

Mirror Descent and Spectral Sparsification

Yunwei Ren
Princeton University
yunwei.ren@princeton.edu

May 3, 2025

Contents

1	Introduction	1
2	Mirror Descent	2
2.1	From gradient descent to mirror descent	2
2.2	Regret minimization and online mirror descent	7
2.3	Zero-sum games	9
3	Spectral Sparsification	11
3.1	From spectral sparsification to regret minimization	11
3.2	Entropy regularization and $O(d \log d)$ sparsification	13
3.3	$l_{1-1/q}$ regularization and $O(d)$ sparsification	15
	References	18

1 Introduction

The purpose of this note is to provide a short introduction to the mirror descent algorithm and its applications in spectral sparsification.

In Section 2, we will start with the gradient descent algorithm, then derive the (offline) mirror descent algorithm as an extension of gradient descent, and prove its convergence guarantee (Section 2.1). Then, in Section 2.2, we extend the discussion to the online setting, where the goal is to minimize the regret w.r.t. a fixed choice. Section 2.3 contains a brief discussion on the minimax theorem in zero-sum games and a mirror descent proof of this theorem. It provides a simple example where online mirror descent can be used to prove a seemingly irrelevant result. This section is primarily based on an optimization course taught by Yuanzhi Li at CMU in 2022 Spring and Chapter 8 of [Haz22].

In Section 3, we consider the spectral sparsification problem. That is, given $v_1, \dots, v_n \in \mathbb{R}^d$ with $n^{-1} \sum_{i=1}^n v_i v_i^\top = I_d$, we wish to find a sparse $s \in \mathbb{R}_+^n$ such that $\sum_{i=1}^n s_i v_i v_i^\top \approx I_d$. In Section 3.1, we reduce this task to a pair of regret minimization tasks, so that online mirror descent can be used. Then, we consider two instantiations of online mirror descent algorithms that lead to an $O(d \log d)$ sparsifier (Section 3.2) and an $O(d)$ sparsifier (Section 3.3). This section is based on [AZLO15]

— the paper in which the connection between spectral sparsification and mirror descent was first established.

2 Mirror Descent

2.1 From gradient descent to mirror descent

We start with the setting where we are given a fixed convex function $f : \mathbb{R}^d \rightarrow \mathbb{R}$. The task here is to find an approximate minimizer of f . We will first consider arguably the simplest iterative method — gradient descent, and then introduce mirror descent as a natural extension of gradient descent.

Formally, the gradient descent algorithm is the update rule:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t), \quad (1)$$

where $\eta > 0$ is the step size and $\mathbf{x}_0 \in \mathbb{R}^d$ is a chosen initial point. When $\eta \rightarrow 0$, we recover the (continuous-time) gradient flow $\dot{\mathbf{x}}_t = -\nabla f(\mathbf{x}_t)$, under which we have $\frac{d}{dt} f(\mathbf{x}_t) = -\|\nabla f(\mathbf{x}_t)\|^2$. In words, under gradient flow, the loss f will decrease as long as the gradient is nonzero and the decrease is proportional to $\|\nabla f(\mathbf{x}_t)\|^2$. This fact is also true for the (discrete-time) gradient descent.

Lemma 2.1 (Gradient descent lemma). *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a C^1 L -smooth¹ function (not necessarily convex) and $(\mathbf{x}_t)_t$ be gradient descent iterates (1). Then, if $\eta \leq 1/L$, we have*

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{\eta}{2} \|\nabla f(\mathbf{x}_t)\|^2.$$

Proof. Since f is L -smooth, we have

$$\begin{aligned} f(\mathbf{x}_{t+1}) &= f(\mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)) \leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), -\eta \nabla f(\mathbf{x}_t) \rangle + \frac{L}{2} \|\eta \nabla f(\mathbf{x}_t)\|^2 \\ &\leq f(\mathbf{x}_t) - \left(1 - \frac{L\eta}{2}\right) \eta \|\nabla f(\mathbf{x}_t)\|^2. \end{aligned}$$

To complete the proof, note that when $\eta < 1/L$, we have $1 - L\eta/2 \geq 1/2$. □

The gradient descent lemma implies that gradient descent can efficiently find an approximate first-order stationary point, that is, a point at which the gradient is close to 0. Moreover, as one may expect, when f is convex, we can ensure gradient descent find an approximate minimizer. To prove this claim, we will use the so-called basic mirror descent lemma.

Lemma 2.2 (Law of cosines). *For any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^d$, we have*

$$\langle \mathbf{z} - \mathbf{x}, \mathbf{y} - \mathbf{x} \rangle = \frac{1}{2} \left(\|\mathbf{y} - \mathbf{x}\|^2 + \|\mathbf{z} - \mathbf{x}\|^2 - \|\mathbf{y} - \mathbf{z}\|^2 \right).$$

Proof. Elementary geometry or calculation. □

Lemma 2.3 (Basic mirror descent lemma). *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a C^1 convex function and let $(\mathbf{x}_t)_t$ be the gradient descent iterates (1). Then, for any $\mathbf{y} \in \mathbb{R}^d$, we have*

$$f(\mathbf{x}_t) \leq f(\mathbf{y}) + \frac{1}{2\eta} \left(\|\mathbf{y} - \mathbf{x}_t\|^2 - \|\mathbf{y} - \mathbf{x}_{t+1}\|^2 + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \right).$$

¹A function f is said to be L -smooth if its gradient is L -Lipschitz w.r.t. the Euclidean norm.

Remark. As the name suggests, there will be a true mirror descent lemma, in which the Euclidean distance is replaced by the Bregman divergence, and it will be proved by the Bregman divergence version of the law of cosines.

If we pretend that $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2$ is small and choose $\mathbf{y} = \mathbf{x}_*$ to be the minimizer of f , then this lemma says $f(\mathbf{x}_t) - f(\mathbf{x}_*) \lesssim \frac{1}{2\eta} \left(\|\mathbf{x}_* - \mathbf{x}_t\|^2 - \|\mathbf{x}_* - \mathbf{x}_{t+1}\|^2 \right)$. In particular, this implies that if $f(\mathbf{x}_t)$ is still far away from $f(\mathbf{x}_*)$, then \mathbf{x}_{t+1} must be much closer to \mathbf{x}_* when compared to \mathbf{x}_t . ♣

Remark. Since f is convex, we have

$$\begin{aligned} f(\mathbf{y}) &\geq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{y} - \mathbf{x}_t \rangle \\ &= f(\mathbf{x}_t) + \frac{1}{\eta} \langle \mathbf{x}_t - \mathbf{x}_{t+1}, \mathbf{y} - \mathbf{x}_t \rangle \\ &= f(\mathbf{x}_t) - \frac{1}{2\eta} \left(\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + \|\mathbf{y} - \mathbf{x}_t\|^2 - \|\mathbf{x}_{t+1} - \mathbf{y}\|^2 \right), \end{aligned}$$

where the last line comes from the law of cosine (Lemma 2.2). Rearrange terms and we complete the proof. ♣

With the gradient descent lemma and the basic mirror descent lemma in hand, we can now estimate the convergence rate of gradient descent.

Proposition 2.4 (Convergence rate of gradient descent). *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a C^1 L -smooth convex function and $(\mathbf{x}_t)_t$ be the gradient descent iterates (1). Suppose that $\eta \leq 1/L$. Then, for any $\mathbf{x}_* \in \mathbb{R}^d$ and $T > 0$, we have*

$$\frac{1}{T} \sum_{t=0}^{T-1} f(\mathbf{x}_t) \leq f(\mathbf{x}_*) + \frac{\|\mathbf{x}_* - \mathbf{x}_0\|^2}{2\eta T} + \frac{f(\mathbf{x}_0) - f(\mathbf{x}_T)}{T}.$$

In particular, if \mathbf{x}_* is the minimizer of f , then we have

$$\frac{1}{T} \sum_{t=0}^{T-1} f(\mathbf{x}_t) \leq f(\mathbf{x}_*) + \frac{\|\mathbf{x}_* - \mathbf{x}_0\|^2}{\eta T}. \quad (2)$$

Proof. First, by the basic mirror descent lemma, we have

$$f(\mathbf{x}_t) \leq f(\mathbf{x}_*) + \frac{1}{2\eta} \left(\|\mathbf{x}_* - \mathbf{x}_t\|^2 - \|\mathbf{x}_* - \mathbf{x}_{t+1}\|^2 + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \right).$$

Sum both sides from 0 to $T - 1$, and we obtain

$$\sum_{t=0}^{T-1} f(\mathbf{x}_t) \leq T f(\mathbf{x}_*) + \frac{1}{2\eta} \left(\|\mathbf{x}_* - \mathbf{x}_0\|^2 - \|\mathbf{x}_* - \mathbf{x}_T\|^2 + \sum_{t=0}^{T-1} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \right).$$

Choose $\eta \leq 1/L$. Then, by the gradient descent lemma, we have

$$\sum_{t=0}^{T-1} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 = \eta^2 \sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \leq 2\eta \sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})) = 2\eta(f(\mathbf{x}_0) - f(\mathbf{x}_T)).$$

Thus,

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} f(\mathbf{x}_t) &\leq f(\mathbf{x}_*) + \frac{1}{2\eta T} \left(\|\mathbf{x}_* - \mathbf{x}_0\|^2 - \|\mathbf{x}_* - \mathbf{x}_T\|^2 + 2\eta(f(\mathbf{x}_0) - f(\mathbf{x}_T)) \right) \\ &\leq f(\mathbf{x}_*) + \frac{\|\mathbf{x}_* - \mathbf{x}_0\|^2}{2\eta T} + \frac{f(\mathbf{x}_0) - f(\mathbf{x}_T)}{T}. \end{aligned}$$

When \mathbf{x}_* is the minimizer, since f is L -smooth, we have

$$f(\mathbf{x}_0) - f(\mathbf{x}_T) \leq f(\mathbf{x}_0) - f(\mathbf{x}_*) \leq \frac{L}{2} \|\mathbf{x}_0 - \mathbf{x}_*\|^2.$$

Hence, we can further rewrite the above as

$$\frac{1}{T} \sum_{t=0}^{T-1} f(\mathbf{x}_t) \leq f(\mathbf{x}_*) + \frac{\|\mathbf{x}_* - \mathbf{x}_0\|^2}{2\eta T} + \frac{L}{2} \frac{\|\mathbf{x}_0 - \mathbf{x}_*\|^2}{T} \leq f(\mathbf{x}_*) + \frac{\|\mathbf{x}_* - \mathbf{x}_0\|^2}{\eta T},$$

where the second inequality comes from $\eta \leq 1/L$ □

One can see from (2) that two factors affect the convergence rate of gradient descent. The first one is the smoothness of f , which determines how large η can be, and the second is the initial distance between \mathbf{x}_* and \mathbf{x}_0 . If one could change the underlying distance, so that under the new “distance”, f is still reasonably smooth and the distance between \mathbf{x}_* and \mathbf{x}_0 is smaller, one should expect that the corresponding version of gradient descent will converge faster. This is (one of) the motivations of the mirror descent algorithm.

