution, the minimum exists almost surely.

5 One notable exception is when the perturbations are drawn independently from standard Gumbel distributions, and the decision set is the d-dimensional simplex: in this case, `FPL` is known to be equivalent with `EWA`—see, e.g., Abernethy et al. (2014) for further discussion.

extend the recurrence-weighting approach developed in the previous sections to construct a reliable proxy to the above loss estimate, there are several downsides to this approach. First, thanks to the fact that one has to obtain estimates of $1/q_{t,i}$ for all i's, one cannot hope to terminate the sampling procedure in reasonable time. Second, reliable estimation requires multiple samples of $K_{t,i}$, where the sample size has to explicitly depend on the desired confidence level.

Thus, we follow a different path: Motivated by the work of Audibert and Bubeck (2010), we propose to use a loss-estimate vector $\widetilde{\boldsymbol{\ell}}_t$ of the form

$$\widetilde{\ell}_{t,i} = \frac{1}{\beta} \log\left(1 + \beta \widehat{\ell}_{t,i}\right) \tag{7}$$

with an appropriately chosen $\beta > 0$. Then, defining $\widetilde{\boldsymbol{L}}_{t-1} = \sum_{s=1}^{t-1} \widetilde{\boldsymbol{\ell}}_s$, we propose a variant of `FPL+RW` that simply replaces $\widehat{\boldsymbol{L}}_{t-1}$ by $\widetilde{\boldsymbol{L}}_{t-1}$ in the rule (6) for choosing $\boldsymbol{V}_t$. We refer to this variant of `FPL+RW` as `FPL+RW.P`. In the next section, we provide performance guarantees for both algorithms.

3.1 Performance guarantees

Now we are ready to state our main results. Proofs will be presented in Section 4. First, we present a performance guarantee for `FPL+RW` in terms of the *expected regret*:

Theorem 1 *The expected regret of `FPL+RW` satisfies*

$$\widehat{R}_T \leq \frac{m\left(\log\left(d/m\right) + 1\right)}{\eta} + 2\eta m d T + \frac{dT}{eM}$$

under semi-bandit information. In particular, with

$$\eta = \sqrt{\frac{\log(d/m) + 1}{dT}} \qquad \text{and} \qquad M = \left\lceil \frac{\sqrt{dT}}{em\sqrt{2\left(\log(d/m) + 1\right)}} \right\rceil,$$

the expected regret of `FPL+RW` is bounded as

$$\widehat{R}_T \leq 3m\sqrt{2dT\left(\log\frac{d}{m} + 1\right)}.$$

Our second main contribution is the following bound on the regret of `FPL+RW.P`.

Theorem 2 *Fix an arbitrary $\delta > 0$. With probability at least $1-\delta$, the regret of `FPL+RW.P` satisfies*

$$\begin{aligned} R_T \leq & \frac{m\left(\log(d/m) + 1\right)}{\eta} + \eta\left(\left(mM\right)^2 \log\frac{5}{\delta} + 2(e-1)mdT\right) + \frac{dT}{eM} \\ & + \beta\left(mM^2 \log\frac{5}{\delta} + 2(e-1)dT\right) + \frac{m\log(5d/\delta)}{\beta} \\ & + m\sqrt{2(e-2)T}\log\frac{5}{\delta} + \sqrt{T\log\frac{5}{\delta}} + \sqrt{2(e-2)T}. \end{aligned}$$

In particular, with

$$M = \left\lceil \sqrt{\frac{dT}{m}} \right\rceil, \quad \beta = \sqrt{\frac{m}{dT}}, \quad \text{and} \quad \eta = \sqrt{\frac{\log(d/m) + 1}{dT}},$$

the regret of `FPL+RW.P` is bounded as

$$
\begin{aligned}
R_T \le{} & \left(\log\frac{5}{\delta} + 4.44\right) m\sqrt{dT\left(\log\frac{d}{m}+1\right)} + \left(\log\frac{5}{\delta} + \log\frac{5d}{\delta} + 3.81\right)\sqrt{mdT} \\
& + 1.44\log\frac{5}{\delta} m\sqrt{T} + \left(\sqrt{\log\frac{5}{\delta}} + 1.44\right)\sqrt{T}
\end{aligned}
$$

with probability at least $1-\delta$.

3.2 Running time

Let us now turn our attention to computational issues. First, we note that the efficiency of `FPL`-type algorithms crucially depends on the availability of an efficient oracle that solves the static combinatorial optimization problem of finding $\arg\min_{\boldsymbol{v}\in\mathcal{S}} \boldsymbol{v}^\top\boldsymbol{\ell}$. Computing the running time of the full-information variant of `FPL` is straightforward: assuming that the oracle computes the solution to the static problem in $O(f(\mathcal{S}))$ time, `FPL` returns its prediction in $O(f(\mathcal{S})+d)$ time (with the d overhead coming from the time necessary to generate the perturbations). Naturally, our loss estimation scheme multiplies these computations by the number of samples taken in each round. While terminating the estimation procedure after M samples helps in controlling the running time with high probability, observe that the naïve bound of MT on the number of samples becomes way too large when setting M as suggested by Theorems 1 and 2. The next proposition shows that the amortized running time of Recurrence Weighting remains as low as $O(d)$ even for large values of M.

Proposition 3 *Let S_t denote the number of sample actions taken by `RW` in round t. Then, $\mathbb{E}[S_t] \le d$. Also, for any $\delta > 0$,*

$$\sum_{t=1}^{T} S_t \le (e-1)dT + M\log\frac{1}{\delta}$$

holds with probability at least $1-\delta$.

Proof For proving the first statement, let us fix a time step t and notice that

$$S_t = \max_{j:V_{t,j}=1} K_{t,j} = \max_{j=1,2,\dots,d} V_{t,j}K_{t,j} \le \sum_{j=1}^{d} V_{t,j}K_{t,j}.$$

Now, observe that $\mathbb{E}\left[K_{t,j}\middle|\mathcal{F}_{t-1},V_{t,j}\right] \le 1/\mathbb{E}\left[V_{t,j}\middle|\mathcal{F}_{t-1}\right]$, which gives $\mathbb{E}[S_t] \le d$, thus proving the first statement. For the second part, notice that $X_t = \left(S_t - \mathbb{E}\left[S_t\middle|\mathcal{F}_{t-1}\right]\right)$ is a martingale-difference sequence with respect to $(\mathcal{F}_t)$ with $X_t \le M$ and with conditional variance

$$
\begin{aligned}
\mathrm{Var}\left[X_t\middle|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\left(S_t - \mathbb{E}\left[S_t\middle|\mathcal{F}_{t-1}\right]\right)^2\middle|\mathcal{F}_{t-1}\right] \le \mathbb{E}\left[S_t^2\middle|\mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\max_j \left(V_{t,j}K_{t,j}\right)^2\middle|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[\sum_{j=1}^{d} V_{t,j}K_{t,j}^2\middle|\mathcal{F}_{t-1}\right] \\
&\le \sum_{j=1}^{d} \min\left\{\frac{2}{q_{t,j}}, M\right\} \le dM,
\end{aligned}
$$

where we used $\mathbb{E}\left[K_{t,i}^2\middle|\mathcal{F}_{t-1}\right] = \frac{2-q_{t,i}}{q_{t,i}^2}$. Then, the second statement follows from applying a version of Freedman's inequality due to Beygelzimer et al. (2011) (stated as Lemma 16 in the appendix) with $B = M$ and $\Sigma_T \le dMT$. ■

Notice that choosing $M = O\big(\sqrt{dT}\big)$ as suggested by Theorems 1 and 2, the above result guarantees that the amortized running time of `FPL+RW` is $O\big((d+\sqrt{d/T})\cdot(f(\mathcal{S})+d)\big)$ with high probability.

