Online Convex Optimization in the Random Order Model

Dan Garber [1] **Gal Korcia** [2] **Kfir Y. Levy** [2]

Abstract

Online Convex Optimization (OCO) is a powerful framework for sequential prediction, portraying the natural uncertainty inherent in data-streams as though the data were generated by an almost omniscient adversary. However, this view, which is often too pessimistic for real-world data, comes with a price. The complexity of solving many important online tasks in this adversarial framework becomes much worse than that of their offline and even stochastic counterparts. In this work we consider a natural random-order version of the OCO model, in which the adversary can choose the set of loss functions, but does not get to choose the order in which they are supplied to the learner; Instead, they are observed in uniformly random order. Focusing on two important families of online tasks, one in which the cumulative loss function is strongly convex (though individual loss functions may not even be convex), and the other being online k-PCA, we show that under standard well-conditioned-data assumptions, standard online gradient descent (OGD) methods become much more efficient in the random-order model. In particular, for the first group of tasks OGD guarantees poly-logarithmic regret. In the case of online k-PCA, OGD guarantees sublinear regret using only a rank-k SVD on each iteration and memory linear in the size of the solution.

1. Introduction

Online Convex Optimization (OCO) (Hazan, 2016; Shalev-Shwartz, 2012) has emerged in the past two decades as a powerful and popular paradigm for modeling sequential prediction problems in face of uncertainty. Its main strength is that the notion of uncertainty considered is very strong and involves worst-case scenarios, i.e., the sequence of loss functions is considered to be selected by an *adversary* who knows the prediction algorithm (perhaps up to potential randomness used for prediction). At the same time, this paradigm lands itself to tractable optimization, and theoretically efficient algorithms that provably minimize the regret are known to exist.

Of course, when considering applications of this paradigm to practical prediction tasks which involve very high dimensional settings and high-throughput rates, classical polynomial-time efficiency may be prohibitive, and there is a need for practically efficient algorithms, usually such that require at most linear (in the dimension) memory and linear runtime per prediction round.

Unfortunately, for several important online tasks such as online linear regression and online PCA, state-of-the-art methods require quadratic memory and at least quadratic runtime per round, which greatly limits their applicability to high dimensional data, see for instance (Garber, 2019a;b) and references therein. Perhaps surprisingly, in both of these cases, this is in clear contrast to their *offline* and even *stochastic* counterparts, which admit significantly more efficient algorithms, under standard well-conditionedness assumptions (e.g., moderate condition number of data matrix in linear regression, or non-negligible eigengap in covariance matrix for PCA). In particular, in both cases these efficient algorithms are simply gradient descent / stochastic gradient descent methods. It is thus natural to ask, if *Online Gradient Descent* (Zinkevich, 2003a), which requires only linear memory and linear runtime (putting aside to gradient computation time) could be shown to be efficient for online variants of such fundamental tasks.

Towards obtaining highly-efficient algorithms for online sequential prediction, in this work we consider a weaker variant of OCO, in which, while we still allow an adversary to choose the set of loss functions (with complete knowledge of the prediction algorithm), we do not allow him to control the order in which they are supplied to the learner. Instead, here we assume that the losses are observed in a random order, according to a uniformly chosen permu-

[1] Department of Industrial Engineering and Management, Technion - Israel Institute of Technology, Haifa, Israel [2] Department of Electrical Engineering, Technion - Israel Institute of Technology, Haifa, Israel. Correspondence to: Dan Garbar <dangar@technion.ac.il>, Gal Korcia <galko@campus.technion.ac.il>, Kfir Y. Levy <kfirylevy@technion.ac.il>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

tation. We refer to this model as *Random-Order Online Convex Optimization* (ROOCO). In particular, we note that ROOCO forms a natural middle ground between the easier and highly popular setting of prediction with a stochastic adversary, who chooses a distribution over loss functions from a certain family of distributions, and on every round a loss function is sampled i.i.d. from this distribution, and the more difficult standard OCO setting with an oblivious adversary, in which the adversary also controls the order of the loss functions.

It is quite clear that the proposed ROOCO model has several major drawbacks in comparison to the standard OCO model, even when focusing on practical scenarios: it clearly does not capture sequences with temporal structure, such as those we might expect to be inherit in applications such as online portfolio selection (Cover, 2011; Agarwal et al., 2006) or time-series forecasting (Anava et al., 2013). Nevertheless, we argue that many scenarios of interest do not depend inherently on time. For instance in prediction tasks in which data is generated by many independent sources (such as a web application that serves many individual and unrelated clients or a router in a highly distributed computer network), or when it is reasonable to assume that data is streamed without particular order. Such scenarios may indeed be faithfully modeled by ROOCO.

For further motivation, in Figure 1 we report results on a simple online linear regression task with the MNIST dataset (LeCun et al., 1998). We show that in the fully adversarial setting of OCO, we can easily construct pathological, yet not practically reasonable, adversarial orders over the data that make OGD incur cumulative loss significantly higher than under the much more plausible random-order scenario.

Our main algorithmic contribution is in showing that for two types of fundamental applications: the first includes prediction with sequences for which the cumulative loss is strongly convex (though each individual loss function need not even be convex), and hence generalizes as a special case the important task of online linear regression in the well-conditioned case, and the second being online k-PCA in the well-conditioned case (i.e., under spectral gap in the cumulative covariance matrix), the standard online gradient descent method minimizes the regret with high-probability while requiring only linear memory and linear runtime per round (assuming the gradient vector is given). In particular, in the first case of sequences with cumulative strong convexity, we show that OGD attains poly-logarithmic regret with high probability. In the second case of k-PCA, we show that with high probability, OGD only requires to maintain a rank-k projection matrix, which is linear in the size of the solution. Also, each iteration requires only a rank-k SVD which requires nearly linear time to compute (in the size of the solution). This significantly improves over the current state-of-the-art complexities for the fully adversarial OCO model which require quadratic memory and at least quadratic runtime, for both online linear regression and k-PCA. For sequences of convex losses, even when the cumulative loss is strongly convex, the only method known to obtain logarithmic regret in the fully-adversarial OCO setting, is the Online Newton Step method (ONS), which also require the losses to be exp-concave (Hazan et al., 2007). ONS requires in general quadratic memory to store the sum of outer products of gradients, and at least quadratic runtime to compute to so-called newton direction. For k-PCA all existing methods for the fully adversarial setting require to store in memory the cumulative covariance which requires quadratic memory and/or to compute SVD of potentially high-rank matrices (Warmuth & Kuzmin, 2006a;b; Garber et al., 2015; Allen-Zhu & Li, 2017; Carmon et al., 2019).

On the technical side, we leverage recent developments in matrix concentration bounds for sampling without replacement (Mackey et al., 2014b) as our main non-standard technical tool. For the k-PCA problem we rely on the ideas introduced in (Garber, 2019a) for $k = 1$ and later also used in (Arora & Marinov, 2019) with arbitrary k, to show that with a "warm-start" initialization (which is not difficult to obtain in our setting), the iterates of OGD on the convex relaxation, are of rank k with high probability.

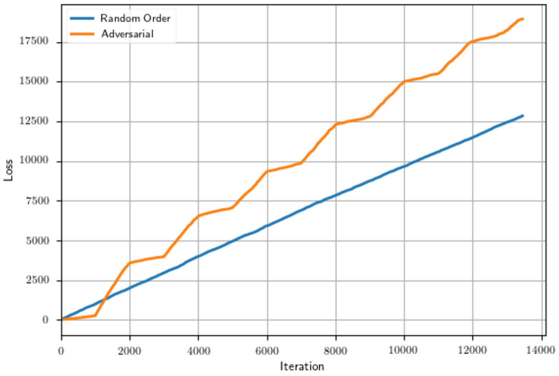


Figure 1. Linear regression with MNIST. This experiment demonstrates how the random order model gives rise to better performance of OGD. We only consider images labeled "3" and "5" (a total of 13454 images). In the adversarial setting, we partition our data into alternating consecutive homogeneous blocks of size 1000, where each block either consists of only "3"-images or only "5"-images. It can be seen that the cumulative loss for the adversarial order case is much worse compared to the random order case.

2. Preliminaries

Given any positive integer m we use the standard notations $[m] \coloneqq \{1, 2, \ldots, m\}$ and $[[m]] \coloneqq \{0, 1, \ldots, m\}$. Matrices are denoted by bold capital letters, vectors are bold lowercase and scalars or entries are not bold. $\mathbf{e}_i$ will denote the ith standard basis vector in $\mathbb{R}^d$, equal to 1 in component i and 0 everywhere else. The dimension of $\mathbf{e}_i$ will always be clear from context. For vectors in $\mathbb{R}^d$ we let $\|\cdot\|$ denote the Euclidean norm. The spectral norm of a matrix is denoted by $\|\mathbf{X}\| = \sigma_{\max}(\mathbf{X})$ where $\sigma_{\max}(\mathbf{X})$ is the maximum singular value of the matrix $\mathbf{X} \in \mathbb{R}^{d\times d}$. The Euclidean inner product between two matrices is $\langle \mathbf{X}, \mathbf{Y}\rangle = \mathrm{Tr}(\mathbf{X}^\top \mathbf{Y})$, and the corresponding Euclidean norm, called the Frobenius norm is denoted $\|\mathbf{X}\|_F$. That is, $\|\mathbf{X}\|_F = \langle \mathbf{X}, \mathbf{X}\rangle^{1/2}$. The symbols $\lambda_{\max}(\mathbf{X})$ and $\lambda_{\min}(\mathbf{X})$ refer to the algebraic maximum and minimum eigenvalues of a matrix $\mathbf{X} \in \mathbb{R}^{d\times d}$.

Definition 2.1. *Given a convex and compact set $\mathcal{K} \subset \mathbb{R}^d$ and a real-valued function f, differentiable over $\mathcal{K}$, we say f is* G-Lipschitz *over $\mathcal{K}$ if for all $\mathbf{x} \in \mathcal{K}$, we have $\|\nabla f(\mathbf{x})\| \le G$.*

In particular, if f is convex, differentiable and G-Lipschitz over a convex and compact $\mathcal{K}$, we have that for any $\mathbf{x}, \mathbf{y} \in \mathcal{K}$, $f(\mathbf{x}) - f(\mathbf{y}) \le (\mathbf{x} - \mathbf{y})^\top \nabla f(\mathbf{x}) \le G\|\mathbf{x} - \mathbf{y}\|$.

Definition 2.2. *Given a convex and compact set $\mathcal{K} \subset \mathbb{R}^d$ and a real-valued function f which is differentiable over $\mathcal{K}$, we say f is* α-strongly convex *over $\mathcal{K}$ if for all $\mathbf{x}, \mathbf{y} \in \mathcal{K}$, $f(\mathbf{x}) \le f(\mathbf{y}) + (\mathbf{x} - \mathbf{y})^\top \nabla f(\mathbf{x}) - \frac{\alpha}{2}\|\mathbf{x} - \mathbf{y}\|^2$.*

Moreover, if f is also twice differentiable over $\mathcal{K}$, then, for any $\mathbf{x} \in \mathcal{K}$, $\nabla^2 f(\mathbf{x}) \succeq \alpha \mathbf{I}$.

Definition 2.3. *Given a convex and compact set $\mathcal{K} \subset \mathbb{R}^d$ and a real-valued function f which is differentiable over $\mathcal{K}$, we say f is* b-smooth *if and only if its gradient is b-Lipschitz, for some $b > 0$ over $\mathcal{K}$, i.e., for all $\mathbf{x}, \mathbf{y} \in \mathcal{K}$, $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le b \cdot \|\mathbf{x} - \mathbf{y}\|$. If the function f is twice differentiable over $\mathcal{K}$, the above condition is equivalent to the following condition on the Hessian, $\nabla^2 f(\mathbf{x}) \preceq b\mathbf{I}$.*

3. Random Order Model

Online Convex Optimization. Online Convex Optimization is posed as a repeated game between a learner and an adversary. First, the adversary chooses T convex loss functions $f_1, \ldots, f_T$ from a fixed convex and compact domain $\mathcal{K} \subset \mathbb{R}^d$ to the reals[1]. On each iteration $t \in [T]$, the learner makes a prediction $\mathbf{w}_t \in \mathcal{K}$ using some online algorithm $\mathcal{A}$, and suffers a loss of $f_t(\mathbf{w}_t)$. In addition, we assume that the Euclidean diameter of the set $\mathcal{K}$ is bounded by D, and that for any $\mathbf{w} \in \mathcal{K}$ and all $t \in [T]$, it holds that $\|\nabla f_t(\mathbf{w})\| \le G$, for some $G > 0$. The goal of the learner is to minimize the regret, which is given by

$$\mathrm{regret}_T(\mathcal{A}) = \sum_{t=1}^{T} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{w}).$$

Perhaps the simplest algorithm for OCO is the Online Gradient Descent (OGD) algorithm, which is an adaptation of the classical projected (sub)gradient descent for offline convex optimization. On each iteration, the algorithm picks the next prediction as follows:

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{K}}(\mathbf{w}_t - \eta_t \nabla f_t(\mathbf{w}_t)), \tag{1}$$

where for any $\mathbf{w}_0 \in \mathbb{R}^d$ we denote

$$\Pi_{\mathcal{K}}(\mathbf{w}_0) = \arg\min_{\mathbf{w}\in\mathcal{K}} \|\mathbf{w} - \mathbf{w}_0\| .$$

Throughout the paper, we consider a variant of this standard setting: online convex optimization in a *random order model*, in which the adversary can choose the set of loss functions, but cannot control the order in which these loss functions are presented to the learner. Instead, these losses are observed in a uniformly random order. This is formally stated in the following assumption.

Assumption 3.1 (random order assumption)**.** *We say that a sequence of T functions satisfies the* random order *assumption if each of the functions is chosen by an adversary, but their order is determined be a random permutation over T elements.*

3.1. Cumulative Strongly Convex Sequences

Our first suite of results concerns a family of online tasks, which generalizes online linear and logistic regression in the random order model (Assumption 3.1). As discussed in the introduction, our aim is to obtain poly-logarithmic regret for a sequence of functions whose average is strongly convex, while each of the functions might not even be convex by itself. In this subsection, the learner's prediction is generally assumed to be a vector in $\mathbb{R}^d$, and is denoted $\mathbf{w}_t$. Note that the Hessian of each loss function is a $d \times d$ real matrix.

Throughout the discussion, we always assume that the following holds for the loss functions. Given a convex and compact set $\mathcal{K} \subset \mathbb{R}^d$, each of the functions $f_1, \ldots f_T$ is twice differentiable, b-smooth and with gradients bounded in ℓ_2-norm over the domain $\mathcal{K}$. Moreover D is the ℓ_2-diameter of $\mathcal{K}$. Next we present the main assumption that refers to the Hessians of the loss functions and define "well-conditioned data".

Assumption 3.2 (cumulative α-strong convexity)**.** *Given $\alpha \in \mathbb{R}_+$ and a convex and compact domain $\mathcal{K} \subset \mathbb{R}^d$, we say that a sequence of T twice differentiable functions*

[1] In this work we focus only on *oblivious* adversaries which choose the entire sequence of losses beforehand, but with knowledge of the algorithm used by the learner.

$f_1, \ldots, f_T$ over the domain $\mathcal{K}$ satisfies the cumulative α-strong convexity *assumption, if*

$$\forall \mathbf{w} \in \mathcal{K}: \quad \frac{1}{T} \sum_{t=1}^{T} \nabla^2 f_t(\mathbf{w}) \succeq \alpha \mathbf{I}.$$

We consider three classes of functions to which our results apply.

