Exponential Convergence in Stochastic k-PCA: Efficient Low-Rank Solutions Without Variance Reduction

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

October 4, 2024

Abstract

We present an efficient algorithm for stochastic Principal Component Analysis (PCA) that converges exponentially fast, bypassing the well-known minimax lower bound for general data distributions. Leveraging the intrinsic low-rank structure of the data, our proposed Matrix Krasulina's method generalizes the classic Krasulina algorithm from $k = 1$ to $k \geq 1$, solving the k-PCA problem with constant learning rates. This method exploits the self-regulating nature of the gradient variance, achieving fast convergence without requiring full data passes or variance reduction techniques. We provide theoretical guarantees and demonstrate the method's practical effectiveness on both synthetic and real-world datasets, validating its robustness even in non-strictly low-rank scenarios. The key contributions are a principled analysis of gradient variance decay and a generalization of Krasulina's method to the matrix setting.

Keywords: Stochastic Optimization, Principal Component Analysis, Krasulina's Method, Exponential Convergence, Low-Rank Data

1 Introduction

Principal Component Analysis (PCA) is a foundational tool in data analysis and dimensionality reduction. It aims to project high-dimensional data into a lower-dimensional subspace while retaining as much variance as possible. For a given dataset, stochastic PCA methods allow for efficient optimization when the data cannot fit into memory or when only a stream of data is available.

Traditional approaches to PCA, such as those based on Singular Value Decomposition (SVD), are computationally expensive for large datasets, requiring full data passes. In contrast, stochastic methods like Oja's algorithm and Krasulina's algorithm offer scalable solutions with iteration-wise runtimes independent of the data size. However, they typically suffer from slower convergence rates, constrained by minimax lower bounds for general data distributions.

In this paper, we propose a novel algorithm, Matrix Krasulina's method, which generalizes the classic Krasulina algorithm for k-PCA. This method exploits low-rank structure in the data to achieve exponential convergence without the need for variance reduction or full data passes. We show that on low-rank datasets, the variance of the stochastic gradient decays naturally as the algorithm progresses, leading to faster convergence compared to existing methods.

Mathematically, for a centered d-dimensional random vector $X \in \mathbb{R}^d$, the k-PCA problem seeks to find the "optimal" projection of X onto a subspace of dimension k that captures the maximum possible variance. Formally, the objective is to find a rank-k matrix W such that:[1]

$$\max_{W \in \mathbb{R}^{k \times d}, WW^\top = I_k} \operatorname{var}(W^\top W X)$$

[1]This ensures that the subspace spanned by the rows of W preserves as much of the variance of X as possible.

In the objective above, $W^\top W = W^\top (WW^\top)^{-1} W$ is an orthogonal projection matrix into the subspace spanned by the rows of W. Thus, the k-PCA problem seeks matrix W whose row-space captures as much variance of X as possible. This is equivalent to finding a projection into a subspace that minimizes variance of data outside of it:

$$\min_{W \in \mathbb{R}^{k \times d}, WW^\top = I_k} \mathbb{E} \, \|X - W^\top W X\|^2 \tag{1.1}$$

Likewise, given a sample of n centered data points $\{X_i\}_{i=1}^n$, the empirical version of problem (1.1) is

$$\min_{W \in \mathbb{R}^{k \times d}, WW^\top = I_k} \frac{1}{n} \sum_{i=1}^{n} \|X_i - W^\top W X_i\|^2 \tag{1.2}$$

The optimal k-PCA solution, the row space of optimal W, can be used to represent high-dimensional data in a low-dimensional subspace ($k \ll d$), since it preserves most variation from the original data. As such, it usually serves as the first step in exploratory data analysis or as a way to compress data before further operation.

The solutions to the nonconvex problems (1.1) and (1.2) are the subspaces spanned by the top k eigenvectors (also known as the *principal subspace*) of the population and empirical data covariance matrix, respectively. Although we do not have access to the population covariance matrix to directly solve (1.1), given a batch of samples $\{x_i\}_{i=1}^n$ from the same distribution, we can find the solution to (1.2), which asymptotically converges to the population k-PCA solution [Loukas(2017)]. Different approaches exist to solve (1.2) depending on the nature of the data and the computational resources available:

SVD-based solvers When data size is manageable, one can find the exact solution to (1.2) via a singular value decomposition (SVD) of the empirical data matrix in $\min\{O(nd^2), O(n^2 d)\}$-time and $O(nd)$-space, or in case of truncated SVD in $O(ndk)$-time ($O(nd \log k)$ for randomized solver [Halko et al.(2011)]).

Power method For large-scale datasets, that is, both n and d are large, the full data may not fit in memory. Power method [Golub and Van Loan(1996), p.450] and its variants are popular alternatives in this scenario; they have less computational and memory burden than SVD-based solvers; power method approximates the principal subspace iteratively: At every iteration, power method computes the inner product between the algorithm's current solution and n data vectors $\{x_i\}_{i=1}^n$, an $O(nd_s)$-time operation, where d_s is the average data sparsity. Power method converges exponentially fast [Shamir(2015)]: To achieve ε accuracy, it has a total runtime of $O(nd_s \log \frac{1}{\varepsilon})$. That is, power method requires multiple passes over the full dataset.

Online (incremental) PCA In real-world applications, datasets might become so large that even executing a full data pass is impossible. Online learning algorithms are developed under an abstraction of this setup: They assume that data come from an "endless stream" and only process one data point (or a constant sized batch) at a time. Online PCA mostly fall under two frameworks: 1. The online worst-case scenario, where the stream of data can have a non-stationary distribution [Nie et al.(2016), Boutsidis et al.(2015), Warmuth and Kuzmin(2006)]. 2. The stochastic scenario, where one has access to i.i.d. samples from an unknown but fixed distribution [Shamir(2015), Balsubramani et al.(2013), Mitliagkas et al.(2013), Arora et al.(2013)].

In this paper, we focus on the stochastic setup: We show that a simple variant of stochastic gradient descent (SGD), which generalizes the classic Krasulina's algorithm from $k = 1$ to general $k \geq 1$, can provably solve the k-PCA problem in Eq. (1.1) with an exponential convergence rate. It is worth noting that stochastic PCA algorithms, unlike batch-based solvers, can be used to optimize both the population PCA objective (1.1) and its empirical counterpart (1.2).

Oja's method and VR-PCA While SGD-type algorithms have iteration-wise runtime independent of the data size, their convergence rate, typically linear in the number of iterations, is significantly slower than that of batch gradient descent (GD). To speed up the convergence of SGD, the seminal work of [Johnson and Zhang(2013)] initiated a line of effort in deriving Variance-Reduced (VR) SGD by cleverly mixing the stochastic gradient updates with occasional batch gradient updates. For convex problems, VR-SGD algorithms have provable exponential convergence rate. Despite the non-convexity of k-PCA problem, [Shamir(2015), Shamir(2016a)] augmented Oja's method [Oja(1982)], a popular stochastic version of power method, with the VR step, and showed both theoretically and empirically that the resulting VR-PCA algorithm achieves exponential convergence. However, since a single VR iteration requires a full-pass over the dataset, VR-PCA is no longer an online algorithm.

Minimax lower bound In general, the tradeoff between convergence rate and iteration-wise computational cost is unavoidable in light of the minimax information lower bound [Vu and Lei(2013), Vu and Lei(2012)]: Let Δ^n (see Definition 3) denote the distance between the ground-truth rank-k principal subspace and the algorithm's estimated subspace after seeing n samples. [Vu and Lei(2013), Theorem 3.1] established that *there exists data distribution (with full-rank covariance matrices)* such that the following lower bound holds:

$$\mathbb{E}\left[\Delta^n\right] \geq \Omega(\frac{\sigma^2}{n}) \quad \text{for } \sigma^2 \geq \frac{\lambda_1 \lambda_{k+1}}{(\lambda_k - \lambda_{k+1})^2}\,, \tag{1.3}$$

Here λ_k denotes the k-th largest eigenvalue of the data covariance matrix. This immediately implies a $\Omega(\frac{\sigma^2}{t})$ lower bound on the convergence rate of online k-PCA algorithms, since for online algorithms the number of iterations t equals the number of data samples n. Thus, sub-linear convergence rate is impossible for online k-PCA algorithms on general data distributions.

1.1 Our result: escaping minimax lower bound on intrinsically low rank data

Despite the discouraging lower bound for online k-PCA, note that in Eq. (1.3), σ equals zero when the data covariance has rank less than or equal to k, and consequently, the lower bound becomes un-informative. Does this imply that data low-rankness can be exploited to overcome the lower bound on the convergence rate of online k-PCA algorithms?

Our result answers the question affirmatively: Theorem 1 suggests that on low-rank data, an online k-PCA algorithm, namely, Matrix Krasulina (Algorithm 1), produces estimates of the principal subspace that locally converges to the ground-truth in order $O(\mathbb{E}\left(-Ct\right))$, where t is the number of iterations (the number of samples seen) and C is a constant. Our key insight is that Krasulina's method [Krasulina(1969)], in contrast to its better-studied cousin Oja's method [Oja(1982)], is stochastic gradient descent with a self-regulated gradient for the PCA problem, and that when the data is of low-rank, the gradient variance vanishes as the algorithm's performance improves.

In a broader context, our result is an example of "learning faster on easy data", a phenomenon widely observed for online learning [Beygelzimer et al.(2015)], clustering [Kumar and Kannan(2010)], and active learning [Wang and Singh(2016)], to name a few. While low-rankness assumption has been widely used to regularize solutions to matrix completion problems [Jain et al.(2013), Keshavan et al.(2010), Candès and Recht(2009)] and to model the related robust PCA problem [Netrapalli et al.(2014), Candès et al.(2011)], we are unaware of previous such methods that exploit data low-rankness to significantly reduce computation.

2 Preliminaries

We consider the following online stochastic learning setting: At time $t \in \mathbb{N} \setminus \{0\}$, we receive a random vector $X^t \in \mathbb{R}^d$ drawn i.i.d from an unknown centered probability distribution with a finite second moment. We denote by X a generic random sample from this distribution. Our goal is to learn $W \in \mathbb{R}^{k' \times d}$ so as to optimize the objective in Eq (1.1).

Notations We let Σ^* denote the covariance matrix of X, $\Sigma^* := \mathbb{E}\left[XX^\top\right]$. We let $\{u_i\}_{i=1}^k$ denote the top k eigenvectors of covariance matrix Σ^*, corresponding to its largest k eigenvalues, $\lambda_1 \geq, \ldots, \geq \lambda_k$. Given that Σ^* has rank k, we can represent it by its top k eigenvectors: $\Sigma^* := \sum_{i=1}^k \lambda_i u_i u_i^\top$. We let $U^* := \sum_{i=1}^k u_i u_i^\top$. That is, U^* is the orthogonal projection matrix into the subspace spanned by $\{u_i\}_{i=1}^k$. For any integer $p > 0$, we let I_p denote the p-by-p identity matrix. We denote by $\|\cdot\|_F$ the Frobenius norm, by $tr(\cdot)$ the trace operator. For two square matrices A and B of the same dimension, we denote by $A \succeq B$ if $A - B$ is positive semidefinite. We use curly capitalized letters such as $\mathcal{G}$ to denote events. For an event $\mathcal{G}$, we denote by $\mathbb{1}_{\mathcal{G}}$ its indicator random variable; that is, $\mathbb{1}_{\mathcal{G}} = 1$ if event $\mathcal{G}$ occurs and 0 otherwise.

Optimizing the empirical objective We remark that our setup and theoretical results apply not only to the optimization of population k-PCA problem (1.1) in the infinite data stream scenario, but also to the empirical version (1.2): Given a finite dataset, we can simulate the stochastic optimization setup by sampling uniformly at random from it. This is, for example, the setup adopted by [Shamir(2016a), Shamir(2015)].

Assumptions In our analysis, we assume that Σ^* has low rank and that the data norm is bounded almost surely; that is, there exits b and k such that

$$\mathbb{P}\left(\sup_X \|X\|^2 > b\right) = 0 \quad \text{and} \quad rank(\Sigma^*) = k \tag{2.4}$$

2.1 Oja and Krasulina

In this section, we introduce two classic online algorithms for 1-PCA, Oja's method and Krasulina's method.

Oja's method Let $w^t \in \mathbb{R}^d$ denote the algorithm's estimate of the top eigenvector of Σ^* at time t. Then letting η^t denote learning rate, and X be a random sample, Oja's algorithm has the following update rule:

$$w^t \leftarrow w^{t-1} + \eta^t (XX^\top w^{t-1}) \quad \text{and} \quad w^t \leftarrow \frac{w^t}{\|w^t\|}$$

We see that Oja's method is a stochastic approximation algorithm to power method. For $k > 1$, Oja's method can be generalized straightforwardly, by replacing w^t with matrix $W^t \in \mathbb{R}^{k\times d}$, and by replacing the normalization step with row orthonormalization, for example, by QR factorizaiton.

Krasulina's method Krasulina's update rule is similar to Oja's update but has an additional term:

$$w^t \leftarrow w^{t-1} + \eta^t(XX^\top w^{t-1} - w^{t-1}(X^\top \frac{w^{t-1}}{\|w^{t-1}\|})^2)$$

In fact, this is stochastic gradient descent on the objective function below, which is equivalent to Eq (1.1):

$$\mathbb{E}\,\|X - \frac{w^t(w^t)^\top}{\|w^t\|^2}X\|^2$$

We are unaware of previous work that generalizes Krasulina's algorithm to $k > 1$.

2.2 Gradient variance in Krasulina's method

Our key observation of Krasulina's method is as follows: Let $\widetilde{w}^t := \frac{w^t}{\|w^t\|}$; Krasulina's update can be re-written as

$$w^t \leftarrow w^{t-1} + \|w^t\|\eta^t(XX^\top \widetilde{w}^{t-1} - \widetilde{w}^{t-1}(X^\top \widetilde{w}^{t-1})^2)$$

Let

$$s^t := (\widetilde{w}^t)^\top X \quad \text{(projection coefficient)}$$

and

$$r^t := X^\top - s^t(\widetilde{w}^t)^\top = X^\top - (\widetilde{w}^t)^\top X(\widetilde{w}^t)^\top \quad \text{(projection residual)}$$

Krasulina's algorithm can be further written as:

$$w^t \leftarrow w^{t-1} + \|w^t\|\eta^t s^{t-1}(r^{t-1})^\top$$

The variance of the stochastic gradient term can be upper bounded as:

$$\|w^t\|^2 \mathrm{Var}\left(s^{t-1}(r^{t-1})^\top\right) \le \|w^t\|^2 \sup_X \|X\|^2\,\mathbb{E}\,\|r^t\|^2$$

Note that

$$\mathbb{E}\,\|r^t\|^2 = \mathbb{E}\,\|X - \frac{w^t(w^t)^\top}{\|w^t\|^2}X\|^2$$

This reveals that the variance of the gradient naturally decays as Krasulina's method decreases the k-PCA optimization objective. Intuitively, as the algorithm's estimated (one-dimensional) subspace w^t gets closer to the ground-truth subspace u_1, $(w^t)^\top X$ will capture more and more of X's variance, and $\mathbb{E}\,\|r^t\|^2$ eventually vanishes.

In our analysis, we take advantage of this observation to prove the exponential convergence rate of Krasulina's method on low rank data.

3 Main results

Generalizing vector $w^t \in \mathbb{R}^d$ to matrix $W^t \in \mathbb{R}^{k'\times d}$ as the algorithm's estimate at time t, we derive *Matrix Krasulina's method* (Algorithm 1), so that the row space of W^t converges to the k-dimensional subspace spanned by $\{u_1, \ldots, u_k\}$.

Algorithm 1 Matrix Krasulina's method

Input: Initial matrix $W^o \in \mathbb{R}^{k' \times d}$; learning rate schedule (η^t); number of iterations, T;
while $t \leq T$ **do**

1. Sample X^t i.i.d. from the data distribution

2. Orthonormalize the rows of W^{t-1} (e.g., via QR factorization)

3. $W^t \leftarrow W^{t-1} + \eta^t W^{t-1} X^t (X^t - (W^{t-1})^\top W^{t-1} X^t)^\top$

end while
Output: $W^\top$

Matrix Krasulina's method Inspired by the original Krasulina's method, we design the following update rule for the Matrix Krasulina's method (Algorithm 1): Let

$$s^t := W^{t-1} X^t \quad \text{and} \quad r^t := X^t - (W^{t-1})^\top (W^{t-1}(W^{t-1})^\top)^{-1} W^{t-1} X^t \,,$$

Since we impose an orthonormalization step in Algorithm 1, r^t is simplified to

$$r^t := X^t - (W^{t-1})^\top W^{t-1} X^t$$

Then the update rule of Matrix Krasulina's method can be re-written as

$$W^t \leftarrow W^{t-1} + \eta^t s^t (r^t)^\top$$

For $k' = 1$, this reduces to Krasulina's update with $\|w^t\| = 1$. The self-regulating variance argument for the original Krasulina's method still holds, that is, we have

$$\mathbb{E}\,\|s^t (r^t)^\top\|^2 \leq b\,\mathbb{E}\,\|r^t\|^2 = b\,\mathbb{E}\,\|X - (W^t)^\top W^t X\|^2$$

where b is as defined in Eq (2.4). We see that the last term coincides with the objective function in Eq. (1.1).

Loss measure Given the algorithm's estimate W^t at time t, we let P^t denote the orthogonal projection matrix into the subspace spanned by its rows, $\{W^t_{i,\star}\}_{i=1}^{k'}$, that is,

$$P^t := (W^t)^\top (W^t (W^t)^\top)^{-1} W^t = (W^t)^\top W^t \,,$$

In our analysis, we use the following loss measure to track the evolvement of W^t: :=[Subspace distance] Let $\mathcal{S}$ and $\widehat{\mathcal{S}}^t$ be the ground-truth principal subspace and its estimate of Algorithm 1 at time t with orthogonal projectors U^* and P^t, respectively. We define the subspace distance between S and $\widehat{S}^t$ as $\Delta^t := tr(U^*(I - P^t)) = k - tr(U^* P^t)$. Note that Δ^t in fact equals the sum of squared canonical angles between $\mathcal{S}$ and $\widehat{\mathcal{S}}^t$, and coincides with the subspace distance measure used in related theoretical analyses of k-PCA algorithms [Allen-Zhu and Li(2017), Shamir(2016a), Vu and Lei(2013)]. In addition, Δ^t is related to the k-PCA objective function defined in Eq. (1.1) as follows (proved in Appendix Eq (B.10)):

$$\lambda_k \Delta^t \leq \mathbb{E}\,\|X - (W^t)^\top (W^t (W^t)^\top)^{-1} W^t X\|^2 \leq \lambda_1 \Delta^t$$

We prove the local exponential convergence of Matrix Krasulina's method measured by Δ^t. Our main contribution is summarized by the following theorem.

Theorem 1 (Exponential convergence with constant learning rate)**.** *Suppose assumption Eq.* (2.4) *holds. Suppose the initial estimate* $W^o \in \mathbb{R}^{k' \times d}$ $(k' \geq k)$ *in Algorithm 1 satisfies that, for some* $\tau \in (0,1)$*,*

$$tr(U^* P^o) \geq k - \frac{1-\tau}{2},$$

Suppose for any $\delta > 0$*, we choose a constant learning rate* $\eta^t = \eta$ *such that*

$$\eta \leq \min \left\{ \frac{\sqrt{2}-1}{b}, \frac{\lambda_k \tau}{\lambda_1 b(k+3)}, \frac{2\lambda_k \tau}{\frac{16}{1-\tau} \ln \frac{1}{\delta}(b + \|\Sigma^*\|_F)^2 + b(k+1)\lambda_1} \right\}$$

Then there exists event $\mathcal{G}_t$ *such that* $\mathbb{P}(\mathcal{G}_t) \geq 1 - \delta$*, and*

$$\mathbb{E}\left[\Delta^t | \mathcal{G}_t\right] \leq \frac{1}{1-\delta} \mathbb{E}\left(-t\eta\tau\lambda_k\right)$$

From Theorem 1, we observe that (a). The convergence rate of Algorithm 1 on strictly low-rank data does not depend on the data dimension d, but only on the intrinsic dimension k. This is verified by our experiments (see Sec. 5). (b). We see that the learning rate should be of order $O(\frac{1}{k\lambda_1})$: Empirically, we found that setting η to be roughly $\frac{1}{10\lambda_1}$ gives us the best convergence result. Note, however, this learning rate setup is not practical since it requires knowledge of eigenvalues.

Comparison between Theorem 1 and [Shamir(2016a), Theorem 1] (1). The result in [Shamir(2016a)] does not rely on the low-rank assumption of Σ^*. Since the variance of update in Oja's method is not naturally decaying, they use VR technique inspired by [Johnson and Zhang(2013)] to reduce the variance of the algorithm's iterate, which is computationally heavy: the block version of VR-PCA converges at rate $O(\mathbb{E}(-CT))$, where T denotes the number of data passes. (2). Our result has a similar learning rate dependence on the data norm bound b as that of [Shamir(2016a), Theorem 1]. (3). The initialization requirement in Theorem 1 is comparable to [Shamir(2016a), Theorem 1]; we note that the factor 1/2 in $\frac{1-\tau}{2}$ in our requirement is not strictly necessary in our analysis, and can be set arbitrarily close to 1. (4). Conditioning on the event of successful convergence, their exponential convergence rate result holds deterministically, whereas our convergence rate guarantee holds in expectation.

3.1 Related Works

Theoretical guarantees of stochastic optimization traditionally require convexity [Shalev-Shwartz et al.(2009)]. However, many modern machine learning problems, especially those arising from deep learning and unsupervised learning, are non-convex; PCA is one of them: The objective in (1.1) is non-convex in W. Despite this, a series of recent theoretical works have proven stochastic optimization to be effective for PCA, mostly variants of Oja's method [Allen-Zhu and Li(2017), Shamir(2016b), Shamir(2016a), Shamir(2015), De Sa et al.(2014), Hardt and Price(2014), Balsubramani et al.(2013)].

Krasulina's method [Krasulina(1969)] was much less studied than Oja's method; a notable exception is the work of [Balsubramani et al.(2013)], which proved an expected $O(1/t)$ rate for both Oja's and Krasulina's algorithm for 1-PCA.

There were very few theoretical analysis of stochastic k-PCA algorithms with $k > 1$, with the exception of [Allen-Zhu and Li(2017), Shamir(2016a), Balcan et al.(2016), Li et al.(2016)]. All had focused on variants of Oja's algorithm, among which [Shamir(2016a)] was the only previous work, to the best of our knowledge, that provided a local exponential convergence rate guarantee of Oja's algorithm for $k \geq 1$. Their result holds for general data distribution, but their variant of Oja's algorithm, VR-PCA, requires several full passes over the datasets, and thus not fully online.

Open questions In light of our result and related works, we have two open questions: (1). While our analysis has focused on analyzing Algorithm 1 with a constant learning rate on low-rank data, we believe it can be easily adapted to show that with a c/t (for some constant $c > 0$) learning rate, the algorithm achieves $O(1/t)$ convergence on *any datasets*. Note for the case $k' = 1$, the linear convergence rate of Algorithm 1 (original Krasulina's method) is already proved by [Balsubramani et al.(2013)]. (2). Many real-world datasets are not strictly low-rank, but *effectively low-rank* (see, for example, Figure 4): Informally, we say a dataset is effectively low-rank **if there exists $k \ll d$ such that $\frac{\sum_{i>k}\lambda_i}{\sum_{j\leq k}\lambda_j}$ is small**, We conjecture that our analysis can be adapted to show theoretical guarantee of Algorithm 1 on effectively low-rank datasets as well. In Section 5, our empirical results support this conjecture. Formally characterizing the dependence of convergence rate on the "effective low-rankness" of a dataset can provide a smooth transition between the worst-case lower bound in [Vu and Lei(2013)] and our result in Theorem 1.

4 Sketch of analysis

In this section, we provide an overview of our analysis and lay out the proofs that lead to Theorem 1 (the complete proofs are deferred to the Appendix). On a high level, our analysis is done in the following steps:

Section 4.1 We show that if the algorithm's iterates, W^t, stay inside the basin of attraction, which we formally define as event $\mathcal{G}_t$,

$$\mathcal{G}_t := \{\Delta^i \leq 1 - \tau, \forall i \leq t\},$$

then a function of random variables Δ^t forms a supermartingale.

Section 4.2 We show that provided a good initialization, it is likely that *the algorithm's outputs $W^1, \ldots, W^t$ stay inside the basin of attraction for every t.*

Section 4.3 We show that at each iteration t, conditioning on $\mathcal{G}_t$, $\Delta^{t+1} \leq \beta\Delta^t$ for some $\beta < 1$ if we set the learning rate η^t appropriately.

Appendix C Iteratively applying this recurrence relation leads to Theorem 1.

Additional notations: Before proceeding to our analysis, we introduce some technical notations for stochastic processes: Let $(\mathcal{F}_t)$ denote the natural filtration (collection of σ-algebras) associated to the stochastic process, that is, the data stream (X^t). Then by the update rule of Algorithm 1, for any t, W^t, P^t, and Δ^t are all $\mathcal{F}_t$-measurable, and $\mathcal{G}_t \in \mathcal{F}_t$.

4.1 A conditional supermartingale

Letting $M_i := \mathbb{1}_{\mathcal{G}_{i-1}} \mathbb{E}\left(s\Delta^i\right)$, Lemma 1 shows that $(M_i)_{i\geq 1}$ forms a supermartingale.

Lemma 1 (Supermartingale construction)**.** *Suppose $\mathcal{G}_0$ holds. Let C^t and Z be as defined in Proposition 2. Then for any $i \leq t$, and for any constant $s > 0$,*

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{G}_i}\mathbb{E}\left(s\Delta^{i+1}\right)|\mathcal{F}_i\right]$$
$$\leq \mathbb{1}_{\mathcal{G}_{i-1}}\mathbb{E}\left(s\Delta^i\left(1 - 2\eta^{i+1}\lambda_k\tau + (\eta^{i+1})^2 C^{i+1}\lambda_1\right) + 2s^2(\eta^{i+1})^2|Z|^2\right).$$

The proof of Lemma 1 utilizes the iteration-wise convergence inequality in Prop. 2 of Section 4.3.

4.2 Bounding probability of bad event $\mathcal{G}_t^c$

Let $\mathcal{G}_0$ denote the good event happening upon initialization of Algorithm 1. Observe that the good events form a nested sequence of subsets through time:

$$\mathcal{G}_0 \supset \mathcal{G}_1 \supset \dots \mathcal{G}_t \supset \dots$$

This implies that we can partition the bad event $\mathcal{G}_t^c$ into a union of individual bad events:

$$\mathcal{G}_t^c = \cup_{i=1}^t \Big(\mathcal{G}_{i-1} \setminus \mathcal{G}_i \Big),$$

The idea behind Proposition 1 is that, we first transform the union of events above into a maximal inequality over a suitable sequence of random variables, which form a supermartingale, and then we apply a type of martingale large-deviation inequality to upper bound $\mathbb{P}\left(\mathcal{G}_t^c\right)$.

Proposition 1 (Bounding probability of bad event)**.** *Suppose the initialization condition in Theorem 1 holds. For any $\delta > 0$, $t \geq 1$, and $i \leq t$, if the learning rate η^i is set such that*

$$\eta^i \leq \min \left\{ \frac{2\lambda_k \tau}{(\frac{16}{1-\tau} \ln \frac{1}{\delta}(b + \|\Sigma^*\|_F)^2 + b(k+1)\lambda_1)}, \frac{\sqrt{2}-1}{b} \right\},$$

Then $\mathbb{P}\left(\mathcal{G}_t^c\right) \leq \delta$.

Proof Sketch. For $i > 1$, we first consider the individual events:

$$\mathcal{G}_{i-1} \setminus \mathcal{G}_i = \mathcal{G}_{i-1} \cap \mathcal{G}_i^c = \{\forall j < i,\ \Delta^j \leq 1 - \tau\} \cap \{\Delta^i > 1 - \tau\}$$

For any strictly increasing positive measurable function g, the above is equivalent to

$$\mathcal{G}_{i-1} \setminus \mathcal{G}_i = \{g(\Delta^i) > g(1-\tau) \ \text{ and } \ \forall j < i,\ g(\Delta^j) \leq g(1-\tau)\}$$

Since event $\mathcal{G}_{i-1}$ occurs is equivalent to $\{\mathbb{1}_{\mathcal{G}_{i-1}} = 1\}$, we can write

$$\mathcal{G}_{i-1} \setminus \mathcal{G}_i = \{g(\Delta^i) > g(1-\tau) \ \text{ and } \ \forall j < i,\ g(\Delta^j) \leq g(1-\tau), \ \text{ and } \ \mathbb{1}_{\mathcal{G}_{i-1}} = 1\}$$

Additionally, since for any $j' < j$, $\mathcal{G}_{j'} \supset \mathcal{G}_j$, that is, $\{\mathbb{1}_{\mathcal{G}_j} = 1\}$ implies $\{\mathbb{1}_{\mathcal{G}_{j'}} = 1\}$, we have

$$\begin{aligned}
&\mathcal{G}_{i-1} \setminus \mathcal{G}_i \\
&= \{g(\Delta^i) > g(1-\tau) \text{ and } \forall j < i,\ g(\Delta^j) \leq g(1-\tau), \mathbb{1}_{\mathcal{G}_{i-1}} = 1, \mathbb{1}_{\mathcal{G}_{j'}} = 1, \forall j' < i-1\} \\
&= \{\mathbb{1}_{\mathcal{G}_{i-1}}\, g(\Delta^i) > g(1-\tau) \text{ and } \forall j < i, \mathbb{1}_{\mathcal{G}_{j-1}}\, g(\Delta^j) \leq g(1-\tau), \text{ and } \mathbb{1}_{\mathcal{G}_j} = 1\} \\
&\subset \{\mathbb{1}_{\mathcal{G}_{i-1}}\, g(\Delta^i) > g(1-\tau) \text{ and } \forall j < i, \mathbb{1}_{\mathcal{G}_{j-1}}\, g(\Delta^j) \leq g(1-\tau)\}
\end{aligned}$$

So the union of the terms $\mathcal{G}_{i-1} \setminus \mathcal{G}_i$ can be upper bounded as

$$\begin{aligned}
&\cup_{i=1}^t \mathcal{G}_{i-1} \setminus \mathcal{G}_i \subset \\
&\cup_{i=2}^t \{\mathbb{1}_{\mathcal{G}_{i-1}}\, g(\Delta^i) > g(1-\tau), \mathbb{1}_{\mathcal{G}_{j-1}}\, g(\Delta^j) \leq g(1-\tau), \forall 1 \leq j < i\} \\
&\cup \{\mathbb{1}_{\mathcal{G}_0}\, g(\Delta^1) > g(1-\tau)\}
\end{aligned}$$

Observe that the event above can also be written as

$$\{ \sup_{1 \le i \le t} \mathbb{1}_{\mathcal{G}_{i-1}} g(\Delta^i) > g(1-\tau) \}.$$

We upper bound the probability of the event above by applying a variant of Doob's inequality. To achieve this, the key step is to find a suitable function g such that the sequence

$$\mathbb{1}_{\mathcal{G}_0} g(\Delta^1), \mathbb{1}_{\mathcal{G}_1} g(\Delta^2), \ldots, \mathbb{1}_{\mathcal{G}_{i-1}} g(\Delta^i), \ldots$$

forms a supermartingale. Via Lemma 1, we show that if we choose $g(x) := \mathbb{E}(sx)$ for any constant $s > 0$, then

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{G}_i} \mathbb{E}\left(s\Delta^{i+1}\right) | \mathcal{F}_i\right] \le \mathbb{1}_{\mathcal{G}_{i-1}} \mathbb{E}\left(s\Delta^i\right), \tag{4.5}$$

provided we choose the learning rate in Algorithm 1 appropriately. Then a version of Doob's inequality for supermartingale [Balsubramani et al.(2013), Durrett(2011), p. 231] implies that

$$\mathbb{P}\left(\sup_i \mathbb{1}_{\mathcal{G}_{i-1}} \mathbb{E}\left(s\Delta^i\right) > \mathbb{E}\left(s(1-\tau)\right)\right) \le \frac{\mathbb{E}\left[\mathbb{1}_{\mathcal{G}_0} \mathbb{E}\left(s\Delta^1\right)\right]}{\mathbb{E}\left(s(1-\tau)\right)},$$

Finally, bounding the expectation on the RHS using our assumption on the initialization condition finishes the proof. □

4.3 Iteration-wise convergence result

Proposition 2 (Iteration-wise subspace improvement)**.** *At the $t+1$-th iteration of Algorithm 1, the following holds:*

(V1) Let $C^t := kb + 2\eta^t b^2 + (\eta^t)^2 b^3$. Then

$$\mathbb{E}\left[tr(U^* P^{t+1}) | \mathcal{F}_t\right] \ge tr(U^* P^t) + 2\eta^{t+1}\lambda_k \Delta^t (1-\Delta^t) - (\eta^{t+1})^2 C^{t+1} \lambda_1 \Delta^t$$

(V2) There exists a random variable Z, with

$$\mathbb{E}\left[Z | \mathcal{F}_t\right] = 0 \quad \textit{and} \quad |Z| \le 2(b + \|\Sigma^*\|_F)\sqrt{\Delta^t}$$

such that

$$tr(U^* P^{t+1}) \ge tr(U^* P^t) + 2\eta^{t+1}\lambda_k \Delta^t (1-\Delta^t) + 2\eta^{t+1} Z - (\eta^{t+1})^2 C^{t+1} \lambda_1 \Delta^t$$

Proof Sketch. By definition,

$$tr(U^* P^{t+1}) = tr(U^* (W^{t+1})^\top (W^{t+1}(W^{t+1})^\top)^{-1} W^{t+1}),$$

where by the update rule of Algorithm 1

$$W^{t+1} = W^t + \eta^{t+1} s^{t+1} (r^{t+1})^\top.$$

We first derive (V1); the proof sketch is as follows:

(i) Since the rows of W^t are orthonormalized, one would expect that a small perturbation of this matrix, W^{t+1}, is also close to orthonormalized, and thus $W^{t+1}(W^{t+1})^\top$ should be close to an identity matrix. Lemma 2 shows this is indeed the case, offsetting by a small term E, which can be viewed as an error/excessive term:

Lemma 2 (Inverse matrix approximation)**.** *Let k' be the number of rows in W^t. Suppose the rows of W^t are orthonormal, that is, $W^t(W^t)^\top = I_{k'}$. Then for $W^{t+1} = W^t + \eta^{t+1} s^{t+1} (r^{t+1})^\top$, we have*

$$(W^{t+1}(W^{t+1})^\top)^{-1} \succeq (1 - \lambda_1(E)) I_{k'} \,,$$

where $\lambda_1(E)$ is the largest eigenvalue of some matrix E, and $\lambda_1(E) = (\eta^{t+1})^2 \|r^{t+1}\|^2 \|s^{t+1}\|^2$.

This implies that

$$\begin{aligned} & tr U^* (W^{t+1})^\top (W^{t+1}(W^{t+1})^\top)^{-1} W^{t+1} \\ \geq\ & (1 - (\eta^{t+1})^2 \|r^{t+1}\|^2 \|s^{t+1}\|^2) tr(U^* (W^{t+1})^\top W^{t+1}) \end{aligned}$$

(ii) We continue to lower bound the conditional expectation of the last term in the previous inequality as

$$\mathbb{E}\left[tr(U^* (W^{t+1})^\top W^{t+1}) | \mathcal{F}_t \right] \geq tr(U^* P^t) + 2\eta^{t+1} tr(U^* P^t \Sigma^* (I_d - P^t))$$

(iii) The last term in the inequality above, $tr(U^* P^t \Sigma^* (I_d - P^t))$, controls the improvement in proximity between the estimated and the ground-truth subspaces. In Lemma 3, we lower bound it as a function of Δ^t:

Lemma 3 (Characterization of stationary points)**.** *Let*

$$\Gamma^t := tr(U^* P^t \Sigma^* (I_d - P^t)) \,,$$

Then the following holds:

(i) $tr(U^* P^t) = tr(U^*)$ implies that $\Gamma^t = 0$.

(ii) $\Gamma^t \geq \lambda_k \Delta^t (1 - \Delta^t)$.

(iv) Finally, combining the results above, we obtain (V1) inequality in the statement of the proposition.

(V2) inequality is derived similarly with the steps above, except that at step 2, instead of considering the conditional expectation of $tr(U^*(W^{t+1})^\top W^{t+1})$, we explicitly represent the zero-mean random variable Z in the inequality. □

5 Experiments

In this section, we present our empirical evaluation of Algorithm 1. We first verified its performance on simulated low-rank data and effectively low-rank data, and then we evaluated its performance on two real-world effectively low-rank datasets.

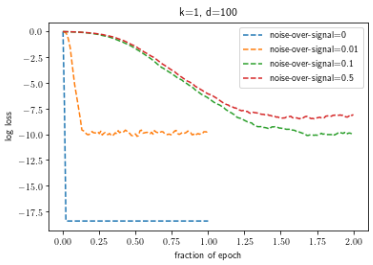

Figure 1: $k = 1$, $d = 100$

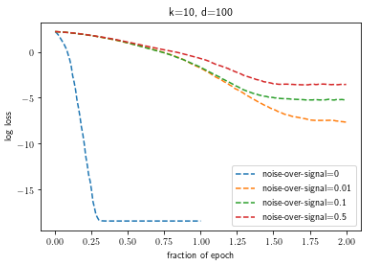

Figure 2: $k = 10$, $d = 100$

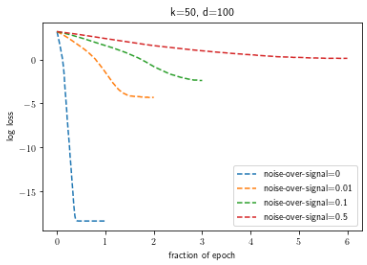

Figure 2: $k = 50$, $d = 100$

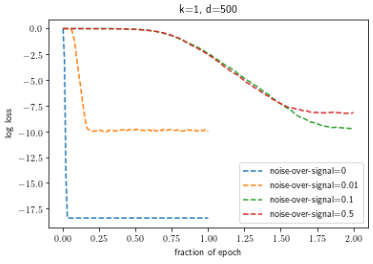

Figure 2: $k = 1$, $d = 500$

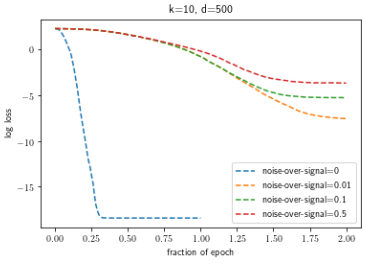

Figure 2: $k = 10$, $d = 500$

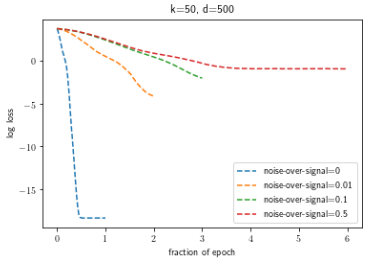

Figure 2: $k = 50$, $d = 500$

Figure 3. log-convergence graph of Algorithm 1: $\ln(\Delta^t)$ vs t at different levels of noise-over-signal ratio ($\frac{\sum_{i>k}\lambda_i}{\sum_{j\le k}\lambda_j}$)

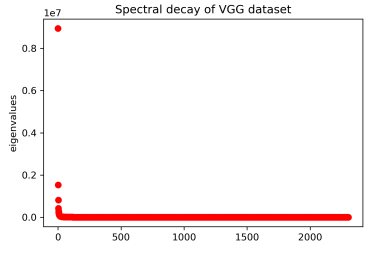

Figure 4. top 6 eigenvalues explains 80% of the data variance.

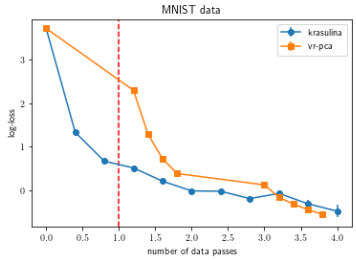

Figure 4. MNIST ($d = 784; k' = 44$)

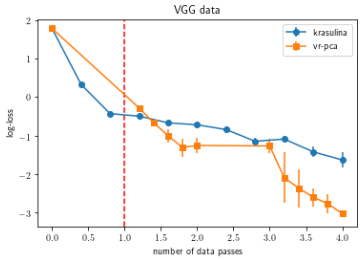

Figure 4. VGG ($d = 2304; k' = 6$); red vertical line marks a full pass over the dataset

5.1 Simulations

The low-rank data is generated as follows: we sample i.i.d. standard normal on the first k coordinates of the d-dimensional data (the rest $d-k$ coordinates are zero), then we rotate all data using a random orthogonal matrix (unknown to the algorithm).

Simulating effectively low-rank data In practice, hardly any dataset is strictly low-rank but many datasets have sharply decaying spectra (recall Figure 4). Although our Theorem 1 is developed under a strict low-rankness assumption, here we empirically test the robustness of our convergence result when data is not strictly low rank but only effectively low rank. Let $\lambda_1, \geq \cdots \geq \lambda_d \geq 0$ be the spectrum of a covariance matrix. For a fixed $k \in [d]$, we let *noise-over-signal* $:= \frac{\sum_{i>k}\lambda_i}{\sum_{j\le k}\lambda_j}$. The *noise-over-signal* ratio intuitively measures how "close" the matrix is to a rank-k

matrix: The smaller the number is, the shaper the spectral decay; when the ratio equals zero, the matrix is of rank at most k. In our simulated data, we perturb the spectrum of a strictly rank-k covariance matrix and generate data with full-rank covariance matrices at the following noise-over-signal ratios, $\{0, 0.01, 0.1, 0.5\}$.

Results Figure 3 shows the *log-convergence graph* of Algorithm 1 on our simulated data: In contrast to the local initialization condition in Theorem 1, we initialized Algorithm 1 with a random matrix W^o and ran it for one or a few epochs, each consists of 5000 iterations. (1). We verified that, on strictly low rank data (noise-over-signal= 0), the algorithm indeed has an exponentially convergence rate (linear in log-error); (2). As we increase the noise-over-signal ratio, the convergence rate gradually becomes slower; (3). The convergence rate is not affected by the actual data dimension d, but only by the intrinsic dimension k, as predicted by Theorem 1.

5.2 Real effectively low-rank datasets

We take a step further to test the performance of Algorithm 1 on two real-world datasets: VGG [Parkhi et al.(2015)] is a dataset of 10806 image files from 2622 distinct celebrities crawled from the web, with $d = 2304$. For MNIST [LeCun and Cortes(2010)], we use the 60000 training examples of digit pixel images, with $d = 784$. Both datasets are full-rank, but we choose k' such that the noise-over-signal ratio at k' is 0.25; that is, the top k' eigenvalues explain 80% of data variance. We compare Algorithm 1 against the exponentially convergent VR-PCA: we initialize the algorithms with the same random matrix and we train (and repeated for 5 times) using the best constant learning rate we found empirically for each algorithm. We see that Algorithm 1 retains fast convergence even if the datasets are not strictly low rank, and that it has a clear advantage over VR-PCA before the iteration reaches a full pass; indeed, VR-PCA requires a full-pass over the dataset before its first iterate.

6 Conclusion

We present *Matrix Krasulina*, an algorithm for online k-PCA, by generalizing the classic Krasulina's method [Krasulina(1969)] from the vector to the matrix case. Our algorithm adapts naturally to low-rank data and converges exponentially fast to the ground-truth principal subspace, both theoretically and empirically. Remarkably, our results show that despite recent efforts to speed up stochastic gradient methods by incorporating an $O(n)$-time variance reduction step, for the k-PCA problem, a purely online stochastic gradient descent variant is sufficient to achieve exponential convergence when the data exhibits low-rank structure.

References

[Krasulina(1969)] T.P. Krasulina. The method of stochastic approximation for the determination of the least eigenvalue of a symmetrical matrix. *USSR Computational Mathematics and Mathematical Physics*, 9 (6):189 – 195, 1969. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(69)90135-9.

[Loukas(2017)] Andreas Loukas. How close are the eigenvectors of the sample and actual covariance matrices? In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2228–2237, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL `http://proceedings.mlr.press/v70/loukas17a.html`.

[Halko et al.(2011)] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, May 2011. ISSN 0036-1445. doi: 10.1137/090771806. URL http://dx.doi.org/10.1137/090771806.

[Golub and Van Loan(1996)] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996. ISBN 0-8018-5414-8.

[Shamir(2015)] Ohad Shamir. A stochastic pca and svd algorithm with an exponential convergence rate. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 144–152. JMLR.org, 2015. URL http://dl.acm.org/citation.cfm?id=3045118.3045135.

[Nie et al.(2016)] Jiazhong Nie, Wojciech Kotlowski, and Manfred K. Warmuth. Online pca with optimal regret. *Journal of Machine Learning Research*, 17(173):1–49, 2016.

[Boutsidis et al.(2015)] Christos Boutsidis, Dan Garber, Zohar Karnin, and Edo Liberty. Online principal components analysis. In *Proceedings of the Twenty-sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, pages 887–901, Philadelphia, PA, USA, 2015. Society for Industrial and Applied Mathematics. URL http://dl.acm.org/citation.cfm?id=2722129.2722190.

[Warmuth and Kuzmin(2006)] Manfred K. Warmuth and Dima Kuzmin. Randomized pca algorithms with regret bounds that are logarithmic in the dimension. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, pages 1481–1488, Cambridge, MA, USA, 2006. MIT Press. URL http://dl.acm.org/citation.cfm?id=2976456.2976642.

[Balsubramani et al.(2013)] 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. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3174–3182, 2013.

[Mitliagkas et al.(2013)] Ioannis Mitliagkas, Constantine Caramanis, and Prateek Jain. Memory limited, streaming pca. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 2886–2894, USA, 2013. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999792.2999934.

[Arora et al.(2013)] Raman Arora, Andrew Cotter, and Nathan Srebro. Stochastic optimization of pca with capped msg. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 1815–1823, USA, 2013. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999792.2999815.

[Johnson and Zhang(2013)] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, pages 315–323, USA, 2013. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999611.2999647.

[Shamir(2016a)] Ohad Shamir. Fast stochastic algorithms for svd and pca: Convergence properties and convexity. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 248–256. JMLR.org, 2016a. URL http://dl.acm.org/citation.cfm?id=3045390.3045418.

[Oja(1982)] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, Nov 1982. ISSN 1432-1416.

[Vu and Lei(2013)] Vincent Q. Vu and Jing Lei. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013. ISSN 00905364. URL http://www.jstor.org/stable/23566753.

[Vu and Lei(2012)] Vincent Vu and Jing Lei. Minimax rates of estimation for sparse pca in high dimensions. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1278–1286, La Palma, Canary Islands, 21–23 Apr 2012. PMLR. URL http://proceedings.mlr.press/v22/vu12.html.

[Beygelzimer et al.(2015)] Alina Beygelzimer, Satyen Kale, and Haipeng Luo. Optimal and adaptive algorithms for online boosting. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2323–2331. JMLR.org, 2015. URL http://dl.acm.org/citation.cfm?id=3045118.3045365.

[Kumar and Kannan(2010)] Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 299–308, 2010. doi: 10.1109/FOCS.2010.35. URL http://dx.doi.org/10.1109/FOCS.2010.35.

[Wang and Singh(2016)] Yining Wang and Aarti Singh. Noise-adaptive margin-based active learning and lower bounds under tsybakov noise condition. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2180–2186. AAAI Press, 2016. URL http://dl.acm.org/citation.cfm?id=3016100.3016203.

[Jain et al.(2013)] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 665–674, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2029-0. doi: 10.1145/2488608.2488693. URL http://doi.acm.org/10.1145/2488608.2488693.

[Keshavan et al.(2010)] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, June 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2046205.

[Candès and Recht(2009)] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, Apr 2009. ISSN 1615-3383. doi: 10.1007/s10208-009-9045-5. URL https://doi.org/10.1007/s10208-009-9045-5.

[Netrapalli et al.(2014)] Praneeth Netrapalli, Niranjan U N, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust pca. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1107–1115. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5430-non-convex-robust-pca.pdf.

[Candès et al.(2011)] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, June 2011. ISSN 0004-5411. doi: 10.1145/1970392.1970395. URL http://doi.acm.org/10.1145/1970392.1970395.

[Allen-Zhu and Li(2017)] Z. Allen-Zhu and Y. Li. First efficient convergence for streaming k-pca: A global, gap-free, and near-optimal rate. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 487–492, Oct 2017. doi: 10.1109/FOCS.2017.51.

[Shalev-Shwartz et al.(2009)] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009. URL http://www.cs.mcgill.ca/~colt2009/papers/018.pdf#page=1.

[Shamir(2016b)] Ohad Shamir. Convergence of stochastic gradient descent for pca. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 257–265. JMLR.org, 2016b. URL http://dl.acm.org/citation.cfm?id=3045390.3045419.

[De Sa et al.(2014)] C. De Sa, K. Olukotun, and C. Ré. Global Convergence of Stochastic Gradient Descent for Some Non-convex Matrix Problems. *ArXiv e-prints*, November 2014.

[Hardt and Price(2014)] Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2861–2869. Curran Associates, Inc., 2014.

[Balcan et al.(2016)] Maria-Florina Balcan, Simon Shaolei Du, Yining Wang, and Adams Wei Yu. An improved gap-dependency analysis of the noisy power method. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 284–309, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL `http://proceedings.mlr.press/v49/balcan16a.html`.

[Li et al.(2016)] Chun-Liang Li, Hsuan-Tien Lin, and Chi-Jen Lu. Rivalry of two families of algorithms for memory-restricted streaming pca. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 473–481, Cadiz, Spain, 09–11 May 2016. PMLR. URL `http://proceedings.mlr.press/v51/li16b.html`.

[Durrett(2011)] Rick Durrett. Probability: Theory and examples, 2011.

[Parkhi et al.(2015)] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[LeCun and Cortes(2010)] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

A Proofs for Proposition 1

Proof of Proposition 1. Recall definition of $\mathcal{G}_t$, $\mathcal{G}_t := \{\Delta^i \le 1-\tau, \forall i \le t\}$. We partition its complement as $\mathcal{G}_t^c = \cup_{i=1}^t \mathcal{G}_{i-1} \setminus \mathcal{G}_i$. For $i > 1$, we first consider the individual events:

$$\mathcal{G}_{i-1} \setminus \mathcal{G}_i = \mathcal{G}_{i-1} \cap \mathcal{G}_i^c = \{\Delta^i > 1-\tau\} \cap \{\forall j < i,\ \Delta^j \le 1-\tau\}$$

For any strictly increasing positive measurable function g, the above is equivalent to

$$\mathcal{G}_{i-1} \setminus \mathcal{G}_i = \{g(\Delta^i) > g(1-\tau) \ \text{ and } \ \forall j < i,\ g(\Delta^j) \le g(1-\tau)\}$$

Since event $\mathcal{G}_{i-1}$ occurs is equivalent to $\{\mathbb{1}_{\mathcal{G}_{i-1}} = 1\}$, we can write

$$\mathcal{G}_{i-1} \setminus \mathcal{G}_i = \{g(\Delta^i) > g(1-\tau) \ \text{ and } \ \forall j < i,\ g(\Delta^j) \le g(1-\tau), \ \text{ and } \ \mathbb{1}_{\mathcal{G}_{i-1}} = 1\}$$

Additionally, since for any $j' < j$, $\mathcal{G}_{j'} \supset \mathcal{G}_j$, that is, $\{\mathbb{1}_{\mathcal{G}_j} = 1\}$ implies $\{\mathbb{1}_{\mathcal{G}_{j'}} = 1\}$, we have

$$\begin{aligned}
&\mathcal{G}_{i-1} \setminus \mathcal{G}_i \\
&= \{g(\Delta^i) > g(1-\tau) \text{ and } \forall j < i,\ g(\Delta^j) \le g(1-\tau), \mathbb{1}_{\mathcal{G}_{i-1}} = 1, \mathbb{1}_{\mathcal{G}_{j'}} = 1, \forall j' < i-1\} \\
&= \{\mathbb{1}_{\mathcal{G}_{i-1}}\, g(\Delta^i) > g(1-\tau) \text{ and } \forall j < i, \mathbb{1}_{\mathcal{G}_{j-1}}\, g(\Delta^j) \le g(1-\tau), \text{ and } \mathbb{1}_{\mathcal{G}_j} = 1\} \\
&\subset \{\mathbb{1}_{\mathcal{G}_{i-1}}\, g(\Delta^i) > g(1-\tau) \text{ and } \forall j < i, \mathbb{1}_{\mathcal{G}_{j-1}}\, g(\Delta^j) \le g(1-\tau)\}
\end{aligned}$$

So the union of the terms $\mathcal{G}_{i-1} \setminus \mathcal{G}_i$ can be upper bounded as

$$\begin{aligned}
&\cup_{i=1}^t \mathcal{G}_{i-1} \setminus \mathcal{G}_i \subset \\
&\cup_{i=2}^t \{\mathbb{1}_{\mathcal{G}_{i-1}}\, g(\Delta^i) > g(1-\tau), \mathbb{1}_{\mathcal{G}_{j-1}}\, g(\Delta^j) \le g(1-\tau), \forall 1 \le j < i\} \\
&\cup \{\mathbb{1}_{\mathcal{G}_0}\, g(\Delta^1) > g(1-\tau)\}
\end{aligned}$$

Observe that the event above can also be written as

$$\{\sup_{1 \le i \le t} \mathbb{1}_{\mathcal{G}_{i-1}}\, g(\Delta^i) > g(1-\tau)\}.$$

Now we upper bound $\mathbb{P}\left(\{\sup_{1\le i\le t} \mathbb{1}_{\mathcal{G}_{i-1}}\, g(\Delta^i) > g(1-\tau)\}\right)$ by applying a martingale large deviation inequality. To achieve this, the key step is to find a suitable function g such that the stochastic process

$$\mathbb{1}_{\mathcal{G}_0}\, g(\Delta^1), \mathbb{1}_{\mathcal{G}_1}\, g(\Delta^2), \ldots, \mathbb{1}_{\mathcal{G}_{i-1}}\, g(\Delta^i), \ldots$$

is a supermartingale. In this proof, we choose $g : \mathbb{R} \to \mathbb{R}_{>0}$ to be $g(x) = \mathbb{E}\,(sx)$ for $s = \frac{2}{1-\tau} \ln \frac{1}{\delta}$. By Lemma 1,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\mathcal{G}_i}\, \mathbb{E}\left(s\Delta^{i+1}\right) | \mathcal{F}_i\right] \\
&\le \mathbb{1}_{\mathcal{G}_{i-1}}\, \mathbb{E}\left(s\Delta^i\big(1 - 2\eta^{i+1}\lambda_k\tau + (\eta^{i+1})^2 C^{i+1}\lambda_1\big) + 2s^2(\eta^{i+1})^2 |Z|^2\right) \\
&\le \mathbb{1}_{\mathcal{G}_{i-1}}\, \mathbb{E}\left(s\Delta^i\big(1 - 2\eta^{i+1}\lambda_k\tau + (\eta^{i+1})^2 C^{i+1}\lambda_1\big)\right) \mathbb{E}\left(s^2(\eta^{i+1})^2 8(b + \|\Sigma^*\|_F)^2 \Delta^i\right) \\
&= \mathbb{1}_{\mathcal{G}_{i-1}}\, \mathbb{E}\left(s\Delta^i\left(1 - 2\eta^{i+1}\lambda_k\tau + (\eta^{i+1})^2 C^{i+1}\lambda_1 + s(\eta^{i+1})^2 8(b + \|\Sigma^*\|_F)^2\right)\right)
\end{aligned}$$

Since we choose the learning rate in Algorithm 1 such that

$$\eta^{i+1} < \frac{2\lambda_k\tau}{b(k+1)\lambda_1 + \frac{16}{1-\tau}\ln\frac{1}{\delta}(b + \|\Sigma^*\|_F)^2} = \frac{2\lambda_k\tau}{b(k+1)\lambda_1 + 8s(b + \|\Sigma^*\|_F)^2}. \tag{A.6}$$

And since $\eta^{i+1} \leq \frac{\sqrt{2}-1}{b}$, it can be seen that

$$C^{i+1} = kb + 2\eta^{i+1}b^2 + (\eta^{i+1})^2 b^3 \leq b(k+1) \tag{A.7}$$

Combining Eq (A.6) and (A.7), we get

$$-2\eta^{i+1}\lambda_k \tau + (\eta^{i+1})^2 C^{i+1}\lambda_1 + s(\eta^{i+1})^2 8(b + \|\Sigma^*\|_F)^2 \leq 0$$

Therefore,

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{G}_i}\mathbb{E}\left(s\Delta^{i+1}\right)|\mathcal{F}_i\right] \leq \mathbb{1}_{\mathcal{G}_{i-1}}\mathbb{E}\left(s\Delta^i\right)$$

Thus, letting $M_i = \mathbb{1}_{\mathcal{G}_{i-1}}\mathbb{E}\left(s\Delta^i\right)$, $(M_i)_{i\geq 1}$ forms a supermartingale. A version of Doob's inequality for supermartingale [Durrett(2011), p. 231] implies that

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{G}_t^c\right) = \mathbb{P}\left(\cup_{i=1}^t \mathcal{G}_{i-1} \setminus \mathcal{G}_i\right) \\
&\leq \mathbb{P}\left(\sup_{i\geq 1}\mathbb{1}_{\mathcal{G}_{i-1}}\mathbb{E}\left(s\Delta^i\right) > \mathbb{E}\left(s(1-\tau)\right)\right) = \mathbb{P}\left(\sup_{i\geq 1} M_i > \mathbb{E}\left(s(1-\tau)\right)\right) \\
&\leq \frac{\mathbb{E}\left[M_1\right]}{\mathbb{E}\left(s(1-\tau)\right)} = \frac{\mathbb{E}\left[\mathbb{1}_{\mathcal{G}_0}\mathbb{E}\left(s\Delta^1\right)\right]}{\mathbb{E}\left(s(1-\tau)\right)}
\end{aligned}$$

We bound the expectation as follows: By Inequality A.8 of Lemma 1,

$$\mathbb{E}\left(s\Delta^1\right)\mathbb{1}_{\mathcal{G}_0} \leq \mathbb{E}\left(s\left(\Delta^0(1-2\eta^1\lambda_k(1-\Delta^0)) - 2\eta^1 Z + (\eta^1)^2 C^1\lambda_1\Delta^0\right)\right)\mathbb{1}_{\mathcal{G}_0}$$

Taking expectation on both sides,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\mathcal{G}_0}\mathbb{E}\left(s\Delta^1\right)\right] \\
&\leq \mathbb{E}\left(s\left(\Delta^0(1-2\eta^1\lambda_k(1-\Delta^0)) + (\eta^1)^2 C^1\lambda_1\Delta^0\right)\right)\mathbb{E}\left[\mathbb{E}\left(s(-2\eta^1 Z)\right)\right] \\
&\leq \mathbb{E}\left(s\left(\Delta^0(1-2\eta^1\lambda_k(1-\Delta^0)) + (\eta^1)^2 C^1\lambda_1\Delta^0\right)\right)\mathbb{E}\left(2s^2(\eta^1)^2|Z|^2\right) \\
&\leq \mathbb{E}\left(s\Delta^0\left(1-2\eta^1\lambda_k\tau + (\eta^1)^2 C^1\lambda_1 + s(\eta^1)^2 8(b+\|\Sigma^*\|_F)^2\right)\right) \\
&\leq \mathbb{E}\left(s\frac{1-\tau}{2}\left(1-2\eta^1\lambda_k\tau + (\eta^1)^2 C^1\lambda_1 + s(\eta^1)^2 8(b+\|\Sigma^*\|_F)^2\right)\right) \\
&\leq \mathbb{E}\left(s\frac{1-\tau}{2}\right)
\end{aligned}$$

where the second inequality holds by Hoeffding's lemma (using the same argument as in Lemma 1), and the third and fourth inequality is by the fact that $\Delta^0 \leq \frac{1-\tau}{2}$ holds by our assumption. Finally,

$$\frac{\mathbb{E}\left[\mathbb{1}_{\mathcal{G}_0}\mathbb{E}\left(s\Delta^1\right)\right]}{\mathbb{E}\left(s(1-\tau)\right)} \leq \mathbb{E}\left(-s(1-\tau)/2\right) \leq \delta\,,$$

since we set $s = \frac{2}{1-\tau}\ln\frac{1}{\delta}$. □

A.1 Auxiliary lemma for Proposition 1

Proof of Lemma 1. By V2 of Proposition 2, for Σ^* with rank k,

$$\begin{aligned} tr(U^*P^{i+1}) &\geq tr(U^*P^i) \\ +2\eta^{i+1}\sum_{\ell=1}^{k}\lambda_\ell(1-u_\ell^\top P^i u_\ell)(u_\ell^\top P^i u_\ell - \sum_{m\neq\ell}[1-u_m^\top P^i u_m]) &+ 2\eta^{i+1}Z \\ -(\eta^{i+1})^2 C^{i+1} tr(\Sigma^* - \Sigma^* P^i) & \end{aligned}$$

From this, we can derive

$$\begin{aligned} \Delta^{i+1} &\leq \Delta^i - 2\eta^{i+1}\sum_{\ell=1}^{k}\lambda_\ell(1-u_\ell^\top P^t u_\ell)(1-\Delta^i) - 2\eta^{i+1}Z + (\eta^{i+1})^2 C^{i+1}\lambda_1\Delta^i \\ &\leq \Delta^i - 2\eta^{i+1}\lambda_k tr(U^* - U^*P^i)(1-\Delta^i) - 2\eta^{i+1}Z + (\eta^{i+1})^2 C^{i+1}\lambda_1\Delta^i \\ &= \Delta^i - 2\eta^{i+1}\lambda_k\Delta^i(1-\Delta^i) - 2\eta^{i+1}Z + (\eta^{i+1})^2 C^{i+1}\lambda_1\Delta^i \\ &= \Delta^i(1-2\eta^{i+1}\lambda_k(1-\Delta^i)) - 2\eta^{i+1}Z + (\eta^{i+1})^2 C^{i+1}\lambda_1\Delta^i \end{aligned}$$

This implies that for any $s > 0$,

$$\mathbb{E}\left(s\Delta^{i+1}\right) \leq \mathbb{E}\left(s\Big(\Delta^i(1-2\eta^{i+1}\lambda_k(1-\Delta^i)) - 2\eta^{i+1}Z + (\eta^{i+1})^2 C^{i+1}\lambda_1\Delta^i\Big)\right)$$

Multiplying both sides of the inequality by $\mathbb{1}_{\mathcal{G}_i}$, we get

$$\begin{aligned} &\mathbb{E}\left(s\Delta^{i+1}\right)\mathbb{1}_{\mathcal{G}_i} \\ &\leq \mathbb{E}\left(s\Big(\Delta^i(1-2\eta^{i+1}\lambda_k(1-\Delta^i)) - 2\eta^{i+1}Z + (\eta^{i+1})^2 C^{i+1}\lambda_1\Delta^i\Big)\right)\mathbb{1}_{\mathcal{G}_i} \end{aligned} \tag{A.8}$$

We can further upper bound the RHS of Inequality (A.8) above as

$$\begin{aligned} &\mathbb{E}\left(s\Big(\Delta^i(1-2\eta^{i+1}\lambda_k(1-\Delta^i)) - 2\eta^{i+1}Z + (\eta^{i+1})^2 C^{i+1}\lambda_1\Delta^i\Big)\right)\mathbb{1}_{\mathcal{G}_i} \\ &\leq \mathbb{E}\left(s\Big(\Delta^i(1-2\eta^{i+1}\lambda_k\tau) - 2\eta^{i+1}Z + (\eta^{i+1})^2 C^{i+1}\lambda_1\Delta^i\Big)\right)\mathbb{1}_{\mathcal{G}_i} \\ &\leq \mathbb{E}\left(s\Big(\Delta^i(1-2\eta^{i+1}\lambda_k\tau) - 2\eta^{i+1}Z + (\eta^{i+1})^2 C^{i+1}\lambda_1\Delta^i\Big)\right)\mathbb{1}_{\mathcal{G}_{i-1}} \\ &\leq \mathbb{1}_{\mathcal{G}_{i-1}}\,\mathbb{E}\left(s\Big(\Delta^i(1-2\eta^{i+1}\lambda_k\tau) + (\eta^{i+1})^2 C^{i+1}\lambda_1\Delta^i\Big)\right)\mathbb{E}\left(s\big(-2\eta^{i+1}Z\big)\right) \end{aligned}$$

The first inequality is due to the fact that "$\{\mathbb{1}_{\mathcal{G}_i} = 1\}$ implies $\{\Delta^i \leq 1-\tau\}$" and the second inequality holds since $\mathcal{G}_i \subset \mathcal{G}_{i-1}$. Incorporating this bound into inequality (A.8) and taking conditional expectation w.r.t. $\mathcal{F}_i$ on both sides, we get

$$\begin{aligned} &\mathbb{1}_{\mathcal{G}_i}\,\mathbb{E}\left[\mathbb{E}\left(s\Delta^{i+1}\right)|\mathcal{F}_i\right] = \mathbb{E}\left[\mathbb{E}\left(s\Delta^{i+1}\right)\mathbb{1}_{\mathcal{G}_i}|\mathcal{F}_i\right] \\ &\leq \mathbb{1}_{\mathcal{G}_{i-1}}\,\mathbb{E}\left(s\Big(\Delta^i(1-2\eta^{i+1}\lambda_k\tau) + (\eta^{i+1})^2 C^{i+1}\lambda_1\Delta^i\Big)\right)\mathbb{E}\left[\mathbb{E}\left(s\big(-2\eta^{i+1}Z\big)\right)|\mathcal{F}_i\right] \end{aligned}$$

Now we upper bound $\mathbb{E}\left[\mathbb{E}\left(s\big(-2\eta^{i+1}Z\big)\right)|\mathcal{F}_i\right]$: Since

$$-2\eta^{i+1}|Z| \le 2\eta^{i+1}(-Z) \le 2\eta^{i+1}|Z|\,,$$

and

$$\mathbb{E}\left[2s\eta^{i+1}(-Z)|\mathcal{F}_i\right] = \mathbb{E}\left[2s\eta^{i+1}Z|\mathcal{F}_i\right] = 0\,,$$

by Hoeffding's lemma

$$\mathbb{E}\left[\mathbb{E}\left(2s\eta^{i+1}(-Z)|\mathcal{F}_i\right)\right] \le \mathbb{E}\left(\frac{s^2(4\eta^{i+1}|Z|)^2}{8}\right) = \mathbb{E}\left(2s^2(\eta^{i+1})^2|Z|^2\right)\,.$$

Combining this with the previous bound, we get

$$\begin{aligned}&\mathbb{1}_{\mathcal{G}_i}\,\mathbb{E}\left[\mathbb{E}\left(s\Delta^{i+1}\right)|\mathcal{F}_i\right]\\ &\le \mathbb{1}_{\mathcal{G}_{i-1}}\,\mathbb{E}\left(s\Big(\Delta^i(1-2\eta^{i+1}\lambda_k\tau) + (\eta^{i+1})^2C^{i+1}\lambda_1\Delta^i\Big)\right)\mathbb{E}\left(2s^2(\eta^{i+1})^2|Z|^2\right)\end{aligned}$$

□

B Proofs for Proposition 2

Proof of Proposition 2. We consider

$$\mathbb{E}\left[trU^*P^{t+1}\big|\mathcal{F}_t\right] = \mathbb{E}\left[trU^*(W^{t+1})^\top(W^{t+1}(W^{t+1})^\top)^{-1}W^{t+1}\big|\mathcal{F}_t\right]\,,$$

Since U^* is positive semidefinite, we can write it as $U^* = ((U^*)^{1/2})^2$. By the proof of Lemma 2,

$$(W^{t+1}(W^{t+1})^\top)^{-1} \succeq (1-(\eta^{t+1})^2\|r^{t+1}\|^2\|s^{t+1}\|^2)I_{k'}$$

Letting $V := W^{t+1}(U^*)^{1/2}$, this implies that

$$V^\top\Big[W^{t+1}(W^{t+1})^\top)^{-1} - (1-(\eta^{t+1})^2\|r^{t+1}\|^2\|s^{t+1}\|^2)I_{k'}\Big]V \succeq 0$$

That is, the matrix on the left-hand-side above is positive semi-definite. Since trace of a positive semi-definite matrix is non-negative, we have

$$tr(V^\top W^{t+1}(W^{t+1})^\top)^{-1}V) \ge tr(V^\top(1-(\eta^{t+1})^2\|r^{t+1}\|^2\|s^{t+1}\|^2)V)$$

By commutative property of trace, we further get

$$\begin{aligned}tr(U^*(W^{t+1})^\top[W^{t+1}(W^{t+1})^\top]^{-1}W^{t+1}) &= tr(V^\top W^{t+1}(W^{t+1})^\top)^{-1}V)\\ &\ge tr(V^\top(1-(\eta^{t+1})^2\|r^{t+1}\|^2\|s^{t+1}\|^2)V)\\ &= (1-(\eta^{t+1})^2\|r^{t+1}\|^2\|s^{t+1}\|^2)tr(U^*(W^{t+1})^\top W^{t+1})\end{aligned}$$

Taking expectation on both sides, we get

$$\mathbb{E}\left[trU^*P^{t+1}\big|\mathcal{F}_t\right] \ge (1-(\eta^{t+1})^2\|r^{t+1}\|^2\|s^{t+1}\|^2)\,\mathbb{E}\left[tr(U^*(W^{t+1})^\top W^{t+1})\big|\mathcal{F}_t\right]$$

Now we in turn lower bound $\mathbb{E}\left[tr[U^*(W^{t+1})^\top W^{t+1}]\middle|\mathcal{F}_t\right]$. First, we have

$$\begin{aligned}(W^{t+1})^\top W^{t+1} &= (W^t + \eta^{t+1}s^{t+1}(r^{t+1})^\top)^\top(W^t + \eta^{t+1}s^{t+1}(r^{t+1})^\top)\\ = P^t + \eta^{t+1}r^{t+1}(s^{t+1})^\top W^t + \eta^{t+1}(W^t)^\top s^{t+1}(r^{t+1})^\top &+ (\eta^{t+1})^2\|s^{t+1}\|^2 r^{t+1}(r^{t+1})^\top\end{aligned}$$

This implies that

$$\begin{aligned}\mathbb{E}\left[tr[U^*(W^{t+1})^\top W^{t+1}]\middle|\mathcal{F}_t\right] &= tr(U^*\,\mathbb{E}\left[(W^{t+1})^\top W^{t+1}\middle|\mathcal{F}_t\right])\\ &= tr(U^*P^t) + \eta^{t+1}tr\,\mathbb{E}\left[U^*r^{t+1}(s^{t+1})^\top\middle|\mathcal{F}_t\right]W^t\\ &\quad +\eta^{t+1}tr(\mathbb{E}\left[U^*(W^t)^\top s^{t+1}(r^{t+1})^\top\middle|\mathcal{F}_t\right])\\ &\quad +(\eta^{t+1})^2\,\mathbb{E}\left[\|s^{t+1}\|^2 tr(U^*r^{t+1}(r^{t+1})^\top)\middle|\mathcal{F}_t\right]\\ &\geq tr(U^*P^t) + \eta^{t+1}trU^*\,\mathbb{E}\left[r^{t+1}(s^{t+1})^\top\middle|\mathcal{F}_t\right]W^t\\ &\quad +\eta^{t+1}tr(U^*\,\mathbb{E}\left[(W^t)^\top s^{t+1}(r^{t+1})^\top\middle|\mathcal{F}_t\right])\\ &\geq tr(U^*P^t) + 2\eta^{t+1}tr(U^*\,\mathbb{E}\left[(W^t)^\top s^{t+1}(r^{t+1})^\top\middle|\mathcal{F}_t\right])\end{aligned}$$

the second to last inequality follows since we can drop the non-negative term, and the last inequality holds since the $tr(A) = tr(A^\top)$ for any square matrix A. Since

$$\mathbb{E}\left[s^{t+1}(r^{t+1})^\top\middle|\mathcal{F}_t\right] = W^t(\Sigma^* - \Sigma^*P^t)\,,$$

we have

$$trU^*\,\mathbb{E}\left[(W^t)^\top s^{t+1}(r^{t+1})^\top\middle|\mathcal{F}_t\right] = trU^*(P^t\Sigma^* - P^t\Sigma^*P^t)\,.$$

By Lemma 3,

$$\begin{aligned}&trU^*\,\mathbb{E}\left[(W^t)^\top s^{t+1}(r^{t+1})^\top\middle|\mathcal{F}_t\right]\\ &\qquad = trU^*(P^t\Sigma^* - P^t\Sigma^*P^t)\\ \geq \sum_{i=1}^k \lambda_i(1-u_i^\top P^t u_i)(u_i^\top P^t u_i - &\sum_{j\neq i, j\in[k]}[1-u_j^\top P^t u_j])\end{aligned}$$

Then we have,

$$\begin{aligned}\mathbb{E}\left[tr[U^*(W^{t+1})^\top W^{t+1}]\middle|\mathcal{F}_t\right] &\geq tr(U^*P^t)\\ +2\eta^{t+1}\sum_{i=1}^k \lambda_i(1-u_i^\top P^t u_i)(u_i^\top P^t u_i - &\sum_{j\neq i, j\in[k]}[1-u_j^\top P^t u_j])\end{aligned}$$

Now we can bound $\mathbb{E}\left[trU^*P^{t+1}\middle|\mathcal{F}_t\right]$ as:

$$\begin{aligned}&\mathbb{E}\left[tr(U^*(W^{t+1})^\top[W^{t+1}(W^{t+1})^\top]^{-1}W^{t+1})\middle|\mathcal{F}_t\right]\\ \geq \mathbb{E}\left[tr(U^*(W^{t+1})^\top W^{t+1})\middle|\mathcal{F}_t\right] &- \mathbb{E}\left[(\eta^{t+1})^2\|r^{t+1}\|^2\|s^{t+1}\|^2 tr[U^*(W^{t+1})^\top W^{t+1}]\middle|\mathcal{F}_t\right]\end{aligned}$$

$$\geq tr(U^*P^t) + 2\eta^{t+1}\sum_{i=1}^{k}\lambda_i(1-u_i^\top P^t u_i)(u_i^\top P^t u_i - \sum_{j\neq i, j\in[k]}[1-u_j^\top P^t u_j])$$
$$- \mathbb{E}\left[(\eta^{t+1})^2\|r^{t+1}\|^2\|s^{t+1}\|^2 tr(U^*(W^{t+1})^\top W^{t+1})\big|\mathcal{F}_t\right] \tag{B.9}$$

Note that the second term in the inequality above can be lower bounded as:

$$\sum_{i=1}^{k}\lambda_i(1-u_i^\top P^t u_i)(u_i^\top P^t u_i - \sum_{j\neq i, j\in[k]}[1-u_j^\top P^t u_j])$$
$$= \sum_{i=1}^{k}\lambda_i(1-u_i^\top P^t u_i)(\sum_{j\in[k]} u_j^\top P^t u_j - (k-1))$$
$$= \sum_{i=1}^{k}\lambda_i(1-u_i^\top P^t u_i)(1-\Delta^t) \geq \lambda_k\Delta^t(1-\Delta^t)$$

Since $k' \leq d$, and rows of W^t are orthonormal, we get

$$\|s^{t+1}\|^2 = \|W^t X^{t+1}\|^2 \leq \|X^{t+1}\|^2 .$$

Similarly, $\|r^{t+1}\|^2 \leq \|X^{t+1}\|^2$. Therefore,

$$\|s^{t+1}\|^2 tr(U^*(W^{t+1})^\top W^{t+1})$$
$$\leq \|X^{t+1}\|^2\Big(trU^*P^t + 2\eta^{t+1}trU^*r^{t+1}(X^{t+1})^\top P^t + (\eta^{t+1})^2\|s^{t+1}\|^2 trU^*r^{t+1}(r^{t+1})^\top\Big)$$
$$= \|X^{t+1}\|^2\Big(trU^*P^t + 2\eta^{t+1}(X^{t+1})^\top P^t U^* r^{t+1} + (\eta^{t+1})^2\|s^{t+1}\|^2(r^{t+1})^\top U^* r^{t+1}\Big)$$
$$\leq \|X^{t+1}\|^2\Big(trU^*P^t + 2\eta^{t+1}\|X^{t+1}\|^2 + (\eta^{t+1})^2\|s^{t+1}\|^2\|r^{t+1}\|^2\Big)$$
$$\leq \|X^{t+1}\|^2\Big(trU^*P^t + 2\eta^{t+1}\|X^{t+1}\|^2 + (\eta^{t+1})^2\|X^{t+1}\|^4\Big)$$

On the other hand, we have

$$\mathbb{E}\left[\|r^{t+1}\|^2\big|\mathcal{F}_t\right] = tr(\Sigma^* - \Sigma^* P^t)$$

Thus, the quadratic term (quadratic in η^{t+1}) in Eq (B.9) can be upper bounded as

$$\mathbb{E}\left[(\eta^{t+1})^2\|r^{t+1}\|^2\|s^{t+1}\|^2 tr(U^*(W^{t+1})^\top W^{t+1})\big|\mathcal{F}_t\right]$$
$$\leq (\eta^{t+1})^2 C_o^t\,\mathbb{E}\left[\|r^{t+1}\|^2\big|\mathcal{F}_t\right] = (\eta^{t+1})^2 C_o^t tr(\Sigma^* - \Sigma^* P^t)$$

where

$$C_o^t := \max_X \|X\|^2\big(trU^*P^t + 2\eta^{t+1}\|X\|^2 + (\eta^{t+1}\|X\|^2)^2\big)$$
$$\leq \max_X \|X\|^2\big(k + 2\eta^{t+1}\|X\|^2 + (\eta^{t+1}\|X\|^2)^2\big)$$
$$= kb + 2\eta^{t+1}b^2 + (\eta^{t+1})^2 b^3$$

Note that by our definition of C^{t+1},

$$C^{t+1} := kb + 2\eta^{t+1}b^2 + (\eta^{t+1})^2 b^3$$

We get that

$$\mathbb{E}\left[(\eta^{t+1})^2 \|r^{t+1}\|^2 \|s^{t+1}\|^2 tr(U^*(W^{t+1})^\top W^{t+1}) \middle| \mathcal{F}_t\right] \leq C^{t+1}(\eta^{t+1})^2 tr(\Sigma^* - \Sigma^* P^t)\,.$$

Since

$$\lambda_1 U^* \succeq \Sigma^* \succeq \lambda_k U^*$$

We have

$$\begin{aligned}(I-P^t)^\top \lambda_1 U^* (I-P^t) &\succeq (I-P^t)^\top \Sigma^* (I-P^t) \\ &\succeq (I-P^t)^\top \lambda_k U^* (I-P^t)\end{aligned}$$

Note that the projection matrix satisfies $(I-P^t)^\top = (I-P^t)$ and $(I-P^t)(I-P^t) = (I-P^t)$. This implies that

$$\lambda_1 tr U^*(I-P^t) \geq tr\Sigma^*(I-P^t) \geq \lambda_k tr U^*(I-P^t) \tag{B.10}$$

Finally, plug the lower bound in Eq. (B.9) completes the proof:

$$\begin{aligned}\mathbb{E}\left[tr(U^* P^{t+1}) | \mathcal{F}_t\right] &\geq tr(U^* P^t) + \lambda_k \Delta^t (1-\Delta^t) - (\eta^{t+1})^2 C^{t+1} \lambda_1 tr(U^*(I-P^t)) \\ &\geq tr(U^* P^t) + \lambda_k \Delta^t (1-\Delta^t) - (\eta^{t+1})^2 C^{t+1} \lambda_1 \Delta^t\end{aligned}$$

The inequality of the statement in Version 2 can be obtained similarly, by setting

$$Z = 2\left(tr(U^*(W^t)^\top s^{t+1}(r^{t+1})^\top) - \mathbb{E}\left[tr(U^*(W^t)^\top s^{t+1}(r^{t+1})^\top) \middle| \mathcal{F}_t\right]\right)$$

It is clear that $\mathbb{E}\left[Z \middle| \mathcal{F}_t\right] = 0$. Now we upper bound $|Z|$: Since

$$tr(U^*(W^t)^\top s^{t+1}(r^{t+1})^\top) = tr U^* P^t X^{t+1}(X^{t+1})^\top (I-P^t)$$

we get (subsequently, we denote P^t by P, X^{t+1} by X)

$$\begin{aligned}|Z| &= |2tr(U^* P X X^\top (I-P)) - 2tr(U^* P \Sigma^* (I-P))| \\ &= 2|tr(XX^\top - \Sigma^*)(I-P)U^* P| \leq 2\sqrt{\|XX^\top - \Sigma^*\|_F^2 \|(I-P)U^* P\|_F^2}\end{aligned}$$

We first bound $\|(I-P)U^* P\|_F^2$,

$$\|(I-P)U^* P\|_F^2 \leq \|(I-P)U^*\|_F^2 = tr(U^* - U^* P)$$

where the first inequality is due to the fact that P is a projection matrix so that norms are at best preserved if not smaller; the second inequality is also due to the fact that both U^* and $I-P$

are projection matrices, and thus $(I-P)(I-P) = I-P$ and $U^*U^* = U^*$. Now we bound $\|XX^\top - \Sigma^*\|_F^2$:

$$\begin{aligned}\|XX^\top - \Sigma^*\|_F^2 &= tr(XX^\top - \Sigma^*)^\top(XX^\top - \Sigma^*)\\ &= \|X\|^4 - 2X^\top\Sigma^* X + \|\Sigma^*\|_F^2 \le \|X\|^4 + \|\Sigma^*\|_F^2\end{aligned}$$

where the last inequality is due to the fact that Σ^* is positive semidefinite, that is, for any x, we have $x^\top\Sigma^* x \ge 0$. Finally,

$$\begin{aligned}|Z| &\le 2\sqrt{\|XX^\top - \Sigma^*\|_F^2\|(I-P)U^*P\|_F^2}\\ &\le 2\sqrt{(\|X\|^4 + \|\Sigma^*\|_F^2)tr(U^* - U^*P)}\\ &\le 2(\|X\|^2 + \|\Sigma^*\|_F)\sqrt{\Delta^t} \le 2(b + \|\Sigma^*\|_F)\sqrt{\Delta^t}\end{aligned}$$

The third inequality is by the following argument: for any $a \ge 0, b \ge 0$, we have $\sqrt{a^2+b^2} \le a+b$, Letting $a = \|X\|^2$ and $b = \|\Sigma^*\|_F$ leads to the inequality. □

B.1 Auxiliary lemmas for Proposition 2

Proof of Lemma 2.

$$\begin{aligned}W^{t+1}(W^{t+1})^\top &= (W^t + \eta^{t+1}s^{t+1}(r^{t+1})^\top)(W^t + \eta^{t+1}s^{t+1}(r^{t+1})^\top)^\top\\ &= (W^t)(W^t)^\top + \eta^{t+1}s^{t+1}(r^{t+1})^\top(W^t)^\top\\ &+\eta^{t+1}W^t r^{t+1}(s^{t+1})^\top + (\eta^{t+1})^2\|r^{t+1}\|^2 s^{t+1}(s^{t+1})^\top\\ &= I_{k'} + (\eta^{t+1})^2\|r^{t+1}\|^2 s^{t+1}(s^{t+1})^\top\end{aligned}$$

where the last equality holds because W^t has orthonormalized rows, and r^{t+1} is orthogonal to rows of W^t. Let

$$E := (\eta^{t+1})^2\|r^{t+1}\|^2 s^{t+1}(s^{t+1})^\top .$$

Note that E is symmetric and positive semidefinite. We can eigen-decompose E as

$$E = Q\Lambda Q^\top$$

where Q is the eigenbasis and Λ is a diagonal matrix with real non-negative diagonal values, with $\Lambda_{11} \ge \Lambda_{22}, \ge, \dots \Lambda_{k'k'}$, corresponding to the non-decreasing eigenvalues of E. Then

$$(I_{k'} + E)^{-1} = (QQ^\top + Q\Lambda Q^\top)^{-1} = Q(I_{k'} + \Lambda)^{-1}Q^\top ,$$

Since $I_{k'} + \Lambda$ is a diagonal matrix, for any $i \in [k']$, we have

$$(I_{k'} + \Lambda)^{-1}_{ii} = \frac{1}{1+\Lambda_{ii}} \ge 1 - \Lambda_{ii} \ge 1 - \Lambda_{11}$$

This implies that the matrix

$$Q[(I_{k'} + \Lambda)^{-1} - (1-\Lambda_{11})I_{k'}]Q^\top$$

is positive semidefinite, that is,

$$Q(I_{k'} + \Lambda)^{-1}Q^\top \succeq Q(1-\Lambda_{11})I_{k'}Q^\top = (1-\Lambda_{11})I_{k'}$$

Thus,

$$(W^{t+1}(W^{t+1})^\top)^{-1} = (I_{k'} + E)^{-1} \succeq (1 - \Lambda_{11}) I_{k'}$$

Finally, we compute the largest eigenvalue of E, $\lambda_1(E) := \Lambda_{11}$:

$$\begin{aligned}\lambda_1(E) &= \max_{\|y\|=1} y^\top E y = \max_{\|y\|=1} (\eta^{t+1})^2 \|r^{t+1}\|^2 (y^\top s^{t+1} (s^{t+1})^\top y) \\ &= (\eta^{t+1})^2 \|r^{t+1}\|^2 \max_{\|y\|=1} (\langle s^{t+1}, y\rangle)^2 = (\eta^{t+1})^2 \|r^{t+1}\|^2 \|s^{t+1}\|^2\end{aligned}$$

This completes the proof. □

Proof of Lemma 3. We first prove statement 1. Since U^* is symmetric and positive semidefinite, we can write it as $U^* = ((U^*)^{1/2})^2$. So we have

$$\begin{aligned}tr(U^* - U^* P^t) &= tr(U^*(I - P^t)) \\ = tr((U^*)^{1/2}(I - P^t)(I - P^t)(U^*)^{1/2}) &= \|(I - P^t)(U^*)^{1/2}\|_F^2\end{aligned}$$

Therefore, $tr(U^*) = tr(U^* P^t)$ implies that

$$tr(U^* - U^* P^t) = \|(I - P^t)(U^*)^{1/2}\|_F^2 = 0$$

which implies

$$(I - P^t)(U^*)^{1/2} = 0$$

where "0" denotes the zero matrix. Thus,

$$\Gamma^t = tr(P^t \Sigma^* (I - P^t) U^*) = tr(P^t \Sigma^* (I - P^t)(U^*)^{1/2}(U^*)^{1/2}) = tr 0 = 0\,.$$

Now we prove statement 2. First, we upper bound $tr(P^t \Sigma^* P^t U^*)$:

$$\begin{aligned}tr(P^t \Sigma^* P^t U^*) &= tr(\sum_{p=1}^{k'} \sum_{i=1}^{k} \sum_{q=1}^{k'} \sum_{j=1}^{k} \lambda_i \langle w_p, u_i\rangle \langle w_q, u_i\rangle \langle w_q, u_j\rangle w_p u_j^\top) \\ &= \sum_i \lambda_i \sum_j \sum_p \langle w_p, u_i\rangle \langle w_p, u_j\rangle \sum_q \langle w_q, u_i\rangle \langle w_q, u_j\rangle \\ &= \sum_i \lambda_i \sum_j (u_i^\top P^t u_j)^2\end{aligned}$$

Note that by Cauchy-Schwarz inequality,

$$(u_i^\top P^t u_j)^2 = (u_i^\top P^t (P^t)^\top u_j)^2 \le \|P^t u_i\|^2 \|P^t u_j\|^2 = (u_i^\top P^t u_i)(u_j^\top P^t u_j)$$

On the other hand, for any i and $j \neq i$ since $u_i \perp u_j$, we have

$$u_i^\top P^t u_j = u_i^\top u_j - u_i^\top (I - P^t) u_j = -u_i^\top (I - P^t) u_j$$

we have

$$\begin{aligned}(u_i^\top P^t u_j)^2 = (u_i^\top (I - P^t) u_j)^2 &= (u_i^\top (I - P^t)(I - P^t) u_j)^2 \\ &\le \|(I - P^t) u_i\|^2 \|(I - P^t) u_j\|^2\end{aligned}$$

$$= (\|u_i\|^2 - \|P^t u_i\|^2)(\|u_j\|^2 - \|P^t u_j\|^2)$$
$$= (1 - u_i^\top P^t u_i)(1 - u_j^\top P^t u_j)$$

where the inequality is by Cauchy-Schwarz inequality, and the third equality is by combining orthogonality of projection P^t and Pythagorean theorem. This implies that

$$tr(P^t \Sigma^* P^t U^*) = \sum_i \lambda_i \sum_j (u_i^\top P^t u_j)^2$$
$$= \sum_i \lambda_i (u_i^\top P^t u_i)^2 + \sum_i \lambda_i \sum_{j \neq i} (u_i^\top P^t u_j)^2$$
$$\leq \sum_i \lambda_i (u_i^\top P^t u_i)^2 + \sum_i \lambda_i \sum_{j \neq i} (1 - u_i^\top P^t u_i)(1 - u_j^\top P^t u_j)$$

Next, we expand $tr(P^t \Sigma^* U^*)$:

$$tr(P^t \Sigma^* U^*) = tr(U^* P^t \Sigma^*) = tr(\sum_i u_i u_i^\top P^t \sum_j \lambda_j u_j u_j^\top)$$
$$= \sum_i \sum_j \lambda_j u_i^\top P^t u_j u_i^\top u_j = \sum_i \lambda_i u_i^\top P^t u_i$$

Combining the upper bound on $tr(P^t \Sigma^* P^t U^*)$, we get,

$$tr(P^t \Sigma^* U^*) - tr(P^t \Sigma^* P^t U^*) = \sum_i \lambda_i u_i^\top P^t u_i - tr(P^t \Sigma^* P^t U^*)$$
$$\geq \sum_i \lambda_i u_i^\top P^t u_i - \sum_i \lambda_i (u_i^\top P^t u_i)^2 - \sum_i \lambda_i \sum_{j \neq i} (1 - u_i^\top P^t u_i)(1 - u_j^\top P^t u_j)$$
$$= \sum_i \lambda_i (1 - u_i^\top P^t u_i)(u_i^\top P^t u_i - \sum_{j \neq i} [1 - u_j^\top P^t u_j])$$

Recall that

$$\Delta^t = k - tr(U^* P^t) = k - \sum_{i=1}^k u_i^\top P^t u_i \,,$$

Therefore, the last term in the inequality above can be further lower bounded by $\lambda_k \Delta^t (1 - \Delta^t)$. □

C Proof of Theorem 1

Proof of Theorem 1. Since by our assumption, $\Delta^o \leq \frac{1-\tau}{2}$, for any $\delta > 0$, and since we choose the learning rate such that

$$\eta \leq \min\{\frac{2\lambda_k \tau}{\frac{16}{1-\tau} \ln \frac{1}{\delta}(b + \|\Sigma^*\|_F)^2 + b(k+1)\lambda_1}, \frac{\sqrt{2}-1}{b}\}\,,$$

we can apply Proposition 1 to bound the probability of bad event, $\mathcal{G}_t^c$ as $\mathbb{P}(\mathcal{G}_t^c) \leq \delta$. By V1 of Proposition 2 (and let C^{t+1} be as denoted therein),

$$\mathbb{E}\left[tr(U^* P^{t+1}) | \mathcal{F}_t\right] \geq tr(U^* P^t) + 2\eta^{t+1} \lambda_k \Delta^t (1 - \Delta^t) - (\eta^{t+1})^2 C^{t+1} \lambda_1 \Delta^t \,,$$

Rearranging the inequality above and adding k to both sides,

$$\begin{aligned}\mathbb{E}\left[\Delta^{t+1}|\mathcal{F}_t\right] &\leq \Delta^t - 2\eta^{t+1}\lambda_k\Delta^t(1-\Delta^t) + (\eta^{t+1})^2C^{t+1}\lambda_1\Delta^t \\ &= \Delta^t\left(1 - 2\eta^{t+1}\lambda_k(1-\Delta^t) + (\eta^{t+1})^2C^{t+1}\lambda_1\right),\end{aligned}$$

Multiplying both sides of the inequality above by $\mathbb{1}_{\mathcal{G}_t}$, we get

$$\mathbb{E}\left[\Delta^{t+1}|\mathcal{F}_t\right]\mathbb{1}_{\mathcal{G}_t} \leq \Delta^t\left(1 - 2\eta^{t+1}\lambda_k(1-\Delta^t) + (\eta^{t+1})^2C^{t+1}\lambda_1\right)\mathbb{1}_{\mathcal{G}_t},$$

Since $\mathcal{G}_t$ is $\mathcal{F}_t$-measurable, we have

$$\mathbb{E}\left[\Delta^{t+1}\,\mathbb{1}_{\mathcal{G}_t}\,|\mathcal{F}_t\right] = \mathbb{E}\left[\Delta^{t+1}|\mathcal{F}_t\right]\mathbb{1}_{\mathcal{G}_t},$$

When $\mathbb{1}_{\mathcal{G}_t} = 1$, we have $1 - \Delta^t \geq \tau$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\Delta^{t+1}\,\mathbb{1}_{\mathcal{G}_t}\,|\mathcal{F}_t\right] &\leq \Delta^t\left(1 - 2\eta^{t+1}\lambda_k\tau + (\eta^{t+1})^2C^{t+1}\lambda_1\right)\mathbb{1}_{\mathcal{G}_t} \\ &\leq \Delta^t\left(1 - 2\eta^{t+1}\lambda_k\tau + (\eta^{t+1})^2C^{t+1}\lambda_1\right)\mathbb{1}_{\mathcal{G}_{t-1}}\end{aligned}$$

where the last inequality holds since $\mathcal{G}_t \subset \mathcal{G}_{t-1}$. Taking expectation over both sides, we get the following recursion relation:

$$\mathbb{E}\left[\Delta^{t+1}\,\mathbb{1}_{\mathcal{G}_t}\right] \leq \mathbb{E}\left[\Delta^t\,\mathbb{1}_{\mathcal{G}_{t-1}}\right]\left(1 - 2\eta^{t+1}\lambda_k\tau + (\eta^{t+1})^2C^{t+1}\lambda_1\right)$$

We further bound $1 - 2\eta^{t+1}\tau\lambda_k + (\eta^{t+1})^2C^{t+1}\lambda_1$. First, note that since we require $\eta^{t+1} \leq \frac{\lambda_k\tau}{\lambda_1 b(k+3)}$, we get

$$\eta^{t+1}b \leq \frac{\lambda_k\tau}{\lambda_1(k+3)} \leq \frac{\tau}{(k+3)} \leq \frac{1}{k+3} \leq \frac{1}{4}.$$

and

$$C^{t+1} = b(k + 2\eta^{t+1}b + (\eta^{t+1})^2b^2) \leq b(k+1).$$

Thus, we get

$$1 - 2\eta^{t+1}\tau\lambda_k + (\eta^{t+1})^2C^{t+1}\lambda_1 \leq 1 - 2\eta^{t+1}\tau\lambda_k + (\eta^{t+1})^2b(k+1)\lambda_1$$

Since our requirement of η^{t+1} also implies that

$$\eta^{t+1} \leq \frac{2\lambda_k\tau}{b(k+1)\lambda_1},$$

it guarantees that

$$0 < 1 - 2\eta^{t+1}\tau\lambda_k + (\eta^{t+1})^2b(k+1)\lambda_1 . < 1$$

For any t, define $\alpha^t := 2\eta^t\tau\lambda_k - (\eta^t)^2b(k+1)\lambda_1$, we have

$$\mathbb{E}\left[\Delta^{t+1}\,\mathbb{1}_{\mathcal{G}_t}\right] \leq \mathbb{E}\left[\Delta^t\,\mathbb{1}_{\mathcal{G}_{t-1}}\right]\left(1 - \alpha^{t+1}\right),$$

Recursively applying this relation, we get

$$\mathbb{E}\left[\Delta^{t+1}\,\mathbb{1}_{\mathcal{G}_t}\right] \leq \Pi_{i=2}^{t+1}(1-\alpha^i)\,\mathbb{E}\left[\Delta^1\,\mathbb{1}_{\mathcal{G}_0}\right]$$

Also note that

$$\Delta^1\,\mathbb{1}_{\mathcal{G}_0} \leq (1-\alpha^1)\Delta^0\,,$$

Therefore,

$$\mathbb{E}\left[\Delta^{t+1}\,\mathbb{1}_{\mathcal{G}_t}\right] \leq \Pi_{i=1}^{t+1}(1-\alpha^i)\Delta^0$$

Since for any $x \in (0,1)$, it holds that $\ln(1-x) \leq -x$; we get

$$\Pi_{i=1}^{t}\big(1-\alpha^i\big) \leq \mathbb{E}\left(-\sum_{i=1}^{t}\alpha^i\right)$$

Plugging in the value of α^i's, we get

$$\mathbb{E}\left[\Delta^{t}\,\mathbb{1}_{\mathcal{G}_{t-1}}\right] \leq \mathbb{E}\left(-\sum_{i=1}^{t}\left(2\eta^i\tau\lambda_k - (\eta^i)^2 b(k+1)\lambda_1\right)\right)$$

Again, by our requirement on learning rate, we have for any t

$$\eta^t \leq \frac{\lambda_k\tau}{\lambda_1 b(k+3)} \leq \frac{\lambda_k\tau}{\lambda_1 b(k+1)}$$

Thus,

$$2\eta^i\tau\lambda_k - (\eta^i)^2 b(k+1)\lambda_1 \geq \eta^i\tau\lambda_k > 0$$

Since we choose a constant learning rate η, this implies that

$$\mathbb{E}\left[\Delta^{t}\,\mathbb{1}_{\mathcal{G}_{t-1}}\right] \leq \mathbb{E}\left(-\sum_{i=1}^{t}\eta\tau\lambda_k\right) = \mathbb{E}\left(-t\eta\tau\lambda_k\right)$$

Finally, since $\mathbb{1}_{\mathcal{G}_t} \leq \mathbb{1}_{\mathcal{G}_{t-1}}$, we get

$$\mathbb{E}\left[\Delta^{t}\,\mathbb{1}_{\mathcal{G}_t}\right] \leq \mathbb{E}\left[\Delta^{t}\,\mathbb{1}_{\mathcal{G}_{t-1}}\right] \leq \mathbb{E}\left(-t\eta\tau\lambda_k\right)$$

Combining this with the definition of conditional expectation, we get

$$\mathbb{E}\left[\Delta^t|\mathcal{G}_t\right] := \frac{\mathbb{E}\left[\Delta^{t}\,\mathbb{1}_{\mathcal{G}_t}\right]}{\mathbb{P}\left(\mathcal{G}_t\right)} \leq \frac{\mathbb{E}\left[\Delta^{t}\,\mathbb{1}_{\mathcal{G}_t}\right]}{1-\delta} \leq \frac{1}{1-\delta}\mathbb{E}\left(-t\eta\tau\lambda_k\right)$$

where the first inequality is by our upper bound on the probability of bad event $\mathcal{G}_t^c$. □

D Canonical (principal) angles between subspaces

:=[[Vu and Lei(2013)]] Let $\mathcal{E}$ and $\mathcal{F}$ be d-dimensional subspaces of $\mathbb{R}^p$ with orthogonal projectors E and F. Denote the singular values of $EF^{\perp}$ by $s_1 \geq s_2 \cdots \geq$. The canonical angles between $\mathcal{E}$ and $\mathcal{F}$ are the numbers

$$\theta_k(\mathcal{E},\mathcal{F}) = \arcsin(s_k)$$

for $k = 1, \ldots, d$ and the angle operator between $\mathcal{E}$ and $\mathcal{F}$ is the $d \times d$ matrix

$$\Theta(\mathcal{E}, \mathcal{F}) = diag(\theta_1, \ldots, \theta_d)\,.$$

subject to

$$\|x\| = \|y\| = 1, x^H x_i = 0, y^H y_i = 0, i = 1, \ldots, k-1.$$

The vectors $\{x_1, \ldots, x_m\}$ and $\{y_1, \ldots, y_m\}$ are called the principal vectors.

Proposition 3. *Let $\mathcal{E}$ and $\mathcal{F}$ be d-dimensional subspaces of $\mathbb{R}^p$ with orthogonal projectors E and F. Then The singular values of $EF^{\perp}$ are*

$$s_1, s_2, \ldots, s_d, 0, \ldots, 0.$$

And

$$\|\sin\Theta(\mathcal{E}, \mathcal{F})\|_F^2 = \|EF^{\perp}\|_F^2\,.$$