Proceedings of Machine Learning Research vol 99:1–25, 2019 32nd Annual Conference on Learning Theory

On the Regret Minimization of Nonconvex Online Gradient Ascent for Online PCA

Dan Garber DANGAR@TECHNION.AC.IL
Technion - Israel Institute of Technology

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

In this paper we focus on the problem of Online Principal Component Analysis in the regret minimization framework. For this problem, all existing regret minimization algorithms for the fully-adversarial setting are based on a positive semidefinite convex relaxation, and hence require quadratic memory and SVD computation (either thin of full) on each iteration, which amounts to at least quadratic runtime per iteration. This is in stark contrast to a corresponding stochastic i.i.d. variant of the problem, which was studied extensively lately, and admits very efficient gradient ascent algorithms that work directly on the natural non-convex formulation of the problem, and hence require only linear memory and linear runtime per iteration. This raises the question: *can non-convex online gradient ascent algorithms be shown to minimize regret in online adversarial settings?*

In this paper we take a step forward towards answering this question. We introduce an *adversarially-perturbed spiked-covariance model* in which, each data point is assumed to follow a fixed stochastic distribution with a non-zero spectral gap in the covariance matrix, but is then perturbed with some adversarial vector. This model is a natural extension of a well studied standard *stochastic* setting that allows for non-stationary (adversarial) patterns to arise in the data and hence, might serve as a significantly better approximation for real-world data-streams. We show that in an interesting regime of parameters, when the non-convex online gradient ascent algorithm is initialized with a "warm-start" vector, it provably minimizes the regret with high probability. We further discuss the possibility of computing such a "warm-start" vector, and also the use of regularization to obtain fast regret rates. Our theoretical findings are supported by empirical experiments on both synthetic and real-world data.

Keywords: online learning, regret minimization, online PCA, online convex optimization

1. Introduction

Nonconvex optimization is ubiquitous in contemporary machine learning, ranging from optimization over sparse vectors or low-rank matrices to training Deep Neural Networks. While traditional (yet still highly active) research on nonconvex optimization focuses mostly on efficient convergence to stationary points, which in general need not even be a local minima, let alone a global one (e.g., Agarwal et al. (2017); Allen-Zhu (2018); Allen-Zhu and Hazan (2016); Carmon et al. (2017)), a more-recent line of work focuses on proving convergence to global minima, usually under certain simplifying assumptions that on one hand make the nonconvex problem tractable, and on the other hand, are sufficiently reasonable in some scenarios of interest (see for instance De Sa et al. (2015); Ge et al. (2016); Bhojanapalli et al. (2016); Arora et al. (2014); Jin et al. (2016) to name only a few). One of the most studied and well known nonconvex optimization problems in machine learning un-

© 2019 D. Garber.

derlies the fundamental task of *Principal Component Analysis* (PCA) Pearson (1901); Hotelling (1933); Jolliffe (2011), in which, given a set of N vectors in $\mathbb{R}^d$, one wishes to find a k-dimensional subspace for $k << d$, such that the projections of these vectors onto this subspace is closest in square-error to the original vectors. It is well known that the optimal subspace corresponds to the span of the top k eigenvectors of the covariance matrix of the data-points. Henceforth, we focus our discussion to the case $k = 1$, i.e., extracting the top principal component. Quite remarkably, while this problem is non-convex (since extracting the top eigenvector amounts to *maximizing* a convex function over the unit Euclidean ball), a well known iterative algorithm known as *Power Method* (or Power Iterations, see for instance Golub and Van Loan (2012)), which simply starts with a random unit vector and repeatedly applies the covariance matrix to it (and then normalizes the result to have unit norm), converges to the global optimal solution rapidly. The convergence guarantee of the PM, can also be shown to imply that the nonconvex projected gradient ascent method with random initialization and a fixed step-size also converges to the top principal component. [1]

In a recent line of work, the convergence of non-convex gradient methods for PCA was extended to a natural online stochastic i.i.d. setting of the problem, in which, given a stream of data points sampled i.i.d. from a fixed (unknown) distribution, the goal is to converge to the top eigenvector of the covariance matrix of the underlying distribution as the sample size increases, yielding algorithms that require only linear memory (i.e., do not need to store the entire sample or large portions of it at any time) and linear runtime to process each data point, see for instance Mitliagkas et al. (2013); Balsubramani et al. (2013); Shamir (2016); Jain et al. (2016); Allen-Zhu and Li (2017a); Li et al. (2018); Xu et al. (2018).

In a second recent line of research, researchers have considered Online PCA as a sequential decision problem in the adversarial framework of regret minimization (aka online learning, see for instance the introductory texts Cesa-Bianchi and Lugosi (2006); Hazan (2016); Shalev-Shwartz (2012)), e.g., Warmuth and Kuzmin (2006a,b); Nie et al. (2013); Dwork et al. (2014); Garber et al. (2015); Allen-Zhu and Li (2017b). In this framework, for each data-point, the online algorithm is required to predict a unit vector (i.e., a subspace of dimension one, recall we are in the case $k = 1$) *before* observing the data-point, and the goal is to minimize regret which is the difference between the square-error of the predictions made and the square-error of the principal component of the entire sequence of data. Different from the i.i.d. stochastic setting, in this framework, the data may be completely arbitrary (though assumed to be bounded in norm), and need not follow a simple generative model. Formally, the regret is given by

$$\text{regret} := \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{w}_i \mathbf{w}_i^\top \mathbf{x}_t\|_2^2 - \min_{\|\mathbf{w}\|_2=1} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{w}\mathbf{w}^\top \mathbf{x}_t\|_2^2,$$

where $\{\mathbf{x}_i\}_{i\in[N]} \subset \mathbb{R}^d$ is the sequence of data points, and $\{\mathbf{w}_i\}_{i\in[N]}$ is the sequence of predictions made by the online algorithm. Using standard manipulations, it can be shown that

$$\text{regret} = \max_{\|\mathbf{w}\|_2=1} \sum_{i=1}^{N} (\mathbf{w}^\top \mathbf{x}_t)^2 - \sum_{i=1}^{N} (\mathbf{w}_i^\top \mathbf{x}_i)^2 = \lambda_1 \left(\sum_{t=1}^{N} \mathbf{x}_i \mathbf{x}_i^\top \right) - \sum_{t=1}^{N} (\mathbf{w}_i^\top \mathbf{x}_i)^2,$$

where $\lambda_1(\cdot)$ denotes the largest (signed) eigenvalue of a real symmetric matrix.

Naturally, the arbitrary nature of the data in the online learning setting, makes the problem much more difficult than the stochastic i.i.d. setting. Notably, all current algorithms which minimize regret in this fully-adversarial setting cannot directly tackle the natural nonconvex formulation of

1. This follows since the steps of PGA can be rewritten as applying PM steps to a scaled and shifted version of the original matrix that preserves the leading eigenvector.

the problem, but consider a well known (tight) convex relaxation, which "lifts" the decision variable from the unit Euclidean ball in $\mathbb{R}^d$ to the set of all $d \times d$ positive semidefinite matrices of unit trace (aka the spectrahedron). While this reformulation allows to obtain regret-minimizing algorithms in the online adversarial settings (since the problem becomes convex), they are dramatically less efficient than the standard nonconvex gradient methods. In particular, all such algorithms require quadratic memory (i.e., $O(d^2)$), and require either a thin or full-rank SVD computation of a full-rank matrix to process each data point, which amounts to at least quadratic runtime per data point (for non trivially-sparse data), see Warmuth and Kuzmin (2006a,b); Nie et al. (2013); Dwork et al. (2014); Garber et al. (2015); Allen-Zhu and Li (2017b). This phenomena naturally raises the question:

Can Nonconvex Online Gradient Ascent be shown to minimize regret for the Online PCA problem?

While in this paper we do not provide a general answer (either positive or negative), we do take a step forward towards understanding the applicability of nonconvex gradient methods to the Online PCA problem. We introduce a "semi-adversarial" setting, which we refer to as *adversarially-perturbed spiked-covariance model*, which assumes the data follows a standard i.i.d. stochastic distribution with a covariance matrix that admits a non-zero spectral gap, however, each data point is then perturbed by some arbitrary, possibly adversarial, vector of non-trivial magnitude. We view this model as a natural extension of the standard stochastic model (which was studied extensively in recent years, see references above) due to its ability to capture arbitrary (adversarial) patterns in the data. Hence, we believe the suggested model might provide a much better approximation for real-world data streams. We formally prove that in a certain regime of parameters, which concerns both the spectral properties of the distribution covariance and the magnitude of adversarial perturbations, given a "warm-start" initialization which is sufficiently correlated with the top principal component of the stochastic distribution, the natural nonconvex online gradient ascent algorithm guarantees an $\tilde{O}(\sqrt{N})$ regret bound with high probability. In particular, the algorithm requires only $O(d)$ memory and $O(d)$ runtime per data point. We further discuss the possibilities of computing such a "warm-start" vector (i.e., initializing from a "cold-start"). Moreover, we explore the possibility of adding regularization to the algorithm, which as we show, under the same assumptions on the data and the same "warm-start" initialization, allows to obtain a $\text{poly}(\log N)$ regret bound, still using only $O(d)$ memory and runtime per data point. [2]

Finally, we present empirical experiments with both synthetic and real-world datasets which complement our theoretical analysis.

1.1. Related work

Besides the works for the fully-adversarial online setting mentioned above, in a very recent work Marinov et al. (2018), the authors have introduced an online PCA setting in which the data-points (the vectors $\{\mathbf{x}_i\}_{i\in[n]}$) are drawn i.i.d. from a fixed (unknown) distribution, however the feedback observed by the algorithm on each round i is a perturbed version $\mathbf{x}_i + \mathbf{y}_i$, where $\mathbf{y}_i$ is an arbitrary noise. The authors show that under the conditions $\max_{i\in[N]} \|\mathbf{x}_i\|_2 \leq 1$ and $\sum_{i=1}^N \|\mathbf{y}_i\|_2 + \|\mathbf{y}_i\|_2^2 \leq \sqrt{N}$, the natural nonconvex online gradient ascent guarantees $O(\sqrt{N})$ regret with high probability (see Theorem 3.2 in Marinov et al. (2018)) w.r.t. the original sequence $\{\mathbf{x}_i\}_{i\in[N]}$ (i.e., without the noise). While this model is somewhat similar to ours, there are three major differences. First, in

2. Note that in the fully-adversarial setting (i.e., there is no stochastic component), there is a $\Omega(\sqrt{N})$ lower bound on the regret, see for instance Warmuth and Kuzmin (2006b).

our setting the adversarial component is considered part of the data, and hence is included in the definition of the regret, as opposed to Marinov et al. (2018). Second, while in Marinov et al. (2018) it is required that the sum of magnitudes of the adversarial components is sublinear in N, which intuitively makes these components negligible in the regret analysis, in this work, we allow each adversarial component to be of constant magnitude, independent of the sequence length. Third, while Marinov et al. (2018) only gives an $O(\sqrt{N})$ regret bound, we show that a regret bound of $\text{poly}(\log N)$ can be obtained.

In another recent work Mianjy and Arora (2018), the authors consider the application of the online mirror decent algorithm to the *stochastic* online PCA problem (i.e., when the data points are sampled i.i.d. from a fixed unknown distribution). This mirror descent algorithm works on the convex semidefinite relaxation of the problem discussed above (and hence requires a potentially-expensive matrix factorization on each iteration). The authors show that under a spectral gap assumption in the distribution covariance matrix, adding strongly-convex regularization (w.r.t. the Euclidean norm) to the algorithm can improve the regret bound from $O(\sqrt{N})$ to $\text{poly}(\log N)$. In this work we draw inspiration from this observation, and show that also in our "semi-adversarial" model, the addition of a regularizing term can improve the regret bound of the nonconvex online gradient method from $\tilde{O}(\sqrt{N})$ to $\text{poly}(\log N)$.

2. Assumptions and Results

In this section we formally introduce our assumptions and main result. As discussed in the introduction, since our aim is make progress on a highly non-trivial problem of providing global convergence guarantees for a non-convex optimization algorithm in an online adversarial setting, our results do not hold for arbitrary (bounded) data, as is usually standard in *convex* online learning settings, but only for a more restricted family of input streams, namely those which follow a model we refer to in this paper as the *adversarially-perturbed spiked-covariance model*. Next we formally introduce this model.

2.1. Adversarially-Perturbed Spiked-Covariance Model

Throughout the paper we assume the data, i.e., the vectors $\{\mathbf{x}_t\}_{t\in[N]}$, satisfy the following assumption.