Definition 3.3 (quadratic functions)**.** *Let $\mathbf{A}_1, \ldots, \mathbf{A}_T$ be a sequence of $d \times d$ real symmetric matrices. We define the loss function $f_t(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top \mathbf{A}_t \mathbf{w}$ for every $t \in [T]$. The sequence $f_1, \ldots, f_T$ satisfies the cumulative α-strong convexity assumption (Assumption 3.2) if $\frac{1}{T}\sum_{t=1}^{T} \mathbf{A}_t \succeq \alpha \mathbf{I}$.*

Definition 3.4 (strongly convex function composed with linear transformation)**.** *Let $\mathbf{A}_1, \ldots, \mathbf{A}_T$ be a sequence of $d \times d$ real matrices, and let g be a β-strongly convex function. We define the loss function $f_t(\mathbf{w}) = g_t(\mathbf{A}_t \mathbf{w})$ for every $t \in [T]$. The sequence $f_1, \ldots, f_T$ satisfies the cumulative α-strong convexity assumption (Assumption 3.2) if $\frac{1}{T}\sum_{t=1}^{T} \mathbf{A}_t^\top \mathbf{A}_t \succeq \frac{\alpha}{\beta} \mathbf{I}$.*

Definition 3.5 (general cumulative strongly convex sequences)**.** *Let $f_1, \ldots, f_T$ be a sequence of functions from $\mathcal{K}$ to $\mathbb{R}$ which are twice-differential over $\mathcal{K}$ (but not necessarily convex). The sequence $f_1, \ldots, f_T$ satisfies the cumulative α-strong convexity assumption (Assumption 3.2) if the averaged function $\frac{1}{T}\sum_{t=1}^{T} f_t(\cdot)$ is α-strongly convex over $\mathcal{K}$.*

Main Results Next we present our main results. In Table 1 we summarize and compare our results to relevant results in two related models. The first is the stochastic model, where loss functions are drawn i.i.d. from a fixed distribution. The second is the adversarial model, where the order of the samples arrives at an arbitrary (possibly adversarial) order. Recall that individual losses are allowed to be non-convex, and we only assume the cumulative loss to be convex.

The following theorem states our main result of this subsection, suggesting that under the random order model (Assumption 3.1), and for a sequence of loss functions that satisfies the cumulative strong convexity assumption (Assumption 3.2), OGD guarantees poly-logarithmic regret.

Theorem 3.6. *Let $\{f_t\}_{t \in [T]}$ be a sequence of loss functions that satisfies the random order model and cumulative α-strong convexity assumptions (Assumptions 3.1 and 3.2 accordingly). Then, for any $\delta \in (0, 1)$, Online Gradient Descent with step sizes $\eta_t = \frac{1}{\alpha t}$ for any $t \in [T]$, gives with probability at least $1 - \delta$ the following for quadratic loss functions (see Definition 3.3)*

$$regret_T = \tilde{O}\left(\frac{b^2(G + Db)G}{\alpha^3} \log\left(\frac{T}{\delta}\right) \log T\right).$$

For strongly convex functions of a linear transformation (see Definition 3.4), OGD with the same step sizes gives with probability at least $1 - \delta$

$$regret_T = \tilde{O}\left(\frac{b(G + Db)G}{\alpha^2} \log\left(\frac{T}{\delta}\right) \log T\right).$$

Finally, for general loss functions with cumulative strong convexity (see Definition 3.5), OGD with the same step sizes gives with probability at least $1 - \delta$

$$regret_T = \tilde{O}\left(\frac{b^2(G + Db)G}{\alpha^3} \left(\log \frac{T}{\delta} + d\right) \log T\right),$$

where in all results we use $\tilde{O}(\cdot)$ to suppress logarithmic dependencies on the dimension and other parameters, but not on T.

In Section 4.1 we provide a proof sketch for Theorem 3.6. The full proof is deferred to the appendix.

The next theorems show that the random order assumption (Assumption 3.1) is crucial for the improved regret bounds in Theorem 3.6. Specifically, there exists a sequences of loss functions that satisfy the cumulative strong convexity assumption (Assumption 3.2), and for which *adversarial order* makes OGD suffer linear regret in T with at least constant probability; this is proved in Theorem 3.7. Furthermore, in Theorem 3.8, we show that even the additional assumption that each of the loss functions is convex by itself (in combination with cumulative strong convexity) cannot yield a polylogarithmic regret without the random order assumption (Assumption 3.1); specifically, there exists such a sequence where adversarial order results in $\Omega(\sqrt{T})$ regret.

Theorem 3.7. *Let $\mathcal{K} = [0, 1]$. There exists a sequence $\{f_t\}_{t \in [T]}$ of quadratic loss functions from $\mathcal{K}$ to $\mathbb{R}$, which are $\Theta(1)$-smooth and satisfy $\Theta(1)$-cumulative strong convexity (Assumption 3.2), for which OGD with arbitrary positive step-sizes $\{\eta_t\}_{t \in [T]}$ and uniformly random initialization, incurs linear regret with probability at least* 0.4.

Theorem 3.8. *There exist a choice of feasible set $\mathcal{K}$ and a sequence of convex loss functions $f_1, \ldots, f_T$ which satisfy $\Theta(1)$-cumulative strong convexity (Assumption 3.2), such that any algorithm suffers $\Omega(\sqrt{T})$ regret.*

3.2. Online Principal Component Analysis

We recall the problem of online Principal Component Analysis (PCA) (Warmuth & Kuzmin, 2006b; Nie et al., 2013). Fix $1 \leq k < d$, where d is the dimension. In the online setting, for each batch of m data points that arrives on round $t \in [T]$ $\mathbf{X}_t \in \mathbb{R}^{d \times m}$ (i.e., data-points as columns), with covariance matrix $\mathbf{M}_t = \mathbf{X}_t \mathbf{X}_t^\top$, the online algorithm is required to predict a projection matrix $\mathbf{P}_t \in \mathbb{R}^{d \times k}$ onto a k-dimensional subspace of $\mathbb{R}^d$ which belongs to the set

Table 1. Regret guarantees for Online Gradient Descent under various models and assumptions. The ✗ symbol indicates that a relevant assumption is not needed. All results for the stochastic model are from (Hazan et al., 2007). The $\tilde{O}$ notation hides logarithmic dependencies in d and other parameters, but not in T. SCLT stands for strongly convex functions of linear transformations (see Definition 3.4). For our result with general loss functions we assume T is sub-exponential in d.

MODEL		STRONGLY CONVEX?	CONVEX?	REGRET
STOCHASTIC		✗	IN EXPECTATION	$\Theta(\sqrt{T})$
		IN EXPECTATION	IN EXPECTATION	$\Theta(\frac{G^2}{\alpha}\log T)$
RANDOM ORDER	QUADRATIC (DEF 3.3)	CUMULATIVE	CUMULATIVE	$O\left(\frac{b^2(G+Db)G}{\alpha^3}\log^2 T\right)$ (THEOREM 3.6)
	SCLT (DEF 3.4)	CUMULATIVE	INDIVIDUAL	$\tilde{O}\left(\frac{b(G+Db)G}{\alpha^2}\log^2 T\right)$ (THEOREM 3.6)
	GENERAL (DEF 3.5)	CUMULATIVE	CUMULATIVE	$\tilde{O}\left(\frac{b^2 d(G+Db)G}{\alpha^3}\log T\right)$ (THEOREM 3.6)
ADVERSARIAL		✗/ CUMULATIVE	✗/ CUMULATIVE	$\Theta\left(T\right)$ (THEOREM 3.7)
		✗/ CUMULATIVE	INDIVIDUAL	$\Theta(\sqrt{T})$ (THEOREM 3.8)
		✗	INDIVIDUAL	$\Theta(\sqrt{T})$
		INDIVIDUAL	INDIVIDUAL	$\Theta(\frac{G^2}{\alpha}\log T)$

$\mathcal{I}_{d,k} \equiv \{\mathbf{P} : \mathbf{P}^\top\mathbf{P} = \mathbf{I}_k\}$ *before* observing $\mathbf{X}_t$. The loss on each round t is then defined as

$$f_t(\mathbf{P}_t) := \frac{1}{2}\|\mathbf{X}_t - \mathbf{P}_t\mathbf{P}_t^\top\mathbf{X}_t\|_F^2.$$

The regret for the online PCA is given by

$$\begin{aligned}\text{regret}_T &= \sum_{t\in[T]}\|\mathbf{X}_t - \mathbf{P}_t\mathbf{P}_t^\top\mathbf{X}_t\|_F^2\\ &\quad - \min_{\mathbf{P}\in\mathcal{I}_{d,k}}\sum_{t\in[T]}\|\mathbf{X}_t - \mathbf{P}\mathbf{P}^\top\mathbf{X}_t\|_F^2\\ &= \max_{\mathbf{P}\in\mathcal{I}_{d,k}}\sum_{t\in[T]}\text{Tr}(\mathbf{P}\mathbf{P}^\top\mathbf{M}_t) - \sum_{t\in[T]}\text{Tr}(\mathbf{P}_t\mathbf{P}_t^\top\mathbf{M}_t).\end{aligned} \tag{2}$$

Since Problem (2) is non-convex, we consider a well known convex relaxation that was proposed in (Warmuth & Kuzmin, 2006b) which relaxes it by taking the convex hull

$$\mathcal{S}_{d,k} = \text{convex-hull}\{\mathbf{P}\mathbf{P}^\top \mid \mathbf{P} \in \mathcal{I}_{d,k}\}.$$

Thus, we introduce the symmetric matrix $\mathbf{W}_t \in \mathbb{R}^{d\times d}$, which belongs to the set $\mathcal{S}_{d,k}$. Belonging to $\mathcal{S}_{d,k}$ amounts to satisfying the constraints $\mathbf{W}_t \in \{\mathbf{0} \preceq \mathbf{W} \preceq \mathbf{I}, \text{Tr}(\mathbf{W}) = k\}$ (Warmuth & Kuzmin, 2006b). Using this, we can reformulate Problem (2) and get the following optimization problem with linear payoff functions, allowing us to use known algorithms such as Online Gradient Ascent (OGA).

$$\text{regret}_T = \max_{\mathbf{W}\in\mathcal{S}_{d,k}}\sum_{t\in[T]}\text{Tr}(\mathbf{W}\mathbf{M}_t) - \sum_{t\in[T]}\text{Tr}(\mathbf{W}_t\mathbf{M}_t). \tag{3}$$

Algorithm 1 batch Online Principal Component Analysis based on Online Gradient Ascent

```
input: $T$, $L$, $\mathbf{W}_0$, step sizes $\{\eta_\tau\}_{\tau\geq 0}$
for $\tau = 0, ..., T/L - 1$ do
    play $\mathbf{W}_\tau$ and observe the block-averaged covariance matrix $\mathbf{M}_\tau = \frac{1}{L}\sum_{t=\tau L+1}^{(\tau+1)L}\mathbf{X}_t\mathbf{X}_t^\top$
    set $\mathbf{W}_{\tau+1} = \Pi_{\mathcal{S}_{d,k}}(\mathbf{W}_\tau + \eta_\tau\mathbf{M}_\tau)$
end for
```

In this paper, we assume that the data $\{\mathbf{X}_t\}_{t\in[T]}$ satisfy the random order model (Assumption 3.1), i.e., the matrices $\mathbf{X}_1, \ldots, \mathbf{X}_T$ are chosen by an oblivious adversary but are supplied to the learner in a uniformly-chosen random order. Our algorithm, Algorithm 1, is a variant of OGA that predicts over blocks of L consecutive covariance matrices. We denote the average over a length-L block τ of such covariance matrices by $\mathbf{M}_\tau = \frac{1}{L}\sum_{t=\tau L+1}^{(\tau+1)L}\mathbf{M}_t = \frac{1}{L}\sum_{t=\tau L+1}^{(\tau+1)L}\mathbf{X}_t\mathbf{X}_t^\top$, and the average of all covariance matrices as $\mathbf{M} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{M}_t = \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t^\top$, where for ease of presentation, throughout the rest of this paper we assume, only slightly loosing generality, that T/L is an integer. This allows us to define the following linear payoff function over blocks:

$$F_\tau(\mathbf{W}) := \text{Tr}(\mathbf{W}\mathbf{M}_\tau).$$

Thus, the corresponding regret minimization task, formulated in Problem (3), becomes

$$\text{regret}_T = L\Big(\max_{\mathbf{W}\in\mathcal{S}_{d,k}}\sum_{\tau}\text{Tr}(\mathbf{W}\mathbf{M}_\tau) - \sum_{\tau}\text{Tr}(\mathbf{W}_\tau\mathbf{M}_\tau)\Big). \tag{4}$$

Algorithm 1 involves the projected-gradient mapping over the set $\mathcal{S}_{d,k}$ and thus requires a full-rank SVD computation in each iteration, at lease in worst-case.

Next, we show that under an eigengap assumption this computation can be replaced by a more efficient rank-k SVD computation. First (in Lemma 3.9), we explore how the projection onto $\mathcal{S}_{d,k}$ is constructed. We demonstrate that an eigengap assumption suffices for the projection to have rank at most k.

Lemma 3.9. *Let $\mathbf{Y}$ be a symmetric matrix in $\mathbb{R}^{d\times d}$ and write its eigen-decomposition as $\mathbf{Y} = \sum_{i=1}^{d} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$. Suppose that $\mathbf{Y}$ admits an eigengap $\lambda_k(\mathbf{Y}) - \lambda_{k+1}(\mathbf{Y}) \geq 1$. Then, it holds that the Euclidean projection of $\mathbf{Y}$ onto the set $\mathcal{S}_{d,k}$ is of the form:*

$$\Pi_{\mathcal{S}_{d,k}}[\mathbf{Y}] = \sum_{i=1}^{k} \mathbf{v}_i \mathbf{v}_i^\top.$$

From this lemma we make the following observation, relating between the SVD of a point to project, and the resulting projection.

Observation 3.10 (Low-rank projection, low-rank SVD)**.** *Let $\mathbf{Y}$ be a symmetric $\mathbb{R}^{d\times d}$ matrix. If rank$(\Pi_{\mathcal{S}_{d,k}}[\mathbf{Y}]) \leq k$, then only the top k components in the SVD of $\mathbf{Y}$ are required to compute the projection. Hence, only a rank-k SVD of $\mathbf{Y}$ is required.*

With the above observation in hand, we present the main results of this section. As reviewed above, it is known that the online PCA problem, when cast as online linear optimization over $\mathcal{S}_{d,k}$, is learnable via a standard variant of Online Gradient Ascent, which achieves an $O(\sqrt{T})$ regret bound, but requires a full SVD computation on each iteration. Next we show that given a sequence of covariance matrices which admits an eigengap with parameter $s > 0$, when partitioning the covariance matrices into length-L blocks with a suitable choice of L, and initializing OGA with a proper ”warm-start” matrix (which is straightforward to obtain in our random-order setting), the full SVD computation can be avoided and replaced by only a rank-k SVD computation per iteration (see Theorem 3.11).

Throughout the rest of this paper we assume that the following holds for the covariance matrices. All matrices $\mathbf{M}_t$ for $t \in [T]$ satisfy $\|\mathbf{M}_t\| \leq b$ for some $b > 0$ and also $\|\mathbf{M}_t\|_F \leq C$ for some $C > 0$. Recall also our assumption that T/L is an integer, where L is the size of the blocks used to group the iterations.