To properly modifying the underlying distance, we need to first introduce a different formulation of the gradient descent update. Let f be the objective and \mathbf{x}_t be the current iterate. In order to (locally) minimize f around \mathbf{x}_t , one reasonable strategy is to minimize its linear approximation $f(\mathbf{x}) \approx f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle$ while restricting \mathbf{x} from moving too far away from \mathbf{x}_t that the approximation becomes inaccurate. Namely, we can try to minimize f via²

$$\mathbf{x}_{t+1} = \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^d} \left\{ f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{x}_t\|^2 \right\}, \quad (3)$$

where $\lambda > 0$ is a parameter that controls the regularization strength. The RHS is a strongly convex function in \mathbf{x} and therefore has a unique minimizer. By setting the gradient to be 0, one can immediately see that (3) is equivalent to

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \lambda^{-1} \nabla f(\mathbf{x}_t).$$

In other words, it is equivalent to gradient descent with step size $\eta = 1/\lambda$. Moreover, in order to modify the underlying geometry in (3), it suffices to change the Euclidean distance in the regularization term to another “distance”. One family of such distances is the Bregman divergences.

Definition 2.5 (Bregman divergence). *Let $\Psi : \Omega \rightarrow \mathbb{R}$ be a C^1 strictly convex function on a convex domain Ω . The **Bregman divergence associated with Ψ** is*

$$D_\Psi(\mathbf{x}, \mathbf{y}) = \Psi(\mathbf{x}) - \Psi(\mathbf{y}) - \langle \nabla \Psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle, \quad \forall \mathbf{x}, \mathbf{y} \in \Omega.$$

²This formulation has far-reaching extensions. For example, one can even use it to define gradient flow in a metric space, where even the gradient of a function is not readily defined.

Remark. Since Ψ is strictly convex, $D_\Psi \geq 0$ with equality holds if and only if $\mathbf{x} = \mathbf{y}$. ♣

Note that D_Ψ is not symmetric in general. As a result, it is usually not a metric. Nevertheless, it covers many well-studied objects as special cases, such as the KL divergence and the usual Euclidean distance, and will turn out to be useful in the design of mirror descent algorithms. First, note that D_Ψ satisfies a generalized version of law of cosines.

Lemma 2.6 (Law of cosines for Bregman divergence). *Let $\Psi : \Omega \rightarrow \mathbb{R}$ be a C^1 strictly convex function on a convex domain Ω . For any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \Omega$, we have*

$$\langle \nabla \Psi(\mathbf{z}) - \nabla \Psi(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle = D_\Psi(\mathbf{y}, \mathbf{x}) + D_\Psi(\mathbf{x}, \mathbf{z}) - D_\Psi(\mathbf{y}, \mathbf{z}).$$

Proof. It suffices to plug in the definition of D_Ψ . □

Before we derive a more explicit formula for the mirror descent update, we need the following classical convex analysis result.

Theorem 2.7 (Theorem 26.5 of [Roc70]). *Let $\Psi : \Omega \subset \mathbb{R}^d \rightarrow \mathbb{R}$ be a C^1 strictly convex function on an open convex domain Ω with $\|\nabla \Psi(\mathbf{x})\| \rightarrow \infty$ as $\mathbf{x} \rightarrow \partial\Omega$. Suppose that $\nabla \Psi(\Omega) = \mathbb{R}^d$. Let $\Psi^*(\mathbf{u}) := \sup_{\mathbf{x} \in \Omega} \{\langle \mathbf{x}, \mathbf{u} \rangle - \Psi(\mathbf{x})\}$ denote the convex conjugate of Ψ . Then, Ψ^* is a C^1 strictly convex function defined on \mathbb{R}^d and $\|\nabla \Psi^*(\mathbf{u})\| \rightarrow \infty$ as $\|\mathbf{u}\| \rightarrow \infty$. Moreover, $\nabla \Psi$ is a bijection from Ω to \mathbb{R}^d and $(\nabla \Psi)^{-1} = \nabla \Psi^*$.*

The proof of this theorem is beyond the scope of this note and one can check [Roc70] for a formal proof. Though this theorem contains some deep observations in convex analysis and offers a different view of mirror descent, the only consequence we will need here is inverting $\nabla \Psi$ is invertible.

Now, we are ready to derive the update rule for mirror descent. As we have discussed earlier, the proximal form of the mirror descent update is

$$\mathbf{x}_{t+1} = \operatorname{argmin}_{\mathbf{x} \in \mathcal{D}} \left\{ \langle \nabla f(\mathbf{x}_t), \mathbf{x} \rangle + \frac{1}{\eta} D_\Psi(\mathbf{x}, \mathbf{x}_t) \right\}, \quad (4)$$

where $\Psi : \Omega \rightarrow \mathbb{R}$ satisfies the condition of Theorem 2.7 and $\mathcal{D} \subset \Omega$ is the constraint set. For simplicity, we have dropped the terms that do not depend on \mathbf{x} in (4). Since Ψ is strictly convex, $D_\Psi(\mathbf{x}, \mathbf{x}_t) = \Psi(\mathbf{x}) - \Psi(\mathbf{x}_t) - \langle \nabla \Psi(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle$ is also strictly convex in \mathbf{x} . Hence, the RHS has a unique minimizer. If $\mathcal{D} = \Omega$, then we can find this minimizer by setting the gradient to be zero. That is,

$$\begin{aligned} \nabla f(\mathbf{x}_t) + \frac{1}{\eta} (\nabla \Psi(\mathbf{x}_{t+1}) - \nabla \Psi(\mathbf{x}_t)) &= 0 \quad \Rightarrow \quad \nabla \Psi(\mathbf{x}_{t+1}) = \nabla \Psi(\mathbf{x}_t) - \eta \nabla f(\mathbf{x}_t) \\ &\Rightarrow \quad \mathbf{x}_{t+1} = \nabla \Psi^* (\nabla \Psi(\mathbf{x}_t) - \eta \nabla f(\mathbf{x}_t)), \end{aligned}$$

where the last line comes from Theorem 2.7. In general, \mathbf{x}_{t+1} can lie outside \mathcal{D} . Similar to the gradient descent case, we will project it back to \mathcal{D} ,³ though this time the Bregman projection will be used.

³One can alternatively view this projection step as a proximal step w.r.t. the convex indicator of \mathcal{D} .

Definition 2.8 (Bregman projection). Let $\Psi : \Omega \subset \mathbb{R}^d \rightarrow \mathbb{R}$ be a function satisfying the conditions in Theorem 2.7 and $\mathcal{D} \subset \Omega$ a convex set. The **Bregman projection** is the mapping

$$\Pi_{\mathcal{D}}^{\Psi} : \mathbb{R}^d \rightarrow \mathcal{D}, \quad \mathbf{x} \mapsto \underset{\mathbf{y} \in \mathcal{D}}{\operatorname{argmin}} D_{\Psi}(\mathbf{y}, \mathbf{x}).$$

The existence and the uniqueness of the minimizer is guaranteed by the strict convexity of Ψ and the condition $\lim_{\mathbf{y} \rightarrow \partial\Omega} \|\nabla\Psi(\mathbf{y})\| = \infty$.

Similar to the usual Euclidean projection, we have the following simple but useful fact.

Lemma 2.9. Let $\Psi, \mathcal{D}, \Pi_{\mathcal{D}}^{\Psi}$ have the same meaning as in Definition 2.8. For any $\mathbf{x} \in \Omega$ and $\mathbf{y} \in \mathcal{D}$, we have

$$\left\langle \nabla\Psi(\mathbf{x}) - \nabla\Psi\left(\Pi_{\mathcal{D}}^{\Psi}(\mathbf{x})\right), \mathbf{y} - \Pi_{\mathcal{D}}^{\Psi}(\mathbf{x}) \right\rangle \leq 0.$$

Proof. For notational simplicity, put $\mathbf{z} = \Pi_{\mathcal{D}}^{\Psi}(\mathbf{x})$ and let $\mathbf{z}_{\theta} = (1 - \theta)\mathbf{z} + \theta\mathbf{y}$, $\theta \in [0, 1]$. Since \mathcal{D} is convex, we have $\mathbf{z}_{\theta} \in \mathcal{D}$. Moreover,

$$\frac{d}{d\theta} D_{\Psi}(\mathbf{z}_{\theta}, \mathbf{x}) = \frac{d}{d\theta} (\Psi(\mathbf{z}_{\theta}) - \langle \nabla\Psi(\mathbf{x}), \mathbf{z}_{\theta} \rangle) = \langle \nabla\Psi(\mathbf{z}_{\theta}) - \nabla\Psi(\mathbf{x}), \mathbf{y} - \mathbf{z} \rangle.$$

Since \mathbf{z} is the minimizer, we must have $\frac{d}{d\theta} D_{\Psi}(\mathbf{z}_{\theta}, \mathbf{x})|_{\theta=0} \geq 0$, i.e., $\langle \nabla\Psi(\mathbf{x}) - \nabla\Psi(\mathbf{z}), \mathbf{y} - \mathbf{z} \rangle \leq 0$. \square

We can now formally define the (projected) mirror descent updates.

Definition 2.10 (Mirror descent). Let $f : \mathcal{D} \rightarrow \mathbb{R}$ be the objective function. Suppose that f is a C^1 and its domain \mathcal{D} is convex. Let $\Psi : \Omega \rightarrow \mathbb{R}$ be a function satisfying the conditions in Theorem 2.7 and $\Omega \supset \mathcal{D}$. Let $\eta > 0$ be the step size and $\mathbf{x}_0 \in \mathcal{D}$ the initial point. The **(projected) mirror descent algorithm** is the following update rule:

$$\hat{\mathbf{x}}_{t+1} = \nabla\Psi^* (\nabla\Psi(\mathbf{x}_t) - \eta\nabla f(\mathbf{x}_t)), \quad \mathbf{x}_{t+1} = \Pi_{\mathcal{D}}^{\Psi}(\hat{\mathbf{x}}_{t+1}).$$

To analyze the convergence rate of mirror descent, we will use the following (true) mirror descent lemma.