4. Analysis

This section presents the proofs of Theorems 1 and 2. In a didactic attempt, we present statements concerning the loss-estimation procedure and the learning algorithm separately: Section 4.1 presents various important properties of the loss estimates produced by Recurrence Weighting, Section 4.2 presents general tools for analyzing Follow-the-Perturbed-Leader methods. Finally, Sections 4.3 and 4.4 put these results together to prove Theorems 1 and 2, respectively.

4.1 Properties of Recurrence Weighting

The basic idea underlying Recurrence Weighting is replacing the importance weights $1/q_{t,i}$ by appropriately defined random variables $K_{t,i}$. As we have seen earlier (Section 2), running `RW` with $M = \infty$ amounts to sampling each $K_{t,i}$ from a geometric distribution with expectation $1/q_{t,i}$, yielding an unbiased loss estimate. In practice, one would want to set M to a finite value to ensure that the running time of the sampling procedure is bounded. Note however that early termination of `RW` introduces a bias in the loss estimates. This section is mainly concerned with the nature of this bias. We emphasize that the statements presented in this section remain valid no matter what randomized algorithm generates the actions $\boldsymbol{V}_t$. Our first lemma gives an explicit expression on the expectation of the loss estimates generated by `RW`.

Lemma 4 *For all j and t such that $q_{t,j} > 0$, the loss estimates* (5) *satisfy*

$$\mathbb{E}\left[\widehat{\ell}_{t,j}\middle|\mathcal{F}_{t-1}\right] = \left(1 - (1-q_{t,j})^M\right)\ell_{t,j}.$$

Proof Fix any j, t satisfying the condition of the lemma. Setting $q = q_{t,j}$ for simplicity, we write

$$\begin{aligned}\mathbb{E}\left[K_{t,j}\left|\mathcal{F}_{t-1}\right.\right] &= \sum_{k=1}^{\infty} k(1-q)^{k-1}q - \sum_{k=M}^{\infty}(k-M)(1-q)^{k-1}q \\ &= \sum_{k=1}^{\infty} k(1-q)^{k-1}q - (1-q)^M \sum_{k=M}^{\infty}(k-M)(1-q)^{k-M-1}q \\ &= \left(1-(1-q)^M\right)\sum_{k=1}^{\infty} k(1-q)^{k-1}q = \frac{1-(1-q)^M}{q}.\end{aligned}$$

The proof is concluded by combining the above with $\mathbb{E}\left[\widehat{\ell}_{t,j}\middle|\mathcal{F}_{t-1}\right] = q_{t,j}\ell_{t,j}\mathbb{E}\left[K_{t,j}\middle|\mathcal{F}_{t-1}\right]$. ■

The following lemma shows two important properties of the recurrence-weighted loss estimates (5). Roughly speaking, the first of these properties ensure that any learning algorithm relying on these estimates will be *optimistic* in the sense that the loss of any *fixed* decision will be underestimated in expectation. The second property ensures that the learner will not be *overly optimistic* concerning its own performance.

Lemma 5 *For all $\boldsymbol{v} \in \mathcal{S}$ and t, the loss estimates* (5) *satisfy the following two properties:*

$$\mathbb{E}\left[\boldsymbol{v}^\top\widehat{\boldsymbol{\ell}}_t\middle|\mathcal{F}_{t-1}\right] \quad\leq\quad \boldsymbol{v}^\top\boldsymbol{\ell}_t, \tag{8}$$

$$\mathbb{E}\left[\sum_{\boldsymbol{u}\in\mathcal{S}} p_t(\boldsymbol{u})\left(\boldsymbol{u}^\top\widehat{\boldsymbol{\ell}}_t\right)\middle|\mathcal{F}_{t-1}\right] \quad\geq\quad \sum_{\boldsymbol{u}\in\mathcal{S}} p_t(\boldsymbol{u})\left(\boldsymbol{u}^\top\boldsymbol{\ell}_t\right) - \frac{d}{eM}. \tag{9}$$

Proof Fix any $\boldsymbol{v} \in \mathcal{S}$ and t. The first property is an immediate consequence of Lemma 4: we have that $\mathbb{E}\left[\widehat{\ell}_{t,k}\middle|\mathcal{F}_{t-1}\right] \leq \ell_{t,k}$ for all k, and thus $\mathbb{E}\left[\boldsymbol{v}^\top\widehat{\boldsymbol{\ell}}_t\middle|\mathcal{F}_{t-1}\right] \leq \boldsymbol{v}^\top\boldsymbol{\ell}_t$. For the second statement, observe that

$$\mathbb{E}\left[\sum_{\boldsymbol{u}\in\mathcal{S}} p_t(\boldsymbol{u})\left(\boldsymbol{u}^{\mathsf{T}}\widehat{\boldsymbol{\ell}}_t\right)\middle|\mathcal{F}_{t-1}\right] = \sum_{i=1}^{d} q_{t,i}\mathbb{E}\left[\widehat{\ell}_{t,i}\middle|\mathcal{F}_{t-1}\right] = \sum_{i=1}^{d} q_{t,i}\left(1-(1-q_{t,i})^M\right)\ell_{t,i}$$

also holds by Lemma 4. To control the bias term $\sum_i q_{t,i}(1-q_{t,i})^M$, note that $q_{t,i}(1-q_{t,i})^M \le q_{t,i}e^{-Mq_{t,i}}$. By elementary calculations, one can show that $f(q) = qe^{-Mq}$ takes its maximum at $q = \frac{1}{M}$ and thus $\sum_{i=1}^{d} q_{t,i}(1-q_{t,i})^M \le \frac{d}{eM}$. ■

Our last lemma concerning the loss estimates (5) bounds the conditional variance of the estimated loss of the learner. This term plays a key role in the performance analysis of `Exp3`-style algorithms (see, e.g., Auer et al. (2002); Uchiya et al. (2010); Audibert et al. (2014)), as well as in the analysis presented in the current paper.

Lemma 6 *For all t, the loss estimates* (5) *satisfy*

$$\mathbb{E}\left[\sum_{\boldsymbol{u}\in\mathcal{S}} p_t(\boldsymbol{u})\left(\boldsymbol{u}^{\mathsf{T}}\widehat{\boldsymbol{\ell}}_t\right)^2\middle|\mathcal{F}_{t-1}\right] \le 2md.$$

Before proving the statement, we remark that the conditional variance can be bounded as md for the standard (although usually infeasible) loss estimates (4). That is, the above lemma shows that, somewhat surprisingly, the variance of our estimates is only twice as large as the variance of the standard estimates.