Theorem 3.11 (convergence of Algorithm 1)**.** *Let $\delta \in (0,1)$ and suppose the data $\{\mathbf{M}_t = \mathbf{X}_t\mathbf{X}_t^\top\}_{t\in[T]}$ satisfy the random order assumption (Assumption 3.1). Additionally, suppose the average $\mathbf{M} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{M}_t$ admits an eigengap $s(\mathbf{M}) := \lambda_k(\mathbf{M}) - \lambda_{k+1}(\mathbf{M}) \geq \rho > 0$. Let $\mathbf{W}^* \in \mathcal{S}_{d,k}$ be the optimal solution of Problem (4). Consider partitioning the matrices $\{\mathbf{M}_t\}_{t\in[T]}$ into length-L blocks with block size $L = \Theta((b^2/\zeta^2)\log(d/\delta))$ [2], and running OGA for T/L iterations with step sizes $\{\eta_\tau\}_{\tau\in[[T/L-1]]}$ with the first iterate $\mathbf{W}_0 \in \mathcal{S}_{d,k}$ satisfying $\|\mathbf{W}_0 - \mathbf{W}^*\|_F \leq \gamma\zeta$, for $\zeta = \rho/(2+6b\gamma)$ and $\gamma = \Theta(\sqrt{k}/\rho)$, and step-sizes $\eta_\tau = \Theta(\sqrt{k}/(C\sqrt{T/L}))$*

For T large enough it holds with probability at least $1-\delta$ that

$$\textit{regret}_T = O((b/\zeta)C\sqrt{kT\log(d/\delta)})$$

and

$$\forall \tau \in [T/L-1] : \textit{rank}(\mathbf{W}_\tau) = k.$$

Note that Theorem 3.11 relies crucially on an *eigengap assumption*, asserting that the k largest eigenvalues of the average of all covariance matrices are strictly larger than all other eigenvalues. This assumption of a non-negligible eigengap in the covariance matrix is very natural for PCA, often observed in practice, and essential for the proof. In Section 4.2 we provide a proof sketch for Theorem 3.11. The full proof is deferred to the appendix.

3.2.1. Computing a ”warm-start” matrix

We now discuss the possibility of satisfying the ”warm-start” requirement in Theorem 3.11. Recall that $\mathbf{W}^*$, the optimizer of Problem (4), is the projection of the whole data sequence $\mathbf{M}$ onto the set $\mathcal{S}_{d,k}$, which corresponds to the top k eigenvectors of $\mathbf{M}$. To compute a warm-start, we can take the covariance matrix over a block $\mathbf{M}'$ of size at least $L_0 := O\left(\frac{b^2}{\zeta_0^2}\log\frac{dT}{\delta}\right)$ from the data, where ζ is a parameter to be determined, and similarly compute the projection of $\mathbf{M}'$ onto the set $\mathcal{S}_{d,k}$, which we denote $\mathbf{W}_0$. That is, $\mathbf{W}_0$ corresponds to the top k eigenvectors of $\mathbf{M}'$.

Using Davis-Kahan $\sin\theta$ Theorem (see for instance Theorem 2 in (Yu et al., 2014)), we have that

$$\|\mathbf{W}_0 - \mathbf{W}^*\|_F \leq \frac{\sqrt{k}\|\mathbf{M}-\mathbf{M}'\|}{\lambda_k(\mathbf{M}) - \lambda_{k+1}(\mathbf{M})} \underset{(a)}{\leq} \frac{\sqrt{k}}{s(\mathbf{M})}\zeta \leq \frac{\sqrt{k}}{\rho}\zeta_0,$$

where (a) follows from the choice of L_0 together with an appropriate concentration bound (see Lemma A.11 in the appendix). In particular, for $\zeta_0 = \frac{\rho\gamma\zeta}{\sqrt{k}}$, $\mathbf{W}_0$ satisfies the requirements of Theorem 3.11 with probability at least $1 - O(\delta)$.

4. Proofs: Main Ideas and Techniques

4.1. Proof Sketch for cumulative strongly convex losses

Here we provide a proof sketch for Theorem 3.6 that discusses the case of cumulative strongly convex losses under the random order assumption (Assumption 3.1).

[2] For ease of presentation we assume T/L is an integer.

Proof Sketch of Theorem 3.6. The first idea in our analysis is to show that the regret of the original loss sequence can be related to a loss of a batched loss sequence. Thus, letting L be block size whose value will be set later on, we can define a sequence of batched losses, $\forall \tau \in [[\lceil T/L \rceil - 1]]$,

$$F_\tau(\mathbf{w}) := \frac{1}{L} \sum_{t=\tau L}^{(\tau+1)L} f_t(\mathbf{w}) \; .$$

Also, for any $\tau \in [[\lceil T/L \rceil - 1]]$ we denote by $\tilde{\mathbf{w}}_\tau$ the value of $\mathbf{w}_t$ at the beginning of each block τ.

Using the above definition together with the OGD update rule, we can show that,

$$\sum_{t=1}^{T} f_t(\mathbf{w}_t) \le L \sum_{\tau=0}^{\lceil T/L \rceil} F_\tau(\tilde{\mathbf{w}}_\tau) + \frac{G^2 L}{\alpha}(1 + \log T) \; ,$$

and note that the second term in the above expression scales with the block size L.

The second part of our analysis shows that for an appropriately large block size $L \ge L_0$, each batched loss function $F_\tau(\cdot)$ is $(\alpha/2)$-strongly convex with high probability. To do so we use concentration inequalities for the Hessian of each batched loss $\nabla^2 F_\tau(\cdot)$. Note that since the losses arrive according to a random order permutation, this requires us to build upon specialized Matrix concentration inequalities (Mackey et al., 2014a) that apply to random matrix permutations in the case of quadratic or general functions and a version of Matrix Chernoff (Tropp, 2011) for the case of strongly convex functions of linear transformations. For quadratic loss functions (but not necessarily convex), the Hessian is constant and it is enough to take $L_0 = \tilde{O}\left(\frac{b^2}{\alpha^2} \log(T/\delta)\right)$. For strongly convex functions of linear transformations, $L_0 = \tilde{O}\left(\frac{b}{\alpha} \log(T/\delta)\right)$ suffices, since the losses are convex and hence the Hessians are positive semidefinite, which allows to apply tighter Matrix-Chernoff bounds. For general losses, for which the Hessians cannot be treated as constant over the feasible set $\mathcal{K}$, we need to take $L_0 = \tilde{O}\left(\frac{db^2}{\alpha^2} \log(T/\delta)\right)$ which follows from using a discretization argument of the set $\mathcal{K}$ and using the union-bound.

The last part of the proof shows that we can use this strong-convexity of the blocks and translate it into a regret bound of $O(\frac{(G+Db)GL}{\alpha} \log T)$ for the original sequence of losses and decision points. Using the appropriate $L := L_0$ concludes the proof.

It is highly important to emphasize that the partitioning of the losses into blocks of length L is only for the purpose of the regret analysis, while our application of OGD is directly on the original sequence of losses. □

4.2. Proof Sketch for online PCA

Here we provide a proof sketch for Theorem 3.11 that discusses the case of Online PCA. In the sketch we focus on showing that with high probability,

$$\forall \tau \in [T/L - 1] : \text{rank}(\mathbf{W}_\tau) = k.$$

Given the above holds, the regret analysis for showing $O(\sqrt{T})$ regret is quite standard (see e.g. (Hazan, 2016)).

Proof Sketch of Theorem 3.11. The first part of the proof shows that with probability at least $1 - \delta$ we can bound the spectral norm distance between the average covariance matrix and every batch of size $L \ge L_0$, i.e., $\forall \tau \in [[\lceil T/L \rceil - 1]]$,

$$\left\| \frac{1}{T} \sum_{t=1}^{T} \mathbf{M}_t - \frac{1}{L} \sum_{t=\tau L+1}^{(\tau+1)L} \mathbf{M}_t \right\| = \|\mathbf{M} - \mathbf{M}_\tau\| \le \zeta \; ,$$

where $L_0 = O\left(\frac{b^2}{\zeta^2} \log \frac{dT}{\delta}\right)$.

Conditioning on the event that the above inequality holds, we are then able to show the following,

$$\|\mathbf{W}^* \mathbf{M} \mathbf{W}^* - \mathbf{W}_\tau \mathbf{M} \mathbf{W}_\tau\| = O(b\gamma\zeta). \tag{5}$$

The proof of the above inequality relies on analyzing the OGA update rule combined with exploiting the eigengap property of the data Matrix $\mathbf{M}$. The intuition is that OGA, when initialized with a "warm-start" matrix, produces with high probability iterates that remain close enough to the optimum.

The next part of the proof works as follows: We assume by induction that for some block τ we have $\text{rank}(\mathbf{W}_\tau) = k$, and use it to show that $\text{rank}(\mathbf{W}_{\tau+1}) = k$ holds as well. This obviously holds for $\tau = 0$ by the assumption of the theorem.

In order to show that the induction step holds we examine the unique structure of the Euclidean projection onto the set $\mathcal{S}_{d,k}$, as captured in Lemma 3.9. We show that $\text{rank}(\mathbf{W}_\tau) = k$ together with Eq. (5) implies that

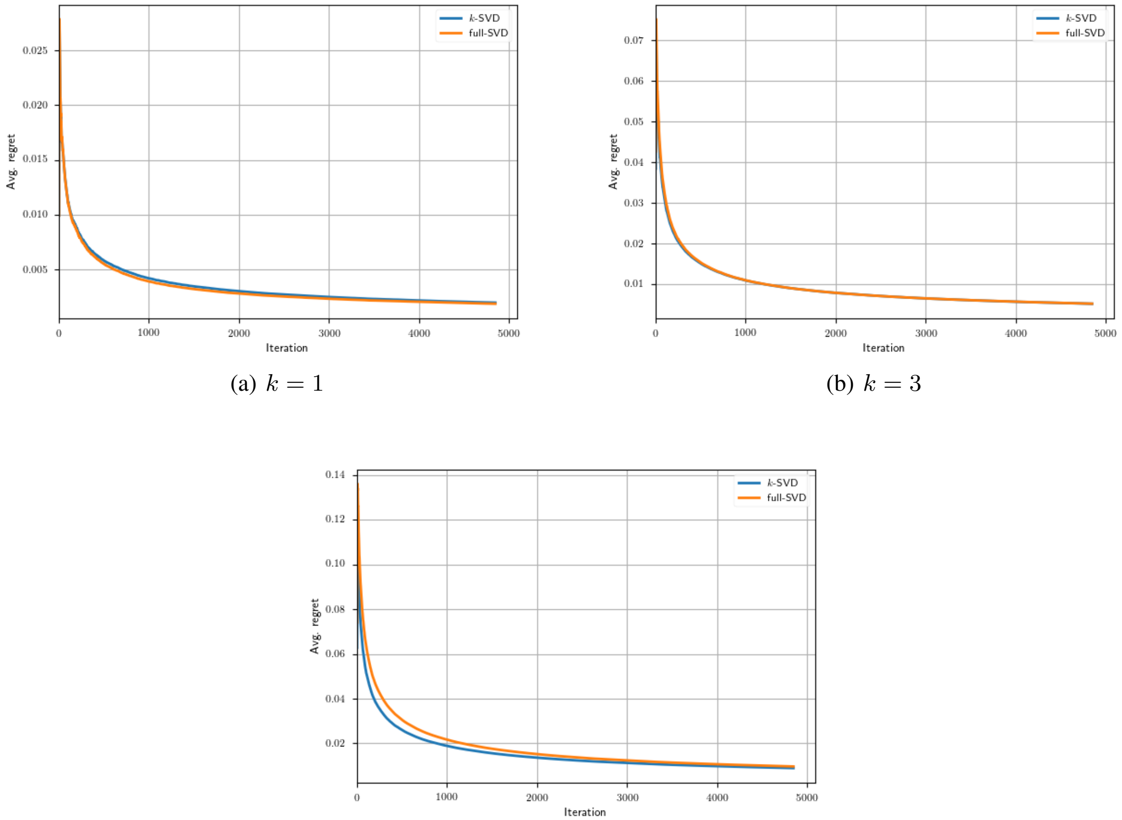
$$\lambda_k(\mathbf{W}_\tau + \eta_\tau \mathbf{M}_\tau) \ge 1 + \lambda_{k+1}(\mathbf{W}_\tau + \eta_\tau \mathbf{M}_\tau). \tag{6}$$

Combining Eq. (6) together with Lemma 3.9, implies that $\text{rank}(\mathbf{W}_{\tau+1}) = k$, and the low-rank property holds also for the next iterate.

□

5. Experiments

We conduct two sets of experiments demonstrating our theoretical findings for the online PCA problem. (Recall also our experiment from the introduction, which considers the online linear regression problem.)



(a) $k = 1$

(b) $k = 3$

(c) $k = 7$

Figure 2. Average regret of full-SVD and k-SVD on synthetic data for $k = \{1, 3, 7\}$. The case of $k = 2$ behaves similarly (see Appendix).

5.1. k-SVD vs. Full-SVD

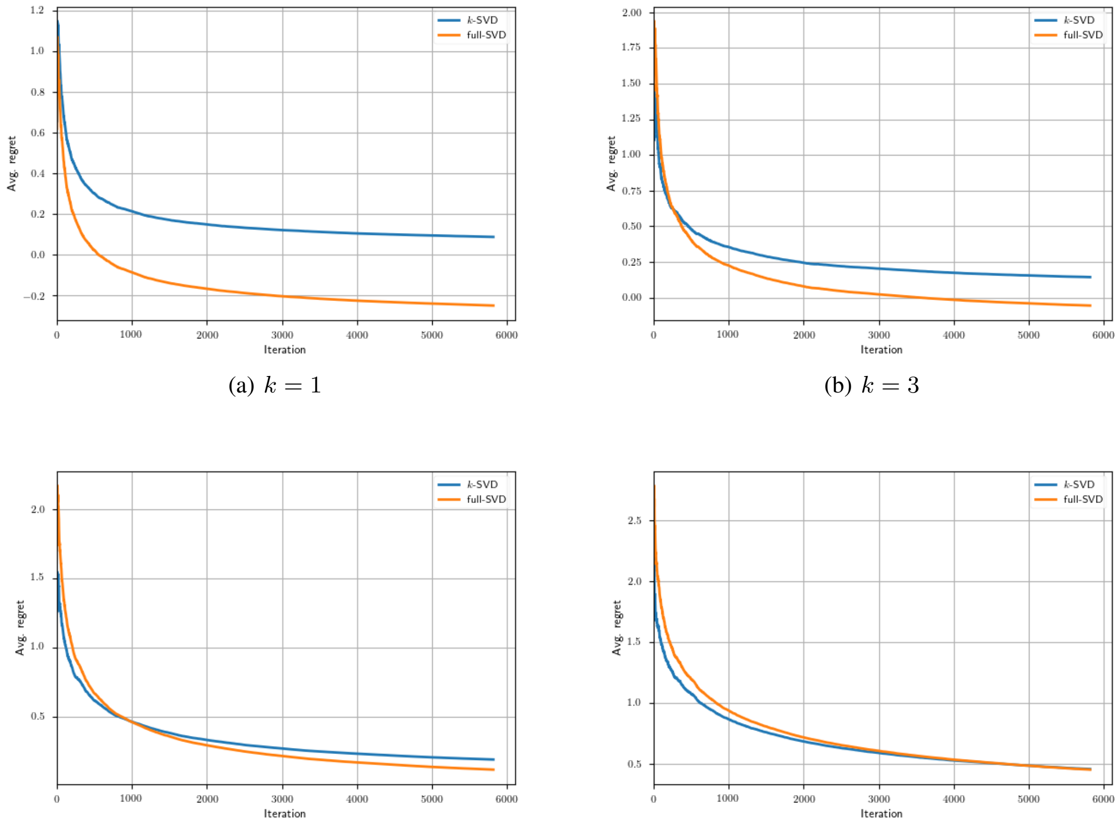
We compare the average regret of two algorithms: the first is Algorithm 1 , where $\mathbf{W}_{\tau+1}$ is computed via rank-k SVD. The second is the same algorithm , but where $\mathbf{W}_{\tau+1}$ is computed via full SVD. We consider the following datasets, where the samples (and therefore also loss functions) arrive in a random order.

