JMLR: Workshop and Conference Proceedings vol (2013) 1–27

Online Learning with Predictable Sequences

Alexander Rakhlin RAKHLIN@WHARTON.UPENN.EDU
Karthik Sridharan SKARTHIK@WHARTON.UPENN.EDU

Abstract

We present methods for online linear optimization that take advantage of benign (as opposed to worst-case) sequences. Specifically if the sequence encountered by the learner is described well by a known "predictable process", the algorithms presented enjoy tighter bounds as compared to the typical worst case bounds. Additionally, the methods achieve the usual worst-case regret bounds if the sequence is not benign. Our approach can be seen as a way of adding *prior knowledge* about the sequence within the paradigm of online learning. The setting is shown to encompass partial and side information. Variance and path-length bounds Hazan and Kale (2010); Chiang et al. (2012) can be seen as particular examples of online learning with simple predictable sequences.

We further extend our methods to include competing with a set of possible predictable processes (models), that is "learning" the predictable process itself concurrently with using it to obtain better regret guarantees. We show that such model selection is possible under various assumptions on the available feedback.

Keywords:

1. Introduction

No-regret methods are studied in a variety of fields, including learning theory, game theory, and information theory Cesa-Bianchi and Lugosi (2006). These methods guarantee a certain level of performance in a sequential prediction problem, irrespective of the sequence being presented. While such "protection" against the worst case is often attractive, the bounds are naturally pessimistic. It is, therefore, desirable to develop algorithms that yield tighter bounds for "more regular" sequences, while still providing protection against worst-case sequences. Some successful results of this type have appeared in Cesa-Bianchi et al. (2007); Hazan and Kale (2010, 2009); Chiang et al. (2012); Bartlett et al. (2007) within the framework of prediction with expert advice and online convex optimization.

In Rakhlin et al. (2011), a general game-theoretic formulation was put forth, with "regularity" of the sequence modeled as a set of restrictions on the possible moves of the adversary. Through a non-constructive analysis, the authors pointed to the *existence* of general regret-minimization strategies for benign sequences, but did not provide a computationally feasible method. In this paper, we present algorithms that achieve some of the regret bounds of Rakhlin et al. (2011) for sequences that can be described as

$$\text{sequence} = \text{predictable process} + \text{adversarial noise}$$

This paper focuses on the setting of online linear optimization, and the results achieved in the full-information case carry over to online *convex* optimization as well. To remind the reader of the setting, the online learning process is modeled as a repeated game with

convex sets $\mathcal{F}$ and $\mathcal{X}$ for the learner and Nature, respectively. At each round $t = 1, \dots, T$, the learners chooses $f_t \in \mathcal{F}$ and observes the move $x_t \in \mathcal{X}$ of Nature. The learner suffers a loss of $\langle f_t, x_t \rangle$ and the goal is to minimize regret, defined as

$$\mathbf{Reg}_T = \sum_{t=1}^{T} \langle f_t, x_t \rangle - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \langle f, x_t \rangle \,.$$

There are a number of ways to model "more regular" sequences. Let us start with the following definition. Fix a sequence of functions $M_t : \mathcal{X}^{t-1} \mapsto \mathcal{X}$, for each $t \in \{1, \dots, T\} \triangleq [T]$. These functions define a predictable process $M_1, \; M_2(x_1), \; \dots, \; M_T(x_1, \dots, x_{T-1})$. If, in fact, $x_t = M_t(x_1, \dots, x_{t-1})$ for all t, one may view the sequence $\{x_t\}$ as a (noiseless) time series, or as an oblivious strategy of Nature. If we knew that the sequence given by Nature follows exactly this evolution, we should suffer no regret.

Suppose that we have a hunch that the actual sequence will be "roughly" given by this predictable process: $x_t \approx M_t(x_1, \dots, x_{t-1})$. In other words, we suspect that the sequence is described as predictable process plus adversarial noise. Can we use this fact to incur smaller regret if our suspicion is correct? Ideally, we would like to "pay" only for the unpredictable part of the sequence.

Information-Theoretic Justification Let us spend a minute explaining why such regret bounds are information-theoretically possible. The key is the following observation, made in Rakhlin et al. (2011). The non-constructive upper bounds on the minimax value of the online game involve a symmetrization step, which we state for simplicity of notation for the linear loss case with $\mathcal{F}$ and $\mathcal{X}$ being dual unit balls:

$$\sup_{x_1, x_1'} \mathbb{E}_{\epsilon_1} \dots \sup_{x_T, x_T'} \mathbb{E}_{\epsilon_T} \left\| \sum_{t=1}^{T} \epsilon_t \Big(x_t' - x_t \Big) \right\|_* \leq 2 \sup_{x_1} \mathbb{E}_{\epsilon_1} \dots \sup_{x_T} \mathbb{E}_{\epsilon_T} \left\| \sum_{t=1}^{T} \epsilon_t x_t \right\|_*$$

If we instead only consider sequences such that at any time $t \in [T]$, x_t and x_t' have to be σ_t-close to the predictable process $M_t(x_1, \dots, x_{t-1})$, we can add and subtract the "center" M_t on the left-hand side of the above equation and obtain tighter bounds ***for free, irrespective of the form of*** $M_t(x_1, \dots, x_{t-1})$. To make this observation more precise, let

$$C_t = C_t(x_1, \dots, x_{t-1}) = \Big\{ x : \|x - M_t(x_1, \dots, x_{t-1})\|_* \leq \sigma_t \Big\} \tag{1}$$

be the set of allowed deviations from the predictable "trend". We then have a bound

$$\sup_{x_1, x_1' \in C_1} \mathbb{E}_{\epsilon_1} \dots \sup_{x_T, x_T' \in C_T} \mathbb{E}_{\epsilon_T} \left\| \sum_{t=1}^{T} \epsilon_t \Big(x_t' - M_t(x_1, \dots, x_{t-1}) + M_t(x_1, \dots, x_{t-1}) - x_t \Big) \right\|_* \leq c \sqrt{\sum_{t=1}^{T} \sigma_t^2}$$

on the value of the game against such "constrained" sequences, where the constant c depends on the smoothness of the norm. This short description only serves as a motivation, and the more precise statements about the value of a game against constrained adversaries can be found in Rakhlin et al. (2011).

The development so far is a good example of how a purely theoretical observation can point to existence of better prediction methods. What is even more surprising, for the methods presented below, the individual σ_t's need not be known ahead of time except for

their total sum $\sum_{t=1}^{T} \sigma_t^2$. Moreover, the latter sum need not be known in advance either, thanks to the standard doubling trick (see Appendix), and one can obtain upper bounds of

$$\sum_{t=1}^{T} \langle f_t, x_t \rangle - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \langle f, x_t \rangle \leq c \sqrt{\sum_{t=1}^{T} \|x_t - M_t(x_1, \ldots, x_{t-1})\|_*^2} \; . \tag{2}$$

Let us now discuss several types of statistics M_t that could be of interest. Regret bounds in terms of $M_t(x_1, \ldots, x_{t-1}) = x_{t-1}$ are known as *path length bounds* Chiang et al. (2012); Rakhlin et al. (2011). Such bounds can be tighter than the pessimistic $O(\sqrt{T})$ bounds when the previous instance x_t is a good proxy for the next move. Regret bounds in terms of $M_t(x_1, \ldots, x_{t-1}) = \frac{1}{t-1} \sum_{s=1}^{t-1} x_s$ are known as *variance bounds* (see Cesa-Bianchi et al. (2007); Hazan and Kale (2009, 2010); Rakhlin et al. (2011)). One may also consider fading memory statistics. That is, $M_t(x_1, \ldots, x_{t-1}) = \sum_{s=1}^{t-1} \alpha_s x_s$ where $\sum_{s=1}^{t-1} \alpha_s = 1$ and $\alpha_s \geq 0$ or even plug in an *auto-regressive model*. If "phases" are expected in the data (e.g., stocks tend to go up in January), one may consider $M_t(x_1, \ldots, x_{t-1}) = x_{t-k}$ for some phase length k. Alternatively, one may consider averaging of the past occurrences $T_j(t) \subset [t]$ of the current phase j to get a better predictive power: $M_t(x_1, \ldots, x_{t-1}) = \sum_{s \in T_t} \alpha_s x_s$.

The use of a predictable process $(M_t)_{t \geq 1}$ can be seen as a way of incorporating *prior knowledge* about the sequence $(x_t)_{t \geq 1}$. Importantly, the bounds still provide the usual worst-case protection if the process M_t does not predict the sequence well. To see this, observe that $\sqrt{\sum_{t=1}^{T} \|x_t - M_t\|_*^2} \leq 2 \max_{x \in \mathcal{X}} \|x\| \sqrt{T}$ which is only a factor of 2 larger than the typical bounds. However when M_t's do indeed predict x_t's well, we have lower regret, a property we get almost for free. Notice that in all our analysis the predictable process $(M_t)_{t \geq 1}$ can be any arbitrary function of the past.

A More General Setting The predictable process M_t has been written so far as a function of $x_1, \ldots, x_{t-1}$, as we assumed the setting of full-information online linear optimization (that is, x_t is revealed to the learner after playing f_t). Whenever our algorithm is deterministic, we may reconstruct the sequence $f_1, \ldots, f_t$ given the sequence $x_1, \ldots, x_{t-1}$, and thus no explicit dependence of M_t of the learner's moves are required. More generally, however, nothing prevents us from defining the predictable process M_t as a function

$$M_t(I_1, \ldots, I_{t-1}, f_1, \ldots, f_{t-1}, q_1, \ldots, q_{t-1}) \tag{3}$$

where I_s is the *information* conveyed to the learner at step $s \in [T]$ (defined on the appropriate information space $\mathcal{I}$) and q_s is the randomized strategy of the learner at time s. For instance, in the well-studied bandit framework, the feedback I_s is defined as the scalar value of the loss $\langle f_s, x_s \rangle$, yet the actual move x_s may remain unknown to the learner. More general partial information structures have also been studied in the literature.

When M_t is written in the form (3), it becomes clear that one can consider scenarios well beyond the partial information models. For instance, the information I_s might contain better or complete information about the past, thus modeling a delayed-feedback setting (see Section 5.1). Another idea is to consider a setting where I_s contains some state information pertinent to the online learning problem.

The paper is organized as follows. In Section 2, we provide a number of algorithms for full-information online linear optimization, taking advantage of a given predictable process

M_t. These methods can be seen as being "optimistic" about the sequence, incorporating M_{t+1} into the calculation of the next decision as if it were the true. We then turn to the partial information scenario in Section 3 and show how to use the full-information bounds together with estimation of the missing information. In Section 4 we turn to the question of *learning* M_t itself during the prediction process. We present several scenarios which differ in the amount of feedback given to the learner. Along the way, we need to prove a bound for nonstochastic multiarmed bandits in terms of the loss of the best arm (proved in the Appendix) – a result that does not appear to be available in the literature. Finally, we consider delayed feedback and other scenarios that fall under the general umbrella.

Remark 1 *We remark that most of the regret bounds we present in this paper are of the form* $A\eta^{-1} + B\eta \sum_{t=1}^{T} \|x_t - M_t\|_*^2$. *If variation around the trend is known in advance, one may choose* η *optimally to obtain the form in* (2). *Otherwise, we employ the standard doubling trick which we provide for completeness in Section B. The doubling trick sets* η *in a data-dependent way to achieve* (2) *with a slightly worse constant.*

Notation: We use the notation $y_{t':t}$ to represent the sequence $y_{t'}, \ldots, y_t$. We also use the notation $x[i]$ to represent the i^{th} element of vector x. We use the notation $x[1:c]$ to represent the vector $(x[1], \ldots, x[c])$. $D_R(f, f')$ is used to represent the Bregman divergence between f and f' w.r.t. function R. We denote the set $\{1, \ldots, T\}$ by $[T]$.

2. Full Information Methods

In this section we assume that the value M_t is known at the beginning of round t: it is either calculated by the learner or conveyed by an external source. The first algorithm we present is a modification of the Follow the Regularized Leader (FTRL) method with a self-concordant regularizer. The advantage of this method is its simplicity and the close relationship to the standard FTRL. Next, we exhibit a Mirror Descent type method which can be seen as a generalization of the recent algorithm of Chiang et al. (2012). Due to lack of space, full-information methods based on the idea of a random playout (Follow the Perturbed Leader) are postponed to the Appendix.