Lemma 2.11 (Mirror descent lemma). Let the symbols have the same meaning as in Definition 2.8. In addition, suppose that f is convex. Then, for any $\mathbf{y} \in \mathcal{D}$, we have

$$f(\mathbf{x}_t) \leq f(\mathbf{y}) + \frac{1}{\eta} (D_{\Psi}(\mathbf{y}, \mathbf{x}_t) - D_{\Psi}(\mathbf{y}, \mathbf{x}_{t+1})) + \frac{1}{\eta} D_{\Psi}(\mathbf{x}_t, \hat{\mathbf{x}}_{t+1}).$$

Proof. Note that we have $\nabla\Psi(\hat{\mathbf{x}}_{t+1}) = \nabla\Psi(\mathbf{x}_t) - \eta\nabla f(\mathbf{x}_t)$. Since f is convex, we have

$$\begin{aligned} f(\mathbf{x}_t) &\leq f(\mathbf{y}) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{y} \rangle \\ &= f(\mathbf{y}) + \frac{1}{\eta} \langle \nabla\Psi(\mathbf{x}_t) - \nabla\Psi(\hat{\mathbf{x}}_{t+1}), \mathbf{x}_t - \mathbf{y} \rangle \\ &= f(\mathbf{y}) + \frac{1}{\eta} \langle \nabla\Psi(\mathbf{x}_{t+1}) - \nabla\Psi(\mathbf{x}_t), \mathbf{y} - \mathbf{x}_t \rangle \\ &\quad + \frac{1}{\eta} \langle \nabla\Psi(\hat{\mathbf{x}}_{t+1}) - \nabla\Psi(\mathbf{x}_{t+1}), \mathbf{y} - \mathbf{x}_{t+1} \rangle - \frac{1}{\eta} \langle \nabla\Psi(\hat{\mathbf{x}}_{t+1}) - \nabla\Psi(\mathbf{x}_{t+1}), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle. \end{aligned}$$

By Lemma 2.9, the first term in the last line is non-positive. Then, by Lemma 2.6, we have

$$\begin{aligned}
f(\mathbf{x}_t) &= f(\mathbf{y}) + \frac{1}{\eta} \langle \nabla \Psi(\mathbf{x}_{t+1}) - \nabla \Psi(\mathbf{x}_t), \mathbf{y} - \mathbf{x}_t \rangle - \frac{1}{\eta} \langle \nabla \Psi(\hat{\mathbf{x}}_{t+1}) - \nabla \Psi(\mathbf{x}_{t+1}), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \\
&= f(\mathbf{y}) + \frac{1}{\eta} (D_\Psi(\mathbf{y}, \mathbf{x}_t) - D_\Psi(\mathbf{y}, \mathbf{x}_{t+1}) + D_\Psi(\mathbf{x}_t, \mathbf{x}_{t+1})) \\
&\quad - \frac{1}{\eta} (D_\Psi(\mathbf{x}_t, \mathbf{x}_{t+1}) - D_\Psi(\mathbf{x}_t, \hat{\mathbf{x}}_{t+1}) + D_\Psi(\mathbf{x}_{t+1}, \hat{\mathbf{x}}_{t+1})) \\
&= f(\mathbf{y}) + \frac{1}{\eta} (D_\Psi(\mathbf{y}, \mathbf{x}_t) - D_\Psi(\mathbf{y}, \mathbf{x}_{t+1})) + \frac{1}{\eta} D_\Psi(\mathbf{x}_t, \hat{\mathbf{x}}_{t+1}).
\end{aligned}$$

□

Corollary 2.12 (Convergence rate of mirror descent). *Let the symbols have the same meaning as in Definition 2.8. In addition, suppose that f is convex. Then, for any $\mathbf{y} \in \mathcal{D}$, we have*

$$\frac{1}{T} \sum_{t=0}^{T-1} f(\mathbf{x}_t) \leq f(\mathbf{y}) + \frac{D_\Psi(\mathbf{y}, \mathbf{x}_0)}{T\eta} + \frac{1}{\eta T} \sum_{t=0}^{T-1} D_\Psi(\mathbf{x}_t, \hat{\mathbf{x}}_{t+1}).$$

Remark. Unlike in Proposition 2.4, we do not explicitly bound the last term using a “mirror descent lemma” (cf. Lemma 2.1). This will allow sharper controls, as we can replace global smoothness conditions with local ones. Nevertheless, one should view the last term as a quantity proportional to η times the average local smoothness under D_Ψ , where the η factor comes from the fact that $D_\Psi(\hat{\mathbf{x}}_{t+1}, \mathbf{x}_t)$ is usually proportional to η^2 . Hence, the second and the third terms on the RHS impose a trade-off between $D_\Psi(\mathbf{x}_0, \mathbf{y})$ and the local smoothness under D_Ψ . ♣

Proof. Sum both sides of the inequality in Lemma 2.11 from 0 to $T - 1$, and we obtain

$$\begin{aligned}
\sum_{t=0}^{T-1} f(\mathbf{x}_t) &\leq T f(\mathbf{y}) + \frac{1}{\eta} (D_\Psi(\mathbf{y}, \mathbf{x}_0) - D_\Psi(\mathbf{y}, \mathbf{x}_T)) + \frac{1}{\eta} \sum_{t=0}^{T-1} D_\Psi(\mathbf{x}_t, \hat{\mathbf{x}}_{t+1}) \\
&\leq T f(\mathbf{y}) + \frac{D_\Psi(\mathbf{y}, \mathbf{x}_0)}{\eta} + \frac{1}{\eta} \sum_{t=0}^{T-1} D_\Psi(\mathbf{x}_t, \hat{\mathbf{x}}_{t+1}).
\end{aligned}$$

Rearrange terms and we complete the proof. □

2.2 Regret minimization and online mirror descent

In this subsection, we extend the above discussion to the online setting, where the objective function can change at each step. The flexibility we get from allowing the objective to be potentially adversarial will lead to surprising ways of applying the mirror descent algorithm.

Consider the following scenario. At step t , the player is allowed to choose an action $\mathbf{x}_t \in \mathcal{D} \subset \mathbb{R}^d$. After that, the adversary will choose a loss function $f_t : \mathcal{D} \rightarrow \mathbb{R}$, which can potentially depend on the player’s choice \mathbf{x}_t . Then, the player will suffer a loss of $f_t(\mathbf{x}_t)$. The average loss after T steps is defined as $\frac{1}{T} \sum_{t=0}^{T-1} f_t(\mathbf{x}_t)$. The player’s goal is to try to beat the best fixed action in hindsight.

Formally, let $\mathbf{x}_* \in \mathcal{D}$ be an arbitrary point. The average loss the player will suffer from always choosing \mathbf{x}_* is $\frac{1}{T} \sum_{t=0}^{T-1} f_t(\mathbf{x}_*)$. The **average regret w.r.t. \mathbf{x}_* after T steps** is defined as

$$\frac{1}{T} R_T(\mathbf{x}_*) := \frac{1}{T} \sum_{t=0}^{T-1} (f_t(\mathbf{x}_t) - f(\mathbf{x}_*)).$$

Note that this covers the offline optimization problem as a special case. In the offline case, the adversary always choose a fixed loss function f . When \mathbf{x}_* is the minimizer of f , the average regret is the average gap between $f(\mathbf{x}_t)$ and the optimal value. If f is convex and reasonably smooth, then by Proposition 2.4 and Corollary 2.12, we know that the player can use the gradient descent algorithm or mirror descent algorithm to ensure the average regret goes to 0 as $T \rightarrow \infty$. We will see that the same is also true for the regret minimization problem.

Definition 2.13 (Online mirror descent). *Let $\Psi : \Omega \subset \mathbb{R}^d \rightarrow \mathbb{R}$ be a function satisfying the conditions in Theorem 2.7. Let $\mathcal{D} \subset \Omega$ be a convex domain and $T \in \mathbb{N}_{>0}$. Suppose that the loss functions the adversary chooses are $(f_t : \Omega \rightarrow \mathbb{R})_{t=0}^T$. Let $\eta > 0$ be the step size and $\mathbf{x}_0 \in \mathcal{D}$ the initial point. The **online mirror descent algorithm** is the following strategy of choosing actions:*

$$\hat{\mathbf{x}}_{t+1} = \nabla \Psi^* (\nabla \Psi(\mathbf{x}_t) - \eta \nabla f_t(\mathbf{x}_t)), \quad \mathbf{x}_{t+1} = \Pi_D^\Psi(\hat{\mathbf{x}}_{t+1}).$$

Remark. Note that the only difference between Definition 2.10 and this definition is that $\nabla f_t(\mathbf{x}_t)$ is used at the t -th step in place of $\nabla f(\mathbf{x}_t)$. ♣

Similar to the previous analysis, we have the following guarantee on the convergence rate of online mirror descent. See the discussion following Corollary 2.12 for an explanation on the two terms in this bound.

Theorem 2.14 (Convergence rate of online mirror descent). *Let the symbols have the same meaning as in Definition 2.8. In addition, suppose that all f_t are convex. Then, for any $\mathbf{y} \in \mathcal{D}$, we have*

$$\frac{1}{T} R_T(\mathbf{y}) \leq \frac{D_\Psi(\mathbf{y}, \mathbf{x}_0)}{T\eta} + \frac{1}{T\eta} \sum_{t=0}^{T-1} D_\Psi(\mathbf{x}_t, \hat{\mathbf{x}}_{t+1}).$$

Proof. First, the proof of Lemma 2.11, *mutatis mutandis*, yields

$$f_t(\mathbf{x}_t) \leq f_t(\mathbf{y}) + \frac{1}{\eta} (D_\Psi(\mathbf{y}, \mathbf{x}_t) - D_\Psi(\mathbf{y}, \mathbf{x}_{t+1})) + \frac{1}{\eta} D_\Psi(\mathbf{x}_t, \hat{\mathbf{x}}_{t+1}).$$

Sum both sides from 0 to $T - 1$, and we obtain

$$\begin{aligned} \sum_{t=0}^{T-1} f_t(\mathbf{x}_t) &\leq \sum_{t=0}^{T-1} f_t(\mathbf{y}) + \frac{1}{\eta} (D_\Psi(\mathbf{y}, \mathbf{x}_0) - D_\Psi(\mathbf{y}, \mathbf{x}_T)) + \frac{1}{\eta} \sum_{t=0}^{T-1} D_\Psi(\mathbf{x}_t, \hat{\mathbf{x}}_{t+1}) \\ &\leq \sum_{t=0}^{T-1} f_t(\mathbf{y}) + \frac{D_\Psi(\mathbf{y}, \mathbf{x}_0)}{\eta} + \frac{1}{\eta} \sum_{t=0}^{T-1} D_\Psi(\mathbf{x}_t, \hat{\mathbf{x}}_{t+1}). \end{aligned}$$

Rearrange terms and we complete the proof. □

2.3 Zero-sum games

In the regret minimization framework, the objective function can be adversarial, which allows us to study minimax optimization problems under this framework. For simplicity, we consider one of the most classical examples of minimax optimization problems — zero-sum games, and prove the celebrated minimax theorem of von Neumann using online mirror descent. The ideas introduced in this subsection are important to our discussion of spectral sparsification in Section 3, as we will partially reformulate spectral sparsification as a zero-sum game.