Table 2. Results for synthetic dataset. "bad projections" percentage, i.e. percentage of number of iterations in which the projection in Algorithm 1 has rank greater than k, for different values of k and with block size $L = 10$.

Syn Dataset	$k = 1$	$k = 2$	$k = 3$	$k = 7$
Bad Proj[%]	0.0082	0.012	0.066	0.37

Synthetic Data A synthetic dataset generated as follows. The data is sampled from a multi-variate Normal distribution with zero mean and diagonal covariance matrix Σ. For each value of k, we have $\Sigma_{i,i} = 1$ for $1 \leq i \leq k$ and $\Sigma_{i,i} = gap \times 2^{-i \times 0.1}$ for $k+1 \leq i \leq d$. In our experiments $gap = 0.1$, $k = \{1, 2, 3, 7\}$, $d = 1000$, and the window size is $L = 10$. We use 3% of the data to compute the initialization $\mathbf{W}_0$, and step sizes $\eta_t = \frac{1}{\sqrt{t}}$. The results on the synthetic data set are the average of 10 experiments and can be found in Figure 2 and in Table 2. It can be seen that even though the dimensionality is high, using low rank-SVD for the projections performs comparably to using full SVD.

MNIST we use the training set of the MNIST handwritten digit recognition dataset, which contains 60,000 28×28 images, which we split into 58200 images for testing, and 1800 images (3%) are used to compute the initialization $\mathbf{W}_0$ and step sizes $\eta_t = \frac{1}{10^3\sqrt{t}}$. We pre-normalized the data by mean centering the feature vectors and scaling each feature by its standard deviation and average the results of 20 experiments. We set $L = 20$ and $k = \{1, 3, 7, 15\}$. The results on the synthetic data set can be found in Figure 3 and in Table 3. Similarly to the experiment on synthetic data, it can be seen that even though the dimensionality is high, low rank SVD for the projections performs comparably to using



(a) $k = 1$

(b) $k = 3$

(c) $k = 7$

(d) $k = 15$

Figure 3. Average regret of full-SVD and k-SVD on MNIST data for $k = \{1, 3, 7, 15\}$.

full SVD.

Table 3. Results for MNIST dataset. ”bad projections” percentage, i.e. percentage of number of iterations in which the projection in Algorithm 1 has rank greater than k, for different values of k and with block size $L = 20$.

MNIST Dataset	$k = 1$	$k = 3$	$k = 7$	$k = 15$
Bad Proj[%]	0	0.23	0.21	0.14

Acknowledgments

Dan Garber is supported by the ISRAEL SCIENCE FOUNDATION (grant No. 1108/18).

References

Agarwal, A., Hazan, E., Kale, S., and Schapire, R. E. Algorithms for portfolio management based on the newton method. In *Proceedings of the 23rd international conference on Machine learning*, pp. 9–16, 2006.

Allen-Zhu, Z. and Li, Y. Follow the compressed leader: Faster online learning of eigenvectors and faster mmwu. In *International Conference on Machine Learning*, pp. 116–125, 2017.

Anava, O., Hazan, E., Mannor, S., and Shamir, O. Online learning for time series prediction. In *Conference on learning theory*, pp. 172–184, 2013.

Arora, R. and Marinov, T. V. Efficient convex relaxations for streaming PCA. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 10496–10505, 2019.

Arora, R., Cotter, A., and Srebro, N. Stochastic optimization of PCA with capped MSG. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, pp. 1815–1823, Red Hook, NY, USA, 2013.

Carmon, Y., Duchi, J. C., Sidford, A., and Tian, K. A rank-1 sketch for matrix multiplicative weights. In Beygelzimer, A. and Hsu, D. (eds.), *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pp. 589–623. PMLR, 2019.

Cover, T. M. Universal portfolios. In *The Kelly Capital Growth Investment Criterion: Theory and Practice*, pp. 181–209. World Scientific, 2011.

Garber, D. On the regret minimization of nonconvex online gradient ascent for online pca. In *Conference on Learning Theory*, pp. 1349–1373, 2019a.

Garber, D. Logarithmic regret for online gradient descent beyond strong convexity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 295–303, 2019b.

Garber, D., Hazan, E., and Ma, T. Online learning of eigenvectors. In *ICML*, pp. 560–568, 2015.

Hazan, E. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

Hazan, E., Agarwal, A., and Kale, S. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Mackey, L., Jordan, M. I., Chen, R. Y., Farrell, B., and Tropp, J. A. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42(3):906–945, 2014a.

Mackey, L., Jordan, M. I., Chen, R. Y., Farrell, B., Tropp, J. A., et al. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42(3):906–945, 2014b.

Nie, J., Kotlowski, W., and Warmuth, M. K. Online PCA with optimal regrets. In *24th International Conference on Algorithmic Learning Theory, ALT*, 2013.

Shalev-Shwartz, S. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

Tropp, J. A. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, 03(01n02):115–126, 2011.

Wang, W. and Lu, C. Projection onto the capped simplex. *ArXiv*, abs/1503.01002, 2015.

Warmuth, M. K. and Kuzmin, D. Online variance minimization. In *19th Annual Conference on Learning Theory, COLT*, 2006a.

Warmuth, M. K. and Kuzmin, D. Randomized PCA algorithms with regret bounds that are logarithmic in the dimension. In *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, NIPS*, 2006b.

Yu, Y., Wang, T., and Samworth, R. J. A Useful Variant of the Davis–Kahan Theorem for Statisticians. *Biometrika*, 102(2):315–323, 2014.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 928–936, 2003a.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pp. 928–936, 2003b.

A. Analysis

A.1. Cumulative Strongly Convex Sequences

In this section, we present the main components for the proof of Theorem 3.6. In the proof, we partition the (randomly permuted) loss functions into consecutive blocks of suitable (large enough) length L, and show that with high probability all of these blocks are strongly convex. Then, we apply appropriate regret analysis to each of the blocks, which crucially relies on the cumulative strong convexity property of all blocks, and get polylogarithmic regret in T.

The next three theorems bound the block size required for the cumulative strong convexity of the whole sequence to propagate to all blocks. Theorem A.1 considers the case of quadratic loss functions (see Definition 3.3), Theorem A.2 concerns the case of strongly convex functions of linear transformations (see Definition 3.4), while Theorem A.3 refers to the most general case (see Definition 3.5).

Theorem A.1. *Let $\{f_t\}_{t\in[T]}$ be a sequence of quadratic functions (see Definition 3.3) that satisfies the random order and the cumulative α-strong convexity assumptions (Assumptions 3.1 and 3.2 accordingly). Then, for every $\delta \in (0,1)$ and $\varepsilon \in (0, \alpha/2]$, with probability at least $1-\delta$ (over the random order of $f_1, \cdots, f_T$), for all $\tau \in [[T/L-1]]$ it holds that*

$$\nabla^2 F_\tau := \frac{1}{L}\sum_{t=\tau L+1}^{(\tau+1)L} \nabla^2 f_t \succeq (\alpha-\varepsilon)\mathbf{I}$$

, for any $L \geq L_0$, where $L_0 = O\left(\frac{b^2}{\varepsilon^2}\log\left(\frac{dT}{\delta}\right)\right)$.

Theorem A.2. *Let $\{f_t\}_{t\in[T]}$ be a sequence of strongly convex functions of linear transformation (see Definition 3.4) that satisfies the random order and the cumulative α-strong convexity assumptions (Assumptions 3.1 and 3.2 accordingly). Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$ (over the random order of $f_1, \cdots, f_T$), for all $\forall \tau \in [[\lceil T/L \rceil - 1]]$ it holds that*

$$\nabla^2 F_\tau := \frac{1}{L}\sum_{t=\tau L+1}^{(\tau+1)L} \nabla^2 f_t \succeq \frac{\alpha}{2}\mathbf{I}$$

for any $L \geq L_0$, where $L_0 = O\left(\frac{b}{\alpha}\log\left(\frac{dT}{\delta}\right)\right)$.

Note that for the case of quadratic functions or strongly convex functions of linear transformations, the Hessian can be treated as if it was constant over the domain $\mathcal{K}$. In order to handle the more general case where the Hessian is not necessarily fixed, we must use additional geometric tools to prove the following result.

Theorem A.3. *Let $\{f_t\}_{t\in[T]} : \mathcal{K} \to \mathbb{R}$ be a sequence of loss functions which satisfies the random order and α-strong convexity assumptions (Assumptions 3.1 and 3.2 accordingly). Assume that for every $\tau \in [[\lceil T/L \rceil - 1]]$ and $\mathbf{w} \in \mathcal{K}$, the Hessian of $F_\tau(\mathbf{w})$ is C_1-Lipschitz over $\mathcal{K}$. Then, for any $\delta > 0$ and $\varepsilon \in (0, \alpha/2)$, with probability at least $1-\delta$ (over the random order of $f_1, \ldots, f_T$), for all $\tau \in [[\lceil T/L \rceil - 1]]$ and $\mathbf{w} \in \mathcal{K}$ it holds that*

$$\nabla^2 F_\tau(\mathbf{w}) := \frac{1}{L}\sum_{t=\tau L+1}^{(\tau+1)L} \nabla^2 f_t(\mathbf{w}) \succeq \left(\alpha - \frac{3}{2}\varepsilon\right)\mathbf{I}.$$

for any $L \geq L_0$, where

$$L_0 = O\left(\frac{6b^2}{\varepsilon^2}\left[\log\left(\frac{dT}{\delta}\right) + d\log\left(C_1 D\sqrt{d}/\epsilon\right)\right]\right).$$

The proof of Theorem 3.6 applies Theorems A.1, A.3 and A.3, taking $\varepsilon = \Theta(\alpha)$.

A.2. Proof of Theorem 3.6

The proof of Theorem 3.6 is based on a standard regret analysis of OGD, which is combined with the cumulative strong convexity property. We first begin with Theorem A.4 and Lemma A.5 which bound the distance between two consecutive OGD steps. Next, Lemma A.6 and Lemma A.7 bound the regret terms and complete the proof. We start by recalling the Hilbert Projection Theorem.

Theorem A.4 (Hilbert Projection Theorem)**.** *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be closed and convex and fix $\mathbf{x} \in \mathbb{R}^d$. It holds that*

$$\forall \mathbf{y} \in \mathcal{K}: \quad \|\mathbf{y} - \Pi_{\mathcal{K}}(\mathbf{x})\|^2 \leq \|\mathbf{y} - \mathbf{x}\|^2.$$

As indicated in the statement of Theorem 3.6, we assume that $\{f_t\}_t \in [T]$ is a sequence of b-smooth and C-Lipschitz loss functions over the set $\mathcal{K}$, given in random order (Assumption 3.1). Observe the following simple fact.

Lemma A.5. *for each $t \in [T]$, it holds that*

$$\|\mathbf{w}_{t+1} - \mathbf{w}_t\| \leq \eta_t G.$$

Proof. Since $\mathbf{w}_{t+1} := \Pi_{\mathcal{K}}(\mathbf{w}_t - \eta_t \nabla f_t(\mathbf{w}_t))$, and using Theorem A.4, we have that $\|\mathbf{w}_{t+1} - \mathbf{w}_t\| \leq \|(\mathbf{w}_t - \eta_t \nabla f_t(\mathbf{w}_t)) - \mathbf{w}_t\| \leq \eta_t G$. □

Recall the definition of $F_\tau(\cdot) = \frac{1}{L}\sum_{t=\tau L}^{(\tau+1)L} f_t(\cdot)$. Then, we can reformulate the regret in the following way

$$\begin{aligned} \text{regret}_T &= \sum_{t=1}^{T} f_t(\mathbf{w}_t) - \sum_{t=1}^{T} f_t(\mathbf{w}^*) \\ &= \sum_{t=1}^{T} f_t(\mathbf{w}_t) - L\sum_{\tau=0}^{T/L-1} F_\tau(\mathbf{w}^*) \end{aligned} \tag{7}$$

For any $\tau \in [[T/L - 1]]$ we let $\tilde{\mathbf{w}}_\tau$ denote the value of $\mathbf{w}_t$ at the beginning of each block τ of size L. Next, we obtain the following bound.

Lemma A.6. *Under the assumptions of Theorem 3.6, it holds that*

$$\sum_{t=1}^{T} f_t(\mathbf{w}_t) \le L \sum_{\tau=0}^{\lceil T/L \rceil} F_\tau(\tilde{\mathbf{w}}_\tau) + \frac{G^2 L}{\alpha}(1 + \log T).$$

Proof. In order to understand the regret analysis of all T samples, we begin by reviewing the regret analysis of a length-L block τ.

$$\begin{aligned}\sum_{t=1+\tau L}^{(\tau+1)L} f_t(\mathbf{w}_t) &= \sum_{t=1+\tau L}^{(\tau+1)L} [f_t(\tilde{\mathbf{w}}_\tau) + f_t(\mathbf{w}_t) - f_t(\tilde{\mathbf{w}}_\tau)] \\ &\underset{(a)}{\le} \sum_{t=1+\tau L}^{(\tau+1)L} f_t(\tilde{\mathbf{w}}_\tau) + G \sum_{t=1+\tau L}^{(1+\tau)L} \|\mathbf{w}_t - \tilde{\mathbf{w}}_\tau\| \\ &\underset{(b)}{\le} L F_\tau(\tilde{\mathbf{w}}_\tau) + G^2 \sum_{t=1+\tau L}^{(1+\tau)L} \sum_{j=1+\tau L}^{t-1} \eta_j,\end{aligned}$$

where (a) follows since $F_\tau(\cdot)$ is convex and G-Lipschitz, and (b) follows from Lemma A.5. Generalizing the aforementioned analysis to all T functions, we have

$$\begin{aligned}\sum_{t=1}^{T} f_t(\mathbf{w}_t) &\le L \sum_{\tau=0}^{T/L-1} F_\tau(\tilde{\mathbf{w}}_\tau) + G^2 L \sum_{t=1}^{T} \eta_t \\ &\le L \sum_{\tau=0}^{T/L-1} F_\tau(\tilde{\mathbf{w}}_\tau) + \frac{G^2 L}{\alpha}(1 + \log T),\end{aligned}$$

where the last step holds from plugging-in our choice of step-sizes and since $\sum_{t=1}^{T} \frac{1}{t} \le 1 + \log T$. □

Plugging the result of Lemma A.6 into Equation (7), we have that

$$\begin{aligned}\text{regret}_T &= \sum_{t=1}^{T} f_t(\mathbf{w}_t) - \sum_{t=1}^{T} f_t(\mathbf{w}^*) \\ &\le L \sum_{\tau=0}^{T/L-1} [F_\tau(\tilde{\mathbf{w}}_\tau) - F_\tau(\mathbf{w}^*)] + \frac{G^2 L}{\alpha}(1 + \log T)\end{aligned} \tag{8}$$

Next, we bound the first term in the RHS (8).