Assumption 1 (Perturbed Spiked Covariance Model) *We say a sequence of N vectors $\{\mathbf{x}_t\}_{t\in[N]} \subset \mathbb{R}^d$ satisfies Assumption 1, if for all $t \in [N]$, $\mathbf{x}_t$ can be written as $\mathbf{x}_t = \mathbf{q}_t + \mathbf{v}_t$, where $\{\mathbf{q}_t\}_{t\in[N]}$ are sampled i.i.d. from a distribution $\mathcal{D}$ and $\{\mathbf{v}_t\}_{t\in[N]}$ is a sequence of arbitrary bounded vectors such that the following conditions hold:*

1. *the vectors $\{\mathbf{v}_t\}_{t\in[N]}$ all lie in a Euclidean ball of radius V centered at the origin, i.e., $\max_{t\in[N]} \|\mathbf{v}_t\|_2 \le V$*
2. *the support of $\mathcal{D}$ is contained in a Euclidean ball of radius R centered at the origin, i.e., $\sup_{\mathbf{q}\in \text{support}(\mathcal{D})} \|\mathbf{q}\|_2 \le R$*
3. *$\mathcal{D}$ has zero mean, i.e., $\mathbb{E}_{\mathbf{q}\sim\mathcal{D}}[\mathbf{q}] = \mathbf{0}$*
4. *the covariance matrix $\mathbf{Q} := \mathbb{E}_{\mathbf{q}\sim\mathcal{D}}[\mathbf{q}\mathbf{q}^\top]$, admits an eigengap $\delta(\mathbf{Q}) := \lambda_1(\mathbf{Q}) - \lambda_2(\mathbf{Q})$ which satisfies $\delta(\mathbf{Q}) \ge V\sqrt{2\lambda_1(\mathbf{Q}) + V^2} + \varepsilon$, for some $\varepsilon > 0$.*

We now make a few remarks regarding Assumption 1. Item (1) assumes that the adversarial perturbations are bounded which is standard in the online learning literature, Item (2) is also a standard assumption, which is used to apply standard concentration arguments for sums of i.i.d random variables. Item (3), i.e., the assumption that the distribution as zero mean, while often standard, is not mandatory in general for our analysis technique to hold, however since it greatly simplifies the analysis we make it.

To better understand Item (4), it helps to think of $\delta(\mathbf{Q}), V^2, \varepsilon$ as quantities proportional to $\lambda_1(\mathbf{Q})$, i.e., consider $\delta(\mathbf{Q}) = c_\delta \lambda_1(\mathbf{Q})$, $V^2 = c_V \lambda_1(\mathbf{Q})$, $\varepsilon = c_\varepsilon \lambda_1(\mathbf{Q})$, for some universal constants $c_\delta, c_V, c_\varepsilon \in (0,1)$. Now, Item (4) in the assumption boils down the the condition $c_\delta \geq \sqrt{2c_V + c_V^2 + c_\varepsilon}$[3]. That is, the eigengap in the covariance $\mathbf{Q}$ needs to dominate the adversarial perturbations in a certain way. Note that in principle, this regime of parameters still allows the ratio V/R (i.e., ratio between maximal magnitude of adversarial component and maximal magnitude of stochastic component) to even be a universal constant. It is also important to note that the assumption of a non-negligeble eigengap in the covariance matrix is natural for PCA and is often observed in practice. We further discuss this assumption after presenting our main theorem - Theorem 1 in the following subsection.

Connection with stochastic i.i.d. models: note that when setting $V = 0$ in Assumption 1 (i.e., there is no adversarial component), our setting reduces to the well studied standard stochastic i.i.d. setting. In particular, in this case item (4) in Assumption 1 simply reduces to the standard assumption in this model that the covariance admits an eigengap bounded away from zero ($\delta(\mathbf{Q}) \geq \varepsilon$).

2.2. Algorithm and Convergence Result

For simplicity of the analysis we consider the data as arriving in blocks of length ℓ, where ℓ is a parameter to be determined later. Towards this end, we assume that $N = T\ell$ for some integer T and we consider prediction in T rounds, such that on each round $t \in [T]$, the algorithm predicts on all ℓ vectors in the tth block, which we denote by $\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(\ell)}$. It is important to emphasize that, while our algorithm considers the original data in blocks, it requires only $O(d)$ memory and $O(d)$ time to process each data point $\mathbf{x}_t^{(i)}$ for any $t \in [T], i \in [\ell]$.

Our algorithm, which we refer to as *nonconvex online gradient ascent*, is given below (see Algorithm 1). Our algorithm comes in two variants, one without additional regularization and one with. As can be seen, the non-regularized version is equivalent to applying the Online Gradient Ascent algorithm Zinkevich (2003); Hazan (2016) with the payoff function $f_t(\mathbf{w}) := \frac{1}{2}\sum_{i=1}^{\ell}(\mathbf{w}^\top \mathbf{x}_t^{(i)})^2$ on each round $t \in [T]$, where $\mathbf{w}$ is constrained to be a unit vector (though here we recall that the payoff function is not concave in $\mathbf{w}$, and the feasible set is not convex). The regularized version is similar, but considers the regularized payoff function $f_t^\alpha(\mathbf{w}) := \frac{1}{2}\sum_{i=1}^{\ell}(\mathbf{w}^\top \mathbf{x}_t^{(i)})^2 - \frac{\alpha}{2}\|\mathbf{w}\|_2^2$. [4]

3. or alternatively, $c_V \leq \sqrt{1 + (c_\delta - c_\varepsilon)^2} - 1 \leq (c_\delta - c_\varepsilon)^2/2$, where the last inequality follows from the standard inequality $\sqrt{a+b} \leq \sqrt{a} + \frac{b}{2\sqrt{a}}$.

4. While at this point it may not be immediately clear how the introduction of the regularizer helps, since we are optimizing over the unit sphere (hence the regularizer has the same value for all feasible points), it will be apparent in the analysis that it allows to obtain faster rates, similarly to the way that adding a strongly convex regularizer enables to obtain faster rates in the standard setting of online convex optimization Hazan (2016).

Algorithm 1 Nonconvex Online Gradient Ascent for Online PCA

1: input: unit vector $\hat{\mathbf{w}}_1$, sequence of learning rates $\{\eta_t\}_{t\geq 1}$, regularization parameter $\alpha \geq 0$.
2: **for** $t = 1 \dots T$ **do**
3: predict vector $\hat{\mathbf{w}}_t$ and observe ℓ vectors $\mathbf{x}_t^{(1)}, \dots, \mathbf{x}_t^{(\ell)}$ and payoff $\sum_{i=1}^{\ell} (\hat{\mathbf{w}}_t^\top \mathbf{x}_t^{(i)})^2$
4: compute the update

$$\hat{\mathbf{w}}_{t+1} \leftarrow \frac{\hat{\mathbf{w}}_t + \eta_t \sum_{i=1}^{\ell} \mathbf{x}_t^{(i)} \mathbf{x}_t^{(i)\top} \hat{\mathbf{w}}_t}{\|\hat{\mathbf{w}}_t + \eta_t \sum_{i=1}^{\ell} \mathbf{x}_t^{(i)} \mathbf{x}_t^{(i)\top} \hat{\mathbf{w}}_t\|_2} \qquad \{\text{without regularization}\}$$

OR

$$\hat{\mathbf{w}}_{t+1} \leftarrow \frac{(1-\eta_t\alpha)\hat{\mathbf{w}}_t + \eta_t \sum_{i=1}^{\ell} \mathbf{x}_t^{(i)} \mathbf{x}_t^{(i)\top} \hat{\mathbf{w}}_t}{\|(1-\eta_t\alpha)\hat{\mathbf{w}}_t + \eta_t \sum_{i=1}^{\ell} \mathbf{x}_t^{(i)} \mathbf{x}_t^{(i)\top} \hat{\mathbf{w}}_t\|_2} \qquad \{\text{with regularization}\}$$

5: **end for**

The following two theorem state our main results.

Theorem 1 *[convergence of Algorithm 1 without regularization ($\alpha = 0$) and constant learning rate] Consider a sequence of vectors $\{\mathbf{x}_t\}_{t\in[N]}$ which follows Assumption 1 and fix $p \in (0,1)$. For N large enough, there exists an integer $\ell = O\left(\frac{R^4\lambda_1(\mathbf{Q})^2}{(\delta(\mathbf{Q})^2 - V^2(2\lambda_1(\mathbf{Q})+V^2))^2} \log\frac{dN}{p}\right)$, such that applying Algorithm 1 with blocks of length ℓ and initialization $\hat{\mathbf{w}}_1$ which satisfies $(\hat{\mathbf{w}}_1^\top \mathbf{x})^2 \geq 1 - \frac{\delta(\mathbf{Q})-V^2}{2\lambda_1(\mathbf{Q})}\left(1 - \frac{\delta(\mathbf{Q})}{9\lambda_1(\mathbf{Q})}\right)$, where $\mathbf{x}$ is the leading eigenvector of $\mathbf{Q}$ (as defined in Assumption 1), and with a constant learning rate $\eta_t = \eta = \frac{1}{\sqrt{T\ell}(R+V)^2}$ for all $t \in [T]$ and without regularization (i.e., $\alpha = 0$), guarantees that with probability at least $1 - p$, the regret is upper-bounded by $O\left(\sqrt{N \log\frac{dN}{p}} \frac{\lambda_1(\mathbf{Q})R^4}{\delta(\mathbf{Q})^2 - V^2(V^2+\lambda_1(\mathbf{Q}))}\right)$.*

Theorem 1 roughly says that when the distribution covariance has a large-enough eigengap with respect to the adversarial perturbations (item 4 in Assumption 1), then non-convex OGA converges from a "warm-start" with $\tilde{O}(\sqrt{N})$ regret. Intuitively, the condition on the eigengap implies that the best-in-hindsight eigenvector cannot be far from $\mathbf{x}$ - the leading eigenvector of the distribution covariance by more than a certain constant. Hence, Theorem 1 can be seen as an online "local" convergence result. Importantly, it is not hard to show that under the conditions of the theorem, the best-in-hindsight eigenvector can also be far from both the initial vector $\hat{\mathbf{w}}_1$ and from $\mathbf{x}$ by a constant (and hence in particular both $\hat{\mathbf{w}}_1$ and $\mathbf{x}$ can incur linear regret). Hence, while our setting is strictly easier than the fully-adversarial online learning setting, it still a highly non-trivial online learning setting. In particular, all previous algorithms that provably minimize the regret under the conditions of Theorem 1 require quadratic memory and quadratic runtime per data-point.

Our second main result shows that the regularized version of Algorithm 1 (i.e., with $\alpha > 0$) can guarantee poly-logarithmic regret in N under Assumption 1.

Theorem 2 *[convergence of Algorithm 1 with regularization ($\alpha > 0$)] Consider a sequence of vectors $\{\mathbf{x}_t\}_{t\in[N]}$ which follows Assumption 1 and fix $p \in (0,1)$. For N large enough, there exists an integer $\ell = O\left(\frac{R^4\lambda_1(\mathbf{Q})^2}{(\delta(\mathbf{Q})^2 - V^2(\lambda_1(\mathbf{Q})+V^2))^2} \log\frac{dN}{p}\right)$, such that applying Algorithm 1 with blocks of length ℓ, regularization parameter which satisfies $\alpha = \Theta\left(\frac{\ell}{\delta(\mathbf{Q})+V^2}\left(\delta(\mathbf{Q})^2 - V^4 - 2V^2\lambda_1(\mathbf{Q})\right)\right)$, and initialization $\hat{\mathbf{w}}_1$ which satisfies $(\hat{\mathbf{w}}_1^\top \mathbf{x})^2 \geq 1 - \frac{\delta(\mathbf{Q})-V^2}{2\lambda_1(\mathbf{Q})}\left(\frac{9}{10} - \frac{\delta(\mathbf{Q})}{9\lambda_1(\mathbf{Q})}\right)$, where $\mathbf{x}$ is the leading eigenvector of $\mathbf{Q}$ (as defined in Assumption 1), and with learning rate $\eta_t = \frac{1}{\alpha t + T_0}$ for some*

$T_0 \geq 0$ large enough, guarantees that with probability at least $1-p$, the regret is upper-bounded by $O\left(\frac{R^8(\delta(\mathbf{Q})+V^2)\lambda_1(\mathbf{Q})^2}{(\delta(\mathbf{Q})^2-V^4-2\lambda_1(\mathbf{Q})V^2)^3}\log(N)\log\left(\frac{dN}{p}\right)\right)$.

It is important to note that the ability to obtain a poly-logarithmic regret bound as Theorem 2 suggests, relies crucially on the eigen-gap assumption in the stochastic covariance matrix in Assumption 1. In particular, for the standard fully-adversarial version of the problem (i.e., there is no stochastic component in the data vectors) it is not possible in worst case to improve over the known $O(\sqrt{N})$ bound (see Warmuth and Kuzmin (2006b)). We see this fact as further motivation for studying intermediate models that bridge between the fully-adversarial and fully-stochastic settings, showing that such models can result in much improved convergence guarantees compared to the possibly over-pessimistic fully-adversarial setting.

2.2.1. COMPUTING A "WARM-START" VECTOR

We now discuss the possibility of satisfying the "warm-start" requirement in Theorem 1.