For all the methods presented below, we assume w.l.o.g. that $M_1 = 0$. Since we assume that M_t can be calculated from the information provided to the learner or the value of M_t is conveyed from outside, we do not write the dependence of M_t on the past explicitly.

2.1. Follow the Regularized Leader with Self-Concordant Barrier

Let $\mathcal{F} \subset \mathbb{R}^d$ be a convex compact set and let $\mathcal{R}$ be a self-concordant function for this set. W.l.o.g. suppose that $\min_{f \in \mathcal{F}} \mathcal{R}(f) = 0$. Given $f \in \mathrm{int}(\mathcal{F})$, define the local norm $\|\cdot\|_f$ with respect to $\mathcal{R}$ by $\|g\|_f \triangleq \sqrt{g^\top (\nabla^2 \mathcal{R}(f)) g}$. The dual norm is then $\|x\|_f^* = \sqrt{x^\top (\nabla^2 \mathcal{R}(f))^{-1} x}$. Given the f_t defined in the algorithm below, we use the shorthand $\|\cdot\|_t = \|\cdot\|_{f_t}$.

Optimistic Follow the Regularized Leader
Input: $\mathcal{R}$ self-concordant barrier, learning rate $\eta > 0$. $f_1 = \arg\min_{f \in \mathcal{F}} \mathcal{R}(f)$.

Update : $f_{t+1} = \arg\min_{f \in \mathcal{F}} \ \eta \left\langle f, \sum_{s=1}^{t} x_s + M_{t+1} \right\rangle + \mathcal{R}(f)$

We notice that for $M_{t+1} = 0$, the method reduces to the FTRL algorithm of Abernethy et al. (2008, 2012). When $M_{t+1} \neq 0$, the algorithm can be seen as "guessing" the next move and incorporating it into the objective. If the guess turns out to be correct, the method should suffer no regret. The following regret bound holds for the proposed algorithm:

Lemma 2 *Let $\mathcal{F} \subset \mathbb{R}^d$ be a convex compact set endowed with a self-concordant barrier $\mathcal{R}$ with $\min_{f\in\mathcal{F}} \mathcal{R}(f) = 0$. For any strategy of Nature, the Optimistic FTRL algorithm yields, for any $f^* \in \mathcal{F}$,*

$$\sum_{t=1}^{T} \langle f_t, x_t\rangle - \sum_{t=1}^{T} \langle f^*, x_t\rangle \le \eta^{-1}\mathcal{R}(f^*) + 2\eta \sum_{t=1}^{T} (\|x_t - M_t\|_t^*)^2 \tag{4}$$

as long as $\eta\|x_t - M_t\|_t^* < 1/4$ for all t.

By the argument of Abernethy et al. (2008, 2012), at the expense of an additive constant in the regret, the comparator f^* can be taken from a smaller set, at a distance $1/T$ from the boundary. For such an f^*, we have $\mathcal{R}(f^*) \le \vartheta \log T$ where ϑ is a self-concordance parameter of $\mathcal{R}$.

2.2. Mirror-Descent algorithm

The next algorithm is a modification of a Mirror Descent (MD) method for regret minimization. Let $\mathcal{R}$ be a 1-strongly convex function with respect to a norm $\|\cdot\|$, and let $D_{\mathcal{R}}(\cdot,\cdot)$ denote the Bregman divergence with respect to $\mathcal{R}$. Let $\nabla\mathcal{R}^*$ be the inverse of the gradient mapping $\nabla\mathcal{R}$. Let $\|\cdot\|_*$ be dual to $\|\cdot\|$ (yet we do not require $\mathcal{F}$ and $\mathcal{X}$ to be dual balls).

Consider the following algorithm:

Optimistic Mirror Descent Algorithm (equivalent form)
Input: $\mathcal{R}$ 1-strongly convex w.r.t. $\|\cdot\|$, learning rate $\eta > 0$, $f_1 = g_1 = \arg\min_g \mathcal{R}(g)$

Update :

$$g_{t+1} = \operatorname*{argmin}_{g\in\mathcal{F}} \eta\langle g, x_t\rangle + D_{\mathcal{R}}(g, g_t)$$
$$f_{t+1} = \operatorname*{argmin}_{f\in\mathcal{F}} \eta\langle f, M_{t+1}\rangle + D_{\mathcal{R}}(f, g_{t+1})$$

Such a two-projection algorithm for the case $M_t = x_{t-1}$ has been exhibited recently in Chiang et al. (2012).

Lemma 3 *Let $\mathcal{F}$ be a convex set in a Banach space $\mathcal{B}$ and $\mathcal{X}$ be a convex set in the dual space $\mathcal{B}^*$. Let $\mathcal{R} : \mathcal{B} \mapsto \mathbb{R}$ be a 1-strongly convex function on $\mathcal{F}$ with respect to some norm $\|\cdot\|$. For any strategy of Nature and any $f^* \in \mathcal{F}$, the Optimistic Mirror Descent yields*

$$\sum_{t=1}^{T} \langle f_t, x_t\rangle - \sum_{t=1}^{T} \langle f^*, x_t\rangle \le \eta^{-1} R_{\max}^2 + \frac{\eta}{2}\sum_{t=1}^{T} \|x_t - M_t\|_*^2$$

where $R_{\max}^2 = \max_{f\in\mathcal{F}} \mathcal{R}(f) - \min_{f\in\mathcal{F}} \mathcal{R}(f)$.

As mentioned before, the sum $\sum_{t=1}^{T} \|x_t - M_t\|_*^2$ need not be known in advance in order to set η, as the usual doubling trick can be employed. Both the Optimistic MD and Optimistic FTRL work in the setting of online convex optimization, where x_t's are now gradients at the points chosen by the learner. Last but not least, notice that if the sequence is not following the trend M_t as we hoped it would, we still obtain the same bounds as for the Mirror Descent (respectively, FTRL) algorithm, up to a constant.

2.2.1. Local Norms for Exponential Weights

For completeness, we also exhibit a bound in terms of local norms for the case of $\mathcal{F} \subset \mathbb{R}^d$ being the probability simplex and $\mathcal{X}$ being the ℓ_∞ ball. In the case of bandit feedback, such bounds serve as a stepping stone to building a strategy that explores according to the local geometry of the set (see Abernethy et al. (2012)). Letting $\mathcal{R}(f) = \sum_{i=1}^{d} f(i) \log f(i) - 1$, the Mirror Descent algorithm corresponds to the well-known Exponential Weights algorithm. We now show that one can also achieve a regret bound in terms of local norms defined through the Hessian $\nabla^2 \mathcal{R}(f)$, which is simply $\mathrm{diag}(f(1)^{-1}, \ldots, f(d)^{-1})$. To this end, let $\|g\|_t = \sqrt{g^\top \nabla^2 \mathcal{R}(f_t) g}$ and $\|x\|_t^* = \sqrt{x \nabla^2 \mathcal{R}(f_t)^{-1} x}$.

Lemma 4 *The Optimistic Mirror Descent on the probability simplex enjoys, for any $f^* \in \mathcal{F}$,*

$$\sum_{t=1}^{T} \langle f_t - f^*, x_t \rangle \le 2\eta \sum_{t=1}^{T} (\|x_t - M_t\|_t^*)^2 + \frac{\log d}{\eta}$$

as long as $\eta \|x_t - M_t\|_\infty \le 1/4$ at each step.

3. Methods for Partial and Bandit Information

We now turn to the setting of partial information and provide a generic estimation procedure along the lines of Hazan and Kale (2009). Here, we suppose that the learner receives only partial feedback I_t which is simply the loss $\langle f_t, x_t \rangle$ incurred at round t. Once again, we suppose to have access to some predictable process M_t. Note the generality of this framework: in some cases we might postulate that M_t needs to be calculated by the learner from the available information (which does not include the actual moves x_t); in other cases, however, we may assume that some statistic M_t (such as some partial information about the past moves) is conveyed to the learner as a side information from an external source. For the methods we present, we simply assume availability of the value M_t.

As in Section 2.1, we assume to have access to a self-concordant function $\mathcal{R}$ for $\mathcal{F}$, with the self-concordance parameter ϑ. Following Abernethy et al. (2008), at time t we define[1] our randomized strategy q_t to be a uniform distribution on the eigenvectors of $\nabla^2 \mathcal{R}(h_t)$ where $h_t \in \mathcal{F}$ is given by a full-information procedure as described below. The full-information procedure is simply Follow the Regularized Leader on the *estimated* moves $\tilde{x}_1, \ldots, \tilde{x}_{t-1}$ constructed from the information $I_1, \ldots, I_{t-1}, f_1, \ldots, f_{t-1}, q_1, \ldots, q_{t-1}$, with $I_s = \langle f_s, x_s \rangle$. The resulting algorithm is dubbed SCRiBLe in (Abernethy et al., 2012). Hazan and Kale (2009)

1. We caution the reader that the roles of f_t and x_t in Abernethy et al. (2008); Hazan and Kale (2009) are exactly the opposite. We decided to follow the notation of Rakhlin et al. (2010, 2012), where in the supervised learning case it is natural to view the move f_t as a function.

observed that this algorithm can be modified by adding and subtracting an estimated mean of the adversarial moves at appropriate steps of the method. We use this idea with a general process M_t:

SCRiBLe for a Predictable Process
Input: $\eta > 0$, ϑ-self-concordant $\mathcal{R}$. Define $h_1 = \arg\min_{f\in\mathcal{F}} \mathcal{R}(f)$.
At time $t = 1$ to T
Let $\Lambda_1, \ldots, \Lambda_n$ and $\lambda_1, \ldots, \lambda_n$ be the eigenvectors and eigenvalues of $\nabla^2\mathcal{R}(h_t)$.
Choose i_t uniformly at random from $\{1, \ldots, n\}$ and $\varepsilon_t = \pm 1$ with probability $1/2$.
Predict $f_t = h_t + \varepsilon_t \lambda_{i_t}^{-1/2} \Lambda_{i_t}$ and observe loss $\langle f_t, x_t\rangle$.
Define $\tilde{x}_t := n\left(\langle f_t, x_t - M_t\rangle\right)\varepsilon_t \lambda_{i_t}^{1/2} \cdot \Lambda_{i_t} + M_t$.
Update : $h_{t+1} = \arg\min_{h\in\mathcal{F}} \left[\eta\left\langle h, \sum_{s=1}^t \tilde{x}_s + M_{t+1}\right\rangle + \mathcal{R}(h)\right]$.

The analysis of the method is based on the bounds for full information predictable processes M_t developed earlier, thus simplifying and generalizing the analysis of Hazan and Kale (2009).

Lemma 5 *Suppose that $\mathcal{F}$ is contained in the ℓ_2 ball of radius 1. The expected regret of the above algorithm (SCRiBLe for a Predictable Process) is*

$$\mathbb{E}\left[\sum_{t=1}^T \langle f_t, x_t\rangle - \sum_{t=1}^T \langle f^*, x_t\rangle\right] \le \eta^{-1}\mathcal{R}(f^*) + 2\eta n^2 \mathbb{E}\left[\sum_{t=1}^T (\langle f_t, x_t - M_t\rangle)^2\right] \le \eta^{-1}\mathcal{R}(f^*) + 2\eta n^2 \sum_{t=1}^T \mathbb{E}\left[\|x_t - M_t\|^2\right] \tag{5}$$

Hence, for any full-information statistic $M'_t = M'_t(x_1, \ldots, x_{t-1})$,

$$\mathbb{E}\left[\sum_{t=1}^T \langle f_t, x_t\rangle - \sum_{t=1}^T \langle f^*, x_t\rangle\right] \le \eta^{-1}\mathcal{R}(f^*) + 4\eta n^2 \sum_{t=1}^T \mathbb{E}\left[\|x_t - M'_t\|^2\right] + 4\eta n^2 \sum_{t=1}^T \mathbb{E}\left[\|M_t - M'_t\|^2\right] \tag{6}$$

Effectively, Hazan and Kale (2009) show that for the full-information statistic $M'_t(x_1, \ldots, x_{t-1}) = \frac{1}{t-1}\sum_{s=1}^{t-1} x_s$, there is a way to construct $M_t = M_t(I_{1:t-1}, f_{1:t-1}, q_{1:t-1})$ in such a way that the third term in (6) is of the order of the second term. This is done by putting aside roughly $O(\log T)$ rounds in order to estimate M'_t, via a process called *reservoir sampling*. However, for more general functions M'_t, the third term might have nothing to do with the second term, and the investigation of which M'_t can be well estimated by M_t is an interesting topic of further research.