Lemma A.7. *Denote the best fixed prediction in hindsight by $\mathbf{w}^* \in \arg\min_{\mathbf{w}\in\mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{w})$, and suppose that for all blocks τ, the function $F_\tau(\cdot)$ is α-strongly convex, for some $\alpha > 0$. Then, under the assumptions of Theorem 3.6, it holds that*

$$\begin{aligned}&L \sum_{\tau=0}^{T/L-1} [F_\tau(\tilde{\mathbf{w}}_\tau) - F_\tau(\mathbf{w}^*)] \\ &\le \frac{1}{\alpha}\left[G^2/2 + (2G + 2Db + 3D\alpha)LG\right](1 + \log T).\end{aligned}$$

Proof. By Theorem A.4, for every $t \in [T]$ it holds that

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 &\le \|\mathbf{w}_t - \eta_t \nabla f_t(\mathbf{w}_t) - \mathbf{w}^*\|^2 \\ &= \|\mathbf{w}_t - \mathbf{w}^*\|^2 + \eta_t^2 \|\nabla f_t(\mathbf{w}_t)\|^2 \\ &\quad - 2\eta_t (\mathbf{w}_t - \mathbf{w}^*)^\top \nabla f_t(\mathbf{w}_t).\end{aligned}$$

Rearranging,

$$\begin{aligned}(\mathbf{w}_t - \mathbf{w}^*)^\top \nabla f_t(\mathbf{w}_t) &\le \frac{1}{2\eta_t}\|\mathbf{w}_t - \mathbf{w}^*\|^2 + \frac{\eta_t}{2}\|\nabla f_t(\mathbf{w}_t)\|^2 \\ &\quad - \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2.\end{aligned}$$

In addition, for every $\tau \in [[\lceil T/L \rceil - 1]]$ we have that

$$\begin{aligned}(\mathbf{w}_t - \mathbf{w}^*)^\top \nabla f_t(\mathbf{w}_t) &= (\tilde{\mathbf{w}}_\tau - \mathbf{w}^*)^\top \nabla f_t(\tilde{\mathbf{w}}_\tau) \\ &\quad + (\mathbf{w}_t - \tilde{\mathbf{w}}_\tau)^\top \nabla f_t(\tilde{\mathbf{w}}_\tau) \\ &\quad + (\mathbf{w}_t - \mathbf{w}^*)^\top (\nabla f_t(\mathbf{w}_t) - \nabla f_t(\tilde{\mathbf{w}}_\tau)).\end{aligned}$$

Combining both equations and using the Cauchy-Schwarz inequality we get

$$\begin{aligned}&(\tilde{\mathbf{w}}_\tau - \mathbf{w}^*)\nabla f_t(\tilde{\mathbf{w}}_\tau) \\ &\le \frac{1}{2\eta_t}\|\mathbf{w}_t - \mathbf{w}^*\|^2 - \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 \\ &+ \frac{\eta_t}{2}\|\nabla f_t(\mathbf{w}_t)\|^2 + \|\mathbf{w}_t - \tilde{\mathbf{w}}_\tau\|\|\nabla f_t(\tilde{\mathbf{w}}_\tau)\| \\ &+ \|\mathbf{w}_t - \mathbf{w}^*\|\|\nabla f_t(\mathbf{w}_t) - \nabla f_t(\tilde{\mathbf{w}}_\tau)\| \\ &\le \frac{1}{2\eta_t}\|\mathbf{w}_t - \mathbf{w}^*\|^2 - \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 \\ &+ \frac{\eta_t}{2}\|\nabla f_t(\mathbf{w}_t)\|^2 + (G + bD)\|\mathbf{w}_t - \tilde{\mathbf{w}}_\tau\|.\end{aligned}$$

Summing over L iterations $(\tau L + 1), \ldots, ((\tau + 1)L)$ we get

$$\begin{aligned}&L(\tilde{\mathbf{w}}_\tau - \mathbf{w}^*)^\top \nabla F_\tau(\tilde{\mathbf{w}}_\tau) \\ &\le \sum_{t=\tau L+1}^{(\tau+1)L} \left[\frac{1}{2\eta_t}\|\mathbf{w}_t - \mathbf{w}^*\|^2 - \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 \right. \\ &\left. + \frac{G^2}{2}\eta_t + (G + Db)\|\mathbf{w}_t - \tilde{\mathbf{w}}_\tau\|\right].\end{aligned}$$

Since all blocks are α-strongly convex,

$$\begin{aligned}&L(F_\tau(\tilde{\mathbf{w}}_\tau) - F_\tau(\mathbf{w}^*)) \\ &\le \sum_{t=\tau L+1}^{(\tau+1)L} \left[\frac{1}{\eta_t}\|\mathbf{w}_t - \mathbf{w}^*\|^2 - \frac{1}{\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 \right. \\ &\left. + \frac{G^2}{2}\eta_t + (G + Db)\|\mathbf{w}_t - \tilde{\mathbf{w}}_\tau\|\right] - \frac{\alpha L}{2}\|\tilde{\mathbf{w}}_\tau - \mathbf{w}^*\|^2.\end{aligned}$$

Summing over all blocks gives

$$
\begin{aligned}
&L\sum_{\tau=0}^{T/L-1}[F_\tau(\tilde{\mathbf{w}}_\tau)-F_\tau(\mathbf{w}^*)]\\
&\le \overbrace{\sum_{\tau=0}^{T/L-1}\sum_{t=\tau L+1}^{(\tau+1)L}\left[\frac{1}{2\eta_t}\|\mathbf{w}_t-\mathbf{w}^*\|^2-\frac{1}{2\eta_t}\|\mathbf{w}_{t+1}-\mathbf{w}^*\|^2\right]}^{\text{(I)}}\\
&-\overbrace{\sum_{\tau=0}^{T/L-1}\frac{\alpha L}{2}\|\tilde{\mathbf{w}}_\tau-\mathbf{w}^*\|^2}^{\text{(II)}}\\
&+\sum_{\tau=0}^{T/L-1}\sum_{t=\tau L+1}^{(\tau+1)L}\left[\frac{G^2}{2}\eta_t+(G+Db)\|\mathbf{w}_t-\tilde{\mathbf{w}}_\tau\|\right]
\end{aligned}
\tag{9}
$$

Using properties of the norm, we get that

$$
\begin{aligned}
\|\mathbf{w}_t-\mathbf{w}^*\|^2 &\le \|\tilde{\mathbf{w}}_\tau-\mathbf{w}^*\|^2+\|\mathbf{w}_t-\tilde{\mathbf{w}}_\tau\|^2\\
&\quad+2\|\tilde{\mathbf{w}}_\tau-\mathbf{w}^*\|\|\mathbf{w}_t-\tilde{\mathbf{w}}_\tau\|\\
&\le \|\tilde{\mathbf{w}}_\tau-\mathbf{w}^*\|^2+3D\|\tilde{\mathbf{w}}_\tau-\mathbf{w}_t\|.
\end{aligned}
$$

Rearranging and plugging this into (I) we get

$$
\begin{aligned}
\text{(I)}-\text{(II)} &\le \sum_{\tau=0}^{T/L-1}\sum_{t=\tau L+1}^{(\tau+1)L}\left[\frac{1}{2\eta_t}\|\mathbf{w}_t-\mathbf{w}^*\|^2\right.\\
&\quad-\frac{1}{2\eta_t}\|\mathbf{w}_{t+1}\mathbf{w}^*\|^2-\frac{\alpha}{2}\|\mathbf{w}_t-\mathbf{w}^*\|^2\\
&\quad\left.+\frac{3D\alpha}{2}\|\mathbf{w}_t-\tilde{\mathbf{w}}_\tau\|\right]\\
&=\frac{1}{2}\sum_{t=1}^{T}\|\mathbf{w}_t-\mathbf{w}^*\|^2\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}-\alpha\right)\\
&\quad+\frac{3D\alpha}{2}\sum_{\tau=0}^{T/L-1}\sum_{t=\tau L+1}^{(\tau+1)L}\|\mathbf{w}_t-\tilde{\mathbf{w}}_\tau\|\\
&\le \frac{3D\alpha}{2}\sum_{\tau=0}^{T/L-1}\sum_{t=\tau L+1}^{(\tau+1)L}\|\mathbf{w}_t-\tilde{\mathbf{w}}_\tau\|,
\end{aligned}
$$

where the last step follows from plugging-in our choice of step-sizes and from defining $\frac{1}{\eta_0}:=0$. Plugging (I) − (II) back to Equation (9), we have

$$
\begin{aligned}
&\sum_{\tau=0}^{\lceil T/L\rceil}\Big[F_\tau(\tilde{\mathbf{w}}_\tau)-F_\tau(\mathbf{w}^*)\Big]\\
&\le \sum_{\tau=0}^{\lceil T/L\rceil}\sum_{t=\tau L+1}^{(\tau+1)L}\left[\frac{G^2}{2}\eta_t+\left(G+Db+\frac{3D\alpha}{2}\right)\|\mathbf{w}_t-\tilde{\mathbf{w}}_\tau\|\right]\\
&\le \frac{1}{2\alpha}\left[G^2+(2G+2Db+3D\alpha)LG\right](1+\log T).
\end{aligned}
$$

□

Back to the regret analysis, combining Lemma A.7 and Equation (8) we get that

$$
\begin{aligned}
\text{regret}_T &\le \frac{G^2L}{\alpha}(1+\log T)+L\sum_{\tau=0}^{T/L-1}F_\tau(\tilde{\mathbf{w}}_\tau)\\
&\quad-L\sum_{\tau=0}^{T/L-1}F_\tau(\mathbf{w}^*)=O\left(\frac{(G+Db)GL}{\alpha}\log T\right),
\end{aligned}
$$

where for the last step, note that $\alpha\le b$ holds by definition. In order for the conditions of Lemma A.7 to hold, we must have that all blocks satisfy the strong convexity assumption (Assumption 3.2) with high probability. We guarantee this using the smallest possible $L=O\left(\frac{b^2}{\alpha^2}\log\frac{dT}{\delta}\right)$ for the case of quadratic functions (Theorem A.1), and $L=O\left(\frac{b}{\alpha}\log\left(\frac{dT}{\delta}\right)\right)$ for the case of strongly convex functions of linear transformations (Theorem A.2). Finally, for general loss functions, we take $L=O\left(\frac{6b^2}{\alpha^2}\left[\log\left(\frac{dT}{\delta}\right)+d\log\left(C_1D\sqrt{d}/\epsilon\right)\right]\right)$ (Theorem A.3).

A.3. Proof of Theorem A.1

In the proof, we make use of the following classical matrix concentration inequality from (Mackey et al., 2014a).

Theorem A.8 (Bernstein's inequality for a combinatorial matrix sum (Mackey et al., 2014a))**.** *Consider an array $(\mathbf{A}_{jk})_{j,k=1}^n$ of matrices in $\mathbb{R}^{d\times d}$ that satisfy*

$$\sum_{j,k=1}^{n}\mathbf{A}_{jk}=\mathbf{0}\quad\textit{and}\quad\|\mathbf{A}_{jk}\|\le R\quad\forall\textit{pair }(j,k)\textit{ of indices.}$$

Define the random matrix $\mathbf{X}:=\sum_{j=1}^n\mathbf{A}_{j\pi(j)}$, where π is a uniformly random permutation on $\{1,\dots,n\}$. Then, for all $t\ge 0$

$$\Pr\{\lambda_{\max}(\mathbf{X})\ge t\}\le d\cdot\exp\left\{\frac{-t^2}{12\sigma^2+4\sqrt{2}Rt}\right\}$$

for

$$\sigma^2:=\frac{1}{n}\left\|\sum_{j,k=1}^{n}\mathbf{A}_{jk}^2\right\|.$$

The following lemma applies Theorem A.8 and gives a tail bound for a certain matrix $\mathbf{X}$ of interest.

Lemma A.9. *Let $\mathbf{M}_1,\dots,\mathbf{M}_T$ be symmetric matrices in $\mathbb{R}^{d\times d}$ that satisfy $\|\mathbf{M}_t\|\le b$ for every $t\in[T]$. Define the random matrix*

$$\mathbf{X}:=\sum_{i=1}^{L}\mathbf{M}_{\pi(i)}-\frac{L}{T}\sum_{i=1}^{T}\mathbf{M}_i,$$

where $1 \leq L \leq T$ and π is a uniformly random permutation on $\{1, \dots, T\}$. Then, for all $t \geq 0$

$$\begin{aligned}&\Pr\{\lambda_{\max}(\mathbf{X}) \geq t \text{ or } \lambda_{\min}(\mathbf{X}) \leq -t\} \\ &= O\left(d \cdot \exp\left\{\frac{-t^2}{12Lb^2 + 8\sqrt{2}bt}\right\}\right).\end{aligned}$$

Proof. Consider an array $(\mathbf{A}_{jk})_{j,k=1}^{T}$ of symmetric matrices in $\mathbb{R}^{d \times d}$ that satisfy

$$\mathbf{A}_{jk} = \left\{ \begin{array}{lr} \mathbf{M}_k - \frac{1}{T}\sum_{t=1}^{T}\mathbf{M}_t, & \forall k \in [T], 1 \leq j \leq L \\ \mathbf{0}, & \text{otherwise} \end{array} \right\}.$$

Therefore, we obtain

$$\sum_{j,k=1}^{T} \mathbf{A}_{jk} = \sum_{j=1}^{L}\sum_{k=1}^{T}\left[\mathbf{M}_k - \frac{1}{T}\sum_{t=1}^{T}\mathbf{M}_t\right] = 0.$$

Since $\|\mathbf{M}_t\| \leq b$ and $\mathbf{M}_t$ is positive semidefinite, the maximum singular value of $\mathbf{A}_{jk}$ is also bounded by $\|\mathbf{A}_{jk}\| \leq b$, where (j,k) are any pair of indices. Define the following matrix

$$\mathbf{X} := \sum_{j=1}^{T} \mathbf{A}_{j\pi(j)} = \sum_{j=1}^{L} \mathbf{M}_{\pi(j)} - \frac{L}{T}\sum_{t=1}^{T}\mathbf{M}_t.$$

Finally, we apply Theorem A.8 from (Mackey et al., 2014a) twice: once for $\mathbf{X}$ itself, which gives an upper bound on the probability that $\lambda_{\max}(\mathbf{X}) \geq t$, and once for $-\mathbf{X}$, which combined with the fact that $\lambda_{\min}(\mathbf{X}) = -\lambda_{\max}(-\mathbf{X})$ gives

$$\begin{aligned}&\mathbb{P}\{\lambda_{\max}(\mathbf{X}) \geq t \text{ or } \lambda_{\min}(\mathbf{X}) \leq -t\} \\ &\leq 2d \ \cdot \ \exp\left\{\frac{-t^2}{12\sigma^2 + 8\sqrt{2}bt}\right\}\end{aligned}$$

for

$$\begin{aligned}\sigma^2 &:= \frac{L}{T}\left\|\sum_{k=1}^{T}\left(\mathbf{M}_k - \frac{1}{T}\sum_{t=1}^{T}\mathbf{M}_t\right)^2\right\| \\ &= \frac{L}{T}\left\|\sum_{k=1}^{T}\mathbf{M}_k^2 - \frac{2}{T}\sum_{k=1}^{T}\mathbf{M}_t\sum_{t=1}^{T}\mathbf{M}_k + \frac{1}{T^2}\left(\sum_{t=1}^{T}\mathbf{M}_t\right)^2\right\|.\end{aligned}$$

Using the fact that $\|\mathbf{M}_t\| \leq b$ gives the desire result. □

Lemma A.10. *Let $\mathbf{M}_1, \dots, \mathbf{M}_T$ be symmetric matrices in $\mathbb{R}^{d \times d}$ that satisfy $\|\mathbf{M}_t\| \leq b$ for every $t \in [T]$ and also $\frac{1}{T}\sum_{t=1}^{T}\mathbf{M}_t \succeq \alpha\mathbf{I}$ for some positive constants α and b and let π be a uniformly-chosen random permutation over $\{1, \dots, T\}$. Then, for every $\delta \in (0,1)$ and $\varepsilon \in (0, \alpha/2]$, with probability at least $1 - \delta$ (over the choice of π), it holds that*

$$\frac{1}{L}\sum_{i=1}^{L}\mathbf{M}_{\pi(i)} \succeq (\alpha - \varepsilon)\mathbf{I},$$

for any $L \geq L_0$, where $L_0 = O\left(\frac{b^2}{\varepsilon^2}\log\left(\frac{dT}{\delta}\right)\right)$.