Proof Fix an arbitrary t. For simplifying notation below, let us introduce $\widetilde{\boldsymbol{V}}$ as an independent copy of $\boldsymbol{V}_t$ such that $\mathbb{P}\left[\widetilde{\boldsymbol{V}} = \boldsymbol{v}\middle|\mathcal{F}_{t-1}\right] = p_t(\boldsymbol{v})$ holds for all $\boldsymbol{v}\in\mathcal{S}$. To begin, observe that for any i

$$\mathbb{E}\left[K_{t,i}^2\middle|\mathcal{F}_{t-1}\right] \le \frac{2-q_{t,i}}{q_{t,i}^2} \le \frac{2}{q_{t,i}^2} \tag{10}$$

holds, as $K_{t,i}$ has a truncated geometric law. The statement is proven as

$$\begin{aligned}
\mathbb{E}\left[\sum_{\boldsymbol{u}\in\mathcal{S}} p_t(\boldsymbol{u})\left(\boldsymbol{u}^{\mathsf{T}}\widehat{\boldsymbol{\ell}}_t\right)^2\middle|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\sum_{i=1}^{d}\sum_{j=1}^{d}\left(\widetilde{V}_i\widehat{\ell}_{t,i}\right)\left(\widetilde{V}_j\widehat{\ell}_{t,j}\right)\middle|\mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{d}\sum_{j=1}^{d}\left(\widetilde{V}_iK_{t,i}V_{t,i}\ell_{t,i}\right)\left(\widetilde{V}_jK_{t,j}V_{t,j}\ell_{t,j}\right)\middle|\mathcal{F}_{t-1}\right] \\
&\qquad \text{(using the definition of } \widehat{\boldsymbol{\ell}}_t\text{)} \\
&\le \mathbb{E}\left[\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{K_{t,i}^2+K_{t,j}^2}{2}\left(\widetilde{V}_iV_{t,i}\ell_{t,i}\right)\left(\widetilde{V}_jV_{t,j}\ell_{t,j}\right)\middle|\mathcal{F}_{t-1}\right] \\
&\qquad \text{(using } 2AB \le A^2+B^2\text{)} \\
&\le 2\mathbb{E}\left[\sum_{i=1}^{d}\frac{1}{q_{t,i}^2}\left(\widetilde{V}_iV_{t,i}\ell_{t,i}\right)\sum_{j=1}^{d}V_{t,j}\ell_{t,j}\middle|\mathcal{F}_{t-1}\right] \\
&\qquad \text{(using symmetry, Eq. (10) and } \widetilde{V}_j \le 1\text{)} \\
&\le 2m\mathbb{E}\left[\sum_{j=1}^{d}\ell_{t,j}\middle|\mathcal{F}_{t-1}\right] \le 2md,
\end{aligned}$$

where the last line follows from using $\left\|\boldsymbol{V}_t\right\|_1 \le m$ and $\mathbb{E}\left[V_{t,i}\middle|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[\widetilde{V}_i\middle|\mathcal{F}_{t-1}\right] = q_{t,i}$. ■

4.2 General tools for analyzing FPL

In this section, we present the key tools for analyzing the FPL-component of our learning algorithm. In some respect, our analysis is a synthesis of previous work on FPL-style methods: we borrow several ideas from Poland (2005) and the proof of Corollary 4.5 in Cesa-Bianchi and Lugosi (2006). Nevertheless, our analysis is the first to directly target combinatorial settings, and yields the tightest known bounds for FPL in this domain. Indeed, the tools developed in this section also permit an improvement for FPL in the full-information setting, closing the presumed performance gap between FPL and EWA in both the full-information and the semi-bandit settings. The statements we present in this section are not specific to the loss-estimate vectors used by FPL+RW.

Like most other known work, we study the performance of the learning algorithm through a *virtual algorithm* that (*i*) uses a time-independent perturbation vector and (*ii*) is allowed to peek one step into the future. Specifically, letting $\widetilde{\boldsymbol{Z}}$ be a perturbation vector drawn independently from the same distribution as $\boldsymbol{Z}_1$, the virtual algorithm picks its t^{th} action as

$$\widetilde{\boldsymbol{V}}_t = \arg\min_{\boldsymbol{v}\in\mathcal{S}} \left\{ \boldsymbol{v}^\mathsf{T} \left(\eta \widehat{\boldsymbol{L}}_t - \widetilde{\boldsymbol{Z}} \right) \right\}. \tag{11}$$

In what follows, we will crucially use that $\widetilde{\boldsymbol{V}}_t$ and $\boldsymbol{V}_{t+1}$ are conditionally independent and identically distributed given $\mathcal{F}_s$ for any $s \le t$. In particular, introducing the notations

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

we will exploit the above property by using $q_{t,i} = \widetilde{q}_{t-1,i}$ and $p_t(\boldsymbol{v}) = \widetilde{p}_{t-1}(\boldsymbol{v})$ numerous times in the proofs below.

First, we show a regret bound on the virtual algorithm that plays the action sequence $\widetilde{\boldsymbol{V}}_1, \widetilde{\boldsymbol{V}}_2, \dots, \widetilde{\boldsymbol{V}}_T$.

Lemma 7 *For any $\boldsymbol{v} \in \mathcal{S}$,*

$$\sum_{t=1}^T \sum_{\boldsymbol{u}\in\mathcal{S}} \widetilde{p}_t(\boldsymbol{u}) \left((\boldsymbol{u} - \boldsymbol{v})^\mathsf{T} \widehat{\boldsymbol{\ell}}_t \right) \le \frac{m\left(\log(d/m) + 1\right)}{\eta}. \tag{12}$$

Although the proof of this statement is rather standard, we include it for completeness. We also note that the lemma slightly improves other known results by replacing the usual $\log d$ term by $\log(d/m)$.

Proof Fix any $\boldsymbol{v} \in \mathcal{S}$. Using Lemma 3.1 of Cesa-Bianchi and Lugosi (2006) (sometimes referred to as the *"follow-the-leader/be-the-leader"* lemma) for the sequence $\left(\eta\widehat{\boldsymbol{\ell}}_1 - \widetilde{\boldsymbol{Z}}, \eta\widehat{\boldsymbol{\ell}}_2, \dots, \eta\widehat{\boldsymbol{\ell}}_T\right)$, we obtain

$$\eta \sum_{t=1}^T \widetilde{\boldsymbol{V}}_t^\mathsf{T} \widehat{\boldsymbol{\ell}}_t - \widetilde{\boldsymbol{V}}_1^\mathsf{T} \widetilde{\boldsymbol{Z}} \le \eta \sum_{t=1}^T \boldsymbol{v}^\mathsf{T} \widehat{\boldsymbol{\ell}}_t - \boldsymbol{v}^\mathsf{T} \widetilde{\boldsymbol{Z}}.$$

Reordering and integrating both sides with respect to the distribution of $\widetilde{\boldsymbol{Z}}$ gives

$$\eta \sum_{t=1}^T \sum_{\boldsymbol{u}\in\mathcal{S}} \widetilde{p}_t(\boldsymbol{u}) \left((\boldsymbol{u} - \boldsymbol{v})^\mathsf{T} \widehat{\boldsymbol{\ell}}_t \right) \le \mathbb{E}\left[\left(\widetilde{\boldsymbol{V}}_T - \boldsymbol{v} \right)^\mathsf{T} \widetilde{\boldsymbol{Z}} \right]. \tag{13}$$

The statement follows from using $\mathbb{E}\left[\widetilde{\boldsymbol{V}}_T^\mathsf{T} \widetilde{\boldsymbol{Z}} \right] \le m(\log(d/m) + 1)$, which is proven in Appendix A as Lemma 14. ∎

The next lemma relates the performance of the virtual algorithm to the actual performance of FPL. The lemma relies on a "sparse-loss" trick similar to the trick used in the proof Corollary 4.5 in Cesa-Bianchi and Lugosi (2006), and is also related to the "unit rule" discussed by Koolen et al. (2010).