First, we note that given the possibility to sample i.i.d. points from the underlying distribution $\mathcal{D}$, it is straightforward to obtain a warm-start vector $\hat{\mathbf{w}}_1$, as required by Theorem 1, by simply initializing $\hat{\mathbf{w}}_1$ to be the leading eigenvector of the empirical covariance of a size-n sample of such points. It is not difficult to show via standard tools such as the Davis-Kahan $\sin\theta$ theorem and a Matrix-Hoeffding concentration inequality (see for instance the proof of the following Lemma 3), that for any $(\epsilon, p) \in (0,1)^2$, a sample of size $n = O\left(\frac{R^4 \ln(d/p)}{\epsilon\delta(\mathbf{Q})^2}\right)$ suffices, so the outcome $\hat{\mathbf{w}}_1$ satisfies: $(\hat{\mathbf{w}}_1^\top \mathbf{x})^2 \geq 1-\epsilon$ with probability at least $1-p$.

If sampling directly from $\mathcal{D}$ is not possible, the following lemma, whose proof is given in the appendix, shows that with a simple additional assumption on the parameters $\delta(\mathbf{Q}), \lambda_1(\mathbf{Q}), V^2$, it is possible to obtain the warm-start initialization directly using data that follows Assumption 1. Moreover, the sample-size n required is independent of the sequence length N, and hence using for instance the first n vectors in the stream to compute such initialization, deteriorates the overall regret bound in Theorems 1, 2 only by a low-order term.

Lemma 3 *[warm-start] Fix some $c \in (0,1]$ and suppose that in addition to Assumption 1 it also holds that $\delta(\mathbf{Q}) \geq \left(32c^{-1}\lambda_1(\mathbf{Q})V^4\right)^{1/3}$. Then, for any $p \in (0,1)$ there exists a sample size $n = O\left(\frac{R^4\lambda_1 \log(d/p)}{c\delta(\mathbf{Q})^3}\right)$, such that initializing $\hat{\mathbf{w}}_1$ to be the leading eigenvector of the empirical covariance $\hat{\mathbf{X}} := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$, where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ follow Assumption 1, guarantees that with probability at least $1-p$: $(\hat{\mathbf{w}}_1^\top \mathbf{x})^2 \geq 1 - c\frac{\delta(\mathbf{Q})-V^2}{2\lambda_1(\mathbf{Q})}$.*

3. Analysis

At a high-level, the proof of Theorems 1, 2 relies on the combination of the following three ideas:

1. We build on the fact that the Online PCA problem, when cast as online *linear* optimization over the spectraedron (i.e., when the decision variable is lifted from a unit vector to a positive semidefinite matrix of unit trace), is online learnable via a standard application of online gradient ascent, which achieves an $O(\sqrt{N})$ regret bound in the non-regularized case and $O(\log N)$ regret bound with additional regularization (note however that as discussed above, this approach requires a full SVD computation on each iteration to compute the projection onto the spectrahedron).

2. We prove, that under Assumption 1, the above "inefficient" algorithm, when initialized with a proper "warm-start" vector, guarantees that the projection onto the spectrahedron is always a rank-one matrix (hence, only a rank-one SVD computation per iteration is required).

3. Finally, we show that the nonconvex online gradient ascent algorithm, Algorithm 1, approximates sufficiently well the steps of the above algorithm (in case the projection is rank-one), avoiding SVD computations all together.

We introduce the following notation that will be used throughout the analysis. For vectors in $\mathbb{R}^d$ we let $\|\cdot\|$ denote the standard Euclidean norm, and for matrices in $\mathbb{R}^{m\times n}$ we let $\|\cdot\|_F$ denote the Frobenius (Euclidean) norm and we let $\|\cdot\|_2$ denote the spectral norm (largest singular value). For all $t \in [T]$, we define $\mathbf{X}_t := \sum_{i=1}^{\ell} \mathbf{x}_t^{(i)}\mathbf{x}_t^{(i)\top}$, $\mathbf{Q}_t := \sum_{i=1}^{\ell} \mathbf{q}_t^{(i)}\mathbf{q}_t^{(i)\top}$, $\mathbf{V}_t := \sum_{i=1}^{\ell} \mathbf{v}_t^{(i)}\mathbf{v}_t^{(i)\top}$, and $\mathbf{M}_t := \sum_{i=1}^{\ell} \mathbf{q}_t^{(i)}\mathbf{v}_t^{(i)\top} + \mathbf{v}_t^{(i)}\mathbf{q}_t^{(i)\top}$. Note that $\mathbf{X}_t = \mathbf{Q}_t + \mathbf{M}_t + \mathbf{V}_t$. Recall that we let $\mathbf{Q}$ denote the covariance matrix associated with the distribution $\mathcal{D}$ (as detailed in Assumption 1), and we let $\lambda_1(\mathbf{Q}), \ldots, \lambda_d(\mathbf{Q})$ denote its eigenvalues in descending order. Also, we let $\mathbf{x}$ denote the leading eigenvector of $\mathbf{Q}$, which under Assumption 1, is unique. We also define $\mathbf{D}_t := \mathbf{Q}_t - \ell \cdot \mathbf{Q} + \mathbf{M}_t$. Note that $\mathbf{X}_t = \ell \cdot \mathbf{Q} + \mathbf{V}_t + \mathbf{D}_t$. Intuitively, under Assumption 1, $\frac{1}{\ell}\mathbf{D}_t$ converges to zero in probability as $\ell \to \infty$.

We denote by $\mathcal{S}$ the spectrahedron, i.e., $\mathcal{S} := \{\mathbf{W} \in \mathbb{R}^{d\times d} \,|\, \mathbf{W} \succeq 0, \mathrm{Tr}(\mathbf{W}) = 1\}$, and we let $\Pi_{\mathcal{S}}[\mathbf{W}]$ denote the Euclidean projection of a symmetric matrix $\mathbf{W} \in \mathbb{R}^{d\times d}$ onto $\mathcal{S}$.

Our main building block towards proving Theorem 1 is to analyze the regret of a different non-convex algorithm for Online PCA. The meta-algorithm, Algorithm 2, builds on the standard convexification scheme for Online PCA, i.e., "lifting" the decision set from the unit ball to the spectrahedron, however, instead of computing exact projections onto the spectrahedron, it follows a nonconvex approach of only approximating the projection via a rank-one solution. We refer to it as a meta-algorithm, since for a given approximation parameter γ, it only requires on each iteration to find an approximate leading eigenvector of the matrix to be projected onto $\mathcal{S}$. As in Algorithm 1, this algorithm also has two variants, one which does not include an additional regularizing term, and one which does.

Note that a straightforward implementation of Algorithm 2 with $\gamma = 0$ corresponds to updating $\hat{\mathbf{w}}_{t+1}$ via accurate SVD of the $d \times (\ell+1)$ matrix $(\sqrt{1-\eta_t\alpha}\hat{\mathbf{w}}_t, \sqrt{\eta_t}\mathbf{x}_t^{(1)}, \ldots, \sqrt{\eta_t}\mathbf{x}_t^{(\ell)})$, which already yields an algorithm with $O(\ell d)$ memory and $O(\ell d)$ amortized runtime per data-point.

Due to lack of space, in the sequel we focus only on the main building-blocks of our analysis. All missing details are given in full detail in the appendix.

Lemma 4 *Let $\mathbf{w} \in \mathbb{R}^d$ be a unit vector and let $\mathbf{X} \in \mathbb{R}^{d\times d}$ be positive semidefinite. Let $\mathbf{w}'$ be the leading eigenvector of the matrix $\mathbf{W} := (1-\eta\alpha)\mathbf{w}\mathbf{w}^\top + \eta\mathbf{X}$, for some $\eta > 0, \alpha \geq 0$ such that $\eta\alpha < 1$. If $\mathbf{w}^\top\mathbf{X}\mathbf{w} \geq \frac{\lambda_1(\mathbf{X})+\lambda_2(\mathbf{X})+\alpha}{2}$, then it follows that $\mathbf{w}'\mathbf{w}'^\top = \Pi_{\mathcal{S}}[\mathbf{W}]$.*

Proof Recall $\mathbf{w}'$ denotes the leading eigenvector of $\mathbf{W}$ and let $\mathbf{y}_2, \ldots, \mathbf{y}_d$ denote the other eigenvectors in non-increasing order (according to the corresponding eigenvalues). It is well known that the projection of $\mathbf{W}$ onto $\mathcal{S}$ is given by

$$\Pi_{\mathcal{S}}[\mathbf{W}] = (\lambda_1(\mathbf{W}) - \lambda)\mathbf{w}'\mathbf{w}'^\top + \sum_{i=2}^{d} \max\{0, \lambda_i(\mathbf{W}) - \lambda\}\mathbf{y}_i\mathbf{y}_i^\top,$$

Algorithm 2 Approximate Non-convex Rank-one Online Gradient Ascent

1: input: unit vector $\hat{\mathbf{w}}_1$, sequence of learning rates $\{\eta_t\}_{t\geq 1}$, sequence of approximation parameters $\{\gamma_t\}_{t\geq 1}$, regularization parameter $\alpha \geq 0$
2: **for** $t = 1 \dots T$ **do**
3: predict vector $\hat{\mathbf{w}}_t$ and observe ℓ vectors $\mathbf{x}_t^{(1)}, \dots, \mathbf{x}_t^{(\ell)}$
4: $\hat{\mathbf{w}}_{t+1} \leftarrow$ some unit vector satisfying: $\|\hat{\mathbf{w}}_{t+1}\hat{\mathbf{w}}_{t+1}^\top - \mathbf{w}_{t+1}\mathbf{w}_{t+1}^\top\|_F \leq \gamma_t$, where $\mathbf{w}_{t+1}$ is the leading eigenvector of either

$$\mathbf{W}_{t+1} := \hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top + \eta_t \textstyle\sum_{i=1}^{\ell} \mathbf{x}_t^{(i)}\mathbf{x}_t^{(i)\top} \qquad \{\text{without regularization}\}$$

OR

$$\mathbf{W}_{t+1} := (1-\eta_t\alpha)\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top + \eta_t \textstyle\sum_{i=1}^{\ell} \mathbf{x}_t^{(i)}\mathbf{x}_t^{(i)\top} \qquad \{\text{with regularization}\}$$

5: **end for**

where λ is a non-negative real scalar such that $\lambda_1(\mathbf{W}) - \lambda + \sum_{i=2}^{d} \max\{\lambda_i(\mathbf{W}) - \lambda, 0\} = 1$. Thus, if we show that $\lambda_1(\mathbf{W}) \geq 1 + \lambda_2(\mathbf{W})$, then it in particular follows that $\Pi_{\mathcal{S}}[\mathbf{W}] = \mathbf{w}'\mathbf{w}'^\top$.

Note that on one hand,

$$\lambda_1(\mathbf{W}) \geq \mathbf{w}^\top\mathbf{W}\mathbf{w} = 1 - \eta\alpha + \eta\mathbf{w}^\top\mathbf{X}\mathbf{w}. \tag{1}$$

On the other-hand, using the last inequality, we can also write

$$\begin{aligned} \lambda_2(\mathbf{W}) &\leq \lambda_1(\mathbf{W}) + \lambda_2(\mathbf{W}) - \mathbf{w}^\top\mathbf{W}\mathbf{w} \\ &= \lambda_1(\mathbf{W}) + \lambda_2(\mathbf{W}) - 1 + \eta\alpha - \eta\mathbf{w}^\top\mathbf{X}\mathbf{w}. \end{aligned} \tag{2}$$

Using Ky Fan's eigenvalue inequality, we have that

$$\begin{aligned} \lambda_1(\mathbf{W}) + \lambda_2(\mathbf{W}) &= \lambda_1((1-\eta\alpha)\mathbf{w}\mathbf{w}^\top + \eta\mathbf{X}) + \lambda_2((1-\eta\alpha)\mathbf{w}\mathbf{w}^\top + \eta\mathbf{X}) \\ &\leq \lambda_1((1-\eta\alpha)\mathbf{w}\mathbf{w}^\top) + \lambda_2((1-\eta\alpha)\mathbf{w}\mathbf{w}^\top) + \lambda_1(\eta\mathbf{X}) + \lambda_2(\eta\mathbf{X}) \\ &= 1 - \eta\alpha + \eta\left(\lambda_1(\mathbf{X}) + \lambda_2(\mathbf{X})\right). \end{aligned} \tag{3}$$

Thus, by combining Eq. (1), (2), (3), we arrive at the following sufficient condition so that $\mathbf{w}'\mathbf{w}'^\top = \Pi_{\mathcal{S}}[\mathbf{W}]$:

$$1 - \eta\alpha + \eta\mathbf{w}^\top\mathbf{X}\mathbf{w} \geq 1 + \left(1 - \eta\alpha + \eta\left(\lambda_1(\mathbf{X}) + \lambda_2(\mathbf{X})\right)\right) - \left(1 - \eta\alpha + \eta\mathbf{w}^\top\mathbf{X}\mathbf{w}\right),$$

which is equivalent to the condition $\mathbf{w}^\top\mathbf{X}\mathbf{w} \geq \frac{\lambda_1(\mathbf{X})+\lambda_2(\mathbf{X})}{2} + \frac{\alpha}{2}$. ■

Lemma 5 *Suppose that on some iteration t of Algorithm 2 it holds that $\eta_t\alpha < 1$ and*

$$(\hat{\mathbf{w}}_t^\top\mathbf{x})^2 \geq 1 - \frac{\delta(\mathbf{Q}) - V^2 - \ell^{-1}\alpha}{2\lambda_1(\mathbf{Q})} + \frac{2\|\mathbf{D}_t\|}{\ell\cdot\lambda_1(\mathbf{Q})}$$

Then, $\mathbf{w}_{t+1}\mathbf{w}_{t+1}^\top = \Pi_{\mathcal{S}}[\mathbf{W}_{t+1}]$.