Proof. Recall the definition of the random matrix $\mathbf{X} := \sum_{i=1}^{L}\mathbf{M}_{\pi(i)} - \frac{L}{T}\sum_{i=1}^{T}\mathbf{M}_i$. Using Lemma A.9, we get that

$$\begin{aligned}&\mathbb{P}\{\lambda_{\min}(\mathbf{X}) \geq -\varepsilon L\} \geq 1 - d \cdot \exp\left\{\frac{-\varepsilon^2 L^2}{12Lb^2 + 8\sqrt{2}b\varepsilon L}\right\} \\ &\geq 1 - \left\{ \begin{array}{lc} d \cdot \exp\left\{\frac{-\varepsilon L}{16\sqrt{2}b}\right\} & \text{for } \varepsilon > \frac{3b}{2\sqrt{2}} \\ d \cdot \exp\left\{\frac{-\varepsilon^2 L}{24b^2}\right\} & \text{otherwise} \end{array} \right\}.\end{aligned} \tag{10}$$

Using Weyl's inequality for the eigenvalues, we get

$$\begin{aligned}\lambda_{\min}\left(\sum_{j=1}^{L}\mathbf{M}_{\pi(j)}\right) &\geq \lambda_{\min}(\mathbf{X}) + \lambda_{\min}\left(\frac{L}{T}\sum_{i=1}^{T}\mathbf{M}_i\right) \\ &\underset{(a)}{\geq} \lambda_{\min}(\mathbf{X}) + \frac{L}{T}\alpha T,\end{aligned} \tag{11}$$

where (a) follows from the first assumption of Theorem A.1. Since $\varepsilon \in (0, \alpha/2]$, the relevant regime in Equation (10) is the second one. Thus, combining Equations (10) and (11) gives

$$\begin{aligned}\mathbb{P}\left\{\lambda_{\min}\left(\sum_{j=1}^{L}\mathbf{M}_{\pi(j)}\right) \geq (\alpha - \varepsilon)\,L\right\}& \\ \geq 1 - d \cdot \exp\left\{\frac{-\varepsilon^2 L}{24b^2}\right\}&.\end{aligned} \tag{12}$$

Hence, we get that Equation (12) holds with probability $1 - \delta$, if L satisfies

$$L \geq \frac{24b^2}{\varepsilon^2}\log\frac{d}{\delta}.$$

□

Since the Hessian $\nabla^2 F_\tau$ of each of the blocks $\tau \in [[\lceil T/L \rceil - 1]]$ is fixed, we can apply Lemma A.10 to all blocks simultaneously. To get that all length-L blocks satisfy the desired property simultaneously with probability $1 - \delta$, we apply the lemma with $\delta' = \delta/\lceil T/L \rceil$ and take a union bound over all $\lceil T/L \rceil$ blocks. Thus, taking

$$L_0 = O\left(\frac{b^2}{\varepsilon^2}\log\frac{d}{\delta'}\right) = O\left(\frac{b^2}{\varepsilon^2}\log\frac{dT}{\delta}\right)$$

completes the proof.

A.4. Proof of Theorem A.2

The proof is simpler than that of Theorem A.1 since here We use the Matrix Chernoff tail bound from (Tropp, 2011) (see Theorem 2.2 there) that considers positive semi-definite matrices sampled without replacement, and get

$$\Pr\left(\lambda_{\min}\left(\sum_{j=1}^{L}\mathbf{M}_{\pi(j)}\right) \le t\cdot L\cdot\lambda_{\min}\left(\frac{1}{T}\sum_{i=1}^{T}\mathbf{M}_i\right)\right) \le d\cdot\exp\left\{-(1-t)^2\cdot\frac{L\cdot\lambda_{\min}(\frac{1}{T}\sum_{i=1}^{T}\mathbf{M}_i)}{2b}\right\}. \tag{13}$$

Taking $t = 1/2$, and using the fact that $\lambda_{\min}\left(\frac{1}{T}\sum_{i=1}^{T}\mathbf{M}_i\right) \ge \alpha$, to get that all length-L blocks satisfy this property simultaneously with probability at least $1-\delta$, we apply this Matrix Chernoff with $\delta' = \delta/{}^{T}\!/_{L}$ and take a union bound over all ${}^{T}\!/_{L}$ blocks. Thus, taking

$$L_0 = O\left(\frac{b}{\alpha}\log\frac{d}{\delta'}\right) = O\left(\frac{b}{\alpha}\log\frac{dT}{\delta}\right)$$

completes the proof.

A.5. Proof of Theorem A.3

The proof applies a standard discretization trick. Since the Euclidean diameter of the feasible set $\mathcal{K}$ is D, it is enclosed within a hypercube in $\mathbb{R}^d$ of side length $2D$, which we denote by K_c. We now apply discretization to the hypercube $\mathcal{K}_c$ by defining the discrete set $\mathcal{N} := \mathcal{K}_c \cap \frac{\epsilon_0}{\sqrt{d}}\mathbb{Z}^d$, for some $\epsilon_0 > 0$ to be determined later. Clearly, $|\mathcal{N}| \le (2D\sqrt{d}/\epsilon_0 + 1)^d$.

Since $\mathcal{N} \subseteq \mathcal{K}$, we can apply the same argument used in Theorem A.1 for each of the points in $\mathcal{N}$, and take a union bound. Note that for each $\mathbf{w} \in \mathcal{K}$, there exists a point $\mathbf{x} \in \mathcal{N}$ such that $\|\mathbf{w}-\mathbf{x}\|_\infty < \varepsilon_0/\sqrt{d}$, which implies in particular that $\|\mathbf{w}-\mathbf{x}\| < \varepsilon_0$. Moreover, each of the Hessians of $F_\tau(\cdot)$ is C_1-Lipschitz continuous over $\mathbb{R}^{d\times d}$ and since $F_\tau(\cdot)$ is convex, we get $\forall\tau \in [[T/L-1]]$

$$\|\nabla^2 F_\tau(\mathbf{w}) - \nabla^2 F_\tau(\mathbf{x})\| \le C_1\|\mathbf{w}-\mathbf{x}\| \le C_1\varepsilon_0.$$

Thus, we can use the same argument of Theorem A.1 to conclude that for any

$$L \ge \left(\frac{6b^2}{\varepsilon^2}\left[\log\left(\frac{dT}{\delta}\right) + d\log\left(2D\sqrt{d}/\epsilon_0 + 1\right)\right]\right).$$

the following holds with probability at least $1-\delta$ for all points $x \in \mathcal{N}$ and for all τ simultaneously:

$$\frac{1}{L}\sum_{t=\tau L+1}^{(\tau+1)L}\nabla^2 f_t(\mathbf{x}) \succeq (\alpha-\varepsilon)\mathbf{I}.$$

Combining all of the above, we get that for all $\mathbf{w} \in \mathcal{K}$ and $\tau \in [[T/L-1]]$, with probability at least $1-\delta$, it holds that

$$\frac{1}{L}\sum_{t=\tau L+1}^{(\tau+1)L}\nabla^2 f_t(\mathbf{w}) \succeq (\alpha - C_1\varepsilon_0 - \varepsilon)\mathbf{I} \succeq (\alpha - \frac{3}{2}\varepsilon)\mathbf{I}.$$

In the last step, we choose $\varepsilon_0 \le \frac{\varepsilon}{2C_1}$ so that $C_1\varepsilon_0 + \varepsilon \le \frac{3\varepsilon}{2}$. Using union bound analysis over the size $|\mathcal{N}|$, we get that the size of block, L, must satisfy

$$L \ge \left(\frac{6b^2}{\varepsilon^2}\left[\log\left(\frac{dT}{\delta}\right) + d\log\left(4C_1D\sqrt{d}/\epsilon + 1\right)\right]\right).$$

A.6. Proof of Theorem 3.7

We consider two loss functions $h_1(w) = -5w^2 + 6w - 1$ and $h_2(w) = 5w^2 - 1$, where the global minimum over $\mathcal{K} = [0,1]$ of both functions is attained at $w^* = 0$. Consider now running OGD for T iterations with arbitrary step sizes $\{\eta_t\}_{t\in[T]}$, where for iterations $t = 1,\ldots,T/4$ we set $f_t(w_t) = h_1(w_t)$, and for $t = T/4+1,\ldots,T$ we set $f_t(w_t) = h_2(w_t)$ (w.l.o.g. assume $T/4$ is an integer). It is immediate to validate that the cumulative loss in this case is 5-strongly-convex.

Now, suppose OGD in initialized with a point in the interval $[0.6,1] \subset \mathcal{K}$, which holds with probability 0.4 under our random initialization assumption. Then, the attained regret is

$$\begin{aligned}\mathrm{regret}_T &= \sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(w^*)\\ &= \underbrace{\sum_{t=1}^{T/4} h_1(w_t) - \sum_{t=1}^{T/4} h_1(w^*)}_{\text{(I)}}\\ &\quad + \underbrace{\sum_{t=T/4+1}^{T} h_2(w_t) - \sum_{t=T/4+1}^{T} h_2(w^*)}_{\text{(II)}}.\end{aligned}$$

Note that (II) is non-negative, since $h_2(w_t) \ge h_2(w^*)$ over the interval $[0,1]$. Importantly, conditioning over the random event that the initialization point is in the interval $[0.6,1]$, we have that for all $t \le T/4+1$, $w_t \in [0.6,1]$. This holds by straightforward induction: on each such iteration t for which $w_t \in [0.6,1]$ (which holds for the initialization) we have that $h_1'(w_t) \le 0$, and thus w_{t+1} is also in $[0.6,1]$. Hence, $h_1(w_t) \ge h_1(1)$ for all $t \le T/4+1$. Therefore,

$$\begin{aligned}\mathrm{regret}_T &\ge \sum_{t=1}^{T/4} h_1(w_t) - \sum_{t=1}^{T/4} h_1(w^*)\\ &\ge \sum_{t=1}^{T/4} h_1(1) - \sum_{t=1}^{T/4} h_1(0) = \frac{T}{4}\cdot 0 - \frac{T}{4}\cdot(-1) = \frac{T}{4}.\end{aligned}$$

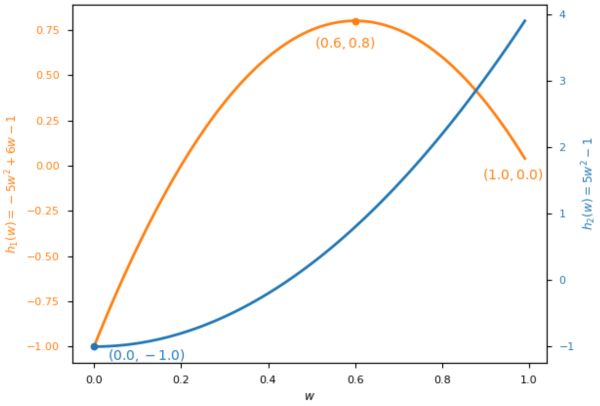


Figure 4. Graphic description of $h_1(w)$ and $h_2(w)$

A.7. Proof of Theorem 3.8

The proof follows in an almost straightforward manner from the standard $\Omega(\sqrt{T})$ lower-bound for OCO with convex and Lipschitz losses (e.g., Theorem 3.2 in (Hazan, 2016)).

Given some algorithm for OCO $\mathcal{A}$, let $\mathcal{K}$ be a convex and compact set and $f_1, \ldots, f_{T/2}$ be loss functions (assume w.l.o.g. that T is even) such that the regret of $\mathcal{A}$ on the sequence $f_1, \ldots, f_{T/2}$ w.r.t. the set $\mathcal{K}$ is $\Omega(\sqrt{T})$. Note that according to the lower bound in e.g., (Hazan, 2016) (Theorem 3.2), such a construction is always possible. Let $\mathbf{w}^*$ be the optimal solution in hindsight w.r.t. to the sequence $f_1, \ldots, f_{T/2}$. Now, extend this sequence of functions by adding the functions $f_t(\mathbf{w}) := \|\mathbf{w} - \mathbf{w}^*\|^2$ for all $T/2 < t \leq T$. Clearly, $\mathbf{w}^*$ is also optimal w.r.t. the entire new sequence and in particular incurs zero loss over $f_{T/2+1}, ..., f_T$. On the other-hand, the algorithm $\mathcal{A}$ incurs non-negative loss on $f_{T/2+1}, ..., f_T$. Thus, the regret of $\mathcal{A}$ w.r.t. the entire sequence $f_1, \ldots, f_T$ is $\Omega(\sqrt{T})$. The proof is completed by noticing that the entire sequence $f_1, \ldots, f_T$ is at least 1-cumulative strongly convex.

A.8. Proof of Theorem 3.11

At a high-level, the proof of Theorem 3.11 consists of the following steps.

Our algorithm runs by taking an OGA step and projecting it onto $\mathcal{S}_{d,k}$. The rest of the proof is dedicated to showing that after an OGA step, the eigengap assumption holds as required. The main lemma which shows that the predictions remain low-rank is Lemma A.13. It relies on two technical tools, Lemma A.11 and A.12, which show that certain relevant expression are kept well-behaved along the algorithm.

The next lemma bounds the spectral distance between the whole sequence of T symmetric positive semi-definite matrices, and a block of L such matrices.

Lemma A.11. *Let $0 < \zeta \leq b$ and let $\{\mathbf{M}_t\}_{t\in[T]}$ be a sequence of T symmetric positive semi-definite matrices in $\mathbb{R}^{d\times d}$, that satisfies the random order assumption (Assumption 3.1) and $\|\mathbf{M}_t\| \leq b$ for any $t \in T$. Also let $L \geq L_0$ for a suitable $L_0 = O\left(\frac{b^2}{\zeta^2}\log\frac{d}{\delta}\right)$. Fix $\tau \in [[\lceil T/L \rceil - 1]]$. Then with probability at least $1 - \delta$ it holds that*

$$\left\| \frac{1}{T}\sum_{t=1}^{T}\mathbf{M}_t - \frac{1}{L}\sum_{t=\tau L+1}^{(\tau+1)L}\mathbf{M}_t \right\| = \|\mathbf{M} - \mathbf{M}_\tau\| \leq \zeta.$$

Proof. The left equality holds by definition. By the definition of the spectral norm for matrices, we have

$$\|\mathbf{M} - \mathbf{M}_\tau\| = \frac{1}{L}\|L\mathbf{M} - L\mathbf{M}_\tau\| = \frac{1}{L}\sigma_{\max}(L\mathbf{M} - L\mathbf{M}_\tau)$$

$$\overset{(a)}{=} \frac{1}{L}\max\{\lambda_{\max}(L\mathbf{M} - L\mathbf{M}_\tau), |\lambda_{\min}(L\mathbf{M} - L\mathbf{M}_\tau)|\},$$

where (a) holds as $L\mathbf{M} - L\mathbf{M}_\tau$ is a symmetric matrix. Applying Lemma A.9 with $t = \zeta L$, we thus get

$$\begin{aligned}
&\Pr(\sigma_{\max}(L\mathbf{M} - L\mathbf{M}_\tau) \leq \zeta L) \\
&\quad \geq 1 - O(d)\cdot\exp\left\{\frac{-\zeta^2 L^2}{12Lb^2 + 8\sqrt{2}b\zeta L}\right\} \\
&\quad \underset{(a)}{\geq} 1 - O(d)\cdot\exp\left\{\frac{-\zeta^2 L}{24b^2}\right\},
\end{aligned}$$

where (a) follows since $\zeta \leq b$. Plugging in the value of $L \geq L_0$ finishes the proof. □

At this point, we set $\zeta = \frac{\rho}{2+6b\gamma}$, for $\gamma = \Theta(\sqrt{k}/\rho)$. We would like to have that for all blocks of length L in our permuted sequence, the bound of Lemma A.11 holds simultaneously. Taking a union bound, we observe that picking block size $L \geq L_0$ for suitable $L_0 = O\left(\frac{b^2}{\zeta^2}\log\frac{dT}{\delta}\right)$ suffices for this to hold with probability at least $1 - \delta$. From here on, we condition on this event, which is henceforth denoted E.