Lemma 8 *For all $t = 1, 2, \dots, T$, assume that $\widehat{\boldsymbol{\ell}}_t$ is such that $\widehat{\ell}_{t,k} \geq 0$ for all $k \in \{1, 2, \dots, d\}$. Then,*

$$\sum_{\boldsymbol{u} \in \mathcal{S}} \left(p_t(\boldsymbol{u}) - \widetilde{p}_t(\boldsymbol{u})\right) \left(\boldsymbol{u}^\intercal \widehat{\boldsymbol{\ell}}_t\right) \leq \eta \sum_{\boldsymbol{u} \in \mathcal{S}} p_t(\boldsymbol{u}) \left(\boldsymbol{u}^\intercal \widehat{\boldsymbol{\ell}}_t\right)^2.$$

Proof Fix an arbitrary t. For all $\boldsymbol{v} \in \mathcal{S}$, we define the "sparse loss vector" $\widehat{\boldsymbol{\ell}}_t^-(\boldsymbol{v})$ with components $\widehat{\ell}_{t,k}^-(\boldsymbol{v}) = v_k \widehat{\ell}_{t,k}$ and

$$\boldsymbol{V}_t^-(\boldsymbol{v}) = \arg\min_{\boldsymbol{u} \in \mathcal{S}} \left\{ \boldsymbol{u}^\intercal \left(\eta \widehat{\boldsymbol{L}}_{t-1} + \eta \widehat{\boldsymbol{\ell}}_t^-(\boldsymbol{v}) - \widetilde{\boldsymbol{Z}} \right) \right\}.$$

Using the notation $p_t^-(\boldsymbol{v}) = \mathbb{P}\left[\boldsymbol{V}_t^-(\boldsymbol{v}) = \boldsymbol{v} \middle| \mathcal{F}_t\right]$, we show in Lemma 15 (stated and proved in Appendix A) that $p_t^-(\boldsymbol{v}) \leq \widetilde{p}_t(\boldsymbol{v})$. Also, define

$$\boldsymbol{U}(\boldsymbol{z}) = \arg\min_{\boldsymbol{u} \in \mathcal{S}} \left\{ \boldsymbol{u}^\intercal \left(\eta \widehat{\boldsymbol{L}}_{t-1} - \boldsymbol{z} \right) \right\}.$$

Letting $f(\boldsymbol{z}) = e^{-\|\boldsymbol{z}\|_1}$ ($\boldsymbol{z} \in \mathbb{R}_+^d$) be the density of the perturbations, we have

$$\begin{aligned}
p_t(\boldsymbol{v}) &= \int\limits_{\boldsymbol{z} \in [0,\infty]^d} \mathbb{1}_{\{\boldsymbol{U}(\boldsymbol{z}) = \boldsymbol{v}\}} f(\boldsymbol{z})\, d\boldsymbol{z} \\
&= e^{\eta \left\|\widehat{\boldsymbol{\ell}}_t^-(\boldsymbol{v})\right\|_1} \int\limits_{\boldsymbol{z} \in [0,\infty]^d} \mathbb{1}_{\{\boldsymbol{U}(\boldsymbol{z}) = \boldsymbol{v}\}} f\left(\boldsymbol{z} + \eta \widehat{\boldsymbol{\ell}}_t^-(\boldsymbol{v})\right) d\boldsymbol{z} \\
&= e^{\eta \left\|\widehat{\boldsymbol{\ell}}_t^-(\boldsymbol{v})\right\|_1} \int \cdots \int\limits_{z_i \in [\widehat{\ell}_{t,i}^-(\boldsymbol{v}), \infty]} \mathbb{1}_{\left\{\boldsymbol{U}\left(\boldsymbol{z} - \eta \widehat{\boldsymbol{\ell}}_t^-(\boldsymbol{v})\right) = \boldsymbol{v}\right\}} f(\boldsymbol{z})\, d\boldsymbol{z} \\
&\leq e^{\eta \left\|\widehat{\boldsymbol{\ell}}_t^-(\boldsymbol{v})\right\|_1} \int\limits_{\boldsymbol{z} \in [0,\infty]^d} \mathbb{1}_{\left\{\boldsymbol{U}\left(\boldsymbol{z} - \eta \widehat{\boldsymbol{\ell}}_t^-(\boldsymbol{v})\right) = \boldsymbol{v}\right\}} f(\boldsymbol{z})\, d\boldsymbol{z} \\
&\leq e^{\eta \left\|\widehat{\boldsymbol{\ell}}_t^-(\boldsymbol{v})\right\|_1} p_t^-(\boldsymbol{v}) \leq e^{\eta \left\|\widehat{\boldsymbol{\ell}}_t^-(\boldsymbol{v})\right\|_1} \widetilde{p}_t(\boldsymbol{v}).
\end{aligned}$$

Now notice that $\left\|\widehat{\boldsymbol{\ell}}_t^-(\boldsymbol{v})\right\|_1 = \boldsymbol{v}^\intercal \widehat{\boldsymbol{\ell}}_t^-(\boldsymbol{v}) = \boldsymbol{v}^\intercal \widehat{\boldsymbol{\ell}}_t$, which gives

$$\widetilde{p}_t(\boldsymbol{v}) \geq p_t(\boldsymbol{v}) e^{-\eta \boldsymbol{v}^\intercal \widehat{\boldsymbol{\ell}}_t} \geq p_t(\boldsymbol{v}) \left(1 - \eta \boldsymbol{v}^\intercal \widehat{\boldsymbol{\ell}}_t\right)$$

for all $\boldsymbol{v} \in \mathcal{S}$. The proof is concluded by reordering and summing for all $\boldsymbol{v}$ as

$$\sum_{\boldsymbol{u} \in \mathcal{S}} p_t(\boldsymbol{u}) \boldsymbol{u}^\intercal \widehat{\boldsymbol{\ell}}_t \leq \sum_{\boldsymbol{u} \in \mathcal{S}} \widetilde{p}_t(\boldsymbol{u}) \boldsymbol{u}^\intercal \widehat{\boldsymbol{\ell}}_t + \eta \sum_{\boldsymbol{u} \in \mathcal{S}} p_t(\boldsymbol{u}) \left(\boldsymbol{v}^\intercal \widehat{\boldsymbol{\ell}}_t\right)^2. \tag{14}$$