Let Δ_n denote the probability simplex in \mathbb{R}^n . Let $\mathbf{A} \in \mathbb{R}^{n \times m}$ be the “loss matrix.” Consider the following game. At each round, the first player choose an action $i \in [n]$ according to a probability distribution $\mathbf{x} \in \Delta_n$. The second player, who is assumed to know \mathbf{x} but not i , will choose an action $j \in [m]$ according to a probability distribution $\mathbf{y} \in \Delta_m$. Then, the first player will receive a loss of $A_{i,j}$. The goal of the first player is to minimize the expected loss, i.e.,

$$\min_{\mathbf{x} \in \Delta_n} \max_{\mathbf{y} \in \Delta_m} \mathbf{x}^\top \mathbf{A} \mathbf{y}.$$

It is easy to see that $\min_{\mathbf{x} \in \Delta_n} \max_{\mathbf{y} \in \Delta_m} \mathbf{x}^\top \mathbf{A} \mathbf{y} \geq \max_{\mathbf{y} \in \Delta_m} \min_{\mathbf{x} \in \Delta_n} \mathbf{x}^\top \mathbf{A} \mathbf{y}$. This is called weak duality and is true for any function. The following minimax theorem says that the reverse, which is called strong duality, can also be true.

Theorem 2.15 (Minimax theorem). *Let $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ be compact convex sets. Let $f : X \times Y \rightarrow \mathbb{R}$ be a continuous function that is convex in the first variable and concave in the second variable. Then, we have*

$$\min_{\mathbf{x} \in X} \max_{\mathbf{y} \in Y} f(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{y} \in Y} \min_{\mathbf{x} \in X} f(\mathbf{x}, \mathbf{y}).$$

In particular, this theorem implies our zero-sum game has strong duality. In the following, we will provide a direct proof of this special case using online mirror descent.

Proposition 2.16. $\min_{\mathbf{x} \in \Delta_n} \max_{\mathbf{y} \in \Delta_m} \mathbf{x}^\top \mathbf{A} \mathbf{y} = \max_{\mathbf{y} \in \Delta_m} \min_{\mathbf{x} \in \Delta_n} \mathbf{x}^\top \mathbf{A} \mathbf{y}$ for any $\mathbf{A} \in \mathbb{R}^{n \times m}$.

Proof. First, consider the weak duality. Let $\mathbf{x}_* \in \Delta_n$ be the minimizer of $\mathbf{x} \mapsto \max_{\mathbf{y} \in \Delta_m} \mathbf{x}^\top \mathbf{A} \mathbf{y}$ on Δ_n (the existence is guaranteed by the compactness of Δ_n and the continuity). Then, we have

$$\min_{\mathbf{x} \in \Delta_n} \max_{\mathbf{y} \in \Delta_m} \mathbf{x}^\top \mathbf{A} \mathbf{y} = \max_{\mathbf{y} \in \Delta_m} \mathbf{x}_*^\top \mathbf{A} \mathbf{y} \geq \max_{\mathbf{y} \in \Delta_m} \min_{\mathbf{x} \in \Delta_n} \mathbf{x}^\top \mathbf{A} \mathbf{y}.$$

For the other direction, consider the following regret minimization problem. Let $\mathbf{x}_t \in \Delta_n$ denote the action the player take at step t . Suppose that the adversary choose the loss function to be $\max_{\mathbf{y} \in \Delta_m} \mathbf{x}_t^\top \mathbf{A} \mathbf{y}$ and let \mathbf{y}_t denote $\arg\max_{\mathbf{y} \in \Delta_m} \mathbf{x}_t^\top \mathbf{A} \mathbf{y}$. Then, we have

$$\min_{\mathbf{x} \in \Delta_n} \max_{\mathbf{y} \in \Delta_m} \mathbf{x}^\top \mathbf{A} \mathbf{y} \leq \max_{\mathbf{y} \in \Delta_m} \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathbf{x}_t \right)^\top \mathbf{A} \mathbf{y} = \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{x}_t^\top \mathbf{A} \mathbf{y}_* \leq \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{x}_t^\top \mathbf{A} \mathbf{y}_t,$$

where $\mathbf{y}_* = \arg\max_{\mathbf{y} \in \Delta_m} \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{x}_t^\top \mathbf{A} \mathbf{y}$. Note that the RHS is exactly the average loss after T steps. Hence, by the definition of the regret, we have

$$\min_{\mathbf{x} \in \Delta_n} \max_{\mathbf{y} \in \Delta_m} \mathbf{x}^\top \mathbf{A} \mathbf{y} \leq \frac{1}{T} \sum_{t=0}^{T-1} \tilde{\mathbf{x}}^\top \mathbf{A} \mathbf{y}_t + \frac{1}{T} R_T(\tilde{\mathbf{x}}), \quad \forall \tilde{\mathbf{x}} \in \Delta_n.$$

Choose $\tilde{\mathbf{x}} = \mathbf{x}_*$ to be the minimizer of $\tilde{\mathbf{x}} \mapsto \frac{1}{T} \sum_{t=0}^{T-1} \tilde{\mathbf{x}}^\top \mathbf{A} \mathbf{y}_t$. Then, we get

$$\min_{\mathbf{x} \in \Delta_n} \max_{\mathbf{y} \in \Delta_m} \mathbf{x}^\top \mathbf{A} \mathbf{y} \leq \min_{\mathbf{x} \in \Delta} \mathbf{x}^\top \mathbf{A} \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathbf{y}_t \right) + \frac{1}{T} R_T(\mathbf{x}_*) \leq \max_{\mathbf{y} \in \Delta_m} \min_{\mathbf{x} \in \Delta} \mathbf{x}^\top \mathbf{A} \mathbf{y} + \frac{1}{T} R_T(\mathbf{x}_*).$$

Thus, to prove strong duality, it suffices to find an algorithm that can solve the regret minimization problem with loss function being $f(\mathbf{x}) = \max_{\mathbf{y} \in \Delta_m} \mathbf{x}^\top \mathbf{A} \mathbf{y}$, in the sense that the average regret goes 0 as $T \rightarrow \infty$. In the rest of this subsection, we will show that online mirror descent with D_Ψ being the (generalized) KL divergence can solve this problem (cf. Lemma 2.17). \square

We now instantiate our online mirror descent algorithm (Definition 2.13) with $\Omega = R_{>0}^n$, $\Psi(\mathbf{x}) = \sum_{i=1}^n x_i \log x_i$, and $\mathcal{D} = \text{Int } \Delta_n$. One can easily verify that

$$\nabla \Psi(\mathbf{x}) = \log \mathbf{x} + \mathbf{1} \quad \text{and} \quad \nabla \Psi^*(\mathbf{g}) = (\nabla \Psi)^{-1}(\mathbf{g}) = \exp(\mathbf{g} - \mathbf{1}).$$

Therefore, the Bregman divergence in this case is

$$D_\Psi(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^n x_i \log \frac{x_i}{z_i} - \langle \mathbf{1}, \mathbf{x} - \mathbf{z} \rangle,$$

which is exactly the generalized KL divergence, and it reduces to the usual KL divergence if we restrict \mathbf{x}, \mathbf{z} to Δ_n . Meanwhile, with $\mathbf{y}_t \in \arg \max_{\mathbf{y} \in \Delta_m} \mathbf{x}_t^\top \mathbf{A} \mathbf{y}$, we can write

$$\nabla_{\mathbf{x}} \left(\mathbf{x} \mapsto \max_{\mathbf{y} \in \Delta_m} \mathbf{x}^\top \mathbf{A} \mathbf{y} \right) \Big|_{\mathbf{x}=\mathbf{x}_t} = \mathbf{A} \mathbf{y}_t,$$

and therefore the mirror descent update can be rewritten as

$$\hat{\mathbf{x}}_{t+1} = \nabla \Psi^* (\nabla \Psi(\mathbf{x}_t) - \eta \mathbf{A} \mathbf{y}_t) = \exp(\log \mathbf{x} - \eta \mathbf{A} \mathbf{y}_t).$$

Lemma 2.17. *Consider the setting described above. Put $M = \max_{i,j} |A_{i,j}|$ and choose $\eta = \sqrt{\frac{D_\Psi(\mathbf{y}, \mathbf{x}_0)}{M^2 T}}$. Then, we have*

$$\frac{1}{T} R_T(\mathbf{y}) \leq \sqrt{\frac{M^2 D_\Psi(\mathbf{y}, \mathbf{x}_0)}{T}} \rightarrow 0 \quad \text{as } T \rightarrow \infty.$$

Proof. Let $\mathbf{a}_i \in \mathbb{R}^m$ denote the i -th row of \mathbf{A} and set $M = \max_i \|\mathbf{a}_i\|_\infty$. Meanwhile, with $i_t = \arg \max_{i \in [m]} (\mathbf{A}^\top \mathbf{x}_t)_i$, we can choose $\mathbf{y}_t = \mathbf{e}_{i_t}$. Then, we have $|\langle \mathbf{a}_i, \mathbf{y}_t \rangle| \leq M$. Therefore, for any $i \in [n]$ and $\eta \leq 1/M$, we have

$$\hat{x}_{t+1,i} = x_{t,i} e^{-\eta \langle \mathbf{a}_i, \mathbf{y}_t \rangle} = x_{t,i} \left(1 - \eta \langle \mathbf{a}_i, \mathbf{y}_t \rangle \pm \eta^2 M^2 \right).$$

Then, we compute

$$\begin{aligned} D_\Psi(\mathbf{x}_t, \hat{\mathbf{x}}_{t+1}) &= \sum_{i=1}^n x_{t,i} \log \left(\frac{x_{t,i}}{\hat{x}_{t+1,i}} \right) + 1 - \sum_{i=1}^n \hat{x}_{t+1,i} \\ &= \eta \sum_{i=1}^n x_{t,i} \langle \mathbf{a}_i, \mathbf{y}_t \rangle + 1 - \sum_{i=1}^n x_{t,i} \left(1 - \eta \langle \mathbf{a}_i, \mathbf{y}_t \rangle \pm \eta^2 M^2 \right) = \pm \eta^2 M^2. \end{aligned}$$

Thus, by Theorem 2.14, we have

$$\frac{1}{T}R_T(\mathbf{y}) \leq \frac{D_\Psi(\mathbf{y}, \mathbf{x}_0)}{T\eta} + \frac{1}{T\eta} \sum_{t=0}^{T-1} \eta^2 M^2 = \frac{D_\Psi(\mathbf{y}, \mathbf{x}_0)}{T\eta} + \eta M^2.$$

Choose $\eta = M^{-1}\sqrt{D_\Psi(\mathbf{y}, \mathbf{x}_0)/T}$ and we complete the proof. \square

3 Spectral Sparsification

Let $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^d$ be such that $n^{-1} \sum_{i=1}^n \mathbf{v}_i \mathbf{v}_i^\top = \mathbf{I}_d$. The task **spectral sparsification** asks us to find a sparse $\mathbf{s} \in \mathbb{R}_+^n$ such that

$$(1 - \varepsilon)\mathbf{I}_d \preceq \sum_{i=1}^n s_i \mathbf{v}_i \mathbf{v}_i^\top \preceq (1 + \varepsilon)\mathbf{I}_d, \quad (5)$$

where $\varepsilon \in (0, 1)$ is the target accuracy.