Proof Recall that $\mathbf{X}_t := \sum_{i=1}^{\ell} \mathbf{x}_t^{(i)}\mathbf{x}_t^{(i)\top}$ and that $\mathbf{W}_{t+1} = (1-\eta_t\alpha)\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top + \eta_t\mathbf{X}_t$.

Using Lemma 4 it suffices to show that

$$\hat{\mathbf{w}}_t^\top\mathbf{X}_t\hat{\mathbf{w}}_t \geq \frac{\lambda_1(\mathbf{X}_t) + \lambda_2(\mathbf{X}_t) + \alpha}{2}. \tag{4}$$

On one hand we have

$$\begin{aligned}\hat{\mathbf{w}}_t^\top \mathbf{X}_t \hat{\mathbf{w}}_t &= \hat{\mathbf{w}}_t^\top (\ell \cdot \mathbf{Q} + \mathbf{V}_t + \mathbf{D}_t)\hat{\mathbf{w}}_t \geq \ell \cdot \hat{\mathbf{w}}_t^\top \mathbf{Q}\hat{\mathbf{w}}_t - \|\mathbf{D}_t\| \\ &\geq \ell(\hat{\mathbf{w}}_t^\top \mathbf{x})^2 \cdot \mathbf{x}^\top \mathbf{Q}\mathbf{x} - \|\mathbf{D}_t\| = \ell(\hat{\mathbf{w}}_t^\top \mathbf{x})^2 \cdot \lambda_1(\mathbf{Q}) - \|\mathbf{D}_t\|. \end{aligned} \tag{5}$$

On the other hand,

$$\begin{aligned}\lambda_1(\mathbf{X}_t) + \lambda_2(\mathbf{X}_t) &= \lambda_1(\ell \cdot \mathbf{Q} + \mathbf{V}_t + \mathbf{D}_t) + \lambda_2(\ell \cdot \mathbf{Q} + \mathbf{V}_t + \mathbf{D}_t) \\ &\underset{(a)}{\leq} \lambda_1(\ell \cdot \mathbf{Q} + \mathbf{V}_t) + \lambda_2(\ell \cdot \mathbf{Q} + \mathbf{V}_t) + 2\|\mathbf{D}_t\| \\ &\underset{(b)}{\leq} \lambda_1(\ell \cdot \mathbf{Q}) + \lambda_2(\ell \cdot \mathbf{Q}) + \lambda_1(\mathbf{V}_t) + \lambda_2(\mathbf{V}_t) + 2\|\mathbf{D}_t\| \\ &\leq \ell(\lambda_1(\mathbf{Q}) + \lambda_2(\mathbf{Q})) + \mathrm{Tr}(\mathbf{V}_t) + 2\|\mathbf{D}_t\| \leq \ell(\lambda_1(\mathbf{Q}) + \lambda_2(\mathbf{Q}) + V^2) + 2\|\mathbf{D}_t\|, \end{aligned} \tag{6}$$

where (a) follows from Weyl's eigenvalue inequality, and (b) follows from Ky Fan's eigenvalue inequality.

Combining Eq. (4), (5), (6), we arrive at the following sufficient condition so that $\mathbf{w}_{t+1}\mathbf{w}_{t+1}^\top = \Pi_{\mathcal{S}}[\mathbf{W}_{t+1}]$:

$$\begin{aligned}(\hat{\mathbf{w}}_t^\top \mathbf{x})^2 &\geq \frac{\lambda_1(\mathbf{Q}) + \lambda_2(\mathbf{Q}) + V^2 + 4\ell^{-1}\|\mathbf{D}_t\| + \ell^{-1}\alpha}{2 \cdot \lambda_1(\mathbf{Q})} = \frac{2\lambda_1(\mathbf{Q}) - \delta(\mathbf{Q}) + V^2 + 4\ell^{-1}\|\mathbf{D}_t\| + \ell^{-1}\alpha}{2\lambda_1(\mathbf{Q})} \\ &= 1 - \frac{\delta(\mathbf{Q}) - V^2 - \ell^{-1}\alpha}{2\lambda_1(\mathbf{Q})} + \frac{2\|\mathbf{D}_t\|}{\ell \cdot \lambda_1(\mathbf{Q})}.\end{aligned}$$

■

Lemma 6 *Suppose that on some iteration t of Algorithm 2 it holds that $(\mathbf{x}^\top \hat{\mathbf{w}}_t)^2 \geq \frac{1}{2}$. Then, for any learning rate $\eta_t > 0$ and $\alpha > 0$ such that $\eta_t \alpha < 1$, it holds that*

$$(\mathbf{x}^\top \mathbf{w}_{t+1})^2 \geq (\mathbf{x}^\top \hat{\mathbf{w}}_t)^2 + \eta_t \ell \frac{(1 - (\mathbf{x}^\top \hat{\mathbf{w}}_t)^2) \cdot \delta(\mathbf{Q}) - (\mathbf{x}^\top \hat{\mathbf{w}}_t)^2 V^2 - 4\ell^{-1}\|\mathbf{D}_t\|}{\lambda_1(\mathbf{W}_{t+1}) - \lambda_2(\mathbf{W}_{t+1})}.$$

Proof Fix some iteration t. We introduce the short notation $\mathbf{w} = \hat{\mathbf{w}}_t$, $\mathbf{w}' = \mathbf{w}_{t+1}$, $\mathbf{W} = \mathbf{W}_{t+1} = (1 - \eta_t\alpha)\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top + \eta_t \mathbf{X}_t$, $\lambda_1 = \lambda_1(\mathbf{W}_{t+1})$, $\lambda_2 = \lambda_2(\mathbf{W}_{t+1})$, and for all $i \geq 2$, $\mathbf{y}_i$ is the eigenvector of $\mathbf{W}_{t+1}$ associated with eigenvalue $\lambda_i = \lambda_i(\mathbf{W}_{t+1})$.

It holds that $\lambda_1 \mathbf{w}'\mathbf{w}' + \sum_{i=2}^d \lambda_i \mathbf{y}_i \mathbf{y}_i^\top = \mathbf{W} = (1 - \eta_t\alpha)\mathbf{w}\mathbf{w}^\top + \eta_t \mathbf{X}_t$. Thus, we have that

$$(\mathbf{x}^\top \mathbf{w}')^2 = \frac{\mathbf{x}^\top \mathbf{W}\mathbf{x} - \sum_{i=2}^d \lambda_i (\mathbf{x}^\top \mathbf{y}_i)^2}{\lambda_1} = \frac{(1 - \eta_t\alpha)(\mathbf{x}^\top \mathbf{w})^2 + \eta_t \mathbf{x}^\top \mathbf{X}_t \mathbf{x} - \sum_{i=2}^d \lambda_i (\mathbf{x}^\top \mathbf{y}_i)^2}{\lambda_1}.$$

Note that $\sum_{i=2}^d \lambda_i (\mathbf{x}^\top \mathbf{y}_i)^2 \leq \lambda_2(\|\mathbf{x}\|^2 - (\mathbf{x}^\top \mathbf{w}')^2) = \lambda_2(1 - (\mathbf{x}^\top \mathbf{w}')^2)$. Thus, we have that

$$(\mathbf{x}^\top \mathbf{w}')^2 \geq \frac{(1 - \eta_t\alpha)(\mathbf{x}^\top \mathbf{w})^2 + \eta_t \mathbf{x}^\top \mathbf{X}_t \mathbf{x} - \lambda_2(1 - (\mathbf{x}^\top \mathbf{w}')^2)}{\lambda_1}.$$

Rearranging we obtain,

$$\begin{aligned}(\mathbf{x}^\top \mathbf{w}')^2 &\geq \frac{(1 - \eta_t\alpha)(\mathbf{x}^\top \mathbf{w})^2 + \eta_t \mathbf{x}^\top \mathbf{X}_t \mathbf{x} - \lambda_2}{\lambda_1 - \lambda_2} \\ &= (\mathbf{x}^\top \mathbf{w})^2 + \frac{\eta_t \mathbf{x}^\top \mathbf{X}_t \mathbf{x} + (1 - \eta_t\alpha - \lambda_1 + \lambda_2)(\mathbf{x}^\top \mathbf{w})^2 - \lambda_2}{\lambda_1 - \lambda_2} \\ &= (\mathbf{x}^\top \mathbf{w})^2 + \frac{\eta_t \mathbf{x}^\top \mathbf{X}_t \mathbf{x} + (1 - \eta_t\alpha - (\lambda_1 + \lambda_2))(\mathbf{x}^\top \mathbf{w})^2 + \lambda_2(2(\mathbf{x}^\top \mathbf{w})^2 - 1)}{\lambda_1 - \lambda_2}.\end{aligned}$$

Note that via Ky Fan's inequality we have that

$$\begin{aligned}\lambda_1+\lambda_2 &\leq \lambda_1((1-\eta_t\alpha)\mathbf{w}\mathbf{w}^\top)+\lambda_2((1-\eta_t\alpha)\mathbf{w}\mathbf{w}^\top)+\lambda_1(\eta_t\mathbf{X}_t)+\lambda_2(\eta_t\mathbf{X}_t)\\ &= 1-\eta_t\alpha+\eta_t(\lambda_1(\mathbf{X}_t)+\lambda_2(\mathbf{X}_t)).\end{aligned}$$

Also, $\lambda_2=\lambda_2((1-\eta_t\alpha)\mathbf{w}\mathbf{w}^\top+\eta_t\mathbf{X}_t)\geq\lambda_2(\eta_t\mathbf{X}_t)=\eta_t\lambda_2(\mathbf{X}_t)$.

Thus, using our assumption that $(\mathbf{x}^\top\mathbf{w})^2\geq 1/2$, we have that

$$(\mathbf{x}^\top\mathbf{w}')^2\geq(\mathbf{x}^\top\mathbf{w})^2+\eta_t\frac{\mathbf{x}^\top\mathbf{X}_t\mathbf{x}-(\lambda_1(\mathbf{X}_t)+\lambda_2(\mathbf{X}_t))(\mathbf{x}^\top\mathbf{w})^2+\lambda_2(\mathbf{X}_t)(2(\mathbf{x}^\top\mathbf{w})^2-1)}{\lambda_1-\lambda_2}.$$

Note that

$$\mathbf{x}^\top\mathbf{X}_t\mathbf{x}=\mathbf{x}^\top(\ell\cdot\mathbf{Q}+\mathbf{V}_t+\mathbf{D}_t)\mathbf{x}\geq\mathbf{x}^\top(\ell\cdot\mathbf{Q}+\mathbf{D}_t)\mathbf{x}\geq\ell\cdot\lambda_1(\mathbf{Q})-\|\mathbf{D}_t\|,$$

$$\lambda_2(\mathbf{X}_t)=\lambda_2(\ell\cdot\mathbf{Q}+\mathbf{V}_t+\mathbf{D}_t)\geq\lambda_2(\ell\cdot\mathbf{Q}+\mathbf{D}_t)\geq\ell\cdot\lambda_2(\mathbf{Q})-\|\mathbf{D}_t\|,$$

$$\begin{aligned}\lambda_1(\mathbf{X}_t)+\lambda_2(\mathbf{X}_t) &\underset{(a)}{\leq} \lambda_1(\ell\cdot\mathbf{Q})+\lambda_2(\ell\cdot\mathbf{Q})+\lambda_1(\mathbf{V}_t+\mathbf{D}_t)+\lambda_2(\mathbf{V}_t+\mathbf{D}_t)\\ &\leq \ell(\lambda_1(\mathbf{Q})+\lambda_2(\mathbf{Q}))+\lambda_1(\mathbf{V}_t)+\lambda_2(\mathbf{V}_t)+2\|\mathbf{D}_t\|\\ &\leq \ell(\lambda_1(\mathbf{Q})+\lambda_2(\mathbf{Q}))+\mathrm{Tr}(\mathbf{V}_t)+2\|\mathbf{D}_t\|\leq\ell(\lambda_1(\mathbf{Q})+\lambda_2(\mathbf{Q})+V^2)+2\|\mathbf{D}_t\|,\end{aligned}$$

where (a) follows again from Ky Fan's inequality.