4. Learning The Predictable Processes

So far we have seen that the learner with an access to an arbitrary predictable process $(M_t)_{t\ge 1}$ has a strategy with a regret bound of $O\left(\sqrt{\sum_{t=1}^T \|x_t - M_t\|_*^2}\right)$. Now if the predictable process is a good predictor of the sequence, then the regret will be low. This raises the

Optimistic Mirror Descent with Learning the Predictable Process
Input: $\mathcal{R}$ 1-strongly convex w.r.t. $\|\cdot\|$, learning rate $\eta > 0$
Initialize $f_1 = g_1 = \arg\min_g \mathcal{R}(g)$ and initialize $q_1 \in \Delta(\Pi)$ as, $\forall \pi \in \Pi, q_1(\pi) = \frac{1}{|\Pi|}$
Set $M_1 = \sum_{\pi\in\Pi} q_1(\pi) M_1^\pi$
At $t = 1, \ldots, T$, predict f_t, observe x_t and update :

$$\forall \pi \in \Pi,\ q_{t+1}(\pi) \propto q_t(\pi)\, e^{-\|M_t^\pi - x_t\|_*^2} \quad \text{and} \quad M_{t+1} = \textstyle\sum_{\pi\in\Pi} q_{t+1}(\pi) M_{t+1}^\pi$$

$$g_{t+1} = \operatorname*{argmin}_{g\in\mathcal{F}} \eta \langle g, x_t\rangle + D_{\mathcal{R}}(g, g_t) \quad , \quad f_{t+1} = \operatorname*{argmin}_{f\in\mathcal{F}} \eta \langle f, M_{t+1}\rangle + D_{\mathcal{R}}(f, g_{t+1})$$

question of model selection: how can the learner *choose* a good predictable process $(M_t)_{t\geq 1}$? Is it possible to learn it *online* as we go, and if so, what does it mean to learn?

To formalize the concept of learning the predictable process, let us consider the case where we have a set Π indexing a set of predictable processes (strategies) we are interested in. That is, each $\pi \in \Pi$ corresponds to predictable process given by $(M_t^\pi)_{t\geq 1}$. Now if we had an oracle which in the start of the game told us which $\pi^* \in \Pi$ predicts the sequence optimally (in hindsight) then we could use the predictable process given by $(M_t^{\pi^*})_{t\geq 1}$ and enjoy a regret bound of $O\left(\sqrt{\inf_{\pi\in\Pi} \sum_{t=1}^T \|x_t - M_t^\pi\|_*^2}\right)$. However we cannot expect to know which $\pi \in \Pi$ is the optimal one from the outset. In this scenario one would like to learn a predictable process that in turn can be used with algorithms proposed thus far to get a regret bound comparable with regret bound one could have obtained knowing the optimal $\pi^* \in \Pi$.

4.1. Learning M_t's : Full Information

To motivate this setting better let us consider an example. Say there are n stock options we can choose to invest in. On each day t, associated with each stock option one has a loss/payoff that occurs upon investing in a single share of that stock. Our goal in the long run is to have a low regret with respect to the single best stock in hindsight. Up to this point, the problem just corresponds to the simple experts setting where each of the n stocks is one expert and on each day we split our investment according to a probability distribution over the n options. However now additionally we allow the learner/investor access to *prediction models* from the set Π. These could be human strategists making forecasts, or outcomes of some hedge-fund model. At each time step the learner can query the prediction made by each $\pi \in \Pi$ as to what the loss on the n stocks would be on that day. Now we would like to have a regret comparable to the regret we can achieve knowing the best model $\pi^* \in \Pi$ that in hind-sight predicted the losses of each stock optimally. We shall now see how to achieve this.

The proof of the following lemma relies on a particular regret bound of (Cesa-Bianchi and Lugosi, 2006, Corollary 2.3) for the exponential weights algorithm that is *in terms of the loss of the best arm*. Such a bound is an improvement over the pessimistic regret bound when the loss of the optimal arm is small.

Lemma 6 *Let $\mathcal{F}$ be a convex subset of a unit ball in a Banach space $\mathcal{B}$ and $\mathcal{X}$ be a convex subset of the dual unit ball. Let $\mathcal{R} : \mathcal{B} \mapsto \mathbb{R}$ be a 1-strongly convex function on $\mathcal{F}$ with respect*

to some norm $\|\cdot\|$*. For any strategy of Nature, the Optimistic Mirror Descent Algorithm yields, for any* $f^* \in \mathcal{F}$*,*

$$\sum_{t=1}^{T} \langle f_t, x_t \rangle - \sum_{t=1}^{T} \langle f^*, x_t \rangle \leq \eta^{-1} R_{\max}^2 + 3.2\,\eta \left(\inf_{\pi \in \Pi} \sum_{t=1}^{T} \|x_t - M_t^{\pi}\|_*^2 + \log |\Pi| \right)$$

where $R_{\max}^2 = \max_{f \in \mathcal{F}} \mathcal{R}(f) - \min_{f \in \mathcal{F}} \mathcal{R}(f)$*.*

Once again, let us discuss what makes this setting different from the usual setting of experts. The forecast given by prediction models is in the form of a *vector*, one for each stock. If we treat each prediction model as an expert with the loss $\|x_t - M_t^{\pi}\|_*^2$, the experts algorithm would guarantee that we achieve the best cumulative loss of this kind. However, this is not the object of interest to us, as we are after the best allocation of our money among the stocks, as measured by $\inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \langle f, x_t \rangle$.

The algorithm stated above can be seen as separating two steps: learning the model (that is, predictable process) and then minimizing regret given the learned process. This is implemented by a general idea of running another (secondary) regret minimizing strategy where loss per round is simply $\|M_t - x_t\|_*^2$ and regret is considered with respect to the best $\pi \in \Pi$. That is, regret of the secondary regret minimizing game is given by

$$\sum_{t=1}^{T} \|x_t - M_t\|_*^2 - \inf_{\pi \in \Pi} \sum_{t=1}^{T} \|x_t - M_t^{\pi}\|_*^2$$

In general, the experts algorithm for minimizing secondary regret can be replaced by any other online learning algorithm.

4.2. Learning M_t's : Partial Information

In the previous section we considered the full information setting where on each round we have access to x_t and for each π we get to see (or compute) M_t^{π}. However one might be in a scenario with only partial access to x_t or M_t^{π}, or both. In fact, there are quite a number of interesting partial-information scenarios, and we consider some of them in this section.

4.2.1. Partial Information about Loss (Bandit Setting)

In this setting at each time step t, we only observe the loss $\langle f_t, x_t \rangle$ and not all of x_t. However, for each $\pi \in \Pi$ we do get access to (or can compute) M_t^{π} for each $\pi \in \Pi$. Consider the following algorithm:

SCRiBLe while Learning the Predictable Process
Input: $\eta > 0$, ϑ-self-concordant $\mathcal{R}$. Define $h_1 = \arg\min_{f\in\mathcal{F}} \mathcal{R}(f)$.
Initialize $q_1 \in \Delta(\Pi)$ as, $\forall \pi \in \Pi, q_1(\pi) = \frac{1}{|\Pi|}$
Set $M_1 = \sum_{\pi\in\Pi} q_1(\pi) M_1^\pi$
At time $t = 1$ to T
Let $\Lambda_1, \dots, \Lambda_n$ and $\lambda_1, \dots, \lambda_n$ be the eigenvectors and eigenvalues of $\nabla^2\mathcal{R}(h_t)$.
Choose i_t uniformly at random from $\{1, \dots, n\}$ and $\varepsilon_t = \pm 1$ with probability $1/2$.
Predict $f_t = h_t + \varepsilon_t \lambda_{i_t}^{-1/2} \Lambda_{i_t}$ and observe loss $\langle f_t, x_t\rangle$.
Define $\tilde{x}_t := n\left(\langle f_t, x_t - M_t\rangle\right) \varepsilon_t \lambda_{i_t}^{1/2} \cdot \Lambda_{i_t} + M_t$.
Update : $q_{t+1}(\pi) \propto q_t(\pi)\, e^{-(\langle f_t, x_t\rangle - \langle f_t, M_t^\pi\rangle)^2}$ and $M_{t+1} = \sum_{\pi\in\Pi} q_{t+1}(\pi) M_{t+1}^\pi$

$$h_{t+1} = \arg\min_{h\in\mathcal{F}} \left[\eta \left\langle h, \sum_{s=1}^{t} \tilde{x}_s + M_{t+1} \right\rangle + \mathcal{R}(h)\right].$$

The following lemma upper bounds regret of this algorithm. Once again, the proof requires a regret bound in terms of the loss of the best arm (Cesa-Bianchi and Lugosi, 2006, Corollary 2.3).

Lemma 7 *Suppose that $\mathcal{F}, \mathcal{X}$ are contained in the ℓ_2 ball of radius 1. The expected regret of* SCRiBLe while Learning the Predictable Process *is*

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle f_t, x_t\rangle - \sum_{t=1}^{T} \langle f^*, x_t\rangle\right] \le \eta^{-1}\mathcal{R}(f^*) + 2\eta n^2 \mathbb{E}\left[\sum_{t=1}^{T} (\langle f_t, x_t - M_t\rangle)^2\right] \tag{7}$$
$$\le \eta^{-1}\mathcal{R}(f^*) + 13\eta n^2 \left(\mathbb{E}\left[\inf_{\pi\in\Pi} \sum_{t=1}^{T} \|x_t - M_t^\pi\|^2\right] + \log|\Pi|\right).$$

4.2.2. Partial Information about Predictable Process

Now let us consider the scenario where on each round we get to see $x_t \in \mathcal{X}$. However, we only see $M_t^{\pi_t}$ for a single $\pi_t \in \Pi$ we select on round t. This scenario is especially useful in the stock investment example provided earlier. While x_t the vector of losses for the stocks on each day can easily be obtained at the end of the trading day, prediction processes might be provided as paid services by various companies. Therefore, we only get to access a limited number of forecasts on each day by paying for them. In this section, we provide an algorithm with corresponding regret bound for this case.

Due to the scarce information about the predictable processes, the proofs of Lemmas 8 and 9 below require an improved regret bound for the multiarmed bandit, an analogue of (Cesa-Bianchi and Lugosi, 2006, Corollary 2.3). Such a bound is proved in Lemma 11 in Section A. We have not seen this result in the literature, and the bound might be of independent interest (note that the bound of Auer et al. (2003) in terms of the largest gain is of a very different nature). Lemma 11 relies on using a self-concordant barrier for the simplex and utilizing the regret bound in terms of local norms.