3.1 From spectral sparsification to regret minimization

First, consider the upper bound $\sum_{i=1}^n s_i \mathbf{v}_i \mathbf{v}_i^\top \preceq (1 + \varepsilon)\mathbf{I}_d$. Here, the goal is to find a sparse $\mathbf{s} \in \mathbb{R}_+^n$ such that $\lambda_{\max}(\sum_{i=1}^n s_i \mathbf{v}_i \mathbf{v}_i^\top) \leq 1 + \varepsilon$. If we do not require \mathbf{s} to be sparse, then this is essentially equivalent to minimizing the function $\mathbf{s} \mapsto \lambda_{\max}(\sum_{i=1}^n s_i \mathbf{v}_i \mathbf{v}_i^\top)$ over $\mathbf{s} \in \Delta_n$. Note that for any PSD matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, we have the following variational characterization of λ_{\max} :

$$\lambda_{\max}(\mathbf{A}) = \max_{\mathbf{X} \in \Delta_{d \times d}} \langle \mathbf{A}, \mathbf{X} \rangle, \quad \Delta_{d \times d} := \{\mathbf{Y} \in \mathbb{R}^{d \times d} : \mathbf{Y} \succeq 0, \text{Tr } \mathbf{Y} = 1\}, \quad (6)$$

where the maximum can be attained at the projection matrix onto the top eigenspace. Matrices in $\Delta_{d \times d}$ are called density matrices, and they can be viewed as a generalization of the probability vectors — it puts weight $\lambda_k(\mathbf{Y})$ on the k -th eigendirection, and the condition $\text{Tr } \mathbf{Y} = 1$ ensures the total mass is 1. With this variational characterization of λ_{\max} , we can rewrite the optimization task as

$$\min_{\mathbf{s} \in \mathbb{R}_+^n} \max_{\mathbf{X} \in \Delta_{d \times d}} \left\langle \sum_{i=1}^n s_i \mathbf{v}_i \mathbf{v}_i^\top, \mathbf{X} \right\rangle,$$

Since the objective equals 1 when $\mathbf{s} = \mathbf{1}_n/n$, it suffices to find a ε -optimal solution that is also sparse. Clear that the objective is (bi)linear in \mathbf{s} and \mathbf{X} , and both Δ_n and $\Delta_{d \times d}$ are convex. Hence, by the minimax theorem (Theorem 2.15), it has strong duality. Therefore, we can alternatively consider the dual problem

$$\min_{\mathbf{X} \in \Delta_{d \times d}} \max_{\mathbf{s} \in \mathbb{R}_+^n} \left\langle -\sum_{i=1}^n s_i \mathbf{v}_i \mathbf{v}_i^\top, \mathbf{X} \right\rangle. \quad (7)$$

Similar to the proof of Proposition 2.16, we consider the regret minimization problem associated with (7). Suppose that the player chooses action $\mathbf{X}_t \in \Delta_{d \times d}$ at time t , the adversary chooses a \mathbf{s}_t that can depend on \mathbf{X}_t , and then the player receives loss $\langle -\sum_{i=1}^n s_{t,i} \mathbf{v}_i \mathbf{v}_i^\top, \mathbf{X}_t \rangle$. Note that for any fixed strategy $\mathbf{U} \in \Delta_{d \times d}$, the average loss is

$$\frac{1}{T} \sum_{t=0}^{T-1} \left\langle -\sum_{i=1}^n s_{t,i} \mathbf{v}_i \mathbf{v}_i^\top, \mathbf{U} \right\rangle = -\left\langle \sum_{i=1}^n \left(\frac{1}{T} \sum_{t=0}^{T-1} s_{t,i} \right) \mathbf{v}_i \mathbf{v}_i^\top, \mathbf{U} \right\rangle = -\left\langle \sum_{i=1}^n \bar{s}_i \mathbf{v}_i \mathbf{v}_i^\top, \mathbf{U} \right\rangle,$$

where $\bar{s} = T^{-1} \sum_{t=0}^{T-1} s_t$. If we take the \mathbf{U}_+ that is optimal in hindsight, then we get

$$\frac{1}{T} \sum_{t=0}^{T-1} \left\langle - \sum_{i=1}^n s_{t,i} \mathbf{v}_i \mathbf{v}_i^\top, \mathbf{U}_+ \right\rangle = -\lambda_{\max} \left(\sum_{i=1}^n \bar{s}_i \mathbf{v}_i \mathbf{v}_i^\top \right).$$

In addition, note that the average regret w.r.t. \mathbf{U}_+ is

$$\frac{1}{T} R_T^X(\mathbf{U}_+) := \frac{1}{T} \sum_{t=0}^{T-1} \left\langle - \sum_{i=1}^n s_{t,i} \mathbf{v}_i \mathbf{v}_i^\top, \mathbf{X}_t \right\rangle + \lambda_{\max} \left(\sum_{i=1}^n \bar{s}_i \mathbf{v}_i \mathbf{v}_i^\top \right).$$

Rearrange terms and we get

$$\lambda_{\max} \left(\sum_{i=1}^n \bar{s}_i \mathbf{v}_i \mathbf{v}_i^\top \right) = \frac{1}{T} R_T^X(\mathbf{U}_+) + \frac{1}{T} \sum_{t=0}^{T-1} \left\langle \sum_{i=1}^n s_{t,i} \mathbf{v}_i \mathbf{v}_i^\top, \mathbf{X}_t \right\rangle. \quad (8)$$

As a result, to solve the upper side of the sparsification problem, it suffices to design sequences $(\mathbf{X}_t)_t$ and $(s_t)_t$ such that \bar{s} is sparse, the regret goes to 0, and $T^{-1} \sum_{t=0}^{T-1} \langle \sum_{i=1}^n s_{t,i} \mathbf{v}_i \mathbf{v}_i^\top, \mathbf{X}_t \rangle \leq 1 + \varepsilon/2$.

For the lower side, the above argument, *mutatis mutandis*, yields following counterpart of (7):

$$\min_{\mathbf{Y} \in \Delta_{d \times d}} \max_{s \in \mathbb{R}_n^+} \left\langle \sum_{i=1}^n s_i \mathbf{v}_i \mathbf{v}_i^\top, \mathbf{Y} \right\rangle.$$

Again, consider the regret minimization problem where the player choose \mathbf{Y}_t and receives loss $\langle \sum_{i=1}^n s_{t,i} \mathbf{v}_i \mathbf{v}_i^\top, \mathbf{Y}_t \rangle$. For fixed \mathbf{U} , the average loss is $\langle \sum_{i=1}^n \bar{s}_i \mathbf{v}_i \mathbf{v}_i^\top, \mathbf{U} \rangle$, whose minimum is the smallest eigenvalue $\lambda_{\min}(\sum_{i=1}^n \bar{s}_i \mathbf{v}_i \mathbf{v}_i^\top)$. Let \mathbf{U}_- denote the minimizer. Then, similar to (8), we have

$$\lambda_{\min} \left(\sum_{i=1}^n \bar{s}_i \mathbf{v}_i \mathbf{v}_i^\top \right) := \frac{1}{T} \sum_{t=0}^{T-1} \left\langle \sum_{i=1}^n s_{t,i} \mathbf{v}_i \mathbf{v}_i^\top, \mathbf{Y}_t \right\rangle - \frac{1}{T} R_T^Y(\mathbf{U}_-).$$

Again, to ensure the smallest eigenvalue is at least $1 - \varepsilon$, it suffices to require $T^{-1} R_T^Y(\mathbf{U}_-) \leq \varepsilon/2$ and $T^{-1} \sum_{t=0}^{T-1} \langle \sum_{i=1}^n s_{t,i} \mathbf{v}_i \mathbf{v}_i^\top, \mathbf{Y}_t \rangle \geq 1 - \varepsilon/2$. Combine the above results, and we obtain the following lemma.

Lemma 3.1 (Sparsification to regret minimization). *Let $\varepsilon > 0$ be given and let $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^d$ be such that $n^{-1} \sum_{i=1}^n \mathbf{v}_i \mathbf{v}_i^\top = \mathbf{I}_d$. Suppose that we can find sequences $(s_t)_t \subset \mathbb{R}_+^n$, and $(\mathbf{X}_t)_t, (\mathbf{Y}_t)_t \subset \Delta_{d \times d}$ and $T > 0$ such that the following hold:*

- (a) $T^{-1} \sum_{t=0}^{T-1} \langle \sum_{i=1}^n s_{t,i} \mathbf{v}_i \mathbf{v}_i^\top, \mathbf{X}_t \rangle \leq 1 + \varepsilon/2$ and $T^{-1} \sum_{t=0}^{T-1} \langle \sum_{i=1}^n s_{t,i} \mathbf{v}_i \mathbf{v}_i^\top, \mathbf{Y}_t \rangle \geq 1 - \varepsilon/2$.
- (b) For any $\mathbf{U} \in \Delta_{d \times d}$, $T^{-1} R_T^X(\mathbf{U}) \leq \varepsilon/2$ and $T^{-1} R_T^Y(\mathbf{U}) \leq \varepsilon/2$, where the loss for $(\mathbf{X}_t)_t$ and $(\mathbf{Y}_t)_t$ are $\langle \mathbf{X}_t, - \sum_{i=1}^n s_{t,i} \mathbf{v}_i \mathbf{v}_i^\top \rangle$ and $\langle \mathbf{Y}_t, \sum_{i=1}^n s_{t,i} \mathbf{v}_i \mathbf{v}_i^\top \rangle$, respectively.