Plugging-in all of the above bounds, we have that

$$\begin{aligned}(\mathbf{x}^\top\mathbf{w}')^2 &\geq (\mathbf{x}^\top\mathbf{w})^2+\eta_t\ell\frac{\lambda_1(\mathbf{Q})-(\lambda_1(\mathbf{Q})+\lambda_2(\mathbf{Q})+V^2)\cdot(\mathbf{x}^\top\mathbf{w})^2+\lambda_2(\mathbf{Q})\cdot(2(\mathbf{x}^\top\mathbf{w})^2-1)-4\ell^{-1}\|\mathbf{D}_t\|}{\lambda_1-\lambda_2}\\ &= (\mathbf{x}^\top\mathbf{w})^2+\eta_t\ell\frac{(1-(\mathbf{w}^\top\mathbf{x})^2)\cdot(\lambda_1(\mathbf{Q})-\lambda_2(\mathbf{Q}))-(\mathbf{x}^\top\mathbf{w})^2V^2-4\ell^{-1}\|\mathbf{D}_t\|}{\lambda_1-\lambda_2}\\ &= (\mathbf{x}^\top\mathbf{w})^2+\eta_t\ell\frac{(1-(\mathbf{x}^\top\mathbf{w})^2)\cdot\delta(\mathbf{Q})-(\mathbf{x}^\top\mathbf{w})^2V^2-4\ell^{-1}\|\mathbf{D}_t\|}{\lambda_1-\lambda_2}.\end{aligned}$$

■

The following lemma gives conditions under-which, for any iteration t of Algorithm 2, the projection $\Pi_{\mathcal{S}}[\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top+\eta_t\mathbf{X}_t]$ is rank-one. This essentially guarantees that Algorithm 2 (up to the approximation errors $\{\gamma_t\}_{t\geq1}$) achieves sublinear regret.

The proof follows via induction, showing with the help of Lemma 6, that the condition in Lemma 5 is satisfied on each iteration.

Lemma 7 *Suppose that when applying Algorithm 2, the following conditions hold for all $t\in[T]$:*

$$\frac{1}{\ell}\|\mathbf{D}_t\|\leq\epsilon\leq\frac{1}{72\lambda_1(\mathbf{Q})}\left(\delta(\mathbf{Q})^2-V^4-2V^2\lambda_1(\mathbf{Q})\right), \tag{7}$$

$$\eta_t\leq\min\left\{\frac{\epsilon}{4\ell\lambda_1(\mathbf{Q})\cdot(V^2+4\epsilon)},\ \frac{1}{\ell(R+V)^2}\right\},\ \gamma_t\leq\min\left\{\frac{\epsilon}{4\lambda_1(\mathbf{Q})},\ 18\epsilon\eta_t\ell\right\},$$

$$\alpha\leq\frac{\ell}{4(\delta(\mathbf{Q})+V^2)}\left(\delta(\mathbf{Q})^2-V^4-2V^2\lambda_1(\mathbf{Q})\right),\ (\mathbf{w}_1^\top\mathbf{x})^2\geq1-\frac{\delta(\mathbf{Q})-V^2-\ell^{-1}\alpha-4\epsilon}{2\lambda_1(\mathbf{Q})}.$$

Then, for all $t\in[T]$, $\mathbf{w}_{t+1}\mathbf{w}_{t+1}^\top=\Pi_{\mathcal{S}}[\mathbf{W}_{t+1}]$.

Finally, the following lemma is used to connect between the regret of Algorithm 2 and Algorithm 1, by essentially showing that Algorithm 1 is an instance of Algorithm 2, and bounding the corresponding sequence of approximation parameters $\{\gamma_t\}_{t\geq 1}$. The proof relies on viewing $\hat{\mathbf{w}}_{t+1}$ as the result of one iteration of the Power Method over the matrix $\mathbf{W}_{t+1}$ and using eigenvector perturbation arguments.

Lemma 8 *Consider some iteration t of Algorithm 1, and let $\mathbf{w}_{t+1}$ denote the leading eigenvector of the matrix $\mathbf{W}_{t+1} := (1+\eta_t\alpha)\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top + \eta_t\mathbf{X}_t$. If $\eta_t \leq \frac{1}{3\ell(R+V)^2}$ and $\alpha \leq \ell(R+V)^2$, then it holds that $\|\hat{\mathbf{w}}_{t+1}\hat{\mathbf{w}}_{t+1} - \mathbf{w}_{t+1}\mathbf{w}_{t+1}^\top\|_F \leq \sqrt{33}(\eta_t\ell(R+V)^2)^2$.*

4. Experiments

We test the following algorithms. Algorithm 2, where $\hat{\mathbf{w}}_{t+1}$ is computed via rank-one SVD, with block-size $\ell = 1$ (R1-OGA), and with non-unit block-size $\ell > 1$ (BR1-OGA), non-convex online gradient ascent, Algorithm 1, with unit block-size $\ell = 1$ (Nonconvex-OGA), and the *convex* online gradient ascent algorithm Zinkevich (2003); Hazan (2016) (equivalent to Algorithm 2, but uses accurate Euclidean projections onto the spectrahedron) with unit block-size (Conv-OGA) [5]. Finally, we record the regret of the initial "warm-start" vector $\hat{\mathbf{w}}_1$ (BaseVec), which serves as initialization for all algorithms. We plot for each iteration t the average-regret up to time t against the leading eigenvector in hindsight (computed w.r.t. all data). For all algorithms introduced in this paper we focus on the non-regularized version (i.e., we set $\alpha = 0$).

We consider the following three datasets: synthetic, MNIST LeCun et al. (1998) and CIFAR10 Krizhevsky (2009) (see appendix for further details).

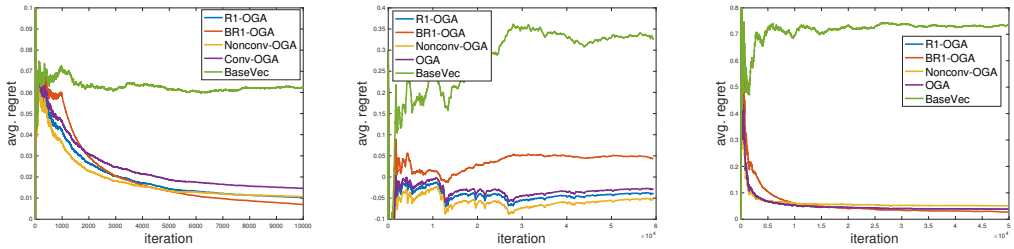

Figure 1: Average-regret on datasets: synthetic (left), MNIST (middle) and CIFAR10 (right).

We see that all algorithms indeed attain low average-regret, and in particular are competitive with OGA which follows a convex approach.

To further examine the applicability of our theoretical approach, for all datasets, we recorded for algorithm BR1-OGA the fraction of projection errors, i.e., the precent of number of iterations t on which the projection of the matrix $\mathbf{W}_{t+1} = \hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top + \eta\mathbf{X}_t$ onto the spectrahedron $\mathcal{S}$ is not a rank-one matrix. The results are 6.24%, 0.26%, 0%, for synthetic, MNIST and CIFAR10, respectively. These low error rates indeed support our theoretical analysis which hinges on showing that under our data model (recall Assumption 1) and given a "warm-start'" initialization, the projections of the matrices $\mathbf{W}_t$ in Algorithm 2 are always rank-one.

5. since computing that exact projection for Conv-OGA via a full SVD is highly time consuming, we approximate it by extracting only the five leading components

References

Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, pages 1195–1199, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4528-6. doi: 10.1145/3055399.3055464. URL http://doi.acm.org/10.1145/3055399.3055464.

Zeyuan Allen-Zhu. Natasha 2: Faster non-convex optimization than sgd. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2680–2691. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7533-natasha-2-faster-non-convex-optimization-than-sgd.pdf.

Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, pages 699–707, 2016.

Zeyuan Allen-Zhu and Yuanzhi Li. First efficient convergence for streaming k-pca: a global, gap-free, and near-optimal rate. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 487–492. IEEE, 2017a.

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

Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. Provable bounds for learning some deep representations. In *International Conference on Machine Learning*, pages 584–592, 2014.

Akshay Balsubramani, Sanjoy Dasgupta, and Yoav Freund. The fast convergence of incremental PCA. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, pages 3174–3182, 2013.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. "convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 654–663, 2017. URL http://proceedings.mlr.press/v70/carmon17a.html.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

Christopher De Sa, Christopher Re, and Kunle Olukotun. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *International Conference on Machine Learning*, pages 2332–2341, 2015.

Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 11–20. ACM, 2014.

Dan Garber and Elad Hazan. Fast and simple pca via convex optimization. *arXiv preprint arXiv:1509.05647*, 2015.

Dan Garber, Elad Hazan, and Tengyu Ma. Online learning of eigenvectors. In *ICML*, pages 560–568, 2015.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

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

Elad Hazan, Satyen Kale, and Shai Shalev-Shwartz. Near-optimal algorithms for online matrix prediction. In *Conference on Learning Theory*, pages 38–1, 2012.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.*, 24, 1933.

Prateek Jain, Chi Jin, Sham M Kakade, Praneeth Netrapalli, and Aaron Sidford. Matching matrix bernstein with little memory: Near-optimal finite sample guarantees for oja's algorithm. *arXiv preprint arXiv:1602.06929*, 2016.

Chi Jin, Sham M Kakade, and Praneeth Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 4520–4528, 2016.

Ian Jolliffe. Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer, 2011.

Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

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

Chris Junchi Li, Mengdi Wang, Han Liu, and Tong Zhang. Near-optimal stochastic approximation for online principal component estimation. *Mathematical Programming*, 167(1):75–97, 2018.

Teodor Vanislavov Marinov, Poorya Mianjy, and Raman Arora. Streaming principal component analysis in noisy setting. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3413–3422, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Poorya Mianjy and Raman Arora. Stochastic PCA with ℓ_2 and ℓ_1 regularization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3531–3539, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/mianjy18a.html.

Ioannis Mitliagkas, Constantine Caramanis, and Prateek Jain. Memory limited, streaming pca. In *Advances in Neural Information Processing Systems*, pages 2886–2894, 2013.

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

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

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

Ohad Shamir. Convergence of stochastic gradient descent for PCA:. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 257–265, 2016.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

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

Manfred K. Warmuth and Dima Kuzmin. 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.

Peng Xu, Bryan He, Christopher De Sa, Ioannis Mitliagkas, and Chris Re. Accelerated stochastic power iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 58–67, 2018.

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

Appendix A. Discussion

In this paper we took a step forward towards understanding the ability of highly-efficient nonconvex online algorithms to minimize regret in adversarial online learning settings. We focused on the particular problem of online principal component analysis with $k = 1$, and showed that under a "semi-adversarial" model, in which the data follows a stochastic distribution with adversarial perturbations, and given a "warm-start" initialization, the natural nonconvex online gradient ascent indeed guarantees sublinear regret. Our theory is further supported by empirical evidence.

We hope this work will motivate further research on online nonconvex optimization with global convergence guarantees. Future directions of interest may include extending our analysis to a wider regime of parameters, and extracting k principal components at once. Also, it is interesting if in the standard adversarial setting, it can be shown that online nonconvex gradient ascent achieves low-regret, or on the other-hand, to show that there exist instances on which it cannot guarantee non-trivial regret. Finally, moving beyond PCA, other online learning problems of interest that may benefit from a non-convex approach include online matrix completion Hazan et al. (2012); Jin et al. (2016), and of course, provable online learning of deep networks.

Appendix B. Additional Information on Experiments

The datasets used for the experiments are as follows.
Synthetic: a random dataset is constructed by generating Gaussian zero-mean data with a random covariance matrix $\mathbf{Q}$ with eigenvalues $\lambda_i = 15 \cdot 0.3^{i-1}$ for all $i \in [d]$, and perturbing them using independent Gaussian zero-mean noise with random covariance matrix $\mathbf{V}$ with eigenvalues $\mu_i = 3 \cdot 0.3^{i-1}$ for all $i \in [d]$, where we use $d = 100$. We set the number of data points to $N = 10000$, and we compute the initialization $\hat{\mathbf{w}}_1$ for all algorithms by computing the leading eigenvector of a sample of size 100 (i.e., 1% of N) based on samples from the covariance $\mathbf{Q}$ only. For the algorithm BR1-OGA we set $\ell = 10$. We average the results of 30 i.i.d. experiments.
MNIST: we use the training set of the MNIST handwritten digit recognition dataset LeCun et al. (1998) which contains 60000 28x28 images, which we split into $N = 59400$ images for testing, while 600 images (i.e., 1% of data) are used to compute the initialization $\hat{\mathbf{w}}_1$. For the algorithm BR1-OGA we set $\ell = 5$.
CIFAR10: we use the CIFAR10 tiny image dataset Krizhevsky (2009) which contains 50000 32x32 images in RGB format. We convert the images to grayscale and use $N = 49900$ images for testing and 100 images (i.e., 0.2% of data) are used to compute the initialization. For BR1-OGA we set $\ell = 5$.