Armed with Lemma 11, we can prove the following:

Optimistic MD with Learning the Pred. Proc. with Partial Information
Input: $\mathcal{R}$ 1-strongly convex w.r.t. $\|\cdot\|$, learning rate $\eta > 0$
Initialize $g_1 = \arg\min_g \mathcal{R}(g)$ and initialize $q_1 \in \Delta(\Pi)$ as, $\forall \pi \in \Pi, q_1(\pi) = \frac{1}{|\Pi|}$
Sample $\pi_1 \sim q_1$ and set $f_1 = \operatorname{argmin}_{f\in\mathcal{F}} \eta \left\langle f, M_1^{\pi_1} \right\rangle + D_{\mathcal{R}}(f, g_1)$
At $t = 1, \ldots, T$, predict f_t and :
Update q_t using SCRiBLe for multi-armed bandit with loss of arm π_t : $\|M_t^{\pi_t} - x_t\|_*^2$ and step-size $1/32|\Pi|^2$.
Sample $\pi_{t+1} \sim q_{t+1}$ and observe $M_{t+1}^{\pi_{t+1}}$. Update
$g_{t+1} = \operatorname{argmin}_{g\in\mathcal{F}} \eta \langle g, x_t \rangle + D_{\mathcal{R}}(g, g_t)$, $f_{t+1} = \operatorname{argmin}_{f\in\mathcal{F}} \eta \left\langle f, M_{t+1}^{\pi_{t+1}} \right\rangle + D_{\mathcal{R}}(f, g_{t+1})$

Lemma 8 *Let $\mathcal{F}$ be a convex set in a Banach space $\mathcal{B}$ and $\mathcal{X}$ be a convex set in the dual space $\mathcal{B}^*$, both contained in unit balls. Let $\mathcal{R} : \mathcal{B} \mapsto \mathbb{R}$ be a 1-strongly convex function on $\mathcal{F}$ with respect to some norm $\|\cdot\|$. For any strategy of Nature, the* Optimistic MD with Learning the Predictable Processes with Partial Information *Algorithm yields, for any $f^* \in \mathcal{F}$,*

$$\mathbb{E}\left[\sum_{t=1}^T \langle f_t, x_t\rangle\right] - \sum_{t=1}^T \langle f^*, x_t\rangle \le \eta^{-1} R_{\max}^2 + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^T \|x_t - M_t^{\pi_t}\|_*^2\right] \tag{8}$$
$$\le \eta^{-1} R_{\max}^2 + \eta\left(\mathbb{E}\inf_{\pi\in\Pi}\sum_{t=1}^T \|x_t - M_t^{\pi}\|_*^2 + 32|\Pi|^3\log(T|\Pi|)\right)$$

where $R_{\max}^2 = \max_{f\in\mathcal{F}} \mathcal{R}(f) - \min_{f\in\mathcal{F}} \mathcal{R}(f)$.

4.2.3. Partial Information about both Loss and Predictable Process

In the third partial information variant, we consider the setting where at time t we only observe the loss $\langle f_t, x_t\rangle$ we suffer at the time step (and not entire x_t) and also only $M_t^{\pi_t}$ corresponding to the predictable process of $\pi_t \in \Pi$ we select at time t. This is a blend of the two partial-information settings considered earlier.

Lemma 9 *Suppose that $\mathcal{F}, \mathcal{X}$ are contained in the ℓ_2 ball of radius 1. The expected regret of* SCRiBLe for Learning the Predictable Process with Partial Feedback *(see next page) is*

$$\mathbb{E}\left[\sum_{t=1}^T \langle f_t, x_t\rangle - \sum_{t=1}^T \langle f^*, x_t\rangle\right] \le \eta^{-1}\mathcal{R}(f^*) + 2\eta n^2 \mathbb{E}\left[\sum_{t=1}^T (\langle f_t, x_t - M_t^{\pi_t}\rangle)^2\right] \tag{9}$$
$$\le \eta^{-1}\mathcal{R}(f^*) + 4\eta n^2\left(\mathbb{E}\left[\inf_{\pi\in\Pi}\sum_{t=1}^T \|x_t - M_t^{\pi}\|^2\right] + 32|\Pi|^3\log(T|\Pi|)\right).$$

SCRiBLe for Learning the Predictable Process with Partial Feedback
Input: $\eta > 0$, ϑ-self-concordant $\mathcal{R}$. Define $h_1 = \arg\min_{f\in\mathcal{F}} \mathcal{R}(f)$.
Initialize $q_1 \in \Delta(\Pi)$ as, $\forall \pi \in \Pi, q_1(\pi) = \frac{1}{|\Pi|}$ and draw $\pi_1 \sim q_1$
At time $t = 1$ to T
Let $\Lambda_1, \dots, \Lambda_n$ and $\lambda_1, \dots, \lambda_n$ be the eigenvectors and eigenvalues of $\nabla^2\mathcal{R}(h_t)$.
Choose i_t uniformly at random from $\{1, \dots, n\}$ and $\varepsilon_t = \pm 1$ with probability $1/2$.
Predict $f_t = h_t + \varepsilon_t \lambda_{i_t}^{-1/2} \Lambda_{i_t}$ and observe loss $\langle f_t, x_t\rangle$.
Define $\tilde{x}_t := n\left(\langle f_t, x_t - M_t^{\pi_t}\rangle\right) \varepsilon_t \lambda_{i_t}^{1/2} \cdot \Lambda_{i_t} + M_t^{\pi_t}$.
Update q_t using SCRiBLe for multi-armed bandit with loss
of arm $\pi_t \in \Pi$: $\left(\langle f_t, x_t\rangle - \langle f_t, M_t^{\pi_t}\rangle\right)^2$ and step size $1/32|\Pi|^2$.
Draw $\pi_{t+1} \sim q_{t+1}$ and update $h_{t+1} = \arg\min_{h\in\mathcal{F}} \left[\eta \left\langle h, \sum_{s=1}^t \tilde{x}_s + M_{t+1}^{\pi_{t+1}}\right\rangle + \mathcal{R}(h)\right]$.

5. Other Examples

5.1. Delayed Feedback

As an example, consider the setting where the information given to the player at round t consists of two parts: the bandit feedback $\langle f_t, x_t\rangle$ about the cost of the chosen action, as well as full information about the past move x_{t-k}. For $t > k$, let $M_t = M_t(I_1, \dots, I_{t-1}) = \frac{1}{t-k-1}\sum_{s=1}^{t-k-1} x_s$. Then

$$\|M_t - M_t'\|^2 = \left\|\tfrac{1}{t-k-1}\textstyle\sum_{s=1}^{t-k-1} x_s - \tfrac{1}{t-1}\sum_{s=1}^{t-1} x_s\right\|^2 \le \left\|\tfrac{k}{(t-1)(t-k-1)}\textstyle\sum_{s=1}^{t-k-1} x_s - \tfrac{1}{t-1}\sum_{s=t-k}^{t-1} x_s\right\|^2 \le \tfrac{4k^2}{(t-1)^2},$$

where $M_t' = \frac{1}{t-1}\sum_{s=1}^{t-1} x_s$ is the full information statistic. It is immediate from Lemma 5 that the expected regret of the algorithm is

$$\mathbb{E}\left[\textstyle\sum_{t=1}^T \langle f_t, x_t\rangle - \sum_{t=1}^T \langle f^*, x_t\rangle\right] \le \eta^{-1}\mathcal{R}(f^*) + 4\eta n^2 \textstyle\sum_{t=1}^T \mathbb{E}\left[\|x_t - M_t'\|^2\right] + 32\eta n^2 k^2$$

This simple argument shows that variance-type bounds are immediate in bandit problems with delayed full information feedback.

5.2. I.I.D. Data

Consider the case of i.i.d. sequence $x_1, \dots, x_T$ drawn from an unknown distribution with mean $\mu \in \mathbb{R}^d$. Let us first discuss the full-information model. Consider the bound of either Lemma 2 or Lemma 3 for $M_t = \frac{1}{t-1}\sum_{s=1}^{t-1} x_s$. For simplicity, let $\|\cdot\|$ be the Euclidean norm (the argument works with any smooth norm). We may write $\|x_t - M_t\|^2 \le \|x_t - \mu\|^2 + \|M_t - \mu\|^2 + 2\langle x_t - \mu, M_t - \mu\rangle$. Taking the expectation over i.i.d. data, the first term in the above bound is variance σ^2 of the distribution under the given norm, while the third term disappears under the expectation. For the second term, we perform the same quadratic expansion to obtain

$$\mathbb{E}\|M_t - \mu\|^2 \le \tfrac{1}{(t-1)^2}\textstyle\sum_{s=1}^{t-1} \mathbb{E}\|x_t - \mu\|^2 \le \frac{\sigma^2}{t-1} \quad \text{and thus} \quad \sum_{t=1}^T \mathbb{E}\|x_t - M_t\|^2 \le T\sigma^2 + \sigma^2(\log T + 1)$$

Coupled with the full-information results of Lemma 2 or Lemma 3, we obtain an $\tilde{O}(\sigma\sqrt{T})$ bound on regret, implying the natural transition from the noisy to deterministically predictable case as the noise level goes to zero.

The same argument works for the case of bandit information, given that M_t can be constructed to estimate M_t' well (e.g. using the arguments of Hazan and Kale (2009)).

Acknowledgements

We gratefully acknowledge the support of NSF under grants CAREER DMS-0954737 and CCF-1116928, as well as Dean's Research Fund.

References

J. Abernethy and A. Rakhlin. Beating the adaptive bandit with high probability. Technical Report UCB/EECS-2009-10, EECS Department, University of California, Berkeley, Jan 2009.

J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, volume 3, page 3, 2008.

J.D. 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, 2012.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2003.

P.L. Bartlett, E. Hazan, and A. Rakhlin. Adaptive online gradient descent. *Advances in Neural Information Processing Systems*, 20:65–72, 2007.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

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

N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2):321–352, 2007.

C.-K. Chiang, T. Yang, C.-J. Lee, M. Mahdavi, C.-J. Lu, R. Jin, and S. Zhu. Online optimization with gradual variations. In *COLT*, 2012.

E. Hazan and S. Kale. Better algorithms for benign bandits. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 38–47. Society for Industrial and Applied Mathematics, 2009.

E. Hazan and S. Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine learning*, 80(2):165–188, 2010.

A.S. Nemirovski and M.J. Todd. Interior-point methods for optimization. *Acta Numerica*, 17(1):191–234, 2008.

A. Rakhlin. Lecture notes on online learning, 2008. Available at http://www-stat.wharton.upenn.edu/~rakhlin/papers/online_learning.pdf.

A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Random averages, combinatorial parameters, and learnability. In *NIPS*, 2010. Available at http://arxiv.org/abs/1006.1138.

A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Stochastic, constrained, and smoothed adversaries. In *NIPS*, 2011. Available at http://arxiv.org/abs/1104.5070.

A. Rakhlin, O. Shamir, and K. Sridharan. Relax and localize: From value to algorithms. *CoRR*, abs/1204.0870, 2012. Submitted.

Appendix A. Improved Bounds for Small Losses

While the regret bound for the original SCRiBLe algorithm follows immediately from the more general Lemma 5, we now state an alternative bound for SCRiBLe in terms of the loss of the optimal decision. The bound holds under the assumption of positivity on the losses. Lemma 10 is of independent interest and will be used as a building block for the analogous result for the multi-armed bandit in Lemma 11. Such bounds in terms of the loss of the best arm are attractive, as they give tighter results whenever the loss of the optimal decision is small. Thanks to this property, Lemma 11 is used in Section 4 in order to obtain bounds in terms of predictable process performance.

Lemma 10 *Consider the case when $\mathcal{R}$ is a self-concordant barrier over $\mathcal{F}$ and sets $\mathcal{F}$ and $\mathcal{X}$ are such that each $\langle f, x\rangle \in [0, s]$. Then for the SCRiBLe algorithm, for any choice of step size $\eta < 1/(2sn^2)$, we have the bound*

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle f_t, x_t\rangle\right] \le \frac{1}{1-(2sn^2)\eta}\left(\sum_{t=1}^{T}\langle f^*, x_t\rangle + \eta^{-1}\mathcal{R}(f^*)\right)$$

We now state and prove a bound in terms of the loss of the best arm for the case of non-stochastic multiarmed bandits. Such a bound is interesting in its own right and, to the best of our knowledge, it does not appear in the literature.[2] Our approach is to use SCRiBLe with a self-concordant barrier for the probability simplex, coupled with the bound of Lemma 10. (We were not able to make this result work with the entropy function, even with the local norm bounds).

Suppose that Nature plays a sequence $x_1, \ldots, x_T \in [0, s]^d$. On each round, we chose an arm j_t and observe $\langle e_{j_t}, x_t\rangle$.