■

4.3 Proof of Theorem 1

Now, everything is ready to prove the bound on the expected regret of `FPL+RW`. Let us fix an arbitrary $\boldsymbol{v} \in \mathcal{S}$. By putting together Lemmas 6, 7 and 8, we immediately obtain

$$\mathbb{E}\left[\sum_{t=1}^T \sum_{\boldsymbol{u} \in \mathcal{S}} p_t(\boldsymbol{u}) \left((\boldsymbol{u} - \boldsymbol{v})^\intercal \widehat{\boldsymbol{\ell}}_t\right)\right] \leq \frac{m\left(\log(d/m) + 1\right)}{\eta} + 2\eta m d T, \tag{15}$$

leaving us with the problem of upper bounding the expected regret in terms of the left-hand side of the above inequality. This can be done by using the properties of the loss estimates (5) stated in Lemma 5:

$$\mathbb{E}\left[\sum_{t=1}^{T} (\boldsymbol{V}_t - \boldsymbol{v})^\mathsf{T} \boldsymbol{\ell}_t\right] \le \mathbb{E}\left[\sum_{t=1}^{T}\sum_{\boldsymbol{u}\in\mathcal{S}} p_t(\boldsymbol{u}) \left((\boldsymbol{u}-\boldsymbol{v})^\mathsf{T} \widehat{\boldsymbol{\ell}}_t\right)\right] + \frac{dT}{eM}.$$

Putting the two inequalities together proves the theorem.

4.4 Proof of Theorem 2

We now turn to prove a bound on the regret of `FPL+RW.P` that holds with high probability. We begin by applying Lemmas 7 and 8 to obtain the central terms in the regret:

$$\sum_{t=1}^{T}\sum_{\boldsymbol{u}\in\mathcal{S}} p_t(\boldsymbol{u}) \left((\boldsymbol{u}-\boldsymbol{v})^\mathsf{T} \widetilde{\boldsymbol{\ell}}_t\right) \le \frac{m(\log(d/m)+1)}{\eta} + \eta \sum_{t=1}^{T}\sum_{\boldsymbol{u}\in\mathcal{S}} p_t(\boldsymbol{u}) \left(\boldsymbol{u}^\mathsf{T}\widetilde{\boldsymbol{\ell}}_t\right)^2.$$

The first challenge posed by the above expression is relating the left-hand side to the true regret with high probability. Once this is done, the remaining challenge is to bound the second term on the right-hand side, as well as the other terms arising from the first step. We begin by showing that the loss estimates used by `FPL+RW.P` consistently underestimate the true losses with high probability.

Lemma 9 *Fix any $\delta' > 0$. For any $\boldsymbol{v} \in \mathcal{S}$,*

$$\boldsymbol{v}^\mathsf{T}\left(\widetilde{\boldsymbol{L}}_T - \boldsymbol{L}_T\right) \le \frac{m\log(d/\delta')}{\beta}$$

holds with probability at least $1-\delta'$.

The simple proof is directly inspired by Appendix C.9 of Audibert and Bubeck (2010).
Proof Fix any t and i. Then,

$$\mathbb{E}\left[\exp\left(\beta\widetilde{\ell}_{t,i}\right)\middle|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[\exp\left(\log\left(1+\beta\widehat{\ell}_{t,i}\right)\right)\middle|\mathcal{F}_{t-1}\right] \le 1+\beta\ell_{t,i} \le \exp(\beta\ell_{t,i}),$$

where we used Lemma 4 in the first inequality and $1+z \le e^z$ that holds for all $z \in \mathbb{R}$. As a result, the process $W_t = \exp\left(\beta\left(\widetilde{L}_{t,i} - L_{t,i}\right)\right)$ is a supermartingale with respect to $(\mathcal{F}_t)$: $\mathbb{E}\left[W_t\middle|\mathcal{F}_{t-1}\right] \le W_{t-1}$. Observe that, since $W_0 = 1$, this implies $\mathbb{E}[W_t] \le \mathbb{E}[W_{t-1}] \le \ldots \le 1$. Markov's inequality implies that

$$\begin{aligned}\mathbb{P}\left[\widetilde{L}_{T,i} > L_{T,i} + \varepsilon\right] &= \mathbb{P}\left[\widetilde{L}_{T,i} - L_{T,i} > \varepsilon\right] \\ &\le \mathbb{E}\left[\exp\left(\beta\left(\widetilde{L}_{T,i} - L_{T,i}\right)\right)\right]\exp(-\beta\varepsilon) \le \exp(-\beta\varepsilon)\end{aligned}$$

holds for any $\varepsilon > 0$. The statement of the lemma follows after using $\|\boldsymbol{v}\|_1 \le m$, applying the union bound for all i, and solving for ε. ■

The following lemma states another key property of the loss estimates.

Lemma 10 *For any t,*

$$\sum_{i=1}^{d} q_{t,i}\widehat{\ell}_{t,i} \le \sum_{i=1}^{d} q_{t,i}\widetilde{\ell}_{t,i} + \frac{\beta}{2}\sum_{i=1}^{d} q_{t,i}\widehat{\ell}_{t,i}^2.$$

Proof The statement follows trivially from the inequality $\log(1+z) \geq z - \frac{z^2}{2}$ that holds for all $z \geq 0$. In particular, for any fixed t and i, we have

$$\log\left(1+\beta\widehat{\ell}_{t,i}\right) \geq \beta\widehat{\ell}_{t,i} - \frac{\beta^2}{2}\widehat{\ell}^2_{t,i}.$$

Multiplying both sides by $q_{t,i}/\beta$ and summing for all i proves the statement. ■

The next lemma relates the total loss of the learner to its total estimated losses.

Lemma 11 *Fix any $\delta' > 0$. With probability at least $1-2\delta'$,*

$$\sum_{t=1}^{T} \boldsymbol{V}_t^\mathsf{T}\boldsymbol{\ell}_t \leq \sum_{t=1}^{T}\sum_{\boldsymbol{u}\in\mathcal{S}} p_t(\boldsymbol{u})\left(\boldsymbol{u}^\mathsf{T}\widehat{\boldsymbol{\ell}}_t\right) + \frac{dT}{eM} + \sqrt{2(e-2)T}\left(m\log\frac{1}{\delta'}+1\right) + \sqrt{T\log\frac{1}{\delta'}}$$

Proof We start by rewriting

$$\sum_{\boldsymbol{u}\in\mathcal{S}} p_t(\boldsymbol{u})\left(\boldsymbol{u}^\mathsf{T}\widehat{\boldsymbol{\ell}}_t\right) = \sum_{i=1}^{d} q_{t,i}K_{t,i}V_{t,i}\ell_{t,i}.$$