Appendix C. Proofs Omitted from section 3

C.1. Proof of Lemma 7

Proof Note that under the assumptions of the lemma it holds on any iteration t that $\eta_t \alpha < 1$. Thus, in light of Lemma 5, it suffices to show that on each iteration t, it holds that

$$(\hat{\mathbf{w}}_t^\top \mathbf{x})^2 \geq 1 - \frac{\delta(\mathbf{Q}) - V^2 - \ell^{-1}\alpha - 4\epsilon}{2\lambda_1(\mathbf{Q})}.$$

We prove this inequality indeed holds for all $t \in [T]$ by induction. Note that for $t = 1$, this clearly holds by our assumption on $\hat{\mathbf{w}}_1$.

Suppose now the assumption holds for some $t \geq 1$. In the following we let λ_i denote the ith largest eigenvalue of the matrix $\mathbf{W}_{t+1} := (1 - \eta_t\alpha)\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top + \eta_t \mathbf{X}_t$. Note that under the induction hypothesis and Assumption 1, it in particular holds that $(\hat{\mathbf{w}}_t^\top \mathbf{x})^2 \geq 1/2$, and hence we can invoke Lemma 6.

We consider two cases. If $(\hat{\mathbf{w}}_t^\top \mathbf{x})^2 \geq 1 - \frac{\delta(\mathbf{Q}) - V^2 - \ell^{-1}\alpha - 5\epsilon}{2\lambda_1(\mathbf{Q})}$, then using Lemma 6 we have that

$$\begin{aligned}(\mathbf{w}_{t+1}^\top \mathbf{x})^2 &\geq 1 - \frac{\delta(\mathbf{Q}) - V^2 - \ell^{-1}\alpha - 5\epsilon}{2\lambda_1(\mathbf{Q})} - \eta_t \ell \frac{V^2 + 4\epsilon}{\lambda_1 - \lambda_2} \\ &\underset{(a)}{\geq} 1 - \frac{\delta(\mathbf{Q}) - V^2 - \ell^{-1}\alpha - 5\epsilon}{2\lambda_1(\mathbf{Q})} - \eta_t \ell \left(V^2 + 4\epsilon\right),\end{aligned}$$

were (a) follows since under the induction hypothesis, we in particular have that $\mathbf{w}_{t+1}\mathbf{w}_{t+1}^\top = \Pi_{\mathcal{S}}[\mathbf{W}_{t+1}]$ (see Lemma 5), which in turn implies that $\lambda_1 \geq 1 + \lambda_2$ (see proof of Lemma 4).

Thus, for any $\eta_t \leq \frac{\epsilon}{4\ell\lambda_1(\mathbf{Q})\cdot(V^2+4\epsilon)}$ we obtain

$$(\mathbf{w}_{t+1}^\top \mathbf{x})^2 \geq 1 - \frac{\delta(\mathbf{Q}) - V^2 - \ell^{-1}\alpha - \frac{9}{2}\epsilon}{2\lambda_1(\mathbf{Q})}.$$

Moreover, we have that

$$\begin{aligned}(\hat{\mathbf{w}}_{t+1}^\top \mathbf{x})^2 &\geq (\mathbf{w}_{t+1}^\top \mathbf{x})^2 - \|\mathbf{w}_{t+1}\mathbf{w}_{t+1}^\top - \hat{\mathbf{w}}_{t+1}\hat{\mathbf{w}}_{t+1}\|_F \geq (\mathbf{w}_{t+1}^\top \mathbf{x})^2 - \gamma_t \\ &\geq 1 - \frac{\delta(\mathbf{Q}) - V^2 - \ell^{-1}\alpha - \frac{9}{2}\epsilon}{2\lambda_1(\mathbf{Q})} - \gamma_t.\end{aligned}$$

Thus, for any $\gamma_t \leq \frac{\epsilon}{4\lambda_1(\mathbf{Q})}$ the claim indeed holds for the first case.

On the other hand, in case $(\hat{\mathbf{w}}_t^\top \mathbf{x})^2 < 1 - \frac{\delta(\mathbf{Q}) - V^2 - \ell^{-1}\alpha - 5\epsilon}{2\lambda_1(\mathbf{Q})}$, by an application of Lemma 6 we have that

$$\begin{aligned}(\mathbf{w}_{t+1}^\top \mathbf{x})^2 &\geq (\hat{\mathbf{w}}_t^\top \mathbf{x})^2 + \eta_t \ell \frac{(1 - (\hat{\mathbf{w}}_t^\top \mathbf{x})^2)\cdot\delta(\mathbf{Q}) - (\hat{\mathbf{w}}_t^\top \mathbf{x})^2 V^2 - 4\epsilon}{\lambda_1 - \lambda_2} \\ &\underset{(a)}{\geq} (\hat{\mathbf{w}}_t^\top \mathbf{x})^2 + \eta_t \ell \frac{\delta(\mathbf{Q}) - (\delta(\mathbf{Q}) + V^2)\left(1 - \frac{\delta(\mathbf{Q}) - V^2 - \ell^{-1}\alpha - 5\epsilon}{2\lambda_1(\mathbf{Q})}\right) - 4\epsilon}{\lambda_1 - \lambda_2} \\ &= (\hat{\mathbf{w}}_t^\top \mathbf{x})^2 + \eta_t \ell \frac{\frac{\delta(\mathbf{Q}) - V^2 - \ell^{-1}\alpha - 5\epsilon}{2\lambda_1(\mathbf{Q})}\cdot\delta(\mathbf{Q}) - V^2\cdot\left(1 - \frac{\delta(\mathbf{Q}) - V^2 - \ell^{-1}\alpha - 5\epsilon}{2\lambda_1(\mathbf{Q})}\right) - 4\epsilon}{\lambda_1 - \lambda_2} \\ &\underset{(b)}{\geq} (\hat{\mathbf{w}}_t^\top \mathbf{x})^2 + \eta_t \ell \frac{\frac{\delta(\mathbf{Q}) - V^2 - \ell^{-1}\alpha}{2\lambda_1(\mathbf{Q})}\cdot\delta(\mathbf{Q}) - V^2\cdot\left(1 - \frac{\delta(\mathbf{Q}) - V^2 - \ell^{-1}\alpha}{2\lambda_1(\mathbf{Q})}\right) - 9\epsilon}{\lambda_1 - \lambda_2} \\ &= (\hat{\mathbf{w}}_t^\top \mathbf{x})^2 + \eta_t \ell \frac{\delta(\mathbf{Q})^2 - V^4 - 2V^2\lambda_1(\mathbf{Q}) - (\delta(\mathbf{Q}) + V^2)\ell^{-1}\alpha - 18\epsilon\lambda_1(\mathbf{Q})}{2\lambda_1(\mathbf{Q})(\lambda_1 - \lambda_2)},\end{aligned}$$

where (a) follows from our assumption on $(\hat{\mathbf{w}}_t^\top \mathbf{x})^2$ in this second case, and (b) follows, since Assumption 1 implies that $\max\{\delta(\mathbf{Q}), V^2\} \leq \lambda_1(\mathbf{Q})$.

Thus, for any

$$\begin{aligned}\epsilon &\leq \frac{1}{72\lambda_1(\mathbf{Q})}\left(\delta(\mathbf{Q})^2 - V^4 - 2V^2\lambda_1(\mathbf{Q})\right) \quad \text{and} \\ \alpha &\leq \frac{\ell}{4(\delta(\mathbf{Q}) + V^2)}\left(\delta(\mathbf{Q})^2 - V^4 - 2V^2\lambda_1(\mathbf{Q})\right)\end{aligned}$$

we have that

$$(\mathbf{w}_{t+1}^\top \mathbf{x})^2 \quad \geq \quad (\hat{\mathbf{w}}_t^\top \mathbf{x})^2 + \eta_t \ell \frac{\delta(\mathbf{Q})^2 - V^4 - 2V^2\lambda_1(\mathbf{Q})}{4\lambda_1(\mathbf{Q})(\lambda_1 - \lambda_2)}.$$

Moreover, as before, we have that

$$\begin{aligned}(\hat{\mathbf{w}}_{t+1}^\top \mathbf{x})^2 &\geq (\mathbf{w}_{t+1}^\top \mathbf{x})^2 - \gamma_t \\ &\geq (\hat{\mathbf{w}}_t^\top \mathbf{x})^2 + \eta_t \ell \frac{\delta(\mathbf{Q})^2 - V^4 - 2V^2\lambda_1(\mathbf{Q})}{4\lambda_1(\mathbf{Q})(\lambda_1 - \lambda_2)} - \gamma_t \\ &\underset{(a)}{\geq} (\hat{\mathbf{w}}_t^\top \mathbf{x})^2 + \eta_t \ell \frac{\delta(\mathbf{Q})^2 - V^4 - 2V^2\lambda_1(\mathbf{Q})}{4\lambda_1(\mathbf{Q})} - \gamma_t,\end{aligned}$$

where (a) follows since Assumption 1 implies that $\delta(\mathbf{Q})^2 - V^4 - 2V^2\lambda_1(\mathbf{Q}) \geq 0$, and since

$$\lambda_1 - \lambda_2 \leq \lambda_1 \leq 1 - \eta_t\alpha + \eta_t\|\mathbf{X}_t\| \leq 1 + \eta_t \sum_{i=1}^{\ell} \|\mathbf{q}_t^{(i)} + \mathbf{v}_t^{(i)}\|^2 \leq 1 + \eta_t \ell (R+V)^2 \leq 2,$$

where the last inequality follows from our assumption that $\eta_t \leq \frac{1}{\ell(R+V)^2}$. Thus, for any $\gamma_t \leq \eta_t \ell \frac{\delta(\mathbf{Q})^2 - V^4 - 2V^2\lambda_1(\mathbf{Q})}{4\lambda_1(\mathbf{Q})}$ (which in particular holds for $\gamma_t \leq 18\eta_t\ell\epsilon$), we have that

$$(\hat{\mathbf{w}}_{t+1}^\top \mathbf{x})^2 \geq (\hat{\mathbf{w}}_t^\top \mathbf{x})^2 \geq 1 - \frac{\delta(\mathbf{Q}) - V^2 - \ell^{-1}\alpha - 4\epsilon}{2\lambda_1(\mathbf{Q})},$$

as needed. ■

C.2. Convergence of Algorithm 2

Lemma 9 (Convergence of Algorithm 2) *Consider applying Algorithm 2 to a sequence of vectors* $\{(\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(\ell)})\}_{t\in[T]}$ *which follow Assumption 1, and suppose that all conditions stated in Lemma 7 hold. Then, for any unit vector* $\mathbf{w}$ *it holds that*

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{i=1}^{\ell}(\mathbf{w}^\top \mathbf{x}_t^{(i)})^2 - \sum_{t=1}^{T}\sum_{i=1}^{\ell}(\hat{\mathbf{w}}_t^\top \mathbf{x}_t^{(i)})^2 \leq \\ &\frac{1}{\eta_1} + \sum_{t=2}^{T}\left(\frac{1-\eta_t\alpha}{2\eta_t} - \frac{1}{2\eta_{t-1}}\right)\|\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top - \mathbf{w}\mathbf{w}^\top\|_F^2 \\ &+ \sum_{t=1}^{T}\left(\frac{3\sqrt{2}}{2}\frac{\gamma_t}{\eta_t} + \frac{\eta_t}{2}(\ell^2(R+V)^4 + \alpha^2)\right).\end{aligned}$$

Proof Fix some unit vector $\mathbf{w}$. By an application of Lemma 7, it holds for all $t \in [T]$ that $\mathbf{w}_{t+1}\mathbf{w}_{t+1}^\top = \Pi_{\mathcal{S}}[\mathbf{W}_{t+1}]$.

Thus, using standard arguments[6], we have that for all $t \in [T]$ it holds that

$$\begin{aligned}\|\mathbf{w}_{t+1}\mathbf{w}_{t+1}^\top - \mathbf{w}\mathbf{w}^\top\|_F^2 &\leq \|\mathbf{W}_{t+1} - \mathbf{w}\mathbf{w}^\top\|_F^2 \\ &= \|(1-\eta_t\alpha)\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top + \eta_t\mathbf{X}_t - \mathbf{w}\mathbf{w}^\top\|_F^2 \\ &= \|\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top - \mathbf{w}\mathbf{w}^\top\|_F^2 \\ &\quad +2\eta_t(\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top - \mathbf{w}\mathbf{w}^\top)\bullet(\mathbf{X}_t - \alpha\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top) \\ &\quad +\eta_t^2\|\mathbf{X}_t - \alpha\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top\|_F^2.\end{aligned}$$

Note that

$$\begin{aligned}\|\hat{\mathbf{w}}_{t+1}\hat{\mathbf{w}}_{t+1}^\top - \mathbf{w}\mathbf{w}^\top\|_F^2 &= \|\hat{\mathbf{w}}_{t+1}\hat{\mathbf{w}}_{t+1}^\top + \mathbf{w}_{t+1}\mathbf{w}_{t+1}^\top - \mathbf{w}_{t+1}\mathbf{w}_{t+1}^\top - \mathbf{w}\mathbf{w}^\top\|_F^2 \\ &\underset{(a)}{\leq} \|\mathbf{w}_{t+1}\mathbf{w}_{t+1}^\top - \mathbf{w}\mathbf{w}^\top\|_F^2 + 3\sqrt{2}\|\mathbf{w}_{t+1}\mathbf{w}_{t+1}^\top - \hat{\mathbf{w}}_{t+1}\hat{\mathbf{w}}_{t+1}^\top\|_F \\ &\leq \|\mathbf{w}_{t+1}\mathbf{w}_{t+1}^\top - \mathbf{w}\mathbf{w}^\top\|_F^2 + 3\sqrt{2}\gamma_t,\end{aligned}$$

where (a) follows since for any two unit vectors $\mathbf{y}, \mathbf{z}$ it holds that $\|\mathbf{y}\mathbf{y}^\top - \mathbf{z}\mathbf{z}^\top\|_F \leq \sqrt{2}$.