SCRiBLe for multi-armed Bandit (Abernethy et al., 2012, 2008)
Input: $\eta > 0$. Let $\mathcal{R}(f) = -\sum_{i=1}^{d-1}\log(f[i]) - \log(1-\sum_{i=1}^{d-1} f[i])$
Initialize q_1 with uniform distribution over arms. Let $h_1 = q_1[1:d-1]$
At time $t = 1$ to T
Let $\{\Lambda_1, \ldots, \Lambda_{d-1}\}$ and $\{\lambda_1, \ldots, \lambda_{d-1}\}$ be the eigenvectors and eigenvalues of $\nabla^2\mathcal{R}(h_t)$.
Choose i_t uniformly at random from $\{1, \ldots, [d-1]\}$ and draw $\varepsilon_t \sim \mathrm{Unif}\{\pm 1\}$.
Set $f_t = h_t + \varepsilon_t \lambda_{i_t}^{-1/2}\Lambda_{i_t}$ and $q_t = (f_t, 1 - \sum_{i=1}^{d-1} f_t[i])$.
Draw arm $j_t \sim q_t$ and suffer loss $\langle e_{j_t}, x_t\rangle$.
Define $\tilde{x}_t := d\left(\langle e_{j_t}, x_t\rangle\right)\varepsilon_t \lambda_{i_t}^{1/2}\cdot\Lambda_{i_t}$.
Update : $h_{t+1} = \arg\min_{h\in\mathbb{R}^{d-1}}\left[\eta\left\langle h, \sum_{s=1}^{t}\tilde{x}_s\right\rangle + \mathcal{R}(h)\right]$.

Lemma 11 *Suppose $x_1, \ldots, x_T \in [0, s]^d$. For any $\eta < 1/(4sd^2)$ the expected regret of the SCRiBLe for multi-armed Bandit algorithm is bounded as :*

$$\mathbb{E}\left\{\sum_{t=1}^{T}\langle e_{j_t}, x_t\rangle\right\} \le \frac{1}{1-4\eta s d^2}\left(\inf_{j\in[d]}\sum_{t=1}^{T}\langle e_j, x_t\rangle + d\eta^{-1}\log(dT)\right)$$

2. The bound of Auer et al. (2003) is in terms of maximal gains, which is very different from a bound in terms of minimal loss. To the best of our knowledge, the trick of redefining losses as negative gains does not work here.

Appendix B. The Doubling Trick

For completeness, we now describe a more or loss standard doubling trick, extending it to the case of partial information. Let $\mathcal{I}$ stand for some information space such that the algorithm receives $I_t \in \mathcal{I}$ at time t, as described in the introduction. Let $\Psi : \cup_s (\mathcal{I} \times \mathcal{F})^s \mapsto \mathbb{R}$ be a (deterministic) function defined for any contiguous time interval of any size $s \in [T]$. By the definition, $\Psi(I_r, \dots, I_t, f_r, \dots, f_t)$ is computable by the algorithm after the t-th step, for any $r \le t$. We make the following monotonicity assumption on Ψ: for any $I_1, \dots, I_t \in \mathcal{I}$ and any $f_1, \dots, f_t \in \mathcal{F}$, $\Psi(I_{1:t-1}, f_{1:t-1}) \le \Psi(I_{1:t}, f_{1:t})$ and $\Psi(I_{2:t}, f_{2:t}) \le \Psi(I_{1:t}, f_{1:t})$.

Lemma 12 *Suppose we have a randomized algorithm that takes a fixed η as input and for some constant A without a priori knowledge of τ, for any $\tau > 0$, guarantees expected regret of the form*

$$\mathbb{E}\left[\sum_{t=1}^{\tau} \mathrm{loss}(f_t, x_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{\tau} \mathrm{loss}(f, x_t)\right] \le A\eta^{-1} + \eta \mathbb{E}\left[\Psi(I_{1:\tau}, f_{1:\tau})\right]$$

where Ψ satisfies the above stated requirements. Then using this algorithm as a black-box for any $T > 0$, we can provide a randomized algorithm with a regret bound

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{loss}(f_t, x_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \mathrm{loss}(f, x_t)\right] \le 16\sqrt{A \mathbb{E}\left[\Psi(I_{1:T}, f_{1:T})\right]}$$

Proof The prediction problem is broken into phases, with a constant learning rate $\eta_i = \eta_0 2^{-i}$ throughout the i-th phase, for some $\eta_0 > 0$. Define for $i \ge 1$

$$s_{i+1} = \min\left\{\tau : \eta_i \Psi(I_{s_i:\tau}, f_{s_i:\tau}) > A\eta_i^{-1}\right\}$$

to be the start of the phase $i+1$, and $s_1 = 1$. Let N be the last phase of the game and let $s_{N+1} = T + 1$. Without loss of generality, assume $N > 1$ (for, otherwise regret is at most $4A/\eta_0$). Then

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{loss}(f_t, x_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \mathrm{loss}(f, x_t)\right] &\le \mathbb{E}\left[\sum_{k=1}^{N} \mathop{\mathbb{E}}_{f_{s_k:s_{k+1}-1}} \left[\sum_{t=s_k}^{s_{k+1}-1} \mathrm{loss}(f_t, x_t) - \inf_{f \in \mathcal{F}} \sum_{t=s_k}^{s_{k+1}-1} \mathrm{loss}(f, x_t)\right]\right] \\
&\le \mathbb{E}\left[\sum_{k=1}^{N} \left(A\eta_k^{-1} + \eta_k \mathop{\mathbb{E}}_{f_{s_k:s_{k+1}-1}} \left[\Psi(I_{s_k:s_{k+1}-1}, f_{s_k:s_{k+1}-1})\right]\right)\right] \\
&\le 2\mathbb{E}\left[\sum_{k=1}^{N} A\eta_k^{-1}\right]
\end{aligned}$$

where the last inequality follows because $\eta_k \Psi(I_{s_k:s_{k+1}-1}, f_{s_k:s_{k+1}-1}) \le A\eta_k^{-1}$ within each phase. Also observe that

$$\eta_{N-1} \Psi(I_{s_{N-1}:s_N}, f_{s_{N-1}:s_N}) > A\eta_{N-1}^{-1},$$

which implies

$$\eta_0^{-1} 2^N = \eta_N^{-1} = 2\eta_{N-1}^{-1} < 2\sqrt{\frac{\Psi(I_{s_{N-1}:s_N}, f_{s_{N-1}:s_N})}{A}} \le 2\sqrt{\frac{\Psi(I_{1:T}, f_{1:T})}{A}}$$

by the monotonicity assumption. Hence, regret is upper bounded by

$$2\sum_{k=1}^{N} A\eta_k^{-1} = 2A\eta_0^{-1}2^N \sum_{k=1}^{N} 2^{k-N} \leq 4A\eta_0^{-1}2^N \leq 8\sqrt{A\ \Psi(I_{1:T}, f_{1:T})}$$

Putting the arguments together,

$$\mathbb{E}\left[\sum_{t=1}^{T} \text{loss}(f_t, x_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \text{loss}(f, x_t)\right] \leq 8\mathbb{E}\left[\sqrt{A\ \Psi(I_{1:T}, f_{1:T})}\right] \leq 8\sqrt{A\ \mathbb{E}\left[\Psi(I_{1:T}, f_{1:T})\right]}$$

Now, observe that the rule for stopping the phase can only be calculated *after* the first time step of the new phase. The easiest way to deal with this is to throw out N time periods and suffer an additional regret of sN (losses are bounded by s). Using $\eta_0 = 4A/s$ this leads to additional factor of $sN \leq s2^N = 4A\eta_0^{-1}2^N \leq 8\sqrt{A\ \Psi(I_{1:T}, f_{1:T})}$, which is a gross over-bound. In conclusion, the overall bound on regret is

$$\mathbb{E}\left[\sum_{t=1}^{T} \text{loss}(f_t, x_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \text{loss}(f, x_t)\right] \leq 16\sqrt{A\mathbb{E}\left[\Psi(I_{1:T}, f_{1:T})\right]}\ .$$

■

We remark that while the algorithm may or may not start each new phase from a cold start (that is, forget about what has been learned), the functions M_t may still contain information about all the past moves of Nature.

With this doubling trick, for any of the full information bounds presented in the paper (for instance Lemmas 2, 3, 4 and 6) we can directly get an algorithm that enjoys a regret bound that is a factor at most 8 from the bound with optimal choice of η.

For Lemmas 5, 7, 8 and 9, we need to apply the doubling trick to an intermediate quantity, as the final bound is given in terms of quantities not computable by the algorithm. Specifically, the doubling trick needs to be applied to Equations (5), (7), (8) and (9), respectively, in order to get bounds that are within a factor 8 from the bounds obtained by optimizing η in the corresponding equations. We can then upper these computable quantities by corresponding unobserved quantities as is done in these lemmas. To see this more clearly let us demonstrate this on the example of Lemma 9. By Equation (9), we have that

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle f_t, x_t\rangle - \sum_{t=1}^{T} \langle f^*, x_t\rangle\right] \leq \eta^{-1}\mathcal{R}(f^*) + 2\eta n^2 \mathbb{E}\left[\sum_{t=1}^{T} (\langle f_t, x_t - M_t^{\pi_t}\rangle)^2\right]$$

Now note that $(\langle f_t, x_t - M_t^{\pi_t}\rangle)^2$ is a quantity computable by the algorithm at each round. Also note that $2\eta n^2 \sum_{t=1}^{T} (\langle f_t, x_t - M_t^{\pi_t}\rangle)^2$ satisfies the condition on Ψ required by Lemma 12, as the sum of squares is monotonic. Hence using the lemma we can conclude that

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle f_t, x_t\rangle - \sum_{t=1}^{T} \langle f^*, x_t\rangle\right] \leq 16\sqrt{2n^2\mathcal{R}(f^*)\mathbb{E}\left[\sum_{t=1}^{T} (\langle f_t, x_t - M_t^{\pi_t}\rangle)^2\right]} \tag{10}$$

The following steps in Lemma 9 (see proof in the Appendix) imply that

$$\mathbb{E}\left[\sum_{t=1}^{T} (\langle f_t, x_t - M_t^{\pi_t}\rangle)^2\right] \leq 2\left(\mathbb{E}\left[\inf_{\pi\in\Pi} \sum_{t=1}^{T} \left\|x_t - \bar{M}_t^{\pi}\right\|^2\right] + 32|\Pi|^3 \log(T|\Pi|)\right)$$

Plugging the above in Equation 10 we can conclude that

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle f_t, x_t\rangle - \sum_{t=1}^{T}\langle f^*, x_t\rangle\right] \le 16\sqrt{4n^2\mathcal{R}(f^*)\left(\mathbb{E}\left[\inf_{\pi\in\Pi}\sum_{t=1}^{T}\left\|x_t - \bar{M}_t^{\pi}\right\|^2\right] + 32|\Pi|^3\log(T|\Pi|)\right)}$$

This is exactly the inequality one would get if the final bound in Lemma 9 is optimized for η, with an additional factor of 8. With similar argument we can get the tight bounds for Lemmas 5, 7 and 8 too, even though they are in the bandit setting.