Now let $k_{t,i} = \mathbb{E}\left[K_{t,i}\middle|\mathcal{F}_{t-1}\right]$ for all i and notice that

$$X_t = \sum_{i=1}^{d} q_{t,i}V_{t,i}\ell_{t,i}\left(k_{t,i}-K_{t,i}\right)$$

is a martingale-difference sequence with respect to $(\mathcal{F}_t)$ with elements upper-bounded by m (as Lemma 4 implies $k_{t,i}q_{t,i} \leq 1$ and $\left\|\boldsymbol{V}_t\right\|_1 \leq m$). Furthermore, the conditional variance of the increments is bounded as

$$\operatorname{Var}\left[X_t\middle|\mathcal{F}_{t-1}\right] \leq \mathbb{E}\left[\left(\sum_{i=1}^{d} q_{t,i}V_{t,i}K_{t,i}\right)^2\middle|\mathcal{F}_{t-1}\right] \leq \mathbb{E}\left[\sum_{j=1}^{d} V_{t,j}\left(\sum_{i=1}^{d} q_{t,i}^2K_{t,i}^2\right)\middle|\mathcal{F}_{t-1}\right] \leq 2m,$$

where the second inequality is Cauchy-Schwarz and the third one follows from $\mathbb{E}\left[K_{t,}i^2\middle|\mathcal{F}_{t-1}\right] \leq 2/q_{t,i}^2$ and $\left\|\boldsymbol{V}_t\right\|_1 \leq m$. Thus, applying Lemma 16 with $B = m$ and $\Sigma_T \leq 2mT$ we get that for any $S \geq m\sqrt{\log\frac{1}{\delta'}/(e-2)}$,

$$\sum_{t=1}^{T}\sum_{i=1}^{d} q_{t,i}\ell_{t,i}V_{t,i}\left(k_{t,i}-K_{t,i}\right) \leq \sqrt{(e-2)\log\frac{1}{\delta'}}\left(\frac{2mT}{S}+S\right)$$

holds with probability at least $1-\delta'$, where we have used $\left\|\boldsymbol{V}_t\right\|_1 \leq m$. After setting $S = m\sqrt{2T\log\frac{1}{\delta'}}$, we obtain that

$$\sum_{t=1}^{T}\sum_{i=1}^{d} q_{t,i}\ell_{t,i}V_{t,i}\left(k_{t,i}-K_{t,i}\right) \leq \sqrt{2\left(e-2\right)T}\left(m\log\frac{1}{\delta'}+1\right) \tag{16}$$

holds with probability at least $1-\delta'$.

To proceed, observe that $q_{t,i}k_{t,i} = 1-(1-q_{t,i})^M$ holds by Lemma 4, implying

$$\sum_{i=1}^{d} q_{t,i}V_{t,i}\ell_{t,i}k_{t,i} \geq \boldsymbol{V}_t^\mathsf{T}\boldsymbol{\ell}_t - \sum_{i=1}^{d} V_{t,i}(1-q_{t,i})^M.$$

Together with Eq. (16), this gives

$$\sum_{t=1}^{T} \boldsymbol{V}_t^\mathsf{T} \boldsymbol{\ell}_t \leq \sum_{t=1}^{T} \sum_{\boldsymbol{u}\in\mathcal{S}} p_t(\boldsymbol{u}) \left(\boldsymbol{u}^\mathsf{T} \widehat{\boldsymbol{\ell}}_t\right) + \sqrt{2\,(e-2)\,T} \left(m \log \frac{1}{\delta'} + 1\right) + \sum_{t=1}^{T} \sum_{i=1}^{d} V_{t,i} (1-q_{t,i})^M.$$

Finally, we use that, by Lemma 5, $(1-q_{t,i})^M \leq 1/(eM)$, and

$$Y_t = \sum_{i=1}^{d} \left(V_{t,i} - q_{t,i}\right) \left(1-q_{t,i}\right)^M$$

is a martingale-difference sequence with respect to $(\mathcal{F}_t)$ with increments bounded in $[-1,1]$. Then, by an application of Hoeffding's inequality, we have

$$\sum_{t=1}^{T} \sum_{i=1}^{d} V_{t,i} (1-q_{t,i})^M \leq \frac{dT}{eM} + \sqrt{T \log \frac{1}{\delta'}}$$

with probability at least $1-\delta'$, thus proving the lemma,. ∎

Finally, our last lemma in this section bounds the second-order terms arising from Lemmas 8 and 10.

Lemma 12 *Fix any $\delta' > 0$. With probability at least $1-2\delta'$, the following hold simultaneously:*

$$\sum_{t=1}^{T} \sum_{\boldsymbol{v}\in\mathcal{S}} p_t(\boldsymbol{v}) \left(\boldsymbol{v}^\mathsf{T} \widehat{\boldsymbol{\ell}}_t\right)^2 \leq (Mm)^2 \log \frac{1}{\delta'} + 2(e-1)mdT$$

$$\sum_{t=1}^{T} \sum_{i=1}^{d} q_{t,i} \widehat{\ell}_{t,i}^2 \leq M^2 m \log \frac{1}{\delta'} + 2(e-1)dT$$

Proof First, recall that

$$\mathbb{E}\left[\sum_{\boldsymbol{v}\in\mathcal{S}} p_t(\boldsymbol{v}) \left(\boldsymbol{v}^\mathsf{T} \widehat{\boldsymbol{\ell}}_t\right)^2 \middle| \mathcal{F}_{t-1} \right] \leq 2md$$

holds by Lemma 8. Now, observe that

$$X_t = \sum_{\boldsymbol{v}\in\mathcal{S}} p_t(\boldsymbol{v}) \left(\left(\boldsymbol{v}^\mathsf{T} \widehat{\boldsymbol{\ell}}_t\right)^2 - \mathbb{E}\left[\left(\boldsymbol{v}^\mathsf{T} \widehat{\boldsymbol{\ell}}_t\right)^2 \middle| \mathcal{F}_{t-1} \right] \right)$$

is a martingale-difference sequence with

$$\begin{aligned} \operatorname{Var}\left[X_t \middle| \mathcal{F}_{t-1} \right] = \mathbb{E}\left[\left(\sum_{\boldsymbol{v}\in\mathcal{S}} p_t(\boldsymbol{v}) \left(\boldsymbol{v}^\mathsf{T} \widehat{\boldsymbol{\ell}}_t\right)^2 \right)^2 \middle| \mathcal{F}_{t-1} \right] &\leq (mM)^2\, \mathbb{E}\left[\sum_{\boldsymbol{v}\in\mathcal{S}} p_t(\boldsymbol{v}) \left(\boldsymbol{v}^\mathsf{T} \widehat{\boldsymbol{\ell}}_t\right)^2 \middle| \mathcal{F}_{t-1} \right] \\ &\leq (mM)^2\, 2md. \end{aligned}$$

and increments bounded by $(mM)^2$. Applying Lemma 16 gives

$$\sum_{t=1}^{T} \sum_{\boldsymbol{v}\in\mathcal{S}} p_t(\boldsymbol{v}) \left(\left(\boldsymbol{v}^\mathsf{T} \widehat{\boldsymbol{\ell}}_t\right)^2 - \mathbb{E}\left[\left(\boldsymbol{v}^\mathsf{T} \widehat{\boldsymbol{\ell}}_t\right)^2 \middle| \mathcal{F}_{t-1} \right] \right) \leq (mM)^2 \log \frac{1}{\delta'} + 2(e-2)mdT,$$

proving the first statement. The second statement is proven analogously, building on the bound

$$\mathbb{E}\left[\sum_{i=1}^{d} q_{t,i} \widehat{\ell}_{t,i}^2 \middle| \mathcal{F}_{t-1} \right] \leq \mathbb{E}\left[\sum_{i=1}^{d} q_{t,i} V_{t,i} K_{t,i}^2 \middle| \mathcal{F}_{t-1} \right] \leq 2d.$$

∎

Theorem 2 follows from combining Lemmas 9 through 12 and applying the union bound.

