Black-Box k-to-1-PCA Reductions: Theory and Applications

Arun Jambulapati   JMBLPATI@GMAIL.COM
University of Michigan

Syamantak Kumar   SYAMANTAK@UTEXAS.EDU
University of Texas at Austin

Jerry Li   JERRL@MICROSOFT.COM
Microsoft Research

Shourya Pandey   SHOURYAP@UTEXAS.EDU
University of Texas at Austin

Ankit Pensia   ANKITP@IBM.COM
IBM Research

Kevin Tian   KJTIAN@CS.UTEXAS.EDU
University of Texas at Austin

Editors: Shipra Agrawal and Aaron Roth

Abstract

The k-principal component analysis (k-PCA) problem is a fundamental algorithmic primitive that is widely-used in data analysis and dimensionality reduction applications. In statistical settings, the goal of k-PCA is to identify a top eigenspace of the covariance matrix of a distribution, which we only have black-box access to via samples. Motivated by these settings, we analyze black-box deflation methods as a framework for designing k-PCA algorithms, where we model access to the unknown target matrix via a black-box 1-PCA oracle which returns an approximate top eigenvector, under two popular notions of approximation. Despite being arguably the most natural reduction-based approach to k-PCA algorithm design, such black-box methods, which recursively call a 1-PCA oracle k times, were previously poorly-understood.

Our main contribution is significantly sharper bounds on the approximation parameter degradation of deflation methods for k-PCA. For a quadratic form notion of approximation we term *ePCA (energy PCA)*, we show deflation methods suffer no parameter loss. For an alternative well-studied approximation notion we term *cPCA (correlation PCA)*, we tightly characterize the parameter regimes where deflation methods are feasible. Moreover, we show that in all feasible regimes, k-cPCA deflation algorithms suffer no asymptotic parameter loss for any constant k. We apply our framework to obtain state-of-the-art k-PCA algorithms robust to dataset contamination, improving prior work in sample complexity by a poly(k) factor.

Keywords: Principal component analysis, deflation, robust statistics, heavy-tailed statistics

1. Introduction

Principal component analysis (PCA) stands as a ubiquitous technique in statistical analysis and dimensionality reduction (see e.g. Pearson (1901) and the popular reference Jolliffe and Cadima (2016)), offering a powerful and general-purpose means to extract information from complex datasets. In one of its most common use cases, applied to a distribution $\mathcal{D}$, the k-PCA problem (Definition 1) seeks to identify k orthogonal *principal components* (PCs) which capture the most variation in $\mathcal{D}$. Succinctly put, the goal is to output the top-k eigenvectors of the covariance $\boldsymbol{\Sigma}$ of $\mathcal{D}$,[1] or an approximation thereof, where the key challenge is that we only have black-box sample access to $\boldsymbol{\Sigma}$.

This statistical k-PCA problem is extremely well-studied, and due to its fundamental nature, many recent works have explored the development of k-PCA algorithms under constraints. These variations include performing PCA in small space Jain et al. (2016), under adversarial corruptions Jambulapati et al. (2020); Diakonikolas et al. (2023), when the distribution is heavy-tailed Minsker (2015); Wei and Minsker (2017), or when the sampling process admits correlations Neeman et al. (2023); Kumar and Sarkar (2023). In some problem settings, additional constraints might be imposed on the data distribution (e.g. sparsity Zou et al. (2006)) or the output of the PCA algorithm (e.g. differential privacy Liu et al. (2022b), fairness Lee et al. (2023)). In all these scenarios, asking for an exact k-PCA is impossible due to our black-box access to the covariance matrix $\boldsymbol{\Sigma}$ in question, so the goal is to output an approximate k-PCA, for appropriate notions of approximation.

Unfortunately, in many of the aforementioned settings, such as Jambulapati et al. (2020); Diakonikolas et al. (2023); Kumar and Sarkar (2023); Liu et al. (2022b), results are only currently known for extracting a single approximate PC, rather than the more general approximate k-PCA problem. However, for most practical applications of PCA such as dimensionality reduction, extracting a single PC is not sufficient and hence these algorithms do not readily apply. Given the sophisticated techniques required to prove correctness of even extracting a single PC in these constrained statistical settings, it is often not immediately clear how to extend the analysis of these works to design a corresponding k-PCA algorithm.