Lemma A.12. *Suppose that the assumptions of Theorem 3.11 hold, and that the event E holds as well. Fix $1 \leq \tau \leq \lceil T/L - 1 \rceil$ and assume that rank$(\mathbf{W}_{\tau-1}) \leq k$. Assume also that $\|\mathbf{W}_{\tau-1} - \mathbf{W}^*\|_F \leq \gamma\zeta$, for $\gamma = \Theta(\sqrt{k}/\rho)$. Then*

$$\begin{aligned}
&\|\mathbf{W}_\tau - \mathbf{W}^*\|_F \leq \gamma\zeta \quad \textit{and} \\
&\|\mathbf{W}^*\mathbf{M}\mathbf{W}^* - \mathbf{W}_\tau\mathbf{M}\mathbf{W}_\tau\| \leq 2b\gamma\zeta.
\end{aligned}$$

Proof. We have

$$
\begin{aligned}
&\|\mathbf{W}^*\mathbf{M}\mathbf{W}^* - \mathbf{W}_\tau\mathbf{M}\mathbf{W}_\tau\| \\
&= \|\mathbf{W}^*\mathbf{M}\mathbf{W}^* - \mathbf{W}^*\mathbf{M}\mathbf{W}_\tau + \mathbf{W}^*\mathbf{M}\mathbf{W}_\tau - \mathbf{W}_\tau\mathbf{M}\mathbf{W}_\tau\| \\
&\underset{(a)}{\le} \|\mathbf{W}^*\mathbf{M}(\mathbf{W}^* - \mathbf{W}_\tau)\| + \|(\mathbf{W}^* - \mathbf{W}_\tau)\mathbf{M}\mathbf{W}_\tau\| \\
&\underset{(b)}{\le} \|\mathbf{M}\|\|\mathbf{W}^*\|\|\mathbf{W}^* - \mathbf{W}_\tau\| + \|\mathbf{M}\|\|\mathbf{W}_\tau\|\|\mathbf{W}^* - \mathbf{W}_\tau\| \\
&\underset{(c)}{\le} 2\|\mathbf{M}\|\|\mathbf{W}^* - \mathbf{W}_\tau\|_F \underset{(d)}{\le} 2b\|\mathbf{W}^* - \mathbf{W}_\tau\|_F,
\end{aligned}
\tag{14}
$$

where (a) follows from the triangle inequality, (b) holds by the fact that $\|\mathbf{A}\mathbf{B}\| \le \|\mathbf{A}\|\|\mathbf{B}\|$ for all matrices $\mathbf{A}$ and $\mathbf{B}$ in $\mathbb{R}^{d\times d}$, (c) holds since $\|\mathbf{W}_\tau\|, \|\mathbf{W}^*\| \le 1$ and also $\|\mathbf{A}\| \le \|\mathbf{A}\|_F$ for any matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, and (d) holds by the fact that $\|\mathbf{M}_t\| \le b$ and thus $\|\mathbf{M}\| \le b$. The rest of the proof is dedicated to getting an upper bound for the term $\|\mathbf{W}^* - \mathbf{W}_\tau\|_F$.

$$
\begin{aligned}
&\|\mathbf{W}_\tau - \mathbf{W}^*\|_F^2 \underset{(a)}{\le} \|\mathbf{W}_{\tau-1} + \eta_{\tau-1}\mathbf{M}_{\tau-1} - \mathbf{W}^*\|_F^2 \\
&= \|\mathbf{W}_{\tau-1} - \mathbf{W}^*\|_F^2 + 2\eta_{\tau-1}\left\langle \mathbf{W}_{\tau-1} - \mathbf{W}^*, \mathbf{M}_{\tau-1}\right\rangle \\
&+ \eta_{\tau-1}^2\|\mathbf{M}_{\tau-1}\|_F^2 \\
&\underset{(b)}{\le} \|\mathbf{W}_{\tau-1} - \mathbf{W}^*\|_F^2 + 2\eta_{\tau-1}\left\langle \mathbf{W}_{\tau-1} - \mathbf{W}^*, \mathbf{M}\right\rangle \\
&+ 2\eta_{\tau-1}\sqrt{2k}\|\mathbf{W}_{\tau-1} - \mathbf{W}^*\|_F\|\mathbf{M} - \mathbf{M}_{\tau-1}\| \\
&+ \eta_{\tau-1}^2\|\mathbf{M}_{\tau-1}\|_F^2 \\
&\underset{(c)}{\le} \|\mathbf{W}_{\tau-1} - \mathbf{W}^*\|_F^2 + 2\eta_{\tau-1}\left\langle \mathbf{W}_{\tau-1} - \mathbf{W}^*, \mathbf{M}\right\rangle \\
&+ 2\zeta\eta_{\tau-1}\sqrt{2k}\|\mathbf{W}_{\tau-1} - \mathbf{W}^*\|_F + \eta_{\tau-1}^2\|\mathbf{M}_{\tau-1}\|_F^2
\end{aligned}
\tag{15}
$$

where (a) follows from Theorem A.4, (b) follows from Hölder's inequality and from the fact that the rank of both $\mathbf{W}^*$ and $\mathbf{W}_{\tau-1}$ is at most k and (c) follows from the event E described earlier. Next we aim to bound the term $\langle \mathbf{W}_{\tau-1} - \mathbf{W}^*, \mathbf{M}\rangle$. In order to do so, we introduce the following construction. Let $\mathbf{W}_\perp^*$ be the matrix in $\{\mathbf{W} \in \mathbb{R}^{d\times d} : \mathrm{Tr}(\mathbf{W}) = d - k, \mathbf{0} \preceq \mathbf{W} \preceq \mathbf{I}\}$ that satisfies $\mathbf{W}_\perp^* + \mathbf{W}^* = \mathbf{I}_d$ and note that $\mathbf{W}_\perp^*\mathbf{W}^* = \mathbf{W}^*\mathbf{W}_\perp^* = 0$. Thus, we have that

$$
\begin{aligned}
&\langle \mathbf{W}_{\tau-1} - \mathbf{W}^*, \mathbf{M}\rangle = \langle \mathbf{W}_{\tau-1}\mathbf{I}_d - \mathbf{W}^*, \mathbf{I}_d\mathbf{M}\rangle \\
&= \langle \mathbf{W}_{\tau-1}\mathbf{W}^* + \mathbf{W}_{\tau-1}\mathbf{W}_\perp^* - \mathbf{W}^*, \mathbf{W}^*\mathbf{M} + \mathbf{W}_\perp^*\mathbf{M}\rangle \\
&= \langle \mathbf{W}_{\tau-1}\mathbf{W}_\perp^*, \mathbf{W}^*\mathbf{M} + \mathbf{W}_\perp^*\mathbf{M}\rangle \\
&+ \langle (\mathbf{W}_{\tau-1} - \mathbf{I}_d)\mathbf{W}^*, \mathbf{W}^*\mathbf{M} + \mathbf{W}_\perp^*\mathbf{M}\rangle \\
&= \langle \mathbf{W}_{\tau-1}\mathbf{W}_\perp^*, \mathbf{W}_\perp^*\mathbf{M}\rangle + \langle (\mathbf{W}_{\tau-1} - \mathbf{I}_d)\mathbf{W}^*, \mathbf{W}^*\mathbf{M}\rangle \\
&= \mathrm{Tr}(\mathbf{W}_{\tau-1}\mathbf{W}_\perp^*(\mathbf{W}_\perp^*\mathbf{M})^\top) \\
&+ \mathrm{Tr}((\mathbf{W}_{\tau-1} - \mathbf{I}_d)\mathbf{W}^*(\mathbf{W}^*\mathbf{M})^\top) \\
&\underset{(a)}{=} \mathrm{Tr}(\mathbf{W}_{\tau-1}\mathbf{W}_\perp^*\mathbf{W}_\perp^*\mathbf{M}) + \mathrm{Tr}((\mathbf{W}_{\tau-1} - \mathbf{I}_d)\mathbf{W}^*\mathbf{W}^*\mathbf{M}) \\
&= \underbrace{\mathrm{Tr}(\mathbf{W}_\perp^*\mathbf{M}\mathbf{W}_{\tau-1})}_{\text{(I)}} - \underbrace{\mathrm{Tr}(\mathbf{M}(\mathbf{W}^* - \mathbf{W}^*\mathbf{W}_{\tau-1}\mathbf{W}^*))}_{\text{(II)}},
\end{aligned}
\tag{16}
$$

where (a) holds since $\mathbf{W}^*\mathbf{M}$ corresponds to the top k eigenvectors of $\mathbf{M}$, and also since $\mathbf{W}_\perp^*\mathbf{M}$ corresponds to the low $n - k$ eigenvectors of $\mathbf{M}$, and therefore both terms are symmetric.

By the fact that $\mathbf{W}_\perp^*\mathbf{M}$ corresponds to the lower $n - k$ eigenvectors of $\mathbf{M}$ and $\mathbf{W}_\perp^*$, it follows that $\mathbf{W}_\perp^*\mathbf{M}$ is positive semidefinite, and since $\mathbf{W}_\perp^*\mathbf{W}_{\tau-1}\mathbf{W}_\perp^*$ is also PSD, the term (I) in the RHS of (16) gives

$$
\begin{aligned}
\text{(I)} &= \mathrm{Tr}(\mathbf{W}_\perp^*\mathbf{M}\mathbf{W}_\perp^*\mathbf{W}_{\tau-1}\mathbf{W}_\perp^*) \\
&\le \lambda_{\max}(\mathbf{W}_\perp^*\mathbf{M})\,\mathrm{Tr}(\mathbf{W}_\perp^*\mathbf{W}_{\tau-1}\mathbf{W}_\perp^*) \\
&\underset{(a)}{=} \lambda_{k+1}(\mathbf{M})\,\mathrm{Tr}(\mathbf{W}_\perp^*\mathbf{W}_{\tau-1}) \\
&\underset{(b)}{\le} (\lambda_k(\mathbf{M}) - s(\mathbf{M}))\,\mathrm{Tr}(\mathbf{W}_\perp^*\mathbf{W}_{\tau-1}),
\end{aligned}
$$

where (a) holds as the non-zero eigenvalues of $\mathbf{W}_\perp^*\mathbf{M}$ are the $n - k$ lower eigenvalues of $\mathbf{M}$, and (b) follows since $\mathbf{M}$ admits an eigengap $s(\mathbf{M})$.

Similarly, since the matrix $\mathbf{W}^* - \mathbf{W}^*\mathbf{W}_{\tau-1}\mathbf{W}^*$ is positive semidefinite, and since the eigenvectors of $\mathbf{W}^*$ (with nonzero eigenvalue) correspond to the top k eigenvectors of $\mathbf{M}$, the term (II) in the RHS of (16) gives

$$
\begin{aligned}
\text{(II)} &= \mathrm{Tr}(\mathbf{M}(\mathbf{W}^* - \mathbf{W}^*\mathbf{W}_{\tau-1}\mathbf{W}^*)) \\
&\ge \lambda_k(\mathbf{M}) \cdot \mathrm{Tr}(\mathbf{W}^* - \mathbf{W}^*\mathbf{W}_{\tau-1}\mathbf{W}^*) \\
&= \lambda_k(\mathbf{M}) \cdot (k - \mathrm{Tr}(\mathbf{W}^*\mathbf{W}_{\tau-1})).
\end{aligned}
$$

Plugging (I) and (II) back to Equation (16), gives

$$\begin{aligned}
&\langle \mathbf{W}_{\tau-1} - \mathbf{W}^*, \mathbf{M} \rangle = \text{(I)} - \text{(II)} \\
&\leq -\lambda_k(\mathbf{M}) \cdot (k - \text{Tr}(\mathbf{W}^*\mathbf{W}_{\tau-1})) \\
&+ (\lambda_k(\mathbf{M}) - s(\mathbf{M})) \cdot \text{Tr}(\mathbf{W}_\perp^*\mathbf{W}_{\tau-1}) \\
&= -\lambda_k(\mathbf{M}) \cdot (k - \text{Tr}(\mathbf{W}^*\mathbf{W}_{\tau-1} + \mathbf{W}_\perp^*\mathbf{W}_{\tau-1})) \\
&- s(\mathbf{M}) \cdot \text{Tr}(\mathbf{W}_\perp^*\mathbf{W}_{\tau-1}) \\
&\underset{(a)}{=} -s(\mathbf{M}) \cdot \text{Tr}((\mathbf{I}_d - \mathbf{W}^*)\mathbf{W}_{\tau-1}) \qquad (17) \\
&\underset{(b)}{=} -s(\mathbf{M})(k - \langle \mathbf{W}_{\tau-1}, \mathbf{W}^* \rangle) \\
&\underset{(c)}{=} -\frac{1}{2}s(\mathbf{M}) \cdot \|\mathbf{W}_{\tau-1} - \mathbf{W}^*\|_F^2 \\
&\underset{(d)}{\leq} -\frac{1}{2}\rho \cdot \|\mathbf{W}_{\tau-1} - \mathbf{W}^*\|_F^2,
\end{aligned}$$

where (a) and (b) hold since $\mathbf{W}_\perp^* + \mathbf{W}^* = \mathbf{I}_d$ and $\text{Tr}(\mathbf{W}_{\tau-1}) = k$, (c) holds since both $\mathbf{W}^*, \mathbf{W}_{\tau-1}$ are rank-k and thus $\|\mathbf{W}_{\tau-1} - \mathbf{W}^*\|_F^2 = \|\mathbf{W}_{\tau-1}\|_F^2 - 2\langle \mathbf{W}_{\tau-1}, \mathbf{W}^* \rangle + \|\mathbf{W}^*\|_F^2 = 2k - 2\langle \mathbf{W}_{\tau-1}, \mathbf{W}^* \rangle$, and (d) follows that $s(\mathbf{M}) \geq \rho$. Plugging Equation (17) back to Equation (15), we have that

$$\begin{aligned}
&\|\mathbf{W}_\tau - \mathbf{W}^*\|_F^2 \\
&\quad \leq \|\mathbf{W}_{\tau-1} - \mathbf{W}^*\|_F^2 - \eta_{\tau-1}\rho \cdot \|\mathbf{W}_{\tau-1} - \mathbf{W}^*\|_F^2 \\
&\quad + 2\eta_{\tau-1}\sqrt{2k}\zeta\|\mathbf{W}_{\tau-1} - \mathbf{W}^*\|_F + \eta_{\tau-1}^2\|\mathbf{M}_{\tau-1}\|_F^2. \qquad (18)
\end{aligned}$$

We now recall that in order to prove the lemma, using Eq. (14), it only remains to upper bound the term $\|\mathbf{W}^* - \mathbf{W}_\tau\|_F$.