5. Improved bounds for learning with full information

Our proof techniques presented in Section 4.2 also enables us to tighten the guarantees for FPL in the full information setting. In particular, consider the algorithm choosing action

$$\boldsymbol{V}_t = \arg\min_{\boldsymbol{v}\in\mathcal{S}} \boldsymbol{v}^\mathsf{T} \left(\eta \boldsymbol{L}_{t-1} - \boldsymbol{Z}_t\right),$$

where $\boldsymbol{L}_t = \sum_{s=1}^{t} \boldsymbol{\ell}_s$ and the components of $\boldsymbol{Z}_t$ are drawn independently from a standard exponential distribution. We state our improved regret bounds concerning this algorithm in the following theorem.

Theorem 13 *For any $\boldsymbol{v} \in \mathcal{S}$, the total expected regret of FPL satisfies*

$$\widehat{R}_T \leq \frac{m\left(\log(d/m) + 1\right)}{\eta} + \eta m \sum_{t=1}^{T} \mathbb{E}\left[\boldsymbol{V}_t^\mathsf{T} \boldsymbol{\ell}_t\right]$$

under full information. In particular, defining $L_T^* = \min_{\boldsymbol{v}\in\mathcal{S}} \boldsymbol{v}^\mathsf{T} L_T$ and setting

$$\eta = \min\left\{ \sqrt{\frac{\log(d/m) + 1}{L_T^*}}, \frac{1}{2} \right\},$$

the regret of FPL satisfies

$$R_T \leq 4m \max\left\{ \sqrt{L_T^* \left(\log\left(\frac{d}{m}\right) + 1\right)}, \left(m^2 + 1\right)\left(\log \frac{d}{m} + 1\right) \right\}.$$

In the worst case, the above bound becomes $2m^{3/2}\sqrt{T\big(\log(d/m) + 1\big)}$, which improves the best known bound for FPL of Kalai and Vempala (2005) by a factor of $\sqrt{d/m}$.

Proof The first statement follows from combining Lemmas 7 and 8, and bounding

$$\sum_{\boldsymbol{u}\in\mathcal{S}}^{N} p_t(\boldsymbol{u}) \big(\boldsymbol{u}^\mathsf{T}\boldsymbol{\ell}_t\big)^2 \leq m \sum_{\boldsymbol{u}\in\mathcal{S}}^{N} p_t(\boldsymbol{u}) \big(\boldsymbol{u}^\mathsf{T}\boldsymbol{\ell}_t\big),$$

while the second one follows from standard algebraic manipulations. ∎

6. Conclusions and open problems

In this paper, we have described the first *general and efficient* algorithm for online combinatorial optimization under semi-bandit feedback. We have proved that the regret of this algorithm is $O\big(m\sqrt{dT\log(d/m)}\big)$ in this setting, and have also shown that FPL can achieve $O\big(m^{3/2}\sqrt{T\log(d/m)}\big)$ in the full information case when tuned properly. While these bounds are off by a factor of $\sqrt{m\log(d/m)}$ and $\sqrt{m}$ from the respective minimax results, they exactly match the best known regret bounds for the well-studied Exponentially Weighted Forecaster (EWA). Whether the remaining gaps can be closed for FPL-style algorithms (e.g., by using more intricate perturbation schemes or a more refined analysis) remains an important open question. Nevertheless, we regard our contribution as a significant step towards understanding the inherent trade-offs between computational efficiency and performance guarantees in online combinatorial optimization and, more generally, in online optimization.

The efficiency of our method rests on a novel loss estimation method called Recurrence Weighting (`RW`). This estimation method is not specific to the proposed learning algorithm. While `RW` has no immediate benefits for `OSMD`-type algorithms where the ideal importance weights are readily available, it is possible to think about problem instances where `EWA` can be efficiently implemented while importance weights are difficult to compute.

The most important open problem left is the case of efficient online linear optimization with *full bandit feedback* where the learner only observes the inner product $\boldsymbol{V}_t^\mathsf{T}\boldsymbol{\ell}_t$ in round t. Learning algorithms for this problem usually require that the (pseudo-)inverse of the covariance matrix $P_t = \mathbb{E}\left[\boldsymbol{V}_t\boldsymbol{V}_t^\mathsf{T}\middle|\,\mathcal{F}_{t-1}\right]$ is readily available for the learner at each time step (see, e.g., McMahan and Blum (2004); Dani et al. (2008); Cesa-Bianchi and Lugosi (2012); Bubeck et al. (2012)). Computing this matrix, however, is at least as challenging as computing the individual importance weights $1/q_{t,i}$. That said, our Recurrence Weighting technique can be directly generalized to this setting by observing that the matrix geometric series $\sum_{n=1}^{\infty}(I-P_t)^n$ converges to P_t^{-1} under certain conditions. This sum can then be efficiently estimated by sampling independent copies of $\boldsymbol{V}_t$, which paves the path for constructing low-bias estimates of the loss vectors. While it seems straightforward to go ahead and use these estimates in tandem with `FPL`, we have to note that the analysis presented in this paper does not carry through directly in this case. The main limitation is that our techniques only apply for loss vectors with *non-negative* elements (cf. Lemma 8). Nevertheless, we believe that Recurrence Weighting should be a crucial component for constructing truly effective learning algorithms for this important problem.

Acknowledgments

The research presented in this paper was supported by French Ministry of Higher Education and Research, by European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°270327 (CompLACS), and by FUI project Hermès. We thank Csaba Szepesvári for thought-provoking discussions.

Appendix A. Further proofs and technical tools

Lemma 14 *Let $Z_1,\dots,Z_d$ be i.i.d. exponential random variables with unit expectation and let $Z_1^*,\dots,Z_d^*$ be their permutation such that $Z_1^* \ge Z_2^* \ge \cdots \ge Z_d^*$. Then, for any $1 \le m \le d$,*

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

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

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

where in the last step, we used that $A = \log\left(\frac{d}{m}\right)$ minimizes $A + de^{-A/m}$ over the real line. ■

Lemma 15 *Fix any $i \in \{1, 2, \ldots, N\}$ and any vectors $\boldsymbol{L} \in \mathbb{R}^d$ and $\boldsymbol{\ell} \in [0, \infty)^d$. Furthermore, define the vector $\boldsymbol{\ell}'$ with components $\ell'_k = v_k(i)\ell_k$ and the perturbation vector $\boldsymbol{Z}$ with independent components. Then,*

$$\begin{aligned}&\mathbb{P}\left[\boldsymbol{v}(i)^\mathsf{T}\left(\boldsymbol{L} + \boldsymbol{\ell}' - \boldsymbol{Z}\right) \le \boldsymbol{v}(j)^\mathsf{T}\left(\boldsymbol{L} + \boldsymbol{\ell}' - \boldsymbol{Z}\right)\ (\forall j \in \{1, 2, \ldots, N\})\right] \\ &\qquad \le \mathbb{P}\left[\boldsymbol{v}(i)^\mathsf{T}\left(\boldsymbol{L} + \boldsymbol{\ell} - \boldsymbol{Z}\right) \le \boldsymbol{v}(j)^\mathsf{T}\left(\boldsymbol{L} + \boldsymbol{\ell} - \boldsymbol{Z}\right)\ (\forall j \in \{1, 2, \ldots, N\})\right].\end{aligned}$$