One of the most commonly-proposed solutions to alleviate the burden of k-PCA algorithm design, assuming existence of a corresponding 1-PCA algorithm, is a reduction-based approach, known as a *deflation method* Mackey (2008); Allen-Zhu and Li (2016) (see Algorithm 1). In this framework, the extraction of a single approximate PC is treated as a subroutine, and the deflation method repeatedly projects out the results of calls to this subroutine from the dataset, k times in total. It is straightforward to see that if the subroutine returned an exact top eigenvector each time, this strategy would indeed succeed in outputting an exact k-PCA of the target matrix. As mentioned previously, under black-box access, the key challenge is to quantify the approximation degradation of this recursive procedure.

Indeed, despite their being perhaps the most natural reduction-based k-PCA strategy, there has been surprisingly-limited work rigorously analyzing the performance of deflation methods. For example, standard analyses require strong gap assumptions on the target matrix (e.g. each of the top-k eigenvalues being separated) Li and Zhang (2015), and this condition has even been quoted in a variety of works as being necessary Li and Zhang (2015); Musco and Musco (2015); Shamir

---

1. The standard definition of an exact k-PCA allows for arbitrary tiebreaking if the dimensionality of the top-k eigenspace is $> k$, see Definition 1 for a formal statement.

(2016); Liu et al. (2022a). The challenge of obtaining provable guarantees for deflation methods (or small-block variants thereof) was an open problem stated in Musco and Musco (2015). Our primary contribution is a direct analysis of the approximation parameter degradation of deflation methods, with no gap assumptions on the target matrix. This refutes the aforementioned conventional wisdom that deflation methods fail without strong spectral assumptions.

Our work is motivated by a closely-related result of Allen-Zhu and Li (2016), which primarily focused on applications in the *white-box* setting where the target matrix is explicitly available.[2] However, the Allen-Zhu and Li (2016) analysis loses large polynomial factors in the approximation quality parameters, under their notion of approximation (Definition 5), which we will shortly compare to our results. Our focus in this work, in comparison, is to give tighter characterizations of when it is possible to obtain lossless or nearly-lossless reductions in the black-box PCA setting, typically of interest in statistical applications.

1.1. Our results

We model black-box access to a target matrix $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$ through the template in Algorithm 1, a black-box k-to-1-PCA reduction.[3] Specifically, the algorithm only interacts with $\mathbf{M}$ through an oracle $\mathcal{O}_{1\text{PCA}}$, which we assume returns an approximate top eigenvector of $\mathbf{PMP}$ for a specified projection matrix $\mathbf{P}$. This also explains why Algorithm 1 is often called a deflation method, as it repeatedly "deflates" directions from consideration as specified by recursively calling $\mathcal{O}_{1\text{PCA}}$.

Algorithm 1: BlackBoxPCA($\mathbf{M}, k, \mathcal{O}_{1\text{PCA}}$)

Input: $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$, $k \in [d]$, $\mathcal{O}_{1\text{PCA}}$, an algorithm which takes as input $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$ and a $d \times d$ orthogonal projection matrix $\mathbf{P}$ and returns a unit vector in $\mathbb{R}^d$ in $\text{span}(\mathbf{P})$; $\mathbf{P}_0 \leftarrow \mathbf{I}_d$

for $i \in [k]$ **do**

    $u_i \leftarrow \mathcal{O}_{1\text{PCA}}(\mathbf{M}, \mathbf{P}_{i-1})$

    $\mathbf{P}_i \leftarrow \mathbf{P}_{i-1} - u_i u_i^\top$

end

Return: $\mathbf{U} \leftarrow \{u_i\}_{i \in [k]} \in \mathbb{R}^{d \times k}$

To motivate our approximation-tolerant analysis, first consider the performance of Algorithm 1 when $\mathcal{O}_{1\text{PCA}}$ returns an exact 1-PCA of $\mathbf{PMP}$ on inputs $(\mathbf{M}, \mathbf{P})$, using the following definition.