We consider two cases. If $\|\mathbf{W}^* - \mathbf{W}_{\tau-1}\|_F \leq \gamma\zeta/2$ then from Eq. (18) we have that

$$\|\mathbf{W}_\tau - \mathbf{W}^*\|_F^2 \leq \gamma^2\zeta^2/4 + 2\eta_{\tau-1}\sqrt{2k}\gamma\zeta^2/2 + \eta_{\tau-1}^2 C^2,$$

where we recall C is an upper-bound on the Frobenius norm of any of the matrices $\mathbf{M}_t$, and so, since $\mathbf{M}_{\tau-1}$ is the average of L such matrices, we have $\|\mathbf{M}_{\tau-1}\|_F^2 \leq C^2$. Since ζ, γ are positive constants independent of the sequence length T, due to our choice of step-size $\eta_{\tau-1}$, for any T sufficiently large, we have that

$$\|\mathbf{W}^* - \mathbf{W}_{\tau-1}\|_F \leq \gamma\zeta/2 \;\Rightarrow\; \|\mathbf{W}_\tau - \mathbf{W}^*\|_F \leq \gamma\zeta.$$

On the other-hand, if $\gamma\zeta \geq \|\mathbf{W}^* - \mathbf{W}_{\tau-1}\|_F \geq \gamma\zeta/2$, using Eq. (18) again we have that,

$$\begin{aligned}
&\|\mathbf{W}_\tau - \mathbf{W}^*\|_F^2 \leq \gamma^2\zeta^2 - \eta_{\tau-1}\rho\gamma^2\zeta^2/4 \\
&+ 2\eta_{\tau-1}\sqrt{2k}\gamma\zeta^2/2 + \eta_{\tau-1}^2 C^2 \\
&\leq \gamma^2\zeta^2 - \eta_{\tau-1}\rho\gamma^2\zeta^2/4 + 2\eta_{\tau-1}\sqrt{2k}\gamma\zeta^2,
\end{aligned}$$

where the last inequality holds again for T sufficiently large. In particular, we see that for $\gamma \geq 8\sqrt{2k}/\rho$ we have that $\|\mathbf{W}_\tau - \mathbf{W}^*\|_F \leq \gamma\zeta$. Thus, the lemma follows.

□

Lemma A.13. *Under the assumptions of Theorem 3.11 and assuming the event E occurs, the following holds for all $\tau \in [[\lceil T/L \rceil - 1]]$ simultaneously:*

$$\textit{rank}(\mathbf{W}_{\tau+1}) = k.$$

Proof. The proof is by a simple induction over τ. For the base case, clearly when initializing $\mathbf{W}_0$ with a projection matrix in $\mathcal{S}_{d,k}$ this holds. Fix now some τ. Recall that $\mathbf{W}_{\tau+1} = \Pi_{\mathcal{S}_{d,k}}(\mathbf{W}_\tau + \eta_\tau\mathbf{M}_\tau)$ and let $\mathbf{Y}_\tau$ denote $\mathbf{W}_\tau + \eta_\tau\mathbf{M}_\tau$. Using Lemma 3.9, it follows that a sufficient condition for having $\text{rank}(\mathbf{W}_{\tau+1}) = k$ is

$$\lambda_k(\mathbf{Y}_\tau) \geq 1 + \lambda_{k+1}(\mathbf{Y}_\tau).$$

Recall that $\mathbf{M}$ denotes the whole data set, i.e. $\mathbf{M} = \frac{1}{T}\sum_{i=1}^T \mathbf{M}_t$. On the one hand,

$$\begin{aligned}
&\lambda_k(\mathbf{Y}_\tau) = \lambda_k(\mathbf{W}_\tau + \eta_\tau\mathbf{M}_\tau) \\
&\underset{(a)}{\geq} \lambda_k(\mathbf{W}_\tau + \eta_\tau\mathbf{M}) + \eta_\tau\lambda_d(\mathbf{M}_\tau - \mathbf{M}) \\
&\geq \lambda_k(\mathbf{W}_\tau + \eta_\tau\mathbf{M}) - \eta_\tau\|\mathbf{M} - \mathbf{M}_\tau\| \\
&\underset{(b)}{\geq} \lambda_k(\mathbf{W}_\tau + \eta_\tau(\mathbf{W}^* + \mathbf{W}_\perp^*)\mathbf{M}(\mathbf{W}^* + \mathbf{W}_\perp^*)) - \eta_\tau\zeta \\
&\underset{(c)}{\geq} \lambda_k(\mathbf{W}_\tau + \eta_\tau\mathbf{W}^*\mathbf{M}\mathbf{W}^*) - \eta_\tau\zeta \\
&\underset{(d)}{\geq} \lambda_k(\mathbf{W}_\tau + \eta_\tau\mathbf{W}_\tau\mathbf{M}\mathbf{W}_\tau) - \eta_\tau\zeta \\
&- \eta_\tau\|\mathbf{W}^*\mathbf{M}\mathbf{W}^* - \mathbf{W}_\tau\mathbf{M}\mathbf{W}_\tau\| \\
&\underset{(e)}{=} \lambda_k(\mathbf{W}_\tau) + \eta_\tau\lambda_k(\mathbf{W}_\tau\mathbf{M}\mathbf{W}_\tau) \\
&- \eta_\tau\|\mathbf{W}^*\mathbf{M}\mathbf{W}^* - \mathbf{W}_\tau\mathbf{M}\mathbf{W}_\tau\| - \eta_\tau\zeta \\
&\underset{(f)}{\geq} 1 + \eta_\tau\lambda_k(\mathbf{W}_\tau\mathbf{M}\mathbf{W}_\tau) - \eta_\tau\|\mathbf{W}^*\mathbf{M}\mathbf{W}^* - \mathbf{W}_\tau\mathbf{M}\mathbf{W}_\tau\| \\
&- \eta_\tau\zeta \\
&\underset{(g)}{\geq} 1 + \eta_\tau\lambda_k(\mathbf{W}^*\mathbf{M}\mathbf{W}^*) - 2\eta_\tau\|\mathbf{W}^*\mathbf{M}\mathbf{W}^* - \mathbf{W}_\tau\mathbf{M}\mathbf{W}_\tau\| \\
&- \eta_\tau\zeta \\
&\underset{(h)}{\geq} 1 + \eta_\tau\lambda_k(\mathbf{M}) - \eta_\tau(\zeta + 4b\gamma\zeta),
\end{aligned}$$

where (a) follows from Weyl's inequality for the eigenvalues, (b) holds under the assumption that the event E occurs and since $\mathbf{W}^* + \mathbf{W}_\perp^* = \mathbf{I}$, (c) holds since $\mathbf{W}^*\mathbf{M}\mathbf{W}_\perp^* = \mathbf{0}$ and since $\mathbf{W}_\perp^*\mathbf{M}\mathbf{W}_\perp^*$ is positive semidefinite, (d) follows from Weyl's inequality, (e) and (f) hold since $\mathbf{W}_\tau$ and $\mathbf{W}_\tau\mathbf{M}\mathbf{W}_\tau$ share the same eigenspace and since under the induction hypothesis $\lambda_1(\mathbf{W}_\tau) = ... = \lambda_k(\mathbf{W}_\tau) = 1$, (g) follows from Weyl's inequality, and (h) follows Lemma A.12 and since $\mathbf{W}^*$ corresponds to the top k eigenvectors of $\mathbf{M}$. On the other hand,

$$\begin{aligned}
&\lambda_{k+1}(\mathbf{Y}_\tau) = \lambda_{k+1}(\mathbf{W}_\tau + \eta_\tau \mathbf{M}_\tau)\\
&\underset{(a)}{\leq} \lambda_{k+1}(\mathbf{W}_\tau + \eta_\tau \mathbf{M}) + \eta_\tau \lambda_1(\mathbf{M}_\tau - \mathbf{M})\\
&\leq \lambda_{k+1}(\mathbf{W}_\tau + \eta_\tau \mathbf{M}) + \eta_\tau \|\mathbf{M} - \mathbf{M}_\tau\|\\
&\underset{(b)}{\leq} \lambda_{k+1}(\mathbf{W}_\tau + \eta_\tau (\mathbf{W}^* + \mathbf{W}^*_\perp)\mathbf{M}(\mathbf{W}^* + \mathbf{W}^*_\perp)) + \eta_\tau \zeta\\
&\underset{(c)}{=} \lambda_{k+1}(\mathbf{W}_\tau + \eta_\tau \mathbf{W}^*\mathbf{M}\mathbf{W}^*) + \eta_\tau \lambda_1(\mathbf{W}^*_\perp \mathbf{M}\mathbf{W}^*_\perp) + \eta_\tau \zeta\\
&\underset{(d)}{=} \lambda_{k+1}(\mathbf{W}_\tau + \eta_\tau \mathbf{W}^*\mathbf{M}\mathbf{W}^*) + \eta_\tau \lambda_{k+1}(\mathbf{M}) + \eta_\tau \zeta\\
&\underset{(e)}{\leq} \lambda_{k+1}(\mathbf{W}_\tau + \eta_\tau \mathbf{W}_\tau\mathbf{M}\mathbf{W}_\tau) + \eta_\tau \lambda_{k+1}(\mathbf{M}) + \eta_\tau \zeta\\
&+ \eta_\tau \|\mathbf{W}^*\mathbf{M}\mathbf{W}^* - \mathbf{W}_\tau\mathbf{M}\mathbf{W}_\tau\|\\
&\underset{(f)}{=} \eta_\tau \lambda_{k+1}(\mathbf{M}) + \eta_\tau \|\mathbf{W}^*\mathbf{M}\mathbf{W}^* - \mathbf{W}_\tau\mathbf{M}\mathbf{W}_\tau\|\\
&+ \eta_\tau \zeta\\
&\underset{(h)}{\leq} \eta_\tau \lambda_{k+1}(\mathbf{M}) + \eta_\tau(\zeta + 2b\gamma\zeta)
\end{aligned}$$

where (a) follows from Weyl's inequality for the eigenvalues, (b) holds under the assumption that the event E occurs and since $\mathbf{W}^* + \mathbf{W}^*_\perp = \mathbf{I}$, (c) follows from the fact that $\mathbf{W}^*\mathbf{M}\mathbf{W}^*_\perp = \mathbf{0}$ and using Weyl's inequality, (d) follows since $\mathbf{W}^*_\perp$ corresponds to the $n-k$ lower eigenvectors of $\mathbf{M}$, (e) follows from Weyl's inequality, (f) follows since under the induction hypothesis, the matrix $\mathbf{W}_\tau + \eta_\tau \mathbf{W}_\tau \mathbf{M}\mathbf{W}_\tau$ is rank-k and so its largest $k+1$ eigenvalue is zero, and (h) follows Lemma A.12. Thus, combining both of the above bounds, we get

$$\begin{aligned}
&\lambda_k(\mathbf{Y}_\tau) - \lambda_{k+1}(\mathbf{Y}_\tau)\\
&\geq 1 + \eta_\tau \lambda_k(\mathbf{M}) - \eta_\tau \lambda_{k+1}(\mathbf{M})\\
&- \eta_\tau(2\zeta + 6b\gamma\zeta)\\
&\geq 1 + \eta_\tau(s(\mathbf{M}) - 2\zeta - 6b\gamma\zeta).
\end{aligned}$$

Thus, for $\zeta \leq \frac{\rho}{2+6b\gamma}$ we have that $\lambda_k(\mathbf{Y}_\tau) - \lambda_{k+1}(\mathbf{Y}_\tau) \geq 1$ as required, thus $\operatorname{rank}(\mathbf{W}_{\tau+1}) = k$, and the induction holds. □

The rest of the proof of Theorem 3.11 follows by a standard regret analysis for Online Gradient Ascent, see e.g. (Hazan, 2016; Zinkevich, 2003b) , which by plugging-in our standard step-size choice and length of the blocks L, directly gives the $O(\sqrt{T})$ regret bound listed in the theorem.

A.9. Proof of Lemma 3.9

The projection of $\mathbf{Y}$ onto $\mathcal{S}_{d,k}$ is defined as the point in $\mathcal{S}_{d,k}$ at minimal distance from $\mathbf{Y}$, and is thus the solution of the following minimization problem.

$$\min_{\mathbf{W}\in\mathcal{S}_{d,k}} \frac{1}{2}\|\mathbf{W} - \mathbf{Y}\|_F^2.$$

Arora et al. show in (Arora et al., 2013) (see Lemma 3.2) that the optimal $\mathbf{W}$, denoted by $\mathbf{W}^*$, differs from $\mathbf{Y}$ only by its eigenvalues, meaning that the projection operates *only* on the eigenvalues. Let $\mathbf{w}, \mathbf{y} \in \mathbb{R}^n$ denote the eigenvalue vectors of $\mathbf{W}$ and $\mathbf{Y}$ (in non-increasing order), respectively. Thus, the aforementioned minimization problem can be reformulated as follows (Wang & Lu, 2015), where $\mathbf{w}$ belongs to the capped simplex set denoted by $S_c = \{\mathbf{y} : \mathbf{0} \leq \mathbf{y} \leq \mathbf{1}, \mathbf{y}^\top \mathbf{1} = k\}$.

$$\min_{\mathbf{w}\in S_c} g(\mathbf{w}) := \frac{1}{2}\|\mathbf{w} - \mathbf{y}\|_2^2.$$

By the definition of a projection and using first-order optimality conditions, $\mathbf{w}^* = [\overbrace{1\ldots1}^{k\text{ times}}0\ldots0]^\top$ is the projection of $\mathbf{y}$ onto S_c if and only if for all $\mathbf{w} \in S_c$ it holds that

$$g(\mathbf{w}) \geq g(\mathbf{w}^*) + \nabla g(\mathbf{w}^*)^\top(\mathbf{w} - \mathbf{w}^*) \geq g(\mathbf{w}^*),$$

which holds if and only if

$$(\mathbf{w}^* - \mathbf{y})^\top(\mathbf{w} - \mathbf{w}^*) \geq 0. \qquad (19)$$

Writing the left hand side of Equation (19) explicitly, we indeed have that

$$\begin{aligned}
\text{LHS of (19)} &\underset{(a)}{=} \sum_{i=1}^{k} \mathbf{w}_i - \mathbf{w}^\top \mathbf{y} + \sum_{i=1}^{k} \mathbf{y}_i - k\\
&= \sum_{i=1}^{k} \mathbf{w}_i(1 - \mathbf{y}_i) + \sum_{i=1}^{k}(\mathbf{y}_i - 1) - \sum_{j=k+1}^{n} \mathbf{w}_i\mathbf{y}_i\\
&= \sum_{i=1}^{k}(\mathbf{y}_i - 1)(1 - \mathbf{w}_i) - \sum_{j=k+1}^{n} \mathbf{w}_i\mathbf{y}_i\\
&\overset{(b)}{\geq} \sum_{i=1}^{k}(\mathbf{y}_{k+1} + 1 - 1)(1 - \mathbf{w}_i) - \mathbf{y}_{k+1} \sum_{j=k+1}^{n} \mathbf{w}_j\\
&= \mathbf{y}_{k+1}(k - \sum_{i=1}^{n} \mathbf{w}_i) = 0,
\end{aligned}$$

where in (a) we use the fact that $\|\mathbf{w}^*\|_2^2 = k$, and where (b) follows since $\mathbf{y}$ admits a gap $\mathbf{y}_k - \mathbf{y}_{k+1} = \lambda_k(\mathbf{Y}) - \lambda_{k+1}(\mathbf{Y}) \geq 1$. Thus, we can conclude that the projection $\mathbf{W}^*$ has at most k non-zero eigenvalues, and hence has rank at most k.

B. Additional Experiments

Here we present additional results for the online PCA experiment of Section 5.1 for the case $k = 2$ with the synthetic dataset.

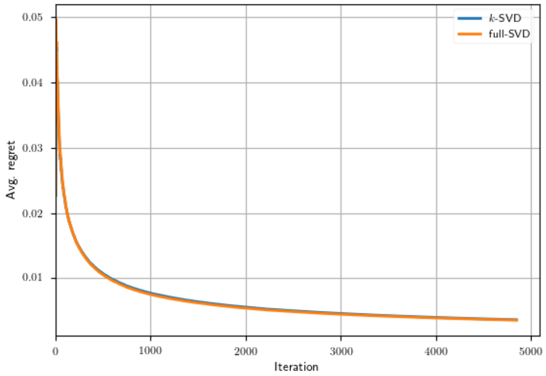


Figure 5. Average regret of full-SVD and k-SVD on synthetic data for $k = 2$.