Proof Fix any $\forall j \in \{1, 2, \ldots, N\} \setminus i$ and define the vector $\boldsymbol{\ell}'' = \boldsymbol{\ell} - \boldsymbol{\ell}'$. Define the events

$$A'_j = \left\{\omega :\ \boldsymbol{v}(i)^\mathsf{T}\left(\boldsymbol{L} + \boldsymbol{\ell}' - \boldsymbol{Z}\right) \le \boldsymbol{v}(j)^\mathsf{T}\left(\boldsymbol{L} + \boldsymbol{\ell}' - \boldsymbol{Z}\right)\right\}$$

and

$$A_j = \left\{\omega :\ \boldsymbol{v}(i)^\mathsf{T}\left(\boldsymbol{L} + \boldsymbol{\ell} - \boldsymbol{Z}\right) \le \boldsymbol{v}(j)^\mathsf{T}\left(\boldsymbol{L} + \boldsymbol{\ell} - \boldsymbol{Z}\right)\right\}.$$

We have

$$\begin{aligned}A'_j &= \left\{\omega :\ \left(\boldsymbol{v}(i) - \boldsymbol{v}(j)\right)^\top \boldsymbol{Z} \ge \left(\boldsymbol{v}(i) - \boldsymbol{v}(j)\right)^\top \left(\boldsymbol{L} + \boldsymbol{\ell}'\right)\right\} \\ &\subseteq \left\{\omega :\ \left(\boldsymbol{v}(i) - \boldsymbol{v}(j)\right)^\top \boldsymbol{Z} \ge \left(\boldsymbol{v}(i) - \boldsymbol{v}(j)\right)^\top \left(\boldsymbol{L} + \boldsymbol{\ell}'\right) - \boldsymbol{v}(j)^\mathsf{T}\boldsymbol{\ell}''\right\} \\ &= \left\{\omega :\ \left(\boldsymbol{v}(i) - \boldsymbol{v}(j)\right)^\top \boldsymbol{Z} \ge \left(\boldsymbol{v}(i) - \boldsymbol{v}(j)\right)^\top \left(\boldsymbol{L} + \boldsymbol{\ell}\right)\right\} = A_j,\end{aligned}$$

where we used $\boldsymbol{v}(i)\boldsymbol{\ell}'' = 0$ and $\boldsymbol{v}(j)\boldsymbol{\ell}'' \ge 0$. Now, since $A'_j \subseteq A_j$, we have $\cap_{j=1}^N A'_j \subseteq \cap_{j=1}^N A_j$, thus proving $\mathbb{P}\left[\cap_{j=1}^N A'_j\right] \le \mathbb{P}\left[\cap_{j=1}^N A_j\right]$ as requested. ■

Lemma 16 (cf. Theorem 1 in Beygelzimer et al. (2011)) *Assume $X_1, X_2, \ldots, X_T$ is a martingale-difference sequence with respect to the filtration $(\mathcal{F}_t)$ with $X_t \le B$ for $1 \le t \le T$. Let $\sigma_t^2 = \operatorname{Var}\left[X_t \middle| \mathcal{F}_{t-1}\right]$ and $\Sigma_t^2 = \sum_{s=1}^t \sigma_s^2$. Then, for any $\delta > 0$,*

$$\mathbb{P}\left[\sum_{t=1}^T \boldsymbol{X}_t > B \log\frac{1}{\delta} + (e-2)\frac{\Sigma_T^2}{B}\right] \le \delta.$$

Furthermore, for any $S > B\sqrt{\log(1/\delta))(e-2)}$,

$$\mathbb{P}\left[\sum_{t=1}^T \boldsymbol{X}_t > \sqrt{(e-2)\log\frac{1}{\delta}}\left(\frac{\Sigma_T^2}{S} + S\right)\right] \le \delta.$$

References

Jacob Abernethy, Chansoo Lee, Abhinav Sinha, and Ambuj Tewari. Online linear optimization via smoothing. In M.-F. Balcan and Cs. Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *JMLR Proceedings*, pages 807–823. JMLR.org, 2014.

Chamy Allenberg, Peter Auer, László Györfi, and György Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In José L. Balcázar, Philip M. Long, and Frank 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. 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.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002. ISSN 0097-5397.

Baruch Awerbuch and Robert D. Kleinberg. Adaptive routing with end-to-end feedback: distributed learning and geometric approaches. In László Babai, editor, *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 45–53. ACM, June 13–16 2004. ISBN 1-58113-852-0.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandit algorithms with supervised learning guarantees. In D.B. Dunson G.J. Gordon, editor, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*, volume 15 of *JMLR Workshop and Conference Proceedings*, pages 19–26, April 11-13 2011.

S. Bubeck, N. Cesa-Bianchi, and S. M. Kakade. Towards minimax policies for online linear optimization with bandit feedback. *Arxiv preprint arXiv:1202.3079*, 2012.

Sébastien Bubeck and Nicolò Cesa-Bianchi. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. Now Publishers Inc, 2012.

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. *Journal of Computer and System Sciences*, 78:1404–1422, 2012.

V. Dani, T. Hayes, and S. Kakade. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, pages 345–352, 2008.

L. Devroye, G. Lugosi, and G. Neu. Prediction by random-walk perturbation. In Steinwart I. Shalev-Shwartz, S., editor, *Proceedings of the 25th Annual Conference on Learning Theory*, pages 460–473, 2013.

Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

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.

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

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

N. Littlestone and M. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

H. Brendan McMahan and Avrim Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In John Shawe-Taylor and Yoram Singer, editors, *Proceedings of the 17th Annual Conference on Learning Theory, COLT 2004*, volume 3120 of *Lecture Notes in Computer Science*, pages 109–123, Berlin, Heidelberg, July 1–4 2004. Springer. ISBN 3-540-22282-0.

Gergely Neu and Gábor Bartók. An efficient algorithm for learning with semi-bandit feedback. In Sanjay Jain, Rémi Munos, Frank Stephan, and Thomas 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.

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

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

Daiki Suehiro, Kohei Hatano, Shuji Kijima, Eiji Takimoto, and Kiyohito Nagano. Online prediction under submodular constraints. In *Algorithmic Learning Theory*, volume 7568 of *Lecture Notes in Computer Science*, pages 260–274. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-34105-2.

E. Takimoto and M. Warmuth. Paths kernels and multiplicative updates. *Journal of Machine Learning Research*, 4:773–818, 2003.

T. Uchiya, A. Nakamura, and M. Kudo. Algorithms for adversarial bandit problems with multiple plays. In *Proceedings of the 21st International Conference on Algorithmic Learning Theory (ALT)*, 2010.

Tim Van Erven, Manfred Warmuth, and Wojciech Kotłowski. Follow the leader with dropout perturbations. In M.-F. Balcan and Cs. Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *JMLR Proceedings*, pages 949–974. JMLR.org, 2014.

V. Vovk. Aggregating strategies. In *Proceedings of the third annual workshop on Computational learning theory (COLT)*, pages 371–386, 1990.