Appendix C. Proofs

Proof [**Proof of Lemma 2**] Define $g_{t+1} = \arg\min_{f\in\mathcal{F}} \eta\left\langle f, \sum_{s=1}^{t} x_s\right\rangle + \mathcal{R}(f)$ to be the (unmodified) Follow the Regularized Leader. Observe that for any $f^* \in \mathcal{F}$,

$$\sum_{t=1}^{T}\langle f_t - f^*, x_t\rangle = \sum_{t=1}^{T}\langle f_t - g_{t+1}, x_t - M_t\rangle + \sum_{t=1}^{T}\langle f_t - g_{t+1}, M_t\rangle + \sum_{t=1}^{T}\langle g_{t+1} - f^*, x_t\rangle \tag{11}$$

We now prove by induction that

$$\sum_{t=1}^{\tau}\langle f_t - g_{t+1}, M_t\rangle + \sum_{t=1}^{\tau}\langle g_{t+1}, x_t\rangle \le \sum_{t=1}^{\tau}\langle f^*, x_t\rangle + \eta^{-1}\mathcal{R}(f^*).$$

The base case $\tau = 1$ is immediate since $M_1 = 0$. For the purposes of induction, suppose that the above inequality holds for $\tau = T-1$. Using $f^* = f_T$ and adding $\langle f_T - g_{T+1}, M_T\rangle + \langle g_{T+1}, x_T\rangle$ to both sides,

$$\begin{aligned}
\sum_{t=1}^{T}\langle f_t - g_{t+1}, M_t\rangle + \sum_{t=1}^{T}\langle g_{t+1}, x_t\rangle &\le \sum_{t=1}^{T-1}\langle f_T, x_t\rangle + \eta^{-1}\mathcal{R}(f_T) + \langle f_T - g_{T+1}, M_T\rangle + \langle g_{T+1}, x_T\rangle \\
&\le \left\langle f_T, \sum_{t=1}^{T-1} x_t + M_T\right\rangle + \eta^{-1}\mathcal{R}(f_T) - \langle g_{T+1}, M_T\rangle + \langle g_{T+1}, x_T\rangle \\
&\le \left\langle g_{T+1}, \sum_{t=1}^{T-1} x_t + M_T\right\rangle + \eta^{-1}\mathcal{R}(g_{T+1}) - \langle g_{T+1}, M_T\rangle + \langle g_{T+1}, x_T\rangle \\
&\le \left\langle g^*, \sum_{t=1}^{T} x_t\right\rangle + \eta^{-1}\mathcal{R}(g^*)
\end{aligned}$$

by the optimality of f_T and g_{T+1}. This concludes the inductive argument, and from Eq. (11) we obtain

$$\sum_{t=1}^{T}\langle f_t - f^*, x_t\rangle \le \sum_{t=1}^{T}\langle f_t - g_{t+1}, x_t - M_t\rangle + \eta^{-1}\mathcal{R}(f^*) \tag{12}$$

Define the Newton decrement for $\Phi_t(f) \triangleq \eta\left\langle f, \sum_{s=1}^{t} x_s + M_{t+1}\right\rangle + \mathcal{R}(f)$ as

$$\lambda(f, \Phi_t) := \|\nabla\Phi_t(f)\|_f^* = \|\nabla^2\Phi_t(f)^{-1}\nabla\Phi_t(f)\|_f.$$

Since $\mathcal{R}$ is self-concordant then so is Φ_t, with their Hessians coinciding. The Newton decrement measures how far a point is from the global optimum. The following result can

be found, for instance, in Nemirovski and Todd (2008): For any self-concordant function $\tilde{\mathcal{R}}$, whenever $\lambda(f, \tilde{\mathcal{R}}) < 1/2$, we have

$$\|f - \arg\min \tilde{\mathcal{R}}\|_f \leq 2\lambda(f, \tilde{\mathcal{R}})$$

where the local norm $\|\cdot\|_f$ is defined with respect to $\tilde{\mathcal{R}}$, i.e. $\|g\|_f := \sqrt{g^\intercal(\nabla^2 \tilde{\mathcal{R}}(f))g}$. Applying this to Φ_t and using the fact that $\nabla\Phi_{t-1}(g_{t+1}) = \eta(M_t - x_t)$,

$$\|f_t - g_{t+1}\|_{f_t} = \|g_{t+1} - \arg\min \Phi_t\|_{f_t} \leq 2\lambda(g_{t+1}, \Phi_t) = 2\eta\|M_t - x_t\|^*_{f_t}. \tag{13}$$

Hence,

$$\begin{aligned}\sum_{t=1}^{T} \langle f_t - f^*, x_t\rangle &\leq \sum_{t=1}^{T} \|f_t - g_{t+1}\|_t \|x_t - M_t\|^*_t + \eta^{-1}\mathcal{R}(f^*) \\ &\leq 2\eta \sum_{t=1}^{T} (\|x_t - M_t\|^*_{f_t})^2 + \eta^{-1}\mathcal{R}(f^*),\end{aligned}$$

which proves the statement. ■

Proof [**Proof of Lemma 3**] For any $f^* \in \mathcal{F}$,

$$\langle f_t - f^*, x_t\rangle = \langle f_t - g_{t+1}, x_t - M_t\rangle + \langle f_t - g_{t+1}, M_t\rangle + \langle g_{t+1} - f^*, x_t\rangle \tag{14}$$

First observe that

$$\langle f_t - g_{t+1}, x_t - M_t\rangle \leq \|f_t - g_{t+1}\| \|x_t - M_t\|_* \leq \frac{\eta}{2}\|x_t - M_t\|^2_* + \frac{1}{2\eta}\|f_t - g_{t+1}\|^2 \ . \tag{15}$$

On the other hand, any update of the form $a^* = \arg\min_{a\in A} \langle a, x\rangle + D_{\mathcal{R}}(a, c)$ satisfies for any $d \in A$ (see e.g. Beck and Teboulle (2003); Rakhlin (2008))

$$\langle a^* - d, x\rangle \leq D_{\mathcal{R}}(d, c) - D_{\mathcal{R}}(d, a^*) - D_{\mathcal{R}}(a^*, c) \ . \tag{16}$$

This yields

$$\langle f_t - g_{t+1}, M_t\rangle \leq \frac{1}{\eta}\left(D_{\mathcal{R}}(g_{t+1}, g_t) - D_{\mathcal{R}}(g_{t+1}, f_t) - D_{\mathcal{R}}(f_t, g_t)\right) \tag{17}$$

and

$$\langle g_{t+1} - f^*, x_t\rangle \leq \frac{1}{\eta}\left(D_{\mathcal{R}}(f^*, g_t) - D_{\mathcal{R}}(f^*, g_{t+1}) - D_{\mathcal{R}}(g_{t+1}, g_t)\right). \tag{18}$$

Using Equations (15), (18) and (17) in Equation (14) we conclude that

$$\begin{aligned}\langle f_t - f^*, x_t\rangle &\leq \frac{\eta}{2}\|x_t - M_t\|^2_* + \frac{1}{2\eta}\|f_t - g_{t+1}\|^2 \\ &+ \frac{1}{\eta}\left(D_{\mathcal{R}}(g_{t+1}, g_t) - D_{\mathcal{R}}(g_{t+1}, f_t) - D_{\mathcal{R}}(f_t, g_t)\right) \\ &+ \frac{1}{\eta}\left(D_{\mathcal{R}}(f^*, g_t) - D_{\mathcal{R}}(f^*, g_{t+1}) - D_{\mathcal{R}}(g_{t+1}, g_t)\right)) \\ &\leq \frac{\eta}{2}\|x_t - M_t\|^2_* + \frac{1}{2\eta}\|f_t - g_{t+1}\|^2 + \frac{1}{\eta}\left(D_{\mathcal{R}}(f^*, g_t) - D_{\mathcal{R}}(f^*, g_{t+1}) - D_{\mathcal{R}}(g_{t+1}, f_t)\right)\end{aligned}$$

By strong convexity of $\mathcal{R}$, $D_{\mathcal{R}}(g_{t+1}, f_t) \geq \frac{1}{2}\|g_{t+1} - f_t\|^2$ and thus

$$\langle f_t - f^*, x_t\rangle \leq \frac{\eta}{2}\|x_t - M_t\|_*^2 + \frac{1}{\eta}\left(D_{\mathcal{R}}(f^*, g_t) - D_{\mathcal{R}}(f^*, g_{t+1})\right)$$

Summing over $t = 1, \dots, T$ yields, for any $f^* \in \mathcal{F}$,

$$\sum_{t=1}^{T}\langle f_t - f^*, x_t\rangle \leq \frac{\eta}{2}\sum_{t=1}^{T}\|x_t - M_t\|_*^2 + \frac{R_{\max}^2}{\eta}$$

where $R_{\max}^2 = \max_{f\in\mathcal{F}}\mathcal{R}(f) - \min_{f\in\mathcal{F}}\mathcal{R}(f)$. ■

Proof [**Proof of Lemma 4**] The proof closely follows the proof of Lemma 3 and together with the technique of Abernethy and Rakhlin (2009). For the purposes of analysis, let g_{t+1} be a projected point at every step (that is, normalized). Then we have the closed form solution for f_t and g_{t+1}:

$$g_{t+1}(i) = \frac{\exp\{-\eta\sum_{s=1}^{t}x_s(i)\}}{\sum_{j=1}^{d}\exp\{-\eta\sum_{s=1}^{t}x_s(j)\}} \text{ and } f_t(i) = \frac{\exp\{-\eta\sum_{s=1}^{t-1}x_s(i) - \eta M_t(i)\}}{\sum_{j=1}^{d}\exp\{-\eta\sum_{s=1}^{t-1}x_s(j) - \eta M_t(j)\}}$$

Hence,

$$\begin{aligned}\frac{g_{t+1}(i)}{f_t(i)} &= \frac{\exp\{-\eta\sum_{s=1}^{t}x_s(i)\}}{\exp\{-\eta\sum_{s=1}^{t-1}x_s(i) - \eta M_t(i)\}}\frac{\sum_{j=1}^{d}\exp\{-\eta\sum_{s=1}^{t-1}x_s(j) - \eta M_t(j)\}}{\sum_{j=1}^{d}\exp\{-\eta\sum_{s=1}^{t}x_s(j)\}} \\ &= \exp\{-\eta(x_t(i) - M_t(i))\}\frac{\sum_{j=1}^{d}\exp\{-\eta\sum_{s=1}^{t-1}x_s(j) - \eta M_t(j)\}}{\sum_{j=1}^{d}\exp\{-\eta\sum_{s=1}^{t}x_s(j)\}\exp\{-\eta(x_t(i) - M_t(i))\}} \\ &= \frac{\exp\{-\eta(x_t(i) - M_t(i))\}}{\sum_{j=1}^{d}f_t(j)\exp\{-\eta(x_t(i) - M_t(i))\}} \end{aligned} \tag{19}$$

For any $f^* \in \mathcal{F}$,

$$\langle f_t - f^*, x_t\rangle = \langle f_t - g_{t+1}, x_t - M_t\rangle + \langle f_t - g_{t+1}, M_t\rangle + \langle g_{t+1} - f^*, x_t\rangle \tag{20}$$

First observe that

$$\langle f_t - g_{t+1}, x_t - M_t\rangle \leq \|f_t - g_{t+1}\|_t\|x_t - M_t\|_t^* \ . \tag{21}$$

Now, since $\nabla^2\mathcal{R}$ is diagonal,

$$\|f_t - g_{t+1}\|_t^2 = \sum_{i=1}^{d}(f_t(i) - g_{t+1}(i))^2/f_t(i) = -1 + \sum_{i=1}^{d}f_t(i)(g_{t+1}(i)/f_t(i))^2$$

using the fact that both f_t and g_{t+1} are probability distributions. In view of (19),

$$\|f_t - g_{t+1}\|_t^2 = -1 + \mathbb{E}\left(\frac{\exp\{-Z\}}{\mathbb{E}\exp\{-Z\}}\right)^2$$

where Z is defined as a random variable taking on values $\eta(x_t(i) - M_t(i))$ with probability $f_t(i)$. Then, if almost surely $\mathbb{E}X - X \le a/2$,

$$\mathbb{E}\left(\frac{\exp\{-Z\}}{\mathbb{E}\exp\{-Z\}}\right)^2 - 1 \le \mathbb{E}\left(\frac{\exp\{-Z\}}{\exp\{-\mathbb{E}Z\}}\right)^2 - 1 = \mathbb{E}\exp\{2(\mathbb{E}Z - Z)\} - 1 \le 4\left(\frac{e^a - a - 1}{a^2}\right)\mathrm{var}(Z)$$

since the function $(e^y - y - 1)/y^2$ is nondecreasing over reals. As long as $|\eta(x_t(i) - M_t(i))| \le 1/4$, we can guarantee that $\mathbb{E}Z - Z < 1/2$, yielding

$$\|f_t - g_{t+1}\|_t \le 2\sqrt{\mathbb{E}Z^2} = 2\sqrt{\sum_{i=1}^d f_t(i)(\eta(x_t(i) - M_t(i)))^2} = 2\eta\|x_t - M_t\|_t^*$$

Combining with (21), we have

$$\langle f_t - g_{t+1}, x_t - M_t\rangle \le 2\eta(\|x_t - M_t\|_t^*)^2 \ . \tag{22}$$

The rest similar to the proof of Lemma 3. We have

$$\langle f_t - g_{t+1}, M_t\rangle \le \frac{1}{\eta}\left(D_{\mathcal{R}}(g_{t+1}, g_t) - D_{\mathcal{R}}(g_{t+1}, f_t) - D_{\mathcal{R}}(f_t, g_t)\right) \ . \tag{23}$$

and

$$\langle g_{t+1} - f^*, x_t\rangle \le \frac{1}{\eta}\left(D_{\mathcal{R}}(f^*, g_t) - D_{\mathcal{R}}(f^*, g_{t+1}) - D_{\mathcal{R}}(g_{t+1}, g_t)\right), \tag{24}$$