Then, (5) holds with s being $\bar{s} = T^{-1} \sum_{t=0}^{T-1} s_t$.

Moreover, if every s_t has at most one nonzero entry, then \bar{s} is T -sparse. Suppose that $s_t = w_t \mathbf{e}_{i_t}$ for some $w_t \geq 0$ and $i_t \in [n]$. Then, we can replace condition (a) with the following condition:

- (a') $T^{-1} \sum_{t=0}^{T-1} \langle w_t \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top, \mathbf{X}_t \rangle \leq 1 + \varepsilon/2$ and $T^{-1} \sum_{t=0}^{T-1} \langle w_t \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top, \mathbf{Y}_t \rangle \geq 1 - \varepsilon/2$.

Remark. When restricting each s_t to have one nonzero entry, we can view each s_t as one step of Frank-Wolfe type update, which is a natural candidate when sparsity is required. To ensure condition (b), we will use online mirror descent with a suitable Bregman divergence. Note that the faster online mirror descent converges, the sparser \bar{s} will be. \clubsuit

3.2 Entropy regularization and $O(d \log d)$ sparsification

We start with the case where the Bregman divergence we use in the mirror descent algorithm is the matrix counterpart of the KL divergence. First, we state without proof some basic properties of this Bregman divergence.

Lemma 3.2. Consider $\Psi : \mathcal{S}_{++}^d \rightarrow \mathbb{R}$, $\mathbf{X} \mapsto \langle \mathbf{X}, \log \mathbf{X} - \mathbf{I}_d \rangle$, where \mathcal{S}_{++}^d is the collection of d -by- d positive definite matrices. The following are true.

- (a) $\nabla \Psi(\mathbf{X}) = \log \mathbf{X}$ and $(\nabla \Psi)^{-1}(\mathbf{G}) = \exp \mathbf{G}$.
- (b) Ψ satisfies the conditions in Theorem 2.7.
- (c) $D_\Psi(\mathbf{X}, \mathbf{Y}) = \langle \mathbf{X}, \log \mathbf{X} - \log \mathbf{Y} \rangle + \langle \mathbf{Y}, \mathbf{I}_d \rangle - \langle \mathbf{X}, \mathbf{I}_d \rangle$.
- (d) $\Pi_{\Delta_{d \times d}}^\Psi(\mathbf{X}) = \mathbf{X} / \text{Tr } \mathbf{X}$.

In the rest of this subsection, Ψ will always denote this function. In addition, let $\mathbf{F}_t \in \mathcal{S}^d$ denote the feedback matrix at step t . That is, if the player chooses \mathbf{X}_t at step t , then the loss they receive is $\langle \mathbf{X}_t, \mathbf{F}_t \rangle$. Recall the definition of online mirror descent from Definition 2.13. In this case, we have

$$\hat{\mathbf{X}}_{t+1} = \exp(\log(\mathbf{X}_t) - \eta \mathbf{F}_t), \quad \mathbf{X}_{t+1} = \hat{\mathbf{X}}_{t+1} / \text{Tr } \hat{\mathbf{X}}_{t+1}. \quad (9)$$

Note that this is exactly the matrix multiplicative weight update method. In addition, by Theorem 2.14, we have the following convergence guarantee.

Lemma 3.3. Let $\Psi(\mathbf{X}) = \langle \mathbf{X}, \log \mathbf{X} - \mathbf{I}_d \rangle$ and \mathbf{F}_t denote the feedback matrix. Let $(\mathbf{X}_t)_t$ denote the mirror descent iterates (9) with $\mathbf{X}_0 = \mathbf{I}_d / d$. Then, for any $\mathbf{U} \in \Delta_{d \times d}$ and $\eta \leq \min_t \|\mathbf{F}_t\|_2^{-1}$, we have

$$\frac{1}{T} R_T(\mathbf{U}) \leq \frac{\log d}{T\eta} + \frac{\eta}{T} \sum_{t=0}^{T-1} \langle \mathbf{X}_t, \mathbf{F}_t^2 \rangle$$

Proof. First, by Theorem 2.14, we have

$$\frac{1}{T} R_T(\mathbf{U}) \leq \frac{D_\Psi(\mathbf{U}, \mathbf{X}_0)}{T\eta} + \frac{1}{T\eta} \sum_{t=0}^{T-1} D_\Psi(\mathbf{X}_t, \hat{\mathbf{X}}_{t+1}).$$

For the initial distance, we compute

$$D_\Psi(\mathbf{U}, \mathbf{X}_0) = \langle \mathbf{U}, \log \mathbf{U} - \log \mathbf{X}_0 \rangle + \langle \mathbf{X}_0, \mathbf{I}_d \rangle - \langle \mathbf{U}, \mathbf{I}_d \rangle = \langle \mathbf{U}, \log \mathbf{U} \rangle + \log d \leq \log d.$$

Then, we compute

$$\begin{aligned} D_\Psi(\mathbf{X}_t, \hat{\mathbf{X}}_{t+1}) &= D_\Psi(\mathbf{X}_t, \exp(\log(\mathbf{X}_t) - \eta \mathbf{F}_t)) \\ &= \langle \mathbf{X}_t, \log \mathbf{X}_t - \log(\exp(\log(\mathbf{X}_t) - \eta \mathbf{F}_t)) \rangle + \langle \exp(\log(\mathbf{X}_t) - \eta \mathbf{F}_t), \mathbf{I}_d \rangle - \langle \mathbf{X}_t, \mathbf{I}_d \rangle \\ &= \eta \langle \mathbf{X}_t, \mathbf{F}_t \rangle + \text{Tr}(\exp(\log(\mathbf{X}_t) - \eta \mathbf{F}_t)) - \langle \mathbf{X}_t, \mathbf{I}_d \rangle \\ &\leq \eta \langle \mathbf{X}_t, \mathbf{F}_t \rangle + \text{Tr}(\mathbf{X}_t \exp(-\eta \mathbf{F}_t)) - 1, \end{aligned}$$

where the inequality comes from the Golden-Thompson inequality. For $\eta \leq 1/\|\mathbf{F}_t\|_2$, we have $\exp(-\eta \mathbf{F}_t) \preceq \mathbf{I} - \eta \mathbf{F}_t + \eta^2 \mathbf{F}_t^2$. Therefore,

$$\text{Tr}(\mathbf{X}_t \exp(-\eta \mathbf{F}_t)) \leq \text{Tr}\left(\mathbf{X}_t \left(\mathbf{I} - \eta \mathbf{F}_t + \eta^2 \mathbf{F}_t^2\right)\right) = 1 - \eta \langle \mathbf{X}_t, \mathbf{F}_t \rangle + \eta^2 \langle \mathbf{X}_t, \mathbf{F}_t^2 \rangle.$$

Thus, we have $D_\Psi(\mathbf{X}_t, \hat{\mathbf{X}}_{t+1}) \leq \eta^2 \langle \mathbf{X}_t, \mathbf{F}_t^2 \rangle$. □

Now, we combine this lemma with Lemma 3.1 to recover the $O(d \log d)$ sparsification result by [SS11].

Theorem 3.4 ($O(d \log d)$ sparsification). *Fix $\varepsilon \in (0, 1)$ and let $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^d$ be such that $n^{-1} \sum_{i=1}^n \mathbf{v}_i \mathbf{v}_i^\top = \mathbf{I}_d$. With high probability, we can efficiently find $\mathbf{s} \in \mathbb{R}_+^n$ with $\|\mathbf{s}\|_0 \leq O(d \log d / \varepsilon^2)$ such that $(1 - \varepsilon)\mathbf{I}_d \preceq \sum_{i=1}^n s_i \mathbf{v}_i \mathbf{v}_i^\top \preceq (1 + \varepsilon)\mathbf{I}_d$.*

Remark. Strictly speaking, to *efficiently* find a desired \mathbf{s} , we have to run the mirror descent updates (9) in an approximate way and adapt the proof accordingly. See Section 6 and the appendix of [AZLO15] for details. ♣

Proof. At each step $t \in [T]$, we choose $i_t = k$ with probability proportional to $\|\mathbf{v}_k\|^2$. That is, with $Z = \sum_{i=1}^n \|\mathbf{v}_i\|^2 = \text{Tr}(\sum_{i=1}^n \mathbf{v}_i \mathbf{v}_i^\top) = nd$, we sample $i_t \in [n]$ according to $\mathbb{P}(i_t = k) = \|\mathbf{v}_k\|^2 / Z$. Set $w_t = d / \|\mathbf{v}_{i_t}\|^2$. Then, we have

$$\mathbb{E} [w_t \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top] = \sum_{i=1}^n \frac{d}{\|\mathbf{v}_i\|^2} \frac{\|\mathbf{v}_i\|^2}{Z} \mathbf{v}_i \mathbf{v}_i^\top = \frac{1}{n} \sum_{i=1}^n \mathbf{v}_i \mathbf{v}_i^\top = \mathbf{I}_d.$$

As a result, $\left\{ \langle \mathbf{X}_t, w_t \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top \rangle - 1 \right\}_t$ is a martingale difference sequence that is uniformly bounded by $2d$. In addition, for its variance, since $\mathbf{v}_{i_t}^\top \mathbf{X}_t \mathbf{v}_{i_t} \leq \|\mathbf{v}_{i_t}\|^2$, we have

$$\mathbb{E} \left(\langle \mathbf{X}_t, w_t \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top \rangle - 1 \right)^2 = \mathbb{E} \left[w_t^2 \left(\mathbf{v}_{i_t}^\top \mathbf{X}_t \mathbf{v}_{i_t} \right)^2 \right] - 2 \langle \mathbf{X}_t, \mathbf{I}_d \rangle + 1 \leq \mathbb{E} [d w_t \mathbf{v}_{i_t}^\top \mathbf{X}_t \mathbf{v}_{i_t}] - 1 \leq d.$$

Thus, by the Bernstein inequality, we have

$$\mathbb{P} \left[\left| \frac{1}{T} \sum_{t=0}^{T-1} \langle w_t \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top, \mathbf{X}_t \rangle - 1 \right| \geq \frac{\varepsilon}{2} \right] \leq 2 \exp \left(-\frac{T \varepsilon^2}{16d} \right).$$

The same bound also holds for $\sum_{t=0}^{T-1} \langle w_t \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top, \mathbf{Y}_t \rangle$. Thus, as long as $T \gtrsim d / \varepsilon^2$, condition (a') of Lemma 3.3 holds with high probability.