Definition 1 (Exact PCA) *For $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$, the exact k-principal component analysis (k-PCA) problem asks to return orthonormal $\mathbf{V} \in \mathbb{R}^{d \times k}$ such that $\langle \mathbf{V}\mathbf{V}^\top, \mathbf{M} \rangle$ is maximal.*

If $\mathcal{O}_{1\text{PCA}}$ is assumed to be an exact 1-PCA algorithm, then it is straightforward to show that Algorithm 1 is an exact k-PCA algorithm (indeed, this follows as a special case of our Theorem 4).

In Sections 3 and 4, our focus is analyzing the behavior of Algorithm 1 when $\mathcal{O}_{1\text{PCA}}$ only has approximate 1-PCA guarantees. We consider two types of approximation which are well-studied in the literature, that we call *energy PCA* (or ePCA, see Definition 2) and *correlation PCA* (or cPCA, see Definition 5). We chose these definitions of approximate PCA because of their prevalence as

---

2. For example, in this white-box setting, one can query matrix-vector products with the target matrix, which is not a realistic assumption if the target matrix is the covariance of a distribution we have sample access to.

3. For notation used throughout the paper, see Section 2; $\mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$ is the set of positive semidefinite $d \times d$ matrices.

error metrics in the PCA literature; to our knowledge, essentially all works on PCA in the statistical setting (i.e., under sample access from a distribution) give guarantees under one of these metrics, including the aforementioned works Minsker (2015); Jain et al. (2016); Wei and Minsker (2017); Jambulapati et al. (2020); Liu et al. (2022b); Kumar and Sarkar (2023); Diakonikolas et al. (2023); Lee et al. (2023). Correspondingly, our main results concern Algorithm 1 when $\mathcal{O}_{1\text{PCA}}$ is assumed to be an approximate 1-ePCA oracle or 1-cPCA oracle, notions formalized in Definitions 3 and 6, and we want Algorithm 1 to respectively return a k-ePCA or a k-cPCA of $\mathbf{M}$.

ePCA. The first approximation notion we consider is *energy k-PCA*, which defines the performance of an approximate PCA algorithm by a sum of quadratic forms over components. Intuitively, this definition quantifies the amount of variance of the data captured by the returned subspace.

Definition 2 (Energy k-PCA) *Let $k \in [d]$, $\mathbf{M} \in \mathbb{S}_{\succeq 0}^{d \times d}$, and $\epsilon \in [0, 1]$. We say orthonormal $\mathbf{U} \in \mathbb{R}^{d \times k}$ is an ϵ-approximate energy k-PCA (or, ϵ-k-ePCA) of $\mathbf{M}$ if*

$$\left\langle \mathbf{U}\mathbf{U}^\top, \mathbf{M} \right\rangle \geq (1 - \epsilon) \|\mathbf{M}\|_k, \text{ where } \|\mathbf{M}\|_k := \max_{\text{orthonormal } \mathbf{V} \in \mathbb{R}^{d \times k}} \left\langle \mathbf{V}\mathbf{V}^\top, \mathbf{M} \right\rangle.$$

Our guarantees for Algorithm 1 as a black-box ePCA reduction are stated in the following result.

Definition 3 (ePCA oracle) *We say $\mathcal{O}_{\text{ePCA}}$ is an ϵ-approximate 1-ePCA oracle (or, ϵ-1-ePCA oracle) if, on inputs $\mathbf{M} \in \mathbb{S}_{\succeq 0}^{d \times d}$, and $\mathbf{P} \in \mathbb{R}^{d \times d}$, where $\mathbf{P}$ is required to be an orthogonal projection matrix, $\mathcal{O}_{\text{ePCA}}$ returns $u \in \mathbb{R}^d$, an ϵ-1-ePCA of $\mathbf{PMP}$, satisfying $u \in \text{span}(\mathbf{P})$.*