We conclude that

$$\begin{aligned}
\langle f_t - f^*, x_t\rangle &\le 2\eta(\|x_t - M_t\|_t^*)^2 \\
&\quad + \frac{1}{\eta}\left(D_{\mathcal{R}}(g_{t+1}, g_t) - D_{\mathcal{R}}(g_{t+1}, f_t) - D_{\mathcal{R}}(f_t, g_t)\right) \\
&\quad + \frac{1}{\eta}\left(D_{\mathcal{R}}(f^*, g_t) - D_{\mathcal{R}}(f^*, g_{t+1}) - D_{\mathcal{R}}(g_{t+1}, g_t)\right)) \\
&\le 2\eta(\|x_t - M_t\|_t^*)^2 + \frac{1}{\eta}\left(D_{\mathcal{R}}(f^*, g_t) - D_{\mathcal{R}}(f^*, g_{t+1}) - D_{\mathcal{R}}(g_{t+1}, f_t)\right)
\end{aligned}$$

Summing over $t = 1, \ldots, T$ yields, for any $f^* \in \mathcal{F}$,

$$\sum_{t=1}^T \langle f_t - f^*, x_t\rangle \le 2\eta \sum_{t=1}^T (\|x_t - M_t\|_t^*)^2 + \frac{\log d}{\eta}$$

■

Proof [**Proof of Lemma 5**] In view of Lemma 2, for any $f^* \in \mathcal{F}$

$$\begin{aligned}
\sum_{t=1}^{T} \langle h_t, \tilde{x}_t \rangle - \sum_{t=1}^{T} \langle f^*, \tilde{x}_t \rangle &\le \eta^{-1}\mathcal{R}(f^*) + 2\eta \sum_{t=1}^{T} (\|\tilde{x}_t - M_t\|_t^*)^2 \\
&= \eta^{-1}\mathcal{R}(f^*) + 2\eta \sum_{t=1}^{T} n^2 (\langle f_t, x_t - M_t \rangle)^2 \left(\left\| \varepsilon_t \lambda_{i_t}^{1/2} \Lambda_{i_t} \right\|_t^* \right)^2 \\
&\le \eta^{-1}\mathcal{R}(f^*) + 2\eta \sum_{t=1}^{T} n^2 (\langle f_t, x_t - M_t \rangle)^2 \\
&\le \eta^{-1}\mathcal{R}(f^*) + 2\eta n^2 \sum_{t=1}^{T} \|x_t - M_t\|^2 \ .
\end{aligned}$$

where for simplicity we use the Euclidean norm and use the assumption $\|f_t\| \le 1$; any primal-dual pair of norms will work here. It is easy to verify that $\tilde{x}_t$ is an unbiased estimate of x_t and $\mathbb{E}[f]_t = h_t$. Thus, by the standard argument and the above upper bound,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \langle f_t, x_t \rangle - \sum_{t=1}^{T} \langle f^*, x_t \rangle \right] &= \mathbb{E}\left[\sum_{t=1}^{T} \langle h_t, x_t \rangle - \sum_{t=1}^{T} \langle f^*, x_t \rangle \right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \langle h_t, \tilde{x}_t \rangle - \sum_{t=1}^{T} \langle f^*, \tilde{x}_t \rangle \right] \\
&\le \eta^{-1}\mathcal{R}(f^*) + 2\eta \sum_{t=1}^{T} n^2 \mathbb{E}\left[(\langle f_t, x_t - M_t \rangle)^2 \right] \\
&\le \eta^{-1}\mathcal{R}(f^*) + 2\eta n^2 \sum_{t=1}^{T} \mathbb{E}\left[\|x_t - M_t\|^2 \right] \ .
\end{aligned}$$

The second statement follows immediately. ■

Proof [**Proof of Lemma 6**] First note that by Lemma 3 we have that for the M_t chosen in the algorithm,

$$\begin{aligned}
\sum_{t=1}^{T} \langle f_t, x_t \rangle - \sum_{t=1}^{T} \langle f^*, x_t \rangle &\le \eta^{-1} R_{\max}^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|x_t - M_t\|_*^2 \\
&\le \eta^{-1} R_{\max}^2 + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{\pi \in \Pi} q_t(\pi) \|x_t - M_t^\pi\|_*^2 \qquad \text{(Jensen's Inequality)} \\
&\le \eta^{-1} R_{\max}^2 + \frac{\eta}{2} \left(\frac{4e}{e-1} \right) \left(\inf_{\pi \in \Pi} \sum_{t=1}^{T} \|x_t - M_t^\pi\|_*^2 + \log |\Pi| \right)
\end{aligned}$$

where the last step is due to Corollary 2.3 of Cesa-Bianchi and Lugosi (2006). Indeed, the updates for q_t's are exactly the experts algorithm with pointwise loss at each round t for expert $\pi \in \Pi$ given by $\|M_t^\pi - x_t\|_*^2$. Also as each $M_t^\pi \in \mathcal{X}$ the unit ball of dual norm, we can conclude that $\|M_t^\pi - x_t\|_*^2 \le 4$ which is why we have a scaling by factor 4. Simplifying leads to the bound in the lemma. ■

Proof [**Proof of Lemma 7**] In view of Lemma 2, for any $f^* \in \mathcal{F}$

$$\begin{aligned}\sum_{t=1}^{T}\langle h_t, \tilde{x}_t\rangle - \sum_{t=1}^{T}\langle f^*, \tilde{x}_t\rangle &\leq \eta^{-1}\mathcal{R}(f^*) + 2\eta\sum_{t=1}^{T}(\|\tilde{x}_t - M_t\|_t^*)^2 \\ &= \eta^{-1}\mathcal{R}(f^*) + 2\eta\sum_{t=1}^{T} n^2(\langle f_t, x_t - M_t\rangle)^2\left(\left\|\varepsilon_t\lambda_{i_t}^{1/2}\Lambda_{i_t}\right\|_t^*\right)^2 \\ &\leq \eta^{-1}\mathcal{R}(f^*) + 2\eta n^2\sum_{t=1}^{T}(\langle f_t, x_t - M_t\rangle)^2\end{aligned}$$

It is easy to verify that $\tilde{x}_t$ is an unbiased estimate of x_t and $\mathbb{E}[f]_t = h_t$. Thus, by the standard argument and the above upper bound we get,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\langle f_t, x_t\rangle - \sum_{t=1}^{T}\langle f^*, x_t\rangle\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\langle h_t, x_t\rangle - \sum_{t=1}^{T}\langle f^*, x_t\rangle\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T}\langle h_t, \tilde{x}_t\rangle - \sum_{t=1}^{T}\langle f^*, \tilde{x}_t\rangle\right] \\ &\leq \eta^{-1}\mathcal{R}(f^*) + 2\eta n^2\mathbb{E}\left[\sum_{t=1}^{T}(\langle f_t, x_t - M_t\rangle)^2\right]\end{aligned}$$

This proves the first inequality of the Lemma. Now by Jensen's inequality, the above bound can be simplified as:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\langle f_t, x_t\rangle - \sum_{t=1}^{T}\langle f^*, x_t\rangle\right] &\leq \eta^{-1}\mathcal{R}(f^*) + 2\eta n^2\mathbb{E}\left[\sum_{t=1}^{T}(\langle f_t, x_t - M_t\rangle)^2\right] \\ &\leq \eta^{-1}\mathcal{R}(f^*) + 2\eta n^2\mathbb{E}\left[\sum_{t=1}^{T}\sum_{\pi\in\Pi} q_t(\pi)(\langle f_t, x_t - M_t^\pi\rangle)^2\right] \\ &\leq \eta^{-1}\mathcal{R}(f^*) + 8\eta n^2\left(\frac{e}{e-1}\right)\left(\mathbb{E}\inf_{\pi\in\Pi}\sum_{t=1}^{T}(\langle f_t, x_t - M_t^\pi\rangle)^2 + \log|\Pi|\right).\end{aligned}$$

where the last step is due to Corollary 2.3 of Cesa-Bianchi and Lugosi (2006). Indeed, the updates for q_t's are exactly the experts algorithm with point-wise loss at each round t for expert $\pi \in \Pi$ given by $(\langle f_t, x_t - M_t^\pi\rangle)^2$. Also as each $M_t^\pi \in \mathcal{X}$ the unit ball of dual norm, hence we can conclude that $(\langle f_t, x_t - M_t^\pi\rangle)^2 \leq 4$ which is why we have a scaling by factor 4. Further since $\|f_t\| \leq 1$ we can conclude that :

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\langle f_t, x_t\rangle - \sum_{t=1}^{T}\langle f^*, x_t\rangle\right] &\leq \eta^{-1}\mathcal{R}(f^*) + 8\eta n^2\left(\frac{e}{e-1}\right)\left(\mathbb{E}\inf_{\pi\in\Pi}\sum_{t=1}^{T}\|x_t - M_t^\pi\|^2 + \log|\Pi|\right) \\ &\leq \eta^{-1}\mathcal{R}(f^*) + 13\eta n^2\left(\mathbb{E}\inf_{\pi\in\Pi}\sum_{t=1}^{T}\|x_t - M_t^\pi\|^2 + \log|\Pi|\right).\end{aligned}$$

This concludes the proof. ■

Proof [**Proof of Lemma 8**] First note that by Lemma 3, since $M_t^{\pi_t}$ is the predictable process we use, we have deterministically that,

$$\sum_{t=1}^{T}\langle f_t, x_t\rangle - \sum_{t=1}^{T}\langle f^*, x_t\rangle \le \eta^{-1}R_{\max}^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|x_t - M_t^{\pi_t}\|_*^2$$

Hence we can conclude that expected regret is bounded as :

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle f_t, x_t\rangle - \sum_{t=1}^{T}\langle f^*, x_t\rangle\right] \le \eta^{-1}R_{\max}^2 + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\|x_t - M_t^{\pi_t}\|_*^2\right] \tag{25}$$

This proves the first inequality in the lemma. However note that the update for q_t's is using SCRiBLe for multiarmed bandit algorithm where the pointwise loss for any $\pi \in \Pi$ at round t given by $\|x_t - M_t^{\pi}\|_*^2$. Also note that maximal value of loss is bounded by $\max_{M_t, x_t}\|x_t - M_t^{\pi}\|_* \le 4$. Hence, using Lemma 11 with $s = 4$ and step size $1/32|\Pi|^2$, we conclude that

$$\mathbb{E}\left[\sum_{t=1}^{T}\|x_t - M_t^{\pi_t}\|_*^2\right] \le 2\inf_{\pi\in\Pi}\sum_{t=1}^{T}\|x_t - M_t^{\pi}\|_*^2 + 64|\Pi|^3\log(T|\Pi|)$$

Using this in Equation (25) we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle f_t, x_t\rangle - \sum_{t=1}^{T}\langle f^*, x_t\rangle\right] \le \eta^{-1}R_{\max}^2 + \eta\left(\inf_{\pi\in\Pi}\sum_{t=1}^{T}\|x_t - M_t^{\pi}\|_*^2 + 32|\Pi|^3\log(T|\Pi|)\right)$$

■

Proof [**Proof of Lemma 9**] In view of Lemma 2, for any $f^* \in \mathcal{F}$

$$\begin{aligned}
\sum_{t=1}^{T}\langle h_t, \tilde{x}_t\rangle - \sum_{t=1}^{T}\langle f^*, \tilde{x}_t\rangle &\le \eta^{-1}\mathcal{R}(f^*) + 2\eta\sum_{t=1}^{T}(\|\tilde{x}_t - M_t^{\pi_t}\|_t^*)^2 \\
&= \eta^{-1}\mathcal{R}(f^*) + 2\eta\sum_{t=1}^{T} n^2(\langle f_t, x_t - M_t^{\pi_t}\rangle)^2\left(\left\|\varepsilon_t\lambda_{i_t}^{1/2}\Lambda_{i_t}\right\|_t^*\right)^2 \\
&\le \eta^{-1}\mathcal{R}(f^*) + 2\eta n^2\sum_{t=1}^{T}(\langle f_t, x_t - M_t^{\pi_t}\rangle)^2 \ .
\end{aligned}$$