For condition (b), let $(\mathbf{X}_t)_t, (\mathbf{Y}_t)_t$ be the entropy-regularized mirror descent iterates (Lemma 2.17) with feedback matrices being $(-w_t \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top)_t$ and $(w_t \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top)_t$, respectively. Since $\left\| w_t \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top \right\|_2 = d$, we can choose any $\eta \leq 1/d$. By Lemma 3.3, for any $\mathbf{U} \in \Delta_{d \times d}$, we have

$$\begin{aligned} \frac{1}{T} R_T(\mathbf{U}) &\leq \frac{\log d}{T\eta} + \frac{\eta}{T} \sum_{t=0}^{T-1} \langle \mathbf{X}_t, w_t^2 \|\mathbf{v}_{i_t}\|^2 \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top \rangle \leq \frac{\log d}{T\eta} + \frac{d\eta}{T} \sum_{t=0}^{T-1} \langle \mathbf{X}_t, w_t \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top \rangle \\ &\leq \frac{\log d}{T\eta} + d\eta\varepsilon, \end{aligned}$$

where the second line comes from condition (a). To complete the proof, it suffices to choose $\eta = 1/d$ and $T = \Theta(d \log d / \varepsilon^2)$. \square

3.3 $l_{1-1/q}$ regularization and $O(d)$ sparsification

In this subsection, we consider a different instantiation of Lemma 3.1. We will use the $l_{1-1/q}$ -regularizer:

$$\Psi : \mathcal{S}_{++}^d \rightarrow \mathbb{R}, \quad \mathbf{X} \mapsto -\frac{q}{q-1} \text{Tr} \mathbf{X}^{1-1/q}, \quad (10)$$

where $q > 1$ is a parameter. This is equivalent to the $l_{1-1/q}$ -norm of the eigenvalues of \mathbf{X} and can be viewed as the matrix version of the Tsallis entropy. Again, we state without proof some basic properties of its associated Bregman divergence.

Lemma 3.5. *Let Ψ be given by (10). The following are true:*

- (a) $\nabla \Psi(\mathbf{X}) = -\mathbf{X}^{-1/q}$ and $(\nabla \Psi)^{-1}(\mathbf{G}) = -\mathbf{G}^{-q}$.
- (b) Ψ satisfies the conditions in Theorem 2.7.
- (c) $D_\Psi(\mathbf{X}, \mathbf{Y}) = \langle \mathbf{Y}^{-1/q}, \mathbf{X} \rangle + \frac{1}{q-1} \text{Tr} \mathbf{Y}^{1-1/q} - \frac{q}{q-1} \text{Tr} \mathbf{X}^{1-1/q}$.
- (d) $\Pi_{\Delta_{d \times d}}^\Psi(\mathbf{X}) = (\mathbf{X}^{-1/q} + \theta \mathbf{I}_d)^{-q}$ where $\theta \in \mathbb{R}$ is the unique real number that ensures the trace of the RHS is 1.

Similar to Lemma 3.3, we have the following convergence guarantee for this choice of Ψ . For simplicity, we will only consider rank-1 feedback matrices and $q = 2$. The proof can be generalized to $q > 2$ and feedback matrices that are positive or negative semidefinite by replacing the Sherman-Morrison formula with the Woodbury formula and using the Araki-Lieb-Thirring inequality.

Lemma 3.6. *Let $\Psi(\mathbf{X}) = -2 \text{Tr} \mathbf{X}^{1/2}$ and \mathbf{F}_t denote the feedback matrix at step t . Suppose that \mathbf{F}_t has rank 1 and let $(\hat{\mathbf{X}}_t)_t$ denote the mirror descent iterates with D_Ψ and $\mathbf{X}_0 = \mathbf{I}_d/d$. Then, for any $\mathbf{U} \in \Delta_{d \times d}$ and η with $\eta \langle \mathbf{X}_t^{1/2}, \mathbf{F}_t \rangle \geq -1/2$ for all t , we have*

$$\frac{1}{T} R_T(\mathbf{U}) \leq \frac{2d^{1/2}}{T\eta} + \frac{2\eta}{T} \sum_{t=0}^{T-1} \left| \langle \mathbf{X}_t, \mathbf{F}_t \rangle \langle \mathbf{X}_t^{1/2}, \mathbf{F}_t \rangle \right|.$$

Proof. Again, recall from Theorem 2.14 that

$$\frac{1}{T} R_T(\mathbf{U}) \leq \frac{D_\Psi(\mathbf{U}, \mathbf{X}_0)}{T\eta} + \frac{1}{T\eta} \sum_{t=0}^{T-1} D_\Psi(\mathbf{X}_t, \hat{\mathbf{X}}_{t+1}).$$

For the initial distance, we have

$$D_\Psi(\mathbf{U}, \mathbf{X}_0) = \langle (\mathbf{I}_d/d)^{-1/2}, \mathbf{U} \rangle + \text{Tr}(\mathbf{I}_d/d)^{1/2} - 2 \text{Tr} \mathbf{U}^{1/2} \leq 2d^{1/2}.$$

Since $\mathbf{F}_t \in \mathbb{R}^{d \times d}$ is of rank-1, we can write $\mathbf{F}_t = \alpha_t \mathbf{f}_t \mathbf{f}_t^\top$ for some $\alpha_t \in \{\pm 1\}$ and $\mathbf{f}_t \in \mathbb{R}^d$. For the mirror descent update, by the Sherman-Morrison formula, we have

$$\hat{\mathbf{X}}_{t+1}^{1/2} = \left(\mathbf{X}_t^{-1/2} + \eta \mathbf{F}_t \right)^{-1} = \mathbf{X}_t^{1/2} - \frac{\alpha_t \eta \mathbf{X}_t^{1/2} \mathbf{f}_t \mathbf{f}_t^\top \mathbf{X}_t^{1/2}}{1 + \alpha_t \eta \mathbf{f}_t^\top \mathbf{X}_t^{1/2} \mathbf{f}_t},$$

as long as $\eta < 1/|\alpha_t \mathbf{f}_t^\top \mathbf{X}_t^{1/2} \mathbf{f}_t| = \left| \langle \mathbf{X}_t^{1/2}, \mathbf{F}_t \rangle \right|^{-1}$. Then, we compute

$$\begin{aligned} D_\Psi(\mathbf{X}_t, \hat{\mathbf{X}}_{t+1}) &= \langle \hat{\mathbf{X}}_{t+1}^{-1/2}, \mathbf{X}_t \rangle + \text{Tr} \hat{\mathbf{X}}_{t+1}^{1/2} - 2 \text{Tr} \mathbf{X}_t^{1/2} \\ &= \langle \mathbf{X}_t^{-1/2} + \eta \mathbf{F}_t, \mathbf{X}_t \rangle + \text{Tr} \left(\mathbf{X}_t^{1/2} - \frac{\alpha_t \eta \mathbf{X}_t^{1/2} \mathbf{f}_t \mathbf{f}_t^\top \mathbf{X}_t^{1/2}}{1 + \alpha_t \eta \mathbf{f}_t^\top \mathbf{X}_t^{1/2} \mathbf{f}_t} \right) - 2 \text{Tr} \mathbf{X}_t^{1/2} \\ &= \eta^2 \frac{\langle \mathbf{X}_t, \mathbf{F}_t \rangle \langle \mathbf{X}_t^{1/2}, \mathbf{F}_t \rangle}{1 + \eta \langle \mathbf{X}_t^{1/2}, \mathbf{F}_t \rangle}. \end{aligned}$$

If η satisfies $\eta \langle \mathbf{X}_t^{1/2}, \mathbf{F}_t \rangle \geq -1/2$, then the last term can be bounded by $2\eta^2 \left| \langle \mathbf{X}_t, \mathbf{F}_t \rangle \langle \mathbf{X}_t^{1/2}, \mathbf{F}_t \rangle \right|$. Combine the above estimations and we complete the proof. \square

In addition, we will need the following simple lemma.

Lemma 3.7. *For any PSD $\mathbf{X} \in \mathbb{R}^{d \times d}$ and $\mathbf{v} \in \mathbb{R}^d$, we have $\langle \mathbf{X}^{1/2}, \mathbf{v} \mathbf{v}^\top \rangle \leq \langle \mathbf{X}, \mathbf{v} \mathbf{v}^\top \rangle^{1/2} \|\mathbf{v}\|$.*

Proof. By the homogeneity, we may assume w.l.o.g. that $\|\mathbf{v}\| = 1$. Let $\mathbf{X} = \sum_{i=1}^d \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$ be the spectral decomposition of \mathbf{X} . Note that $\sum_{i=1}^d \langle \mathbf{u}_i, \mathbf{v} \rangle^2 = \|\mathbf{v}\|^2 = 1$. Hence, by Jensen's inequality,

$$\langle \mathbf{X}^{1/2}, \mathbf{v} \mathbf{v}^\top \rangle = \sum_{i=1}^d \lambda^{1/2} \langle \mathbf{u}_i, \mathbf{v} \rangle^2 \leq \left(\sum_{i=1}^d \lambda \langle \mathbf{u}_i, \mathbf{v} \rangle^2 \right)^{1/2} = \langle \mathbf{X}, \mathbf{v} \mathbf{v}^\top \rangle^{1/2}.$$

\square

Now, we are ready to combine Lemma 3.1 and mirror descent with $\Psi(\mathbf{X}) = -2 \text{Tr} \mathbf{X}^{1/2}$ to recover the $O(d)$ sparsification result of [BSS12].

Theorem 3.8 (Linear sparsification). *Given $\varepsilon \in (0, 1)$ and let $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^d$ with $n^{-1} \sum_{i=1}^n \mathbf{v}_i \mathbf{v}_i^\top = \mathbf{I}_d$, we can efficiently find $\mathbf{s} \in \mathbb{R}_+^n$ with $\|\mathbf{s}\|_0 \leq O(d/\varepsilon^2)$ such that $(1-\varepsilon)\mathbf{I}_d \preceq \sum_{i=1}^n s_i \mathbf{v}_i \mathbf{v}_i^\top \preceq (1+\varepsilon)\mathbf{I}_d$.*

Proof. Let $(\mathbf{X}_t)_t$ and $(\mathbf{Y}_t)_t$ be the mirror descent iterates with Ψ given by (10). Since $n^{-1} \sum_{i=1}^n \mathbf{v}_i \mathbf{v}_i^\top = \mathbf{I}_d$ and $\text{Tr} \mathbf{X}_t = \text{Tr} \mathbf{Y}_t = 1$, we can always choose some \mathbf{v}_t such that $\langle \mathbf{X}_t, \mathbf{v}_t \mathbf{v}_t^\top \rangle \leq \langle \mathbf{Y}_t, \mathbf{v}_t \mathbf{v}_t^\top \rangle$. Let the feedback matrices for $(\mathbf{X}_t)_t$ and $(\mathbf{Y}_t)_t$ being $-w_t \mathbf{v}_t \mathbf{v}_t^\top$ and $w_t \mathbf{v}_t \mathbf{v}_t^\top$, respectively, where $w_t \geq 0$ is a parameter to be chosen later.