Theorem 4 (k-to-1-ePCA reduction) *Let $\epsilon \in (0, 1)$, let $\mathbf{M} \in \mathbb{S}_{\succeq 0}^{d \times d}$, and let $\mathcal{O}_{1\text{PCA}}$ be an ϵ-1-ePCA oracle (Definition 3). Then, Algorithm 1, when run on $\mathbf{M}$, returns $\mathbf{U} \in \mathbb{R}^{d \times k}$, an ϵ-k-ePCA of $\mathbf{M}$.*

Perhaps surprisingly, Theorem 4 states there is no loss in approximation parameters via the reduction in Algorithm 1, if our approximation notion is ePCA. This is optimal; for example, if $d \geq 2k$, and $\mathbf{M} = \mathbf{diag}(\{\mathbb{1}_k, \mathbb{0}_{d-k}\})$, then leting $\mathcal{O}_{\text{ePCA}}$ repeatedly return $\sqrt{1 - \epsilon}e_i + \sqrt{\epsilon}e_{k+i}$ in each iteration $i \in [k]$, the result is an ϵ-k-cPCA of $\mathbf{M}$. Our proof of Theorem 4 in Section 3 is a simple application of Cauchy's interlacing theorem, but to our knowledge it was not previously known.

cPCA. The other notion of approximation we consider is *correlation k-PCA*, a popular definition in the literature Shamir (2016); Allen-Zhu and Li (2017) which defines the performance of an approximate PCA algorithm in terms of the correlation with the small eigenspace of $\mathbf{M}$, allowing for a gap in the definition of "small." Let $\mathbf{V}^{<\lambda}(\mathbf{M})$ denote the orthonormal matrix spanning the eigenspace of $\mathbf{M}$ corresponding to eigenvalues $< \lambda$. We define cPCA as follows, parameterized by a gap Γ and a correlation Δ.

Definition 5 (Correlation k-PCA) *Let $k \in [d]$, $\mathbf{M} \in \mathbb{S}_{\succeq 0}^{d \times d}$, $\Gamma \in [0, 1]$, and $\Delta \in [0, k]$. We say orthonormal $\mathbf{U} \in \mathbb{R}^{d \times k}$ is a (Δ, Γ)-approximate correlation k-PCA (or, (Δ, Γ)-k-cPCA) of $\mathbf{M}$ if*

$$\left\| \left(\mathbf{V}^{<(1-\Gamma)\lambda_k(\mathbf{M})}(\mathbf{M}) \right)^\top \mathbf{U} \right\|_{\mathrm{F}}^2 \leq \Delta.$$

Definition 5 requires that $\mathbf{U}$ be almost entirely-contained in an eigenspace of $\mathbf{M}$ corresponding to eigenvalues which are, at worst, barely outside the top-k space. For example, in the gapped setting

when $\lambda_2(\mathbf{M}) < (1 - \Gamma)\lambda_1(\mathbf{M})$ and $k = 1$ (so $\mathbf{V}^{<(1-\Gamma)\lambda_k(\mathbf{M})}$ is simply the subspace orthogonal to the top PC of $\mathbf{M}$), Definition 5 implies recovery of the top PC in the $\sin^2$ error metric. In general, cPCA guarantees are more geometrically explicit than ePCA counterparts (as they bound correlation with specific eigenspaces), and hence may be preferable when they are available.[4]

The only analysis of Algorithm 1 we are aware of is due to Allen-Zhu and Li (2016), who gave a black-box k-to-1-cPCA reduction under the following definition.

Definition 6 (cPCA oracle) *We say $\mathcal{O}_{\mathrm{cPCA}}$ is a (δ, γ)-approximate[5] 1-cPCA oracle (or, (δ, γ)-1-cPCA oracle) if, on inputs $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$, and $\mathbf{P} \in \mathbb{R}^{d \times d}$, where $\mathbf{P}$ is required to be an orthogonal projection matrix, $\mathcal{O}_{\mathrm{cPCA}}$ returns $u \in \mathbb{R}^d$, a (δ, γ)-1-cPCA of $\mathbf{PMP}$, satisfying $u \in \mathrm{span}(\mathbf{P})$.*

Under Definitions 5, 6, the main result Theorem 4.1(a) of Allen-Zhu and Li (2016) can be rephrased as follows: Algorithm 1 returns a (Δ, Γ)-k-cPCA if $\mathcal{O}_{\mathrm{1PCA}}$ is a (δ, γ)-1-cPCA oracle, for

$$\gamma = \frac{\Gamma}{2}, \quad \delta = \Theta\left(\frac{\Delta^2\Gamma^2}{k^4\kappa_k(\mathbf{M})^2}\right), \quad \text{where } \kappa_k(\mathbf{M}) := \frac{\lambda_1(\mathbf{M})}{\lambda_k(\mathbf{M})}. \tag{1}$$

In the white-box setting, (δ, γ)-1-cPCA runtimes typically scale polynomially in $\frac{1}{\gamma}$ and logarithmically in $\frac{1}{\delta}$ (e.g., Theorem 1, Musco and Musco (2015)), so Allen-Zhu and Li (2016) focused on maintaining a low overhead in γ. However, known sample complexity lower bounds in the black-box statistical setting (e.g., Theorem 6, Allen-Zhu and Li (2017)) show $\gtrsim \frac{1}{\gamma^2\delta}$ samples are required to obtain a (δ, γ)-1-cPCA, even under strong tail bounds such as sub-Gaussianity. Hence, the Allen-Zhu and Li (2016) reduction in the black-box setting uses

$$\approx \frac{k^4\kappa_k(\mathbf{M})^2}{\Delta^2\Gamma^4}, \tag{2}$$

samples for obtaining a (Δ, Γ)-k-cPCA via Algorithm 1, times problem-dependent factors such as the dimension, which is a $\mathrm{poly}(k, \Delta, \Gamma)$ factor worse than the lower bound. Our goal is to understand whether this loss is inherent for black-box cPCA reductions.[6] As a starting point, in Section 2, we prove the following simple results transferring guarantees between Definitions 2 and 5.

Lemma 7 *If $\mathbf{U} \in \mathbb{R}^{d \times k}$ is an ϵ-k-ePCA of $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$, it is a $(\frac{\epsilon\|\mathbf{M}\|_k}{\Gamma\lambda_k(\mathbf{M})}, \Gamma)$-$k$-cPCA of $\mathbf{M}$ for any $\Gamma \in (0, 1)$.*

Lemma 8 *If $u \in \mathbb{R}^d$ is a (Δ, Γ)-1-cPCA of $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$, it is a $(\Gamma + \Delta)$-1-ePCA of $\mathbf{M}$.*

By combining Lemmas 7 and 8 with our black-box ePCA reduction in Theorem 4, it is straightforward to show Algorithm 1 returns a (Δ, Γ)-k-cPCA if $\mathcal{O}_{\mathrm{1PCA}}$ is a (δ, γ)-1-cPCA oracle, for

$$\delta = \gamma = \Theta\left(\frac{\Delta\Gamma}{k\kappa_k(\mathbf{M})}\right). \tag{3}$$

---

4. However, in some black-box settings e.g. robust PCA (Section C.2), certificates on specific eigenspaces are harder to achieve, and only weaker certificates such as quadratic forms have found algorithmic use (giving ePCA guarantees).

5. We sometimes use lowercase Greek letters to denote parameters for the 1-cPCA oracle, to differentiate from capital Greek letters used to parameterize overall k-PCA guarantees.

6. Note that Theorem 4 shows no such loss is necessary, if our approximation metric is instead ePCA.

Our transfer lemmas and Theorem 4 therefore already yield a sample complexity scaling as

$$\approx \frac{k^3 \kappa_k(\mathbf{M})^3}{\Delta^3 \Gamma^3} \tag{4}$$

in standard statistical settings, which is incomparable to (2) in general but already improves upon it in, e.g., the dependences on k and Γ. However, both (2) and (4) yield polynomial overheads in various approximation factors. We hence ask the natural question: can we characterize when Algorithm 1 serves as an asymptotically *lossless* cPCA reduction, even when k is a constant?

In Section 4, we give our main result analyzing the use of Algorithm 1 as a black-box k-to-1-cPCA reduction. We show that if $\Delta \cdot \kappa_k(\mathbf{M})^2 \leq \Gamma^2$ for a target matrix $\mathbf{M}$, then Algorithm 1 indeed is an asymptotically lossless reduction for $k = \Theta(1)$. More precisely, we prove the following result.

Theorem 9 (k-to-1-cPCA reduction) *Let $(\Delta, \Gamma) \in (0,1)$ and $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$ satisfy $\Delta \cdot \kappa_k(\mathbf{M})^2 \leq \Gamma^2$, and let $\mathcal{O}_{1\text{PCA}}$ be a (δ, γ)-1-cPCA oracle (Definition 6), where $\delta := \frac{1}{k^{\Theta(\log k)}} \cdot \Delta$, $\gamma := \frac{1}{\Theta(k^3)} \cdot \Gamma$, for appropriate constants. Then, Algorithm 1 returns $\mathbf{U} \in \mathbb{R}^{d \times k}$, a (Δ, Γ)-k-cPCA of $\mathbf{M}$.*

That is, when $\Delta \cdot \kappa_k(\mathbf{M})^2 \leq \Gamma^2$, it suffices to take $\mathcal{O}_{1\text{PCA}}$ to be a (δ, γ)-1-cPCA oracle, for $\delta = \Theta_k(\Delta)$, $\gamma = \Theta_k(\Gamma)$. We complement Theorem 4 with an impossibility result (see Proposition 32), showing that if $\Delta \cdot \kappa_k(\mathbf{M})^2 > \Gamma^2$, Algorithm 1 fails to have this lossless reduction property, even when $k = 2$ and $d = 3$. In conclusion, Proposition 32 and Theorem 9 completely classify regimes where Algorithm 1 acts as a lossless cPCA reduction, up to the dependence on k.

An immediate question in light of Theorem 9 is to obtain a tighter characterization of the k dependence. We view our quasipolynomial $k^{\Theta(\log k)}$ factor overhead in the parameter degradation, which is e.g. $d^{o(1)}$ for $k = \Theta(\log d)$, to be a proof-of-concept of Algorithm 1's utility as a k-cPCA algorithm design template; we conjecture it can be improved to be $\text{poly}(k)$, which we leave open.

Applications. To showcase our reductions, in Section C we directly apply Theorem 4 to obtain new algorithms for k-ePCA in the statistical setting, which are robust to adversarial corruptions to the dataset. Our first such result extends a recent 1-PCA algorithm of Diakonikolas et al. (2023).

Theorem 10 (Robust sub-Gaussian k-ePCA) *Let $\mathcal{D}$ be an $O(1)$-sub-Gaussian distribution on $\mathbb{R}^d$ with covariance $\mathbf{\Sigma}$. Let $\epsilon \in (0, \epsilon_0)$ for an absolute constant ϵ_0, and $\delta \in (0,1)$. Let T be an ϵ-corrupted set of samples from $\mathcal{D}$ with $|T| = \Theta(\frac{d + \log(1/\delta)}{\epsilon^2 \log^2(1/\epsilon)})$ for an appropriate constant. Algorithm $\mathcal{A}_k$ (Corollary 43) run on inputs T, ϵ, $\gamma = \Theta(\epsilon \log(\frac{1}{\epsilon}))$, δ, and $k \in [d]$ outputs orthonormal $\mathbf{U} \in \mathbb{R}^{d \times k}$ such that, with probability $\geq 1 - \delta$, $\mathbf{U}$ is an $O(\epsilon \log(\frac{1}{\epsilon}))$-$k$-ePCA of $\mathbf{\Sigma}$, in time $O(\frac{ndk}{\epsilon^2} \text{polylog}(\frac{d}{\epsilon \delta}))$.*

We then consider a heavy-tailed variant of the robust statistical estimation problem in Theorem 10; rather than requiring our distribution to be sub-Gaussian, we ask that it is hypercontractive (Definition 35), the de facto finite moment bound assumption used in the literature (see e.g. Mendelson and Zhivotovskiy (2018)). Under this assumption, we design a new robust 1-ePCA algorithm following a *stability* framework of Diakonikolas et al. (2023) (see Definition 40), and prove the following.

Theorem 11 (Robust hypercontractive k-ePCA) *For an even integer $p \geq 4$, let $\mathcal{D}$ be (p, C_p)-hypercontractive on $\mathbb{R}^d$ with mean $\mathbb{0}_d$ and covariance $\mathbf{\Sigma}$. Let $\epsilon \in (0, \epsilon_0)$, $\delta \in (0,1)$, and $\gamma = \Theta(C_p^2 \epsilon^{1 - \frac{2}{p}})$ such that $\gamma \in (0, \gamma_0)$ for absolute constants ϵ_0, γ_0. Let T be an ϵ-corrupted set of*

samples from $\mathcal{D}$ with $|T| = \Theta(\beta(\frac{d \log d + \log(1/\delta)}{\gamma^2}))$ for an appropriate constant, where $\beta := C_p^6 \epsilon^{-\frac{2}{p}}$. Algorithm $\mathcal{A}_k$ (Corollary 43) run on inputs T, ϵ, γ, δ, and $k \in [d]$ outputs orthonormal $\mathbf{U} \in \mathbb{R}^{d \times k}$ such that, with probability $\geq 1 - \delta$, $\mathbf{U}$ is an $O(\gamma)$-k-ePCA of $\mathbf{\Sigma}$, in time $O(\frac{ndk}{\gamma^2}\text{polylog}(\frac{d}{\epsilon\delta}))$.

We show our approximation factor γ in Theorem 11 is tight up to constants in Lemma 48. Moreover, we mention that even for uncorrupted data and $k = 1$, we are unaware of any other polynomial-time PCA algorithm with a similar additive dependence on $\log(\frac{1}{\delta})$ in the sample complexity under a heavy-tailed assumption (rather than sub-Gaussianity), improving on Minsker (2015).

Interestingly, both Theorems 10 and 11 incur no overhead in k in their sample complexities, since our stability-based approach composes with deflation (see Lemma 42). The only other robust k-PCA algorithm we are aware of (under the well-studied definition of robustness in Definition 39) is due to Proposition 2.6, Kong et al. (2020), which applies to distributions satisfying an assumption which is a variant of $(4, C_4)$-hypercontractivity, but is not directly comparable. We give a comparison between our two results in the case $p = 4$ at the end of Section C.2, where we show that under the more standard assumption of hypercontractivity, our recovery rate matches the rate obtained by Kong et al. (2020), while our sample complexity improves by a factor of up to k^2.

Finally, to show how to apply Theorem 9 to statistical settings, we design an online k-cPCA algorithm for heavy-tailed distributions in Theorem 52 by using Oja's algorithm Oja (1982). While Theorem 52 does not match the state-of-the-art due to Allen-Zhu and Li (2017), it follows a more black-box approach rather than requiring a custom sophisticated analysis. As new 1-cPCA algorithms emerge in the literature, e.g. under the aforementioned input or output constraints, we are optimistic the roadmap in Theorem 52 will serve as a useful template for solving the corresponding k-cPCA problems.

1.2. Our techniques

We now give a high-level overview of our proofs for Theorems 4 and 9. Letting $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$, the centerpiece of both our proofs is analyzing the approximation parameter degradation of composing $\mathbf{U}_1 \in \mathbb{R}^{d \times k_1}$, an approximate k_1-PCA of $\mathbf{M}$, and $\mathbf{U}_2 \in \mathbb{R}^{d \times k_2}$, an approximate k_2-PCA of $\widetilde{\mathbