Shifting Regret, Mirror Descent, and Matrices

András György A.GYORGY@IMPERIAL.AC.UK
Dept. of Electrical and Electronic Engineering, Imperial College London, London, SW7 2BT, UK

Csaba Szepesvári SZEPESVA@UALBERTA.CA
Dept. of Computing Science, University of Alberta, Edmonton, AB, T6G 2E8 CANADA

Abstract

We consider the problem of online prediction in changing environments. In this framework the performance of a predictor is evaluated as the loss relative to an arbitrarily changing predictor, whose individual components come from a base class of predictors. Typical results in the literature consider different base classes (experts, linear predictors on the simplex, etc.) separately. Introducing an arbitrary mapping inside the mirror decent algorithm, we provide a framework that unifies and extends existing results. As an example, we prove new shifting regret bounds for matrix prediction problems.

1. Introduction

In the standard online learning framework, the goal of the forecaster is to compete with a set of static reference predictors. However, this goal is only meaningful if a static predictor can be expected to perform well on the given problem. When the environment changes over time, it makes more sense to consider dynamic, time-varying reference predictors. In this paper we consider the problem where the goal of the forecaster is to compete with switching predictors that can switch between elements of a base predictor class and for each prediction mimic the forecast of the actually selected base predictor.

This problem received substantial attention in both learning and information theory, resulting in several algorithms that can compete with switching predictors. Most of these algorithms are based on variants of the exponentially weighted average prediction method, bearing different computational advantages depending on the base predictor class: variants of the fixed-share algorithm of Herbster & Warmuth (1998) are used when the base class is small, while variants of the transition diagram based method of Willems (1996) are applied for large base reference classes that admit efficient solutions for the static prediction problem. While the algorithms for small expert classes achieve near-optimal behavior in complexity that is linear both in the time horizon T and the number of experts N, the algorithms for large classes can typically be implemented with $O(T^2 \log N)$ or $O(T^2 d)$ complexity, where d is the dimension of the expert set when it is infinite. Computationally efficient combinations of the two methods have been proposed for large base predictor classes (Willems & Krom, 1997; Hazan & Seshadhri, 2009; György et al., 2012) whose computational complexity is almost linear in T and maintains the mild dependence on the size of the expert class, while only slightly deteriorating performance (see György et al. 2012 for a general overview of tracking algorithms.) In fact, these methods are general reduction methods that can transform any low-regret algorithm to one with low *switching* (or *tracking*) regret: here the regret scales with the complexity of the comparator sequence, measured by the number of switches. Another measure introduced by Hazan & Seshadhri (2009) considers regret over contiguous time intervals, called the adaptive regret, while recently a stronger version of the same concept was introduced by Daniely et al. (2015) (see, also, Adamskiy et al. 2012). Although strongly adaptive regret and adaptive regret are stronger measures than switching regret, the algorithms developed for these problems are essentially identical, and can be showed to perform well under all of these criteria.

A notable case when (near-) optimal performance relative to switching predictors is achievable with computational complexity that is linear both in time and the dimension of the predictors is the case of online linear and convex optimization: Here, Herbster & Warmuth (2001) unified earlier methods and combined gradient-descent type algorithms with projections, while Zinkevich (2003) showed that the mirror descent algorithm (Nemirovski & Yudin, 1998; Beck & Teboulle, 2003, e.g.) with a quadratic regularizer also enjoys favorable performance guarantees. In

To appear in the proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

these problems the complexity of the reference predictors is typically measured by some norm of the variation of its predictions, measuring how much the predictor *shifts* over time; hence the resulting bounds are usually called *shifting* bounds (when the prediction set is discrete, the complexity measure usually reduces to the number of switches). Note that the general wrapper algorithms derived for switching regret (see Hazan & Seshadhri, 2009; György et al., 2012; Adamskiy et al., 2012; Daniely et al., 2015) are not directly applicable to obtain shifting bounds. Recently, Cesa-Bianchi et al. (2012) combined the tracking results of Herbster & Warmuth (1998); Bousquet & Warmuth (2002) with the projection ideas of (Herbster & Warmuth, 2001) to obtain a projected exponential weighting scheme for linear/convex optimization that improves upon previous results. Finally, Hall & Willett (2013) included models of the temporal behavior of the optimal predictor in the mirror descent algorithm when the regularizer Bregman-divergence is bounded.

In this paper we present a unified view and analysis of algorithms derived for online learning with changing environments (including tracking, shifting, adaptive and strongly adaptive regret), and extend the results of Cesa-Bianchi et al. (2012) to cover any instantiation of the mirror descent algorithm. In particular, after the projection step in the mirror descent algorithm, we allow another, arbitrary transformation of the prediction (this can also be viewed as a generalization of the algorithm of Hall & Willett (2013), although their transformation has different semantics and, as a result, they give regret bounds of different kind). We give sufficient conditions when this twisted predictor achieves good shifting regret bounds. We extend these results by providing shifting bounds for contiguous time intervals, extending the recently introduced strongly adaptive regret notion of Daniely et al. (2015). As an example, we apply the results to prove shifting bounds for matrix prediction problems, which is the first result for the matrix case with non-stationary comparators.

2. Preliminaries

We consider the following standard set-up of online convex optimization. Let X be a finite dimensional vector space X that is equipped with an inner product $\langle \cdot, \cdot \rangle$. For simplicity, the reader may think of X as the d-dimensional Euclidean space $\mathbb{R}^d$ where $\langle x, y \rangle = \sum_{i=1}^d x_i y_i$. Online optimization is a repeated game. In each round $t = 1, 2, \ldots$ of the game, the forecaster chooses a prediction w_t from some decision set $K \subset X$, the environment chooses a loss function $\ell_t : K \to \mathbb{R}$ from a class $\mathcal{L}$ of functions mapping from $K \to \mathbb{R}$. At the end of round t, the loss function ℓ_t is revealed to the forecaster and the forecaster incurs loss $\ell_t(w_t)$. In this paper we consider online convex optimization where K is convex and compact, and the loss functions ℓ_t are convex and bounded.

The goal of the forecaster is to keep its cumulative loss

$$\widehat{L}_T = \sum_{t=1}^T \ell_t(w_t)$$

for some (or any) time horizon T as small as possible. While minimizing the loss $\widehat{L}_T$ is clearly not possible in general, we aim at comparing the loss of the forecaster to the loss of an arbitrary predictor sequence $u_1^T = (u_1, \ldots, u_T) \in K^T$, defined as

$$L_T(u_1^T) = \sum_{t=1}^T \ell_t(u_t).$$

The regret of the forecaster against u_1^T is defined as

$$\mathcal{R}(u_1^T) = \widehat{L}_T - L_T(u_1^T).$$

Instead of considering the regret on the whole time interval $[1, T]$, we will consider the strongly adaptive notion of regret (Daniely et al., 2015), which bounds the regret of the algorithm on any interval $[q, s]$ for any $1 \le q \le s \le T$. More precisely, we will consider this interval regret against a changing predictor sequence $u_q^s = (u_q, \ldots, u_s)$ defined as

$$\mathcal{R}(u_q^s) = \widehat{L}_{q:s} - L_T(u_q^s)$$

where $\widehat{L}_{q:s} = \sum_{t=q}^s \ell_t(w_t)$ (note that to simplify the notation, the time interval in the regret is only denoted through the index of u_q^s). By convexity of the losses, for any $u \in K$,

$$\ell_t(w_t) - \ell_t(u) \le \langle \nabla \ell_t(w_t), w_t - u \rangle, \tag{1}$$

where $\nabla \ell_t$ denotes a subgradient of ℓ_t, hence we will focus on bounding $\langle f_t, w_t - u \rangle$. We will use the notation $f_t = \nabla \ell_t(w_t)$, and assume that f_t is bounded.

Our method is based on the mirror descent (MD) algorithm (Nemirovski & Yudin, 1998; Beck & Teboulle, 2003). To define the mirror descent algorithm, we need some extra definitions. Let $A \subset \mathbb{R}^d$ be a convex set, and we will consider competing with predictors taking values in $K \cap A$, which is assumed to be non-empty. Let $R : A \to \mathbb{R}$ be a Legendre function: R is strictly convex, its derivative, $\nabla R(u)$, exists for any $u \in A^\circ$ (A° denotes the interior of A) and $\| \nabla R(u) \| \to \infty$ as u approaches the boundary of A. The Bregman divergence $D_R : A \times A^\circ \to \mathbb{R}$ with respect to R is defined, for any $(u, v) \in A \times A^\circ$ as

$$D_R(u, v) = R(u) - R(v) - \langle \nabla R(v), u - v \rangle .$$

The dual norm $\| \cdot \|_*$ of $\| \cdot \|$ is defined as $\|u\|_* = \sup_{v \in X, \|v\| \le 1} |\langle u, v \rangle|$.

3. The twisted mirror descent algorithm

Starting at a point $w_1 \in K \cap A^\circ$, the mirror descent algorithm recursively predicts, at time $t+1$,

$$w_{t+1} = \operatorname*{argmin}_{u \in K \cap A} \left[\eta_t \langle f_t, u \rangle + D_R(u, w_t) \right]$$

where $\eta_t > 0$ (recall that $f_t = \nabla \ell_t(w_t)$ denotes a subgradient of ℓ_t at w_t). We consider a generalization of the mirror descent algorithm given in Algorithm 1. We call this algorithm the *twisted mirror descent (TMD)* algorithm. The main point is that once the standard minimization step in the mirror descent algorithm is performed, the resulting value v_t is transformed using some function $\phi_t : K \times \mathcal{L}^{t-1} \to A^\circ$ to get the final prediction $w_t = \phi_t(v_t, \ell_1, \ldots, \ell_{t-1})$. In what follows we will omit the notation showing the dependence of ϕ_t on the loss functions, and will simply write $\phi_t(v_t)$. Obviously, when ϕ_t is identity, the TMD algorithm becomes the standard mirror descent algorithm. Several other algorithms can also be put into this form (Herbster & Warmuth, 2001; Bousquet & Warmuth, 2002; Cesa-Bianchi et al., 2012; Hall & Willett, 2013). In particular, Hall & Willett (2013) propose a variant where ϕ_t is time-invariant and depends on v_t only. We will discuss specific instances later. In the analysis of the algorithm we will also use the unconstrained minimum

$$\tilde{v}_{t+1} = \operatorname*{argmin}_{u \in A} \left[\eta_t \langle \nabla \ell_t(w_t), u \rangle + D_R(u, w_t) \right].$$

It is well-known that, due to our assumptions, both v_t and $\tilde{v}_{t+1}$ are unique and belong to A°, and $v_t = \operatorname{argmin}_{v \in K \cap A} D_R(v, \tilde{v}_{t+1})$ (in practice, usually first $\tilde{v}_{t+1}$ is computed and then it is projected to $K \cap A$ to obtain v_{t+1}).

Algorithm 1 Twisted mirror descent.

1. Set $w_1 \in K \cap A^\circ$.
2. At time $t = 1, 2, \ldots$ predict w_t, and compute

$$\begin{aligned} v_{t+1} &= \operatorname*{argmin}_{u \in K \cap A} \left[\eta_t \langle \nabla \ell_t(w_t), u \rangle + D_R(u, w_t) \right] \\ w_{t+1} &= \phi_{t+1}(v_{t+1}, \ell_1, \ldots, \ell_t) \end{aligned}$$

Similarly to the two standard analyses of the mirror descent algorithm, we will analyze the TMD algorithm based on the following lemma (Herbster & Warmuth, 2001):

Lemma 1. *Let $w \in A^\circ$, $g \in X$, and define $v = \operatorname{argmin}_{w' \in K \cap A} [\langle g, w' \rangle + D_R(w', w)]$ and $\tilde{v} = \operatorname{argmin}_{w' \in A} [\langle g, w' \rangle + D_R(w', w)]$. Then for any $u \in K \cap A$,*

$$\langle g, w - u \rangle \le D_R(u, w) - D_R(u, v) + D_R(v, \tilde{v}).$$

Note that if, in addition R is σ-strongly convex with respect to the norm $\|\cdot\|$, that is, $D_r(u, v) \ge \frac{\sigma}{2}\|u - v\|^2$ for all $u \in A, v \in A^\circ$, then

$$D_R(w, \tilde{v}) \le \langle g, w - \tilde{v} \rangle \le \frac{\|g\|_*^2}{2\sigma}. \tag{2}$$

This yields the so-called prox-lemma (Beck & Teboulle, 2003; Nemirovski et al., 2009)

$$\langle g, w - u \rangle \le D_R(u, w) - D_R(u, v) + \frac{\|g\|_*^2}{2\sigma}.$$

Below we will usually state the general bound of Lemma 1 in the statements, and use the nicer-looking prox-lemma bounds in the examples, except for the matrix prediction case in Section 4 where the divergence form gives qualitatively better results.

3.1. Shifting regret

In this section we generalize the methods of (Cesa-Bianchi et al., 2012) who only considered prediction on the simplex, that is, when $K = \mathcal{P}_d$ is the d-dimensional probability simplex, $R(v) = \sum_{i=1}^d v_i \log v_i - v_i$ for any vector $v = (v_1, \ldots, v_d) \in K$, and the resulting Bregman divergence is $D_R(v, v') = \sum_{i=1}^d v_i \log \frac{v_i}{v'_i}$ when $v' = (v'_1, \ldots, v'_d) \in K$. The results show that the regret of the algorithm relative to u_1^T scales with the speed of change of u_1^T.

We start with a few simple reformulations of Lemma 1 that helps writing telescoping terms and define meaningful mappings ϕ_t. The first result is a generalization of the decomposition given by (Cesa-Bianchi et al., 2012) for the simplex.

Lemma 2. *Assuming A contains the 0 vector, the following bound holds for the TMD algorithm for any $t = 1, 2, \ldots,$ and any $u_t, u_{t+1} \in K \cap A$:*

$$\begin{aligned} &\langle f_t, w_t - u_t \rangle \\ &\le \frac{1}{\eta_t} \Big(D_R(u_t, w_t) - D_R(u_{t+1}, w_{t+1}) + R(u_{t+1}) - R(u_t) \\ &\quad + D_R(0, w_{t+1}) - D_R(0, v_{t+1}) \\ &\quad + \langle \nabla R(v_{t+1}) - \nabla R(w_{t+1}), u_t \rangle \\ &\quad + \langle \nabla R(w_{t+1}), u_t - u_{t+1} \rangle + D_R(w_t, \tilde{v}_{t+1}) \Big) \end{aligned}$$

Proof. The lemma is an easy application of Lemma 1 for the TMD algorithm with $g = \eta_t f_t$, $w = w_t$, followed by the decomposition

$$\begin{aligned} &D(u_t, w_t) - D(u_t, v_{t+1}) \\ &= D(u_t, w_t) - D(u_{t+1}, w_{t+1}) + D(u_{t+1}, w_{t+1}) - D(u_t, v_{t+1}) \end{aligned}$$

and some algebra:

$$\begin{aligned}
&D(u_{t+1}, w_{t+1}) - D(u_t, v_{t+1}) \\
&\quad = \quad R(u_{t+1}) - R(w_{t+1}) - \langle \nabla R(w_{t+1}), u_{t+1} - w_{t+1} \rangle \\
&\qquad -R(u_t) + R(v_{t+1}) + \langle \nabla R(v_{t+1}), u_t - v_{t+1} \rangle \\
&\quad = \quad R(u_{t+1}) - R(u_t) + \langle \nabla R(v_{t+1}) \\
&\qquad - \nabla R(w_{t+1}), u_t \rangle + \langle \nabla R(w_{t+1}), u_t - u_{t+1} \rangle \\
&\qquad +R(0) - R(w_{t+1}) - \langle \nabla R(w_{t+1}), -w_{t+1} \rangle \\
&\qquad -R(0) + R(v_{t+1}) + \langle \nabla R(v_{t+1}), -v_{t+1} \rangle.
\end{aligned}$$

□

To make use of the above result, one needs to define the mappings ϕ_t such that the following three terms be small

- $D_R(0, w_{t+1}) - D_R(0, v_{t+1})$;
- $\nabla R(v_{t+1}) - \nabla R(w_{t+1})$;
- $\nabla R(w_{t+1})$.

The following theorem shows that controlling these quantities by an appropriate choice of ϕ_t indeed results in a meaningful bound.

Theorem 3. *Assume that, for all t, TMD is run with $\eta_t = \eta > 0$ and with a choice of the mappings ϕ_t guaranteeing*

$$\begin{aligned}
D_R(0, w_{t+1}) - D_R(0, v_{t+1}) &\leq L_t \\
\sup_{u \in K \cap A \setminus \{0\}} \frac{\langle \nabla R(v_{t+1}) - \nabla R(w_{t+1}), u \rangle}{\|u\|} &\leq M_t \\
\| \nabla R(w_{t+1}) \|_* &\leq N_t
\end{aligned}$$

for some $L_t, M_t, N_t \in \mathbb{R}$. Then, for any interval $[q, s] \subset [1, T]$, the regret $\mathcal{R}(u_q^s)$ is bounded from above by

$$\begin{aligned}
&\frac{1}{\eta}\Big(D_R(u_q, w_q) - D_R(u_{s+1}, w_{s+1}) + R(u_{s+1}) - R(u_q) \\
&\quad + \sum_{t=q}^{s} \left(L_t + M_t \|u_t\| + N_t \|u_t - u_{t+1}\| \right) \\
&\quad + \sum_{t=q}^{s} D_R(w_t, \tilde{v}_{t+1}) \Big).
\end{aligned} \tag{3}$$

Proof. The Cauchy-Schwartz inequality implies that the second inner product in Lemma 2 can be bounded as

$$\langle \nabla R(w_{t+1}), u_t - u_{t+1} \rangle \leq \| \nabla R(w_{t+1}) \|_* \|u_t - u_{t+1}\|,$$

while the conditions on ϕ_t imply

$$\langle \nabla R(v_{t+1}) - \nabla R(w_{t+1}), u_t \rangle \leq M_t \|u_t\|.$$

Applying these results in Lemma 2 shows

$$\begin{aligned}
&\langle f_t, w_t - u_t \rangle \\
&\leq \frac{1}{\eta}\Big(D_R(u_t, w_t) - D_R(u_{t+1}, w_{t+1}) + R(u_{t+1}) - R(u_t) \\
&\quad + L_t + M_t \|u_t\| + N_t \|u_t - u_{t+1}\| + D_R(w_t, \tilde{v}_{t+1}) \Big)
\end{aligned}$$

Summing this inequality for all $q \leq t \leq s$, the statement of the theorem follows immediately by (1). □

The above result is a typical example of a regret bound with respect to a time-varying reference sequence u_1^T, as it depends on the variations of u_q^s: assuming $L_t = L$ and $M_t = M$ for all T, the dependence is on the total norm $\sum_{t=q}^{s} \|u_t\|$ and the variation $D_V(u_q^s) = \sum_{t=q}^{s-1} \|u_t - u_{t+1}\|$ of the sequence (note that u_{s+1} can always be chosen to be equal to u_s when we express the bound in the theorem).

Example 4. The simplest example when TMD, and actually the pure MD, works is the case when we use a p-norm regularizer with $p \in (1, 2]$ over a ball, that is, $A = X = \mathbb{R}^d$, $R(u) = \frac{1}{2}\|u\|_p^2$, $K = \{u \in \mathbb{R}^d : \|u\|_p \leq D/2\}$. In this case the dual norm is the q-norm with $q = p/(p-1)$. Furthermore, D_R is known to be $(p-1)$-strongly convex with respect to the p-norm. Thus, assuming $\|f_t\|_q \leq G$, (2) implies that $D_R(w_t, \tilde{v}_{t+1}) \leq \eta^2 G^2 / (2(p-1))$. It is easy to see that in this setup the identity mapping $\phi_t(v) = v$ is a good choice (reducing TMD to MD), giving $L_t = M_t = 0$, and $N_t = D/2$ since $\| \nabla R(u) \|_q = \|u\|_p$. Selecting $w_1 = 0$, we have $D_R(u, w_1) = R(u_1) \leq D^2/8$ for any $u \in K$, and setting $u_{T+1} = u_T$ yields $R(u_{T+1}) \leq D^2/8$, giving the following regret bound

$$\mathcal{R}(u_1^T) \leq \frac{D^2 + 2D \cdot D_V(u_1^T)}{4\eta} + \frac{\eta T G^2}{2(p-1)}$$

where $D_V(u_1^T) = \sum_{t=1}^{T-1} \|u_t - u_{t+1}\|_p$ (for simplicity and illustrational purposes, we consider the regret only over the whole interval $[1, T]$). Optimizing η independently of $D_V(u_1^T)$ gives $\eta = \frac{D}{G}\sqrt{\frac{p-1}{2T}}$ and results in the bound

$$\mathcal{R}(u_1^T) \leq G \left(D + \frac{D_V(u_1^T)}{2} \right) \sqrt{\frac{T}{2(p-1)}}.$$

Optimizing η also as a function of an a priori known upper bound $D_V \geq D_V(u_1^T)$, we get

$$\mathcal{R}(u_1^T) \leq G \sqrt{\frac{T(D^2 + 2D \cdot D_V)}{2(p-1)}}.$$

A slightly different (sometimes improved) version of Theorem 3 can be obtained if we can give coordinatewise conditions for the gradients of R in the theorem. In the following we consider the case when $X = \mathbb{R}^d$ and all coordinates

of the subgradients of R are non-positive, and the predictors are taken from the non-negative orthant. In what follows we make $\nabla_i R$ denote the ith coordinate of the subgradient of R and $D^+_{TV}(u,v) = \sum_{i=1}^d \max\{u_i - v_i, 0\}$ for all $u, v \in \mathbb{R}^d$; note that when $\|u\|_1 = \|v\|_1$ then $D^+_{TV} = \frac{1}{2}\|u-v\|_1$ equals the total variation distance.

Theorem 5. *Assume $K \subset [0,\infty)^d$, and $\nabla_i R(u) \le 0$ for all $u \in K \cap A$. Suppose that, for all t, TMD is run with $\eta_t = \eta > 0$ and R that is σ-strongly convex with respect to $\|\cdot\|$, and with a choice of the mappings ϕ_t guaranteeing*

$$\begin{aligned} D_R(0, w_{t+1}) - D_R(0, v_{t+1}) &\le L_t \\ \nabla_i R(v_{t+1}) - \nabla_i R(w_{t+1}) &\le M_t \\ -\nabla_i R(w_{t+1}) &\le N_t \end{aligned}$$

for some $L_t, M_t, N_t \in \mathbb{R}$. Then, for any interval $[q,s] \subset [1,T]$, the regret $\mathcal{R}(u_q^s)$ can be bounded from above by

$$\begin{aligned} &\frac{1}{\eta}\Big(D_R(u_q, w_q) - D_R(u_{s+1}, w_{s+1}) + R(u_{s+1}) - R(u_q) \\ &\quad + \sum_{t=q}^{s}\big(L_t + M_t(\|u_t\|_1 - D^+_{TV}(u_{t+1}, u_t)) \\ &\quad + N_t D^+_{TV}(u_{t+1}, u_t)\big)\Big) + \frac{\eta}{2\sigma}\sum_{t=q}^{s}\|f_t\|_*^2. \end{aligned}$$

Proof. The proof of the theorem follows the same lines as that of Theorem 3. The slight difference is in how the inner products in Lemma 2 are bounded. We will use the fact that $\nabla_i R(v_{t+1}) \le 0$:

$$\begin{aligned} &\langle \nabla R(v_{t+1}) - \nabla R(w_{t+1}), u_t\rangle + \langle \nabla R(w_{t+1}), u_t - u_{t+1}\rangle \\ &\le \sum_{i:u_{t,i}\le u_{t+1,i}} \Big[(\nabla_i R(v_{t+1}) - \nabla_i R(w_{t+1}))u_{t,i} \\ &\qquad - \nabla_i R(w_{t+1})(u_{t+1,i} - u_{t,i})\Big] \\ &\quad + \sum_{i:u_{t,i}> u_{t+1,i}} \Big[(\nabla_i R(v_{t+1}) - \nabla_i R(w_{t+1}))u_{t,i} \\ &\qquad - \nabla_i R(w_{t+1})(u_{t+1,i} - u_{t,i}) \\ &\qquad + \nabla_i R(v_{t+1})(u_{t+1,i} - u_{t,i})\Big] \\ &= \sum_{i:u_{t,i}\le u_{t+1,i}} \Big[(\nabla_i R(v_{t+1}) - \nabla_i R(w_{t+1}))u_{t,i} \\ &\qquad - \nabla_i R(w_{t+1})(u_{t+1,i} - u_{t,i})\Big] \\ &\quad + \sum_{i:u_{t,i}> u_{t+1,i}} (\nabla_i R(v_{t+1}) - \nabla_i R(w_{t+1}))u_{t+1,i} \\ &= N_t \sum_{i:u_{t,i}\le u_{t+1,i}} \Bigg[(u_{t+1,i} - u_{t,i}) \\ &\quad + M_t\Big(\sum_{i=1}^d u_{t+1,i} - \sum_{i:u_{t,i}\le u_{t+1,i}} (u_{t+1,i} - u_{t,i})\Big)\Bigg] \\ &= N_t D^+_{TV}(u_{t+1}, u_t) + M_t(\|u\|_1 - D^+_{TV}(u_{t+1}, u_t)). \end{aligned}$$

The proof can be finished in the same way as in Theorem 3. □

Example 6. The above theorem is very useful when one works on the simplex, as in (Cesa-Bianchi et al., 2012). Then $K = \mathcal{P}_d$ is the d-dimensional probability simplex, $R(v) = \sum_{i=1}^d v_i \log v_i - v_i$ for any vector $v = (v_1, \ldots, v_d) \in K$, and the resulting Bregman divergence is $D_R(v, v') = \sum_{i=1}^d v_i \log \frac{v_i}{v'_i}$ when $v' = (v'_1, \ldots, v'_d) \in K$. Note that in this case the norm is the 1-norm and $\sigma = 1$ (by Pinsker's inequality). Then selecting ϕ_t as the fixed share update of Herbster & Warmuth (2001) satisfies the assumptions of the theorem, and gives rise to the bound in Proposition 1 in (Cesa-Bianchi et al., 2012). That is, ϕ_t is defined as

$$w_{t+1} = \phi_t(v_{t+1}) = (1-\alpha)v_{t+1} + \tfrac{\alpha}{d}\mathbf{1}$$

for some $\alpha > 0$, where $\mathbf{1}$ denotes a d-dimensional vector whose entries are all 1. Then $L_t = 0, M_t = \log\frac{1}{1-\alpha}, N_t = \log\frac{d}{\alpha}$. Let $m(u_1^T) = \sum_{t=1}^{T-1}\|u_{t+1} - u_t\|_1$. Assuming each $f_t \in [0,1]^d$ and, starting the algorithm from the uniform distribution $w_1 = \mathbf{1}/d$, the bound becomes

$$\begin{aligned} &\mathcal{R}(u_1^T) \\ &\le \frac{1}{\eta}\left(\log d + (T - m(u_1^T))\log\tfrac{1}{1-\alpha} + m(u_1^T)\log\tfrac{d}{\alpha}\right) + \frac{\eta T}{8}. \end{aligned}$$

The $1/8$ factor instead of $1/2$ in the last term can be obtained by shifting f_t to $[-1/2, 1/2]^d$, which does not change the linearized regret. This result exactly recovers the corresponding results in (Herbster & Warmuth, 2001; Cesa-Bianchi et al., 2012). The slight improvement compared to Theorem 3 is the appearence of the $-m(u_1^T)$ in multiplying $\log \frac{d}{\alpha}$.

4. Application to linear prediction over trace-bounded positive definite matrices

In this section we consider the application of the previous result to a natural online matrix-prediction problem, taken from Hazan et al. (2012), who showed that a number of matrix-valued prediction problems, such as collaborative filtering, gambling and max-cut can be reduced to this common problem. Here we show how TMD can be applied to this problem to compete with a changing sequence of matrices, thereby extending the scope of results of Hazan et al. (2012).

In order to define the problem, we need some notation. We let $\mathbb{S}$ denote the vector space of $N \times N$ real-valued symmetric matrices equipped with the inner product $\langle X, Y\rangle = \mathrm{tr}(X^\top Y)$. Further, we let $\mathbb{S}^{++} \subset \mathbb{S}$ (and $\mathbb{S}^{+} \subset \mathbb{S}$) denote the set of $N \times N$ real-valued positive definite (respectively, semi-definite) matrices.

Let τ, β be positive numbers. The competitor set is chosen to be

$$K_{\tau,\beta} = \left\{ X \in \mathbb{S}^+ \,:\, \|X\| \le \tau, X_{i,i} \le \beta, 1 \le i \le N \right\} ,$$

where $\|X\|$ is the trace-norm: $\|X\| = \|X\|_{\mathrm{tr}}$, where $\|X\|_{\mathrm{tr}} = \sum_{i=1}^N |\lambda_i(X)|$, $\lambda_i(X)$ being the ith eigenvalue of matrix X. We also introduce $K_\tau = \cup_{\beta>0} K_{\tau,\beta}$. The loss is assumed to be linear: $\ell_t(X) = \langle F_t, X\rangle$. Here, $F_t \in \mathbb{S}$ is constrained to belong to the set

$$L_\gamma = \left\{ F \in \mathbb{S} \,:\, \|F^2\|_* \le \gamma, F^2 \text{ diagonal} \right\} .$$

The (β, τ, γ) online matrix prediction problem is to compete with the best matrix from $K_{\tau,\beta}$ in hindsight given a sequence of loss matrices $F_1, \dots, F_T \in L_\gamma$.

Note that the dual norm of the trace-norm is the spectral (or operator) norm: $\|X\|_* = \max_{1\le i\le N} |\lambda_i(X)|$. Further, by duality, Hölder's inequality holds: $\langle X, Y\rangle \le \|X\|\|Y\|_*$, $X, Y \in \mathbb{S}$.

Let us now consider how TMD can be applied to this setting. Unsurprisingly, we choose the unnormalized negative entropy regularizer to instantiate TMD. To introduce this define the application of a function $f : \mathbb{R} \to \mathbb{R}$ to a symmetric matrix X as $f(X) = \sum_{i=1}^N f(\lambda_i) u_i u_i^\top$, where $X = \sum_{i=1}^N \lambda_i u_i u_i^\top$ is an eigendecomposition of X. Note that $f(X)$ is well defined.

For X positive definite, we let $R(X)$ denote the unnormalized negative entropy of X:

$$R(X) = \mathrm{tr}(X \log(X) - X).$$

It is well-known that R is a Legendre function over $A = \mathbb{S}^+$. In particular, the derivative of R exists on $A^\circ = \mathbb{S}^{++}$ and satisfies $\nabla R(X) = \log(X)$. Thus, the underlying Bregman divergence is equal to

$$D_R(X, Y) = \mathrm{tr}(X \log X - X \log Y - X + Y) .$$

For brevity, we will call $D_R(X, Y)$ the *relative entropy* of X with respect to Y.

It remains to choose the mappings $\phi_t : K_\tau \to K_\tau$. For $0 \le \alpha \le 1$, $c_\phi > 0$ to be chosen later, let $f_\alpha(\lambda) = (1-\alpha)\lambda + \alpha c_\phi / N$. With this we let

$$\phi_t(X) = f_\alpha(X) . \tag{4}$$

Let $I_{N\times N}$ denote the $N \times N$ identity matrix. From Theorem 3, we get the following result:

Theorem 7. *Choose ϕ_t as in the previous paragraph and let $\eta_t = \eta > 0$ for all $t > 0$ such that $\eta\sqrt{\gamma} \le 1$. Let $N^* = \max(\log((1-\alpha)\tau + \alpha\frac{c_\phi}{N}), |\log(\alpha\frac{c_\phi}{N})|)$, $c_\phi \ge \tau$, F_1^T, U_1^T be sequences such that $U_t \in K_{\tau,\beta}$ and $F_t \in F_\gamma$. Let $W_1 = \frac{\tau}{N} I_{N\times N}$ and W_2^T be the sequence of vectors chosen by TMD beginning from $t = 2$ when the adversary's choices are F_1^T, and let $\mathcal{R}(U_1^T) = \sum_{t=1}^T \langle F_t, W_t\rangle - \langle F_t, U_t\rangle$ be the regret of TMD against U_1^T on this sequence. Then,*

$$\begin{aligned}
&\mathcal{R}(U_1^T) \\
&\le \frac{1}{\eta}\Big\{ c_\phi \log(N) + c_\phi - \|U_1\| + \alpha c_\phi T \\
&\qquad + \log\left(\tfrac{1}{1-\alpha}\right) \sum_{t=1}^T \|U_t\| + N^* \sum_{t=1}^T \|U_t - U_{t+1}\| \Big\} \\
&\qquad + \eta \left\{ (1-\alpha)\beta + \alpha\tfrac{c_\phi}{N} \right\} \gamma T .
\end{aligned}$$

The result follows from Theorem 3. Note that Theorem 13 of Hazan et al. (2012) follows from this result (when U_t is the constant sequence). Furthermore, by appropriately tuning the parameters, one can easily obtain from this result the matrix analogues of the results discussed beforehand for the vector case, obtaining the first tracking/shifting regret result for the matrix case (which is the exact analogue of the vector case, hence left out). This clearly demonstrates the power of Theorem 3.

Proof. The result follows from Theorem 3 once we choose the appropriate values for the parameters of this theorem and verify its conditions. Let us first choose values for L_t, N_t and M_t. Fix a value for t.

The sequence (L_t) must be selected so that

$$D_R(0, W_{t+1}) - D_R(0, v_{t+1}) \quad \leq \quad L_t\,.$$

Since $D_R(0, Y) = \operatorname{tr}(Y)$, we have $D_R(0, W_{t+1}) - D_R(0, V_{t+1}) = \operatorname{tr}(W_{t+1} - V_{t+1}) = \alpha(c_\phi - \operatorname{tr}(V_{t+1})) \leq \alpha c_\phi$, thanks to $V_{t+1} \in \mathbb{S}^+$. Hence, we choose $L_t = \alpha c_\phi$.

Now, consider the condition

$$\sup_{U \in K_{\tau,\beta} \cap \mathbb{S}^{++}} \frac{\langle \nabla R(V_{t+1}) - \nabla R(W_{t+1}), U\rangle}{\|U\|} \quad \leq \quad M_t$$

(the domain of ∇R is $\mathbb{S}^{++}$). Fix some $U \in K \cap \mathbb{S}^{++}$ and let $V_{t+1} = \sum_{i=1}^N \lambda_i z_i z_i^\top$ be an eigendecomposition of V_{t+1}. By the definition of ϕ, $W_{t+1} = \sum_{i=1}^N ((1-\alpha)\lambda_i + \alpha c_\phi/N) z_i z_i^\top$. As noted earlier, $\nabla R(X) = \log(X)$. Hence,

$$\begin{aligned} Z &\doteq \nabla R(V_{t+1}) - \nabla R(W_{t+1}) \\ &= \sum_{i=1}^N \log\left(\frac{\lambda_i}{(1-\alpha)\lambda_i + \alpha c_\phi/N}\right) z_i z_i^\top \end{aligned}$$

and so

$$\langle Z, U\rangle = \sum_{i=1}^N \log\left(\frac{\lambda_i}{(1-\alpha)\lambda_i + \alpha c_\phi/N}\right) \langle z_i z_i^\top, U\rangle.$$

Now, since both $z_i z_i^\top$ and U are nonnegative definite, $\langle z_i z_i^\top, U\rangle \geq 0$. Therefore,

$$\langle Z, U\rangle \leq C \sum_{i=1}^N \langle z_i z_i^\top, U\rangle$$

where $C = \max_{1\leq j\leq N} \log\left(\frac{\lambda_j}{(1-\alpha)\lambda_j + \alpha c_\phi/N}\right)$. Therefore,

$$\begin{aligned} &\log\left(\frac{\lambda_j}{(1-\alpha)\lambda_j + \alpha c_\phi/N}\right) \\ &\leq \log\left(\frac{\lambda_j}{(1-\alpha)\lambda_j}\right) = \log\left(\frac{1}{1-\alpha}\right) \end{aligned}$$

and hence $C \leq \log(\frac{1}{1-\alpha})$. Introduce $Z' = \sum_{i=1}^N \log(\frac{1}{1-\alpha}) z_i z_i^\top$. Thus, $\langle Z, U\rangle \leq \langle Z', U\rangle$ and from Hölder's inequality we get

$$\langle Z, U\rangle \leq \|Z'\|_* \|U\| \leq \log\left(\frac{1}{1-\alpha}\right) \|U\|$$

and so we choose $M_t = \log \frac{1}{1-\alpha}$.

Let us now turn to the choice of N_t. We need to choose N_t such that

$$\|\nabla R(W_{t+1})\|_* \quad \leq \quad N_t. \tag{5}$$

We have

$$\|\nabla R(W_{t+1})\|_* = \max_{1\leq i\leq N} \left|\log\left((1-\alpha)\lambda_i + \alpha\frac{c_\phi}{N}\right)\right|\,.$$

A simple case analysis gives that this is upper bounded by $N^* = \max(\log((1-\alpha)\tau + \alpha\frac{c_\phi}{N}), |\log(\alpha\frac{c_\phi}{N})|)$, which can be chosen to be the value of N_t.

Now, let us bound $D_R(W_t, \tilde{V}_{t+1})$. For this, we use the following lemma, which can be extracted from K. Tsuda & Warmuth (2006); Arora & Kale (2007) or Hazan et al. (2012):

Lemma 8. *Let R be the negentropy regularizer, $F \in \mathbb{S}$, $\|F\|_* \leq 1$, $W \in \mathbb{S}^{++}$, $\tilde{V} = \operatorname{argmin}_{V\in\mathbb{S}^+} \langle F, V\rangle + D_R(V, W)$. Then $D_R(W, \tilde{V}) \leq \langle W, F^2\rangle$.*

Proof. Note that $\tilde{V} = \nabla R^{-1}(\nabla R(W) - F) = \exp(\log(W) - F)$. Plugging this into the definition of D_R we get

$$\begin{aligned} D_R(W, \tilde{V}) &= \operatorname{tr}(W\log W - W\log\tilde{V} - W + \tilde{V}) \\ &= \operatorname{tr}(W\log W - W(\log W - F) - W + \tilde{V}) \\ &= \operatorname{tr}(WF - W + \exp(\log(W) - F))\,. \end{aligned}$$

By the Golden-Thompson inequality, $\operatorname{tr}(\exp(\log(W) - F)) \leq \operatorname{tr} W\exp(-F)$. Now, for any $A \in \mathbb{S}$, $\|A\|_* \leq 1$, $\exp(A) \leq I_{N\times N} + A + A^2$. Further, for any $W, A, B \in \mathbb{S}^+$, $A \prec B$ implies $\langle W, A\rangle \leq \langle W, B\rangle$. Hence, $\operatorname{tr} W\exp(-F) \leq \operatorname{tr} W(I_{N\times N} - F + F^2)$. Putting the inequalities together, cancelling terms we get the claimed inequality. □

Using this lemma with $F = \eta F_t$, $W = W_t$, since $\eta\sqrt{\gamma} \leq 1$ by assumption, we get $D_R(W_t, \tilde{V}_{t+1}) \leq \eta^2\langle W_t, F_t^2\rangle$. Now, $W_t = f_\alpha(V_t)$ where $V_t \in K_{\tau,\beta}$. Then, if $V_t = \sum_i \lambda_i z_i z_i^\top$ is the eigendecomposition of V_t, using that F_t^2 is diagonal, it is not hard to see that $\langle W_t, F_t^2\rangle = (1-\alpha)\langle V_t, F_t^2\rangle + \alpha\frac{c_\phi}{N}\langle \sum_i z_i z_i^\top, F^2\rangle \leq ((1-\alpha)\beta + \alpha\frac{c_\phi}{N})\gamma$.

To finish, we choose $U_{T+1} = U_1$ and so it remains to bound $D_R(U_1, W_1)$. Let us first consider $D_R(U_1, W_1)$ for some $U_1 \in K \cap \mathbb{S}^{++}$. Let $U_1 = \sum_{i=1}^N \lambda_i z_i z_i^\top$ be the eigendecomposition of U_1. Since W_1 is the matrix of scaling all vectors by a factor of c_ϕ/N, we can write $W_1 = c_\phi/N \sum_{i=1}^N z_i z_i^\top$. Hence, $D_R(U_1, W_1) = c_\phi + \sum_{i=1}^N \lambda_i \log\frac{\lambda_i}{c_\phi/N} - \lambda_i = c_\phi + c_\phi\log(N) + \sum_{i=1}^N \lambda_i\left(\log\left(\frac{\lambda_i}{c_\phi}\right) - 1\right) \leq c_\phi\log(N) + c_\phi - \|U_1\|$, where the last inequality follows if we assume that $\tau \leq c_\phi$, so that $\lambda_i/\tau \leq 1$ and thus $\log\left(\frac{\lambda_i}{\tau}\right) - 1 \leq -1$.

Plugging in the bounds obtained into (3), we get

$$\begin{aligned}
&\mathcal{R}(U_1^T) \\
&\leq \frac{1}{\eta}\Big\{c_\phi \log(N) + c_\phi - \|U_1\| + \alpha c_\phi T \\
&\qquad + \log\left(\tfrac{1}{1-\alpha}\right)\sum_{t=1}^{T}\|U_t\| + N^*\sum_{t=1}^{T}\|U_t - U_{t+1}\|\Big\} \\
&\qquad\qquad + \left\{(1-\alpha)\beta + \alpha\tfrac{c_\phi}{N}\right\}\gamma T\,.
\end{aligned}$$

which is the desired bound. □

5. Conclusion

We presented a unifying framework for deriving mirror-descent based algorithms for online learning in changing environments. A generic result was provided that indicated how mirror descent algorithms can be modified to obtain shifting regret bounds and shifting regret bounds over intervals. As corollaries, we derived existing variants of the mirror descent algorithm (for various problems), and recovered their shifting regret bounds, as well as derived a new matrix prediction algorithm and the first shifting bound for matrix prediction problems.

Acknowledgements

This work was supported in part by the Alberta Innovates Technology Futures through the Alberta Ingenuity Centre for Machine Learning and by NSERC.

References

Adamskiy, Dmitry, Koolen, Wouter M., Chernov, Alexey V., and Vovk, Vladimir. A closer look at adaptive regret. In *Algorithmic Learning Theory - 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings*, pp. 290–304, 2012.

Arora, S. and Kale, S. A combinatorial, primal-dual approach to semidefinite programs. In *STOC*, pp. 227—236. 2007.

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

Bousquet, O. and Warmuth, M. K. Tracking a small set of experts by mixing past posteriors. *Journal of Machine Learning Research*, 3:363–396, Nov. 2002.

Cesa-Bianchi, N., Gaillard, P., Lugosi, G., and Stoltz, G. Mirror descent meets fixed share (and feels no regret). In Bartlett, P. L., Pereira, F.C.N., Burges, C.J.C., Bottou, L., and Weinberger, K.Q. (eds.), *Advances in Neural Information Processing Systems 25*, pp. 989–997. 2012.

Daniely, Amit, Gonen, Alon, and Shalev-Shwartz, Shai. Strongly adaptive online learning. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pp. 1405–1411, 2015.

György, A., Linder, T., and Lugosi, G. Efficient tracking of large classes of experts. *IEEE Transactions on Information Theory*, IT-58(11):6709–6725, Nov. 2012.

Hall, E. C. and Willett, R. M. Dynamical models and tracking regret in online convex programming. In *Proc. 20th International Conference on Machine Learning (ICML2013)*, volume 28 of *JMLR Workshop and Conference Proceedings*, Atlanta, GA, June 2013.

Hazan, E. and Seshadhri, C. Efficient learning algorithms for changing environments. In *Proc. 26th Annual International Conference on Machine Learning*, pp. 393–400. ACM, 2009.

Hazan, E., Kale, S., and Shalev-Shwartz, S. Near-optimal algorithms for online matrix prediction. In *COLT*. 2012.

Herbster, M. and Warmuth, M. K. Tracking the best expert. *Machine Learning*, 32(2):151–178, 1998.

Herbster, M. and Warmuth, M. K. Tracking the best linear predictor. *Journal of Machine Learning Research*, 1: 281–309, 2001.

K. Tsuda, G. Ratsch and Warmuth, M.K. Matrix exponentiated gradient updates for on-line learning and bregman projection. *Journal of Machine Learning Research*, 6(1): 995–1018, 2006.

Nemirovski, A. and Yudin, D. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1998.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optimization*, 4:1574–1609, 2009.

Willems, F. M. J. Coding for a binary independent piecewise-identically-distributed source. *IEEE Transactions on Information Theory*, IT-42:2210–2217, Nov. 1996.

Willems, F. M. J. and Krom, M. Live-and-die coding for binary piecewise i.i.d. sources. In *Proc. 1997 IEEE Int. Symp. Inform. Theory*, pp. 68, Ulm, Germany, June–July 1997.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proc. 20th International Conference on Machine Learning (ICML-2003)*, Washington, DC, 2003.