Note also that

$$2(\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top - \mathbf{w}\mathbf{w}^\top)\bullet(-\alpha\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top) = -\alpha(2 - 2(\mathbf{w}^\top\hat{\mathbf{w}}_t)^2) = -\alpha\|\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top - \mathbf{w}\mathbf{w}^\top\|_F^2.$$

Combining the three bounds above, we obtain

$$\begin{aligned}&(\mathbf{w}\mathbf{w}^\top - \hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top)\bullet\mathbf{X}_t \leq \\ &\frac{1}{2\eta_t}\left((1-\eta_t\alpha)\|\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top - \mathbf{w}\mathbf{w}^\top\|_F^2 - \|\hat{\mathbf{w}}_{t+1}\hat{\mathbf{w}}_{t+1}^\top - \mathbf{w}\mathbf{w}^\top\|_F^2\right) \\ &+\frac{3\sqrt{2}\gamma_t}{2\eta_t} + \frac{\eta_t}{2}\|\mathbf{X}_t - \alpha\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top\|_F^2.\end{aligned}$$

Note that $\|\mathbf{X}_t - \alpha\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top\|_F^2 \leq \|\mathbf{X}_t\|_F^2 + \alpha^2$. Summing over all iterations we obtain the bound

$$\begin{aligned}\sum_{t=1}^T \mathbf{w}^\top\mathbf{X}_t\mathbf{w} - \sum_{t=1}^T \hat{\mathbf{w}}_t^\top\mathbf{X}_t\hat{\mathbf{w}}_t &\leq \frac{1-\eta_1\alpha}{2\eta_1}\|\hat{\mathbf{w}}_1\hat{\mathbf{w}}_1^\top - \mathbf{w}\mathbf{w}^\top\|_F^2 \\ &\quad +\sum_{t=2}^T\left(\frac{1-\eta_t\alpha}{2\eta_t} - \frac{1}{2\eta_{t-1}}\right)\|\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top - \mathbf{w}\mathbf{w}^\top\|_F^2 \\ &\quad +\sum_{t=1}^T\left(\frac{3\sqrt{2}}{2}\frac{\gamma_t}{\eta_t} + \frac{\eta_t}{2}(\|\mathbf{X}_t\|_F^2 + \alpha^2)\right).\end{aligned}$$

Finally, note that $\|\hat{\mathbf{w}}_1\hat{\mathbf{w}}_1^\top - \mathbf{w}\mathbf{w}^\top\|_F^2 \leq 2$ and that under Assumption 1, it holds for all $t \in [T]$ that

$$\begin{aligned}\|\mathbf{X}_t\|_F &= \|\sum_{i=1}^\ell \mathbf{x}_t^{(i)}\mathbf{x}_t^{(i)\top}\|_F \leq \sum_{i=1}^\ell \|\mathbf{x}_t^{(i)}\mathbf{x}_t^{(i)\top}\|_F = \sum_{i=1}^\ell \|\mathbf{x}_t^{(i)}\|^2 \\ &= \sum_{i=1}^\ell \|\mathbf{q}_t^{(i)} + \mathbf{v}_t^{(i)}\|^2 \leq \ell(R+V)^2. \qquad (8)\end{aligned}$$

6. see for instance the analysis of Online Gradient Descent in Hazan (2016)

Hence, the lemma follows. ■

As an example, if we invoke Lemma 9 with a fixed learning rate $\eta > 0$ (i.e, $\eta_t = \eta$ for all $t \in [T]$), zero regularization ($\alpha = 0$), and $\gamma_t = 0$ for all $t \in [T]$, we get a regret bound

$$\frac{1}{\eta} + \frac{\eta}{2} T \ell^2 (R+V)^4,$$

which by an appropriate choice of η and treating ℓ as a constant, yields the familiar $O(\sqrt{T})$ regret bound for Online Gradient Ascent with linear payoff functions Hazan (2016). A Similar treatment with $\alpha > 0$ and vanishing step-size can be shown (as we indeed show in the sequel) to yield a $O(\log T)$ regret bound.

C.3. Proof of Lemma 8

Proof Let us denote by $\mathbf{y}_2, \ldots, \mathbf{y}_d$ the $(d-1)$ non-leading eigenvectors of the matrix $\mathbf{W}_{t+1}$. Since both $\mathbf{w}_{t+1}, \hat{\mathbf{w}}_{t+1}$ are unit vectors, we have that

$$\|\hat{\mathbf{w}}_{t+1}\hat{\mathbf{w}}_{t+1} - \mathbf{w}_{t+1}\mathbf{w}_{t+1}^\top\|_F^2 = 2\left(1 - (\hat{\mathbf{w}}_{t+1}^\top \mathbf{w}_{t+1})^2\right) = 2\sum_{i=2}^{d} (\hat{\mathbf{w}}_{t+1}^\top \mathbf{y}_i)^2. \tag{9}$$

Note that by the update rule of Algorithm 1 and since $\hat{\mathbf{w}}_t$ is a unit vector, the vector $\hat{\mathbf{w}}_{t+1}$ could be written as

$$\begin{aligned}
\hat{\mathbf{w}}_{t+1} &= \frac{\hat{\mathbf{w}}_t + \eta_t(\mathbf{X}_t\hat{\mathbf{w}}_t - \alpha\hat{\mathbf{w}}_t)}{\|\hat{\mathbf{w}}_t + \eta_t(\mathbf{X}_t\hat{\mathbf{w}}_t - \alpha\hat{\mathbf{w}}_t)\|} \\
&= \frac{\left(\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top + \eta_t(\mathbf{X}_t - \alpha\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top)\right)\hat{\mathbf{w}}_t}{\|\left(\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top + \eta_t(\mathbf{X}_t - \alpha\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top)\right)\hat{\mathbf{w}}_t\|} = \frac{\mathbf{W}_{t+1}\hat{\mathbf{w}}_t}{\|\mathbf{W}_{t+1}\hat{\mathbf{w}}_t\|}.
\end{aligned}$$

Hence, $\hat{\mathbf{w}}_{t+1}$, is the result of applying a single iteration of the well known Power Method for leading eigenvector computation, initialized with the vector $\hat{\mathbf{w}}_t$, to the matrix $\mathbf{W}_{t+1}$. Let us denote by $\mathbf{y}_2, \ldots, \mathbf{y}_d$ the $(d-1)$ non-leading eigenvectors of $\mathbf{W}_{t+1}$. Using standard arguments, see for instance Eq. (18) in Garber and Hazan (2015), we have that

$$\sum_{i=2}^{d} (\hat{\mathbf{w}}_{t+1}^\top \mathbf{y}_i)^2 \leq \frac{\sum_{i=2}^{d}(\hat{\mathbf{w}}_t^\top \mathbf{y}_i)^2}{(\hat{\mathbf{w}}_t^\top \mathbf{w}_{t+1})^2} \left(\frac{\lambda_2(\mathbf{W}_{t+1})}{\lambda_1(\mathbf{W}_{t+1})}\right)^2.$$

Since $\hat{\mathbf{w}}_t$ is a unit vector, we have that $\sum_{i=2}^{d}(\hat{\mathbf{w}}_t^\top \mathbf{y}_i)^2 = 1 - (\hat{\mathbf{w}}_t^\top \mathbf{w}_{t+1})^2$. Moreover, we can bound

$$\begin{aligned}
\frac{\lambda_2(\mathbf{W}_{t+1})}{\lambda_1(\mathbf{W}_{t+1})} &= \frac{\lambda_2((1-\eta_t\alpha)\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top + \eta_t\mathbf{X}_t)}{\lambda_1((1-\eta_t\alpha)\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top + \eta_t\mathbf{X}_t)} \\
&\leq \frac{\lambda_2((1-\eta_t\alpha)\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top) + \lambda_1(\eta_t\mathbf{X}_t)}{\lambda_1((1-\eta_t\alpha)\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top)} \leq \frac{\eta_t\|\mathbf{X}_t\|}{1-\eta_t\alpha},
\end{aligned}$$

where the first inequality follows from Weyl's inequality for the eigenvalues. Thus, plugging-in into Eq. (9), we have that

$$\|\hat{\mathbf{w}}_{t+1}\hat{\mathbf{w}}_{t+1} - \mathbf{w}_{t+1}\mathbf{w}_{t+1}^\top\|_F^2 \leq 2\frac{1-(\hat{\mathbf{w}}_t^\top \mathbf{w}_{t+1})^2}{(\hat{\mathbf{w}}_t^\top \mathbf{w}_{t+1})^2}\frac{\eta_t^2\|\mathbf{X}_t\|^2}{(1-\eta_t\alpha)^2}. \tag{10}$$

Using the Davis-Kahan $\sin\theta$ theorem (see for instance Theorem 4 in Garber et al. (2015)), we have that

$$\begin{aligned}
1-(\hat{\mathbf{w}}_t^\top \mathbf{w}_{t+1})^2 &= \frac{1}{2}\|\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top - \mathbf{w}_{t+1}\mathbf{w}_{t+1}^\top\|_F^2 \\
&\leq \frac{4\|\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top - \mathbf{W}_{t+1}\|^2}{(\lambda_1(\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top) - \lambda_2(\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top))^2} \\
&= \frac{4\|\eta_t\alpha\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top - \eta_t\mathbf{X}_t\|^2}{\lambda_1(\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top)^2} \leq 4\eta_t^2\max\{\|\mathbf{X}_t\|^2,\alpha^2\}.
\end{aligned}$$

Plugging back into Eq. (10) and using the fact that $\eta_t \leq \frac{1}{3\max\{\|\mathbf{X}_t\|,\alpha\}}$ (see bound on $\|\mathbf{X}_t\|$ in Eq. (8)), we can conclude that

$$\begin{aligned}
\|\hat{\mathbf{w}}_{t+1}\hat{\mathbf{w}}_{t+1} - \mathbf{w}_{t+1}\mathbf{w}_{t+1}^\top\|_F^2 &\leq \frac{8\eta_t^4\|\mathbf{X}_t\|^2\max\{\|\mathbf{X}_t\|^2,\alpha^2\}}{(1-4\eta_t^2\max\{\|\mathbf{X}_t\|^2,\alpha^2\})(1-\eta_t\alpha)^2} \\
&\leq \frac{8\eta_t^4\|\mathbf{X}_t\|^2\max\{\|\mathbf{X}_t\|^2,\alpha^2\}}{\left(1-\frac{4}{9}\right)\left(1-\frac{1}{3}\right)^2} \\
&\leq 33(\eta_t\ell(R+V)^2)^4,
\end{aligned}$$

where all inequalities follow from our assumptions on η_t, α and the bound (8). ■

C.4. Proof of main theorems - Theorems 1 and 2

Before we prove the theorems, we need the following lemma.

Lemma 10 (Matrix Hoeffding) *Under the conditions of Assumption 1, it holds for all $t \in [T]$ and for all $\epsilon > 0$ that*

$$\Pr\left(\|\frac{1}{\ell}\mathbf{D}_t\| \geq \epsilon\right) \leq 2d\cdot\exp\left(-\frac{\epsilon^2\ell}{128R^4}\right).$$

Proof By a straightforward application of the Matrix Hoeffding inequality (see for instance Tropp (2012)), we have for any fixed $t \in [T]$ that

$$\begin{aligned}
\Pr\left(\|\tfrac{1}{\ell}(\mathbf{Q}_t - \ell\cdot\mathbf{Q})\| \geq \epsilon\right) &\leq d\cdot\exp\left(-\tfrac{\epsilon^2\ell}{32R^4}\right), \\
\Pr\left(\|\tfrac{1}{\ell}\mathbf{M}_t\| \geq \epsilon\right) &\leq d\cdot\exp\left(-\tfrac{\epsilon^2\ell}{32V^2R^2}\right).
\end{aligned}$$

Thus, the lemma follows from applying both of the above bounds with parameter $\epsilon/2$ and noting that $V^2 \leq R^2$. ■

We can now finally prove Theorem 1 and Theorem 2.

Proof [Proof of Theorem 1]

The proof follows from straightforward application of the tools we have developed thus-far.

We assume for simplicity that $N = T \cdot \ell$ for our choice of ℓ. Note this is without loss of generality, since the remainder $(N - \ell \cdot \lfloor N/\ell \rfloor)$ affects the bound in the theorem only via lower-order terms.

Let us define $\epsilon := \frac{\delta(\mathbf{Q})^2 - V^2(V^2+2\lambda_1(\mathbf{Q}))}{72\lambda_1(\mathbf{Q})}$, and note this choice corresponds to Eq. (7). Thus, for a certain $\ell = O(R^4\epsilon^{-2}\log\frac{dT}{p})$, we have by an application of Lemma 10 that with probability at least $1-p$ it holds for all $t \in [T]$ that $\frac{1}{\ell}\|\mathbf{D}_t\| \leq \epsilon$. Define a constant approximation parameter $\gamma_t = \gamma = \sqrt{33}(\eta\ell(R+V)^2)^2$ (which corresponds to the bound in Lemma 8), where $\eta = \frac{1}{\sqrt{T}\ell(R+V)^2}$ is the fixed chosen learning rate stated in the theorem. Note that for N large enough, all parameters $\epsilon, \eta, \gamma, \hat{\mathbf{w}}_1$ satisfy the conditions of Lemma 7 and Lemma 8 with probability at least $1-p$, and thus, by invoking Lemma 9 (with $\alpha = 0$ and constant γ, η), we have that with probability at least $1-p$ that

$$\begin{aligned}
\lambda_1\left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^\top\right) - \sum_{t=1}^{T}\sum_{i=1}^{\ell}(\hat{\mathbf{w}}_t^\top\mathbf{x}_t^{(i)})^2 &\leq \frac{1}{\eta} + T\left(\frac{3\sqrt{2}}{2}\frac{\gamma}{\eta} + \frac{\eta}{2}\ell^2(R+V)^4\right) \\
&\underset{(a)}{=} \frac{1}{\eta} + T\left(\frac{3\sqrt{66}}{2}\eta\ell^2(R+V)^4 + \frac{\eta}{2}\ell^2(R+V)^4\right) \\
&= \frac{1}{\eta} + \frac{3\sqrt{66}+1}{2}T\eta\ell^2(R+V)^4 \\
&\underset{(b)}{=} O\left(\sqrt{T}\ell(R+V)^2\right) \\
&\underset{(c)}{=} O\left(\sqrt{N\ell}(R+V)^2\right),
\end{aligned}$$

where (a) follows from plugging the value of γ, (b) follows form plugging the value of η and (c) follows since $N = T \cdot \ell$. The theorem now follows from plugging-in the bound on ℓ. ■

Proof [Proof of Theorem 2] As in the proof of Theorem 1 we assume for simplicity that $N = T \cdot \ell$ for our choice of ℓ. We define ϵ and choose block-length ℓ exactly as in the proof of Theorem 2, which implies that with probability at least $1-p$ it holds for all $t \in [T]$ that $\frac{1}{\ell}\|\mathbf{D}_t\| \leq \epsilon$. We set the regularization parameter to $\alpha = \frac{\ell}{10(\delta(\mathbf{Q})+V^2)}\left(\delta(\mathbf{Q})^2 - V^4 - 2V^2\lambda_1(\mathbf{Q})\right)$ which agrees with the requirements of Lemma 7 and Lemma 8. We set the approximation parameter γ_t to $\gamma_t = \sqrt{33}(\eta_t\ell(R+V)^2)^2$ (which corresponds to the bound in Lemma 8). Finally, we set the learning rate on each iteration t to $\eta_t = \frac{1}{\alpha t + T_0}$, for

$$T_0 = \max\{\frac{4\ell\lambda_1(\mathbf{Q})(V^2+4\epsilon)}{\epsilon},\ \ell(R+V)^2,\ 72\ell\lambda_1(\mathbf{Q}),\ \frac{\ell(R+V)^4}{\epsilon}\}.$$

Note that this choice agrees with the requirements of Lemma 7 and Lemma 8 with respect to both sequences $\{\eta_t\}_{t\geq 1}$ and $\{\gamma_t\}_{t\geq 1}$.

Thus, with the choice of initialization $\hat{\mathbf{w}}_1$ stated in the theorem, it follows all parameters ϵ,α, $\{\eta_t\}_{t\in[T]}$,$\{\gamma_t\}_{t\in[T]}$,$\hat{\mathbf{w}}_1$ satisfy the conditions of Lemma 7 and Lemma 8 with probability at least

$1-p$, and thus, by invoking Lemma 9, we have that with probability at least $1-p$ that

$$\begin{aligned}
&\lambda_1\left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^\top\right)-\sum_{t=1}^{T}\sum_{i=1}^{\ell}(\hat{\mathbf{w}}_t^\top\mathbf{x}_t^{(i)})^2 \underset{(a)}{\leq}\\
&\frac{1}{2}\sum_{t=2}^{T}\left((\alpha t+T_0)-\alpha-((t-1)\alpha+T_0)\right)\|\hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^\top-\mathbf{w}\mathbf{w}^\top\|_F^2\\
&+\frac{1}{\eta_1}+\sum_{t=1}^{T}\left(\frac{3\sqrt{2}}{2}\frac{\gamma_t}{\eta_t}+\frac{\eta_t}{2}(\ell^2(R+V)^4+\alpha^2)\right)=\\
&\frac{1}{\eta_1}+\sum_{t=1}^{T}\left(\frac{3\sqrt{2}}{2}\frac{\gamma_t}{\eta_t}+\frac{\eta_t}{2}(\ell^2(R+V)^4+\alpha^2)\right)\underset{(b)}{=}\\
&(\alpha+T_0)+\left(\frac{3\sqrt{66}\ell^2(R+V)^4}{2}+\frac{\ell^2(R+V)^4+\alpha^2}{2}\right)\sum_{t=1}^{T}\eta_t=\\
&(\alpha+T_0)+\frac{(3\sqrt{66}+1)\ell^2(R+V)^4+\alpha^2}{2}\sum_{t=1}^{T}\frac{1}{\alpha t+T_0}\leq\\
&(\alpha+T_0)+\frac{(3\sqrt{66}+1)\ell^2(R+V)^4+\alpha^2}{2\alpha}\sum_{t=1}^{T}\frac{1}{t}\leq\\
&T_0+\frac{(3\sqrt{66}+1)\ell^2(R+V)^4+3\alpha^2}{2\alpha}\left(1+\log T\right),
\end{aligned}$$

where (a) follows from Lemma 9 and our choice of learning rate, and (b) follows from our choice of $\{\gamma_t\}_{t\in[T]}$.

Note that

$$\alpha=O\left(\frac{\ell}{\delta(\mathbf{Q})+V^2}\left(\delta^2(\mathbf{Q})-V^4\right)\right)=O\left(\ell(\delta(\mathbf{Q})+V^2)\right)=O(\ell(R+V)^2),$$

where in the last equality we have used the fact that $\delta(\mathbf{Q})\leq\lambda_1(\mathbf{Q})\leq R^2$. Also, by simple calculations we can see that $T_0=O\left(\frac{\ell(R+V)^4}{\epsilon}\right)$. Thus, we have that

$$\lambda_1\left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^\top\right)-\sum_{t=1}^{T}\sum_{i=1}^{\ell}(\hat{\mathbf{w}}_t^\top\mathbf{x}_t^{(i)})^2=O\left(\frac{\ell(R+V)^4}{\epsilon}+\frac{\ell^2(R+V)^4}{\alpha}\log T\right).$$

Note that $\alpha=\Theta\left(\frac{\ell\epsilon\lambda_1(\mathbf{Q})}{\delta(\mathbf{Q})+V^2}\right)$ and recall that $\ell=O\left(R^4\epsilon^{-2}\log\frac{dT}{p}\right)$. Plugging these values we obtain

$$\begin{aligned}
\lambda_1\left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^\top\right)-\sum_{t=1}^{T}\sum_{i=1}^{\ell}(\hat{\mathbf{w}}_t^\top\mathbf{x}_t^{(i)})^2&=O\left(\frac{R^4(R+V)^4\log\frac{dT}{p}}{\epsilon^3}\left(1+\frac{\delta(\mathbf{Q})+V^2}{\lambda_1(\mathbf{Q})}\log T\right)\right)\\
&=O\left(\frac{R^4(R+V)^4(\delta(\mathbf{Q})+V^2)}{\epsilon^3\lambda_1(\mathbf{Q})}\log(T)\log\left(\frac{dT}{p}\right)\right).
\end{aligned}$$

Finally, plugging the value of ϵ we obtain the regret bound:

$$O\left(\frac{R^4(R+V)^4(\delta(\mathbf{Q})+V^2)\lambda_1(\mathbf{Q})^2}{\left(\delta(\mathbf{Q})^2-V^4-2\lambda_1(\mathbf{Q})V^2\right)^3}\log(T)\log\left(\frac{dT}{p}\right)\right),$$

and the theorem follows. ■

Appendix D. Proof of Lemma 3 ("warm-start")

Proof Let $\hat{\mathbf{w}}_1$ be the leading eigenvector of the normalized covariance $\hat{\mathbf{X}}=\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$, where for all $i\in[n]$ $\mathbf{x}_i=\mathbf{q}_i+\mathbf{v}_i$. Clearly, $\mathbb{E}[\hat{\mathbf{X}}]=\mathbf{Q}+\frac{1}{n}\sum_{i=1}^n \mathbf{v}_i\mathbf{v}_i^\top$, and thus

$$\|\mathbf{Q}-\hat{\mathbf{X}}\|^2\le 2\|\mathbf{Q}-\mathbb{E}[\hat{\mathbf{X}}]\|^2+2\|\hat{\mathbf{X}}-\mathbb{E}[\hat{\mathbf{X}}]\|^2=2\|\frac{1}{n}\sum_{i=1}^n \mathbf{v}_i\mathbf{v}_i^\top\|^2+2\Delta^2,$$

where we use the notation $\Delta=\|\hat{\mathbf{X}}-\mathbb{E}[\hat{\mathbf{X}}]\|$. Via the Davis-Kahan $\sin\theta$ theorem (see for instance Theorem 4 in Garber et al. (2015)) and using the short notation $\delta=\delta(\mathbf{Q})$, we have that

$$\begin{aligned}(\hat{\mathbf{w}}_1^\top\mathbf{x})^2=1-\frac{1}{2}\|\hat{\mathbf{w}}_1\hat{\mathbf{w}}_1^\top-\mathbf{x}\mathbf{x}^\top\|_F^2 &\ge 1-4\frac{\|\hat{\mathbf{X}}-\mathbf{Q}\|^2}{\delta^2}\\ &\ge 1-8\frac{V^4}{\delta^2}-8\frac{\Delta^2}{\delta^2}.\end{aligned}$$

Now, using the short notation $\lambda_1=\lambda_1(\mathbf{Q})$, the requirement

$$(\hat{\mathbf{w}}_1^\top\mathbf{x})^2\ge 1-c\frac{\delta-V^2}{2\lambda_1} \tag{11}$$

boils down to the condition

$$8\frac{V^4+\Delta^2}{\delta^2}\le c\frac{\delta-V^2}{2\lambda_1}\iff 16\lambda_1V^4+c\delta^2V^2+16\lambda_1\Delta^2-c\delta^3\le 0.$$

Solving the above inequality on the right for V^2, we obtain that (11) holds when V^2 is in the interval:

$$0\le V^2\le\frac{-c\delta^2+\sqrt{c^2\delta^4+64\lambda_1c\delta^3-1024\lambda_1^2\Delta^2}}{32\lambda_1}.$$

In particular, for

$$\Delta\le\frac{\sqrt{19c}\delta^{3/2}}{32\sqrt{\lambda_1}} \tag{12}$$

we obtain that (11) holds for

$$0\le V^2\le\frac{-c\delta^2+\sqrt{c^2\delta^4+45\lambda_1c\delta^3}}{32\lambda_1}.$$

Note that since $c \in (0,1]$ and $\lambda_1 \geq \delta$, we have that $\lambda_1 c\delta^3 \geq c^2\delta^4$. Note also that $\sqrt{45} > \sqrt{32}+1$. Thus, we have that (11) holds for V^2 in the interval:

$$0 \leq V^2 \leq \frac{\sqrt{32c\lambda_1\delta^3}}{32\lambda_1} = \frac{\sqrt{c}\delta^{3/2}}{4\sqrt{2}\sqrt{\lambda_1}}.$$

We conclude the proof with the simple observation that using a standard Matrix Hoeffding concentration bound (see for instance Lemma 10), it suffices to take $n = O\left(\frac{R^4\lambda_1 \log(d/p)}{c\delta^3}\right)$ for the bound in (12) to hold with probability at least $1-p$.

■