To apply Lemma 3.6, we need $\eta \langle \mathbf{X}_t^{1/2}, -w_t \mathbf{v}_t \mathbf{v}_t^\top \rangle \geq -1/2$ and $\eta \langle \mathbf{Y}_t^{1/2}, w_t \mathbf{v}_t \mathbf{v}_t^\top \rangle \geq -1/2$. The second condition always holds as the LHS is non-negative. For the first condition, we choose

$$\eta = \frac{\tilde{\eta}}{\beta} \quad \text{and} \quad w_t = \frac{\beta}{\langle \mathbf{Y}_t, \mathbf{v}_t \mathbf{v}_t^\top \rangle \|\mathbf{v}_t\|},$$

where $\tilde{\eta} \in (0, 1/2]$, $\beta > 0$ are parameters to be chosen later. In particular, β will be chosen so that condition (a') of Lemma 3.1 holds. Note that by Lemma 3.7 and the fact $\langle \mathbf{X}_t, \mathbf{v}_t \mathbf{v}_t^\top \rangle \leq \langle \mathbf{Y}_t, \mathbf{v}_t \mathbf{v}_t^\top \rangle$, we do have

$$\eta \langle \mathbf{X}_t^{1/2}, w_t \mathbf{v}_t \mathbf{v}_t^\top \rangle = \frac{\tilde{\eta}}{\beta} \frac{\beta \langle \mathbf{X}_t^{1/2}, \mathbf{v}_t \mathbf{v}_t^\top \rangle}{\langle \mathbf{Y}_t, \mathbf{v}_t \mathbf{v}_t^\top \rangle^{1/2} \|\mathbf{v}_t\|} \leq \tilde{\eta} \leq \frac{1}{2}.$$

To determine β , consider condition (a') of Lemma 3.1. We have

$$\begin{aligned}\frac{1}{T} \sum_{t=0}^{T-1} \langle w_t \mathbf{v}_t \mathbf{v}_t^\top, \mathbf{X}_t \rangle &= \frac{1}{T} \sum_{t=0}^{T-1} \frac{\langle \beta \mathbf{v}_t \mathbf{v}_t^\top, \mathbf{X}_t \rangle}{\langle \mathbf{Y}_t, \mathbf{v}_t \mathbf{v}_t^\top \rangle^{1/2} \|\mathbf{v}_t\|} \leq \frac{\beta}{T} \sum_{t=0}^{T-1} \frac{\langle \mathbf{v}_t \mathbf{v}_t^\top, \mathbf{Y}_t \rangle^{1/2}}{\|\mathbf{v}_t\|}, \\ \frac{1}{T} \sum_{t=0}^{T-1} \langle w_t \mathbf{v}_t \mathbf{v}_t^\top, \mathbf{Y}_t \rangle &= \frac{\beta}{T} \sum_{t=0}^{T-1} \frac{\langle \mathbf{v}_t \mathbf{v}_t^\top, \mathbf{Y}_t \rangle^{1/2}}{\|\mathbf{v}_t\|}.\end{aligned}$$

Thus, to ensure condition (a') of Lemma 3.1, it suffices to choose⁴

$$\beta := \left(\frac{1}{T} \sum_{t=0}^{T-1} \frac{\langle \mathbf{v}_t \mathbf{v}_t^\top, \mathbf{Y}_t \rangle^{1/2}}{\|\mathbf{v}_t\|} \right)^{-1}$$

Then, by Lemma 3.6, for any $\mathbf{U} \in \Delta_{d \times d}$, we have

$$\begin{aligned}\frac{1}{T} R_T^X(\mathbf{U}) &\leq \frac{2\beta d^{1/2}}{\tilde{\eta}T} + \frac{2\tilde{\eta}}{T} \sum_{t=0}^{T-1} \left| \frac{\langle \mathbf{X}_t, \mathbf{v}_t \mathbf{v}_t^\top \rangle \langle \mathbf{X}_t^{1/2}, \mathbf{v}_t \mathbf{v}_t^\top \rangle}{\langle \mathbf{Y}_t, \mathbf{v}_t \mathbf{v}_t^\top \rangle \|\mathbf{v}_t\|} \right| \\ &\leq \frac{2\beta d^{1/2}}{\tilde{\eta}T} + \frac{2\tilde{\eta}}{T} \sum_{t=0}^{T-1} \frac{\langle \mathbf{X}_t, \mathbf{v}_t \mathbf{v}_t^\top \rangle^{1/2}}{\|\mathbf{v}_t\|} \\ &\leq \frac{2\beta d^{1/2}}{\tilde{\eta}T} + 2\tilde{\eta},\end{aligned}\tag{11}$$

where the last line comes from $\mathbf{Y} \preceq \mathbf{I}_d$. Similarly, we also have

$$\frac{1}{T} R_T^Y(\mathbf{U}) \leq \frac{2\beta d^{1/2}}{\tilde{\eta}T} + \frac{2\tilde{\eta}}{T} \sum_{t=0}^{T-1} \frac{\langle \mathbf{Y}_t, \mathbf{v}_t \mathbf{v}_t^\top \rangle^{1/2}}{\|\mathbf{v}_t\|} \leq \frac{2\beta d^{1/2}}{\tilde{\eta}T} + 2\tilde{\eta}, \quad \forall \mathbf{U} \in \Delta_{d \times d}.$$

Therefore, for the regrets to be bounded by $\varepsilon/2$, it suffices to choose

$$\tilde{\eta} = \frac{\eta}{8} \quad \text{and} \quad T \geq \frac{64\beta d^{1/2}}{\varepsilon^2}.$$

To complete the proof, it suffices to show $\beta \lesssim d^{1/2}$, or, equivalently, $\frac{1}{T} \sum_{t=0}^{T-1} \frac{\langle \mathbf{v}_t \mathbf{v}_t^\top, \mathbf{Y}_t \rangle^{1/2}}{\|\mathbf{v}_t\|} \gtrsim d^{-1/2}$.

By Lemma 3.6 with $\mathbf{U} = \mathbf{X}_0 = \mathbf{I}_d/d$ and the definition of R_T^X , we have

$$-\frac{1}{T} \sum_{t=0}^{T-1} \frac{\langle \mathbf{Y}_t, \mathbf{v}_t \mathbf{v}_t^\top \rangle^{1/2}}{\|\mathbf{v}_t\|} + \frac{1}{dT} \sum_{t=0}^{T-1} \frac{\|\mathbf{v}_t\|}{\langle \mathbf{Y}_t, \mathbf{v}_t \mathbf{v}_t^\top \rangle^{1/2}} \leq \frac{2\tilde{\eta}}{T} \sum_{t=0}^{T-1} \frac{\langle \mathbf{X}_t, \mathbf{v}_t \mathbf{v}_t^\top \rangle^{1/2}}{\|\mathbf{v}_t\|}.$$

⁴Note that since we choose $\eta = \tilde{\eta}/\beta$, the iterates $(\mathbf{X}_t)_t, (\mathbf{Y}_t)_t$ do not depend on β , whose only purpose is to normalize the final sparsifier. Hence, one may choose an arbitrary β , run mirror descent to obtain the RHS, and then set β .

Rearrange terms, recall $\langle \mathbf{X}_t, \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top \rangle \leq \langle \mathbf{Y}_t, \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top \rangle$, and we obtain

$$\frac{1}{dT} \sum_{t=0}^{T-1} \frac{\|\mathbf{v}_{i_t}\|}{\langle \mathbf{Y}_t, \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top \rangle^{1/2}} \leq \frac{1 + 2\tilde{\eta}}{T} \sum_{t=0}^{T-1} \frac{\langle \mathbf{Y}_t, \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top \rangle^{1/2}}{\|\mathbf{v}_{i_t}\|}.$$

Multiply both sides with $\frac{1}{T} \sum_{t=0}^{T-1} \frac{\langle \mathbf{Y}_t, \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top \rangle^{1/2}}{\|\mathbf{v}_{i_t}\|}$, and we get

$$\frac{1 + 2\tilde{\eta}}{T^2} \left(\sum_{t=0}^{T-1} \frac{\langle \mathbf{Y}_t, \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top \rangle^{1/2}}{\|\mathbf{v}_{i_t}\|} \right)^2 \geq \frac{1}{dT^2} \sum_{t=0}^{T-1} \frac{\|\mathbf{v}_{i_t}\|}{\langle \mathbf{Y}_t, \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top \rangle^{1/2}} \sum_{t=0}^{T-1} \frac{\langle \mathbf{Y}_t, \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top \rangle^{1/2}}{\|\mathbf{v}_{i_t}\|} \geq \frac{1}{d},$$

where the second inequality comes from the fact $T^2 = \left(\sum_{t=0}^{T-1} a_t^{1/2} a_t^{-1/2} \right)^2 \leq \sum_{t=0}^{T-1} a_t \sum_{t=0}^{T-1} 1/a_t$ for any $a_t > 0$. As a result, we have

$$\frac{1}{T} \sum_{t=0}^{T-1} \frac{\langle \mathbf{Y}_t, \mathbf{v}_{i_t} \mathbf{v}_{i_t}^\top \rangle^{1/2}}{\|\mathbf{v}_{i_t}\|} \geq \sqrt{\frac{1}{1 + 2\tilde{\eta}} \frac{1}{d}} \geq \sqrt{\frac{1}{2d}}.$$

Thus, $\beta \leq \sqrt{2d}$ and the total number of steps is bounded $O(d/\varepsilon^2)$. \square

References

- [AZLO15] Zeyuan Allen-Zhu, Zhenyu Liao, and Lorenzo Orecchia. Spectral Sparsification and Regret Minimization Beyond Matrix Multiplicative Updates. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, pages 237–245, New York, NY, USA, 2015. Association for Computing Machinery. event-place: Portland, Oregon, USA.
- [BSS12] Joshua Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-Ramanujan Sparsifiers. *SIAM Journal on Computing*, 41(6):1704–1721, January 2012.
- [Haz22] Elad Hazan. *Introduction to online convex optimization*. Adaptive Computation and Machine Learning. The MIT Press, Cambridge, Massachusetts London, England, second edition edition, 2022.
- [Roc70] R. Tyrrell Rockafellar. *Convex analysis*. Number 28 in Princeton mathematical series. Princeton University Press, Princeton, N.J, 1970.
- [SS11] Daniel A. Spielman and Nikhil Srivastava. Graph Sparsification by Effective Resistances. *SIAM Journal on Computing*, 40(6):1913–1926, January 2011.