We can bound expected regret of the algorithm as:

$$\begin{aligned}
\mathbb{E}_{\pi_{1:T}, i_{1:T}}\left[\sum_{t=1}^{T}\langle f_t, x_t\rangle - \sum_{t=1}^{T}\langle f^*, x_t\rangle\right] &= \sum_{t=1}^{T}\mathbb{E}_{i_{1:t-1}, \pi_{1:t}}\left[\langle h_t, x_t\rangle\right] - \sum_{t=1}^{T}\langle f^*, x_t\rangle \\
&= \sum_{t=1}^{T}\mathbb{E}_{i_{1:t}, \pi_{1:t}}\left[\langle h_t, \tilde{x}_t\rangle\right] - \sum_{t=1}^{T}\mathbb{E}_{i_t}\left[\langle f^*, \tilde{x}_t\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\langle h_t, \tilde{x}_t\rangle - \sum_{t=1}^{T}\langle f^*, \tilde{x}_t\rangle\right] \\
&\le \eta^{-1}\mathcal{R}(f^*) + 2\eta n^2\mathbb{E}\left[\sum_{t=1}^{T}(\langle f_t, x_t - M_t^{\pi_t}\rangle)^2\right]
\end{aligned} \tag{26}$$

This gives the first inequality of the Lemma. However note that the update for q_t's the distribution over set Π is obtained by running the SCRiBLe for multi-armed bandit algorithm where pointwise loss for any $\pi \in \Pi$ at round t given by $(\langle f_t, x_t - M_t^\pi \rangle)^2$. Also note that maximal value of loss is bounded by 4. Hence using Lemma 11 with $s = 4$ and step size $1/32|\Pi|^2$ we conclude by the regret bound in that lemma that

$$\mathbb{E}\left[\sum_{t=1}^{T}(\langle f_t, x_t - M_t^{\pi_t}\rangle)^2\right] \le 2\mathbb{E}\left[\inf_{\pi\in\Pi}\sum_{t=1}^{T}(\langle f_t, x_t - M_t^{\pi}\rangle)^2 + 64|\Pi|^3\log(T|\Pi|)\right]$$

Plugging this back in Equation (26) we conclude that

$$\begin{aligned}\mathbb{E}\left[\mathbf{Reg}_T\right] &\le \eta^{-1}\mathcal{R}(f^*) + 4\eta n^2\left(\mathbb{E}\left[\inf_{\pi\in\Pi}\sum_{t=1}^{T}(\langle f_t, x_t - M_t^{\pi}\rangle)^2\right] + 32|\Pi|^3\log(T|\Pi|)\right)\\ &\le \eta^{-1}\mathcal{R}(f^*) + 4\eta n^2\left(\mathbb{E}\left[\inf_{\pi\in\Pi}\sum_{t=1}^{T}\|x_t - M_t^{\pi}\|^2\right] + 32|\Pi|^3\log(T|\Pi|)\right)\ .\end{aligned}$$

■

Proof [**Proof of Lemma 10**] In view of Lemma 2, for any $f^* \in \mathcal{F}$

$$\begin{aligned}\sum_{t=1}^{T}\langle h_t, \tilde{x}_t\rangle - \sum_{t=1}^{T}\langle f^*, \tilde{x}_t\rangle &\le \eta^{-1}\mathcal{R}(f^*) + 2\eta\sum_{t=1}^{T}(\|\tilde{x}_t\|_t^*)^2\\ &= \eta^{-1}\mathcal{R}(f^*) + 2\eta\sum_{t=1}^{T}n^2(\langle f_t, x_t\rangle)^2\left(\left\|\varepsilon_t\lambda_{i_t}^{1/2}\Lambda_{i_t}\right\|_t^*\right)^2\\ &\le \eta^{-1}\mathcal{R}(f^*) + 2s\ \eta n^2\sum_{t=1}^{T}\langle f_t, x_t\rangle\left(\left\|\varepsilon_t\lambda_{i_t}^{1/2}\Lambda_{i_t}\right\|_t^*\right)^2\\ &\le \eta^{-1}\mathcal{R}(f^*) + 2s\ \eta n^2\sum_{t=1}^{T}\langle f_t, x_t\rangle\ .\end{aligned}$$

It is easy to verify that $\tilde{x}_t$ is an unbiased estimate of x_t and $\mathbb{E}\left[f\right]_t = h_t$. Thus,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\langle f_t, x_t\rangle - \sum_{t=1}^{T}\langle f^*, x_t\rangle\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\langle h_t, x_t\rangle - \sum_{t=1}^{T}\langle f^*, x_t\rangle\right]\\ &= \mathbb{E}\left[\sum_{t=1}^{T}\langle h_t, \tilde{x}_t\rangle - \sum_{t=1}^{T}\langle f^*, \tilde{x}_t\rangle\right]\\ &\le \eta^{-1}\mathcal{R}(f^*) + 2s\ \eta n^2\mathbb{E}\left[\sum_{t=1}^{T}\langle f_t, x_t\rangle\right]\ .\end{aligned}$$

Hence we can conclude that

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle f_t, x_t\rangle\right] \le \frac{1}{1-(2sn^2)\eta}\left(\inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\langle f, x_t\rangle + \eta^{-1}\mathcal{R}(f^*)\right)$$

■

Proof [**Proof of Lemma 11**] We are interested in solving the multi-armed bandit problem using the self-concordant barrier method so we can get a regret bound in terms of the loss of the optimal arm. We do this in two steps, first we provide an algorithm for linear bandit problem over the simplex. That is we provide an algorithm for the case when learner plays on each round $q_t \in \Delta([d])$, adversary plays loss vector $x_t \in [0,s]^d$ and learner observes $\langle q_t, x_t\rangle$ at the end of the round. Next we show that this bandit algorithm over the simplex can be converted into a multi-armed bandit algorithm. To this end let us first develop a linear bandit algorithm over the simplex based on self-concordant barrier algorithm (SCRiBLe).

Bandit algorithm over simplex: Note that one can rewrite the loss of any $q \in \Delta([d])$ over any $x \in [0,s]^d$ as

$$\begin{aligned}\langle q, x\rangle &= \langle q[1:d-1], x[1:d-1]\rangle + (1 - \langle q[1:d-1], \mathbf{1}\rangle)x[d] \\ &= \langle q[1:d-1], x[1:d-1] - \mathbf{1}x[d]\rangle + x[d] \\ &= \langle (q[1:d-1], 1), (x[1:d-1] - \mathbf{1}x[d], x[d])\rangle\end{aligned}$$

Since the above we have for any distribution over the d arms q, and any loss vector x, we see that solving the linear bandit problem where learner picks from simplex and adversary picks from $[0,s]^d$ is equivalent to the linear bandit game where learner picks vectors from set $\mathcal{F}'$ and adversary picks vectors from set $\mathcal{X}'$ where

$$\mathcal{F}' = \left\{(f,1) : f \in \mathbb{R}^{d-1} \text{ s.t. } \forall i \in [d-1], f[i] \geq 0, \sum_{i=1}^{d-1} f[i] \leq 1\right\}$$

and $\mathcal{X}' = \{(x[1:d-1] - \mathbf{1}x[d], x[d]) : x \in \mathcal{X}\}$. Now we claim that the function $\mathcal{R}(f) = -\sum_{i=1}^{d-1}\log(f[i]) - \log(1 - \sum_{i=1}^{d-1} f[i])$ is a self-concordant barrier of the set $\mathcal{F}'$. To see this first note that the function $\tilde{\mathcal{R}}(f[1:d-1]) = -\sum_{i=1}^{d-1}\log(f[i]) - \log(1 - \sum_{i=1}^{d-1} f[i])$ is a self-concordant barrier on the set $\{f \in \mathbb{R}^{d-1} : \forall i \in [d-1] f[i] \geq 0, \sum_{i=1}^{d-1} f[i] \leq 1\}$. Now since the function $\mathcal{R}$ is simply the same as the function $\tilde{\mathcal{R}}$ applied only on the first $d-1$ coordinates of the input it is easy to see that $\mathcal{R}$ is a self concordant barrier on $\mathcal{F}'$. Hence using Lemma 10 we can conclude that for the SCRiBLe algorithm with this reduction with any choice of $\eta > 0$ and any $q^* \in \Delta([d])$,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\langle q_t, x_t\rangle\right] &\leq \frac{1}{1-(2sd^2)\eta}\left(\sum_{t=1}^{T}\langle q^*, x_t\rangle + d\eta^{-1}\max_{i\in[d]}\log(1/q^*[i])\right) \\ &\leq \frac{1}{1-(2sd^2)\eta}\left(\inf_{q\in\Delta([d])}\sum_{t=1}^{T}\langle q, x_t\rangle + 1 + d\eta^{-1}\log(dT)\right) \\ &= \frac{1}{1-(2sd^2)\eta}\left(\inf_{j\in[d]}\sum_{t=1}^{T}\langle e_j, x_t\rangle + 1 + d\eta^{-1}\log(dT)\right) \qquad (27)\end{aligned}$$

where the last step obtained by picking $q^* = (1-1/T)e_{j^*} + \sum_{i\neq j^*}(1/(d-1)T)e_i$ with $j^* = \operatorname{argmin}_{j\in[d]} \sum_{t=1}^{T}\langle e_j, x_t\rangle$.

Thus we have a linear bandit algorithm over the simplex with the bound given in Equation (27). Now we claim that this algorithm can be used for solving multi-armed bandit problem.

Using linear bandit algorithm over simplex for multi-armed bandit problem: We claim that the algorithm we have developed for the simplex case can be used for the multi-armed bandit problem. To see this note first that for any choice of $q_1, \ldots, q_T \in \Delta([d])$ and any choice of $x_1, \ldots, x_T$,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\langle q_t, x_t\rangle\right] - \inf_{q\in\Delta([d])}\langle q, x_t\rangle &= \mathbb{E}\left[\sum_{t=1}^{T}\mathop{\mathbb{E}}_{j_t\sim q_t}\left[\langle e_{j_t}, x_t\rangle\right]\right] - \inf_{i\in[d]}\langle e_i, x_t\rangle \\ &= \mathbb{E}\left[\sum_{t=1}^{T}\langle e_{j_t}, x_t\rangle - \inf_{i\in[d]}\langle e_i, x_t\rangle\right]\end{aligned}$$

Hence this shows that if we have an algorithm that outputs $q_1, \ldots, q_T$ then on each round by sampling the arm to pick from q_t we get the same regret bound. However note that to run a bandit algorithm over the simplex we needed to be able to observe $\langle q_t, x_t\rangle$, while in reality we only observe $\langle e_{j_t}, x_t\rangle$. There is an easy remedy for this. Note that we needed to observe $\langle q_t, x_t\rangle$ only to produce the unbiased estimate $\tilde{x}_t := d\left(\langle q_t, x_t\rangle\right)\varepsilon_t\lambda_{i_t}^{1/2}\cdot\Lambda_{i_t}$. However, $d\left(\langle q_t, x_t\rangle\right)\varepsilon_t\lambda_{i_t}^{1/2}\cdot\Lambda_{i_t} = \mathbb{E}_{j_t\sim q_t}\left[d\left(\langle e_{j_t}, x_t\rangle\right)\varepsilon_t\lambda_{i_t}^{1/2}\cdot\Lambda_{i_t}\right]$. Hence, $d\left(\langle e_{j_t}, x_t\rangle\right)\varepsilon_t\lambda_{i_t}^{1/2}\cdot\Lambda_{i_t}$ is also an unbiased estimate of $\tilde{x}_t$ and so the algorithm can simply use $\langle e_{j_t}, x_t\rangle$ to build the estimates while enjoying the same bound in expectation. Thus, SCRiBLe for multi-armed bandit enjoys the bound

$$\mathbb{E}\left\{\sum_{t=1}^{T}\langle e_{j_t}, x_t\rangle\right\} \le \frac{1}{1-4\eta s d^2}\left(\inf_{j\in[d]}\sum_{t=1}^{T}\langle e_j, x_t\rangle + d\eta^{-1}\log(dT)\right)$$

which concludes the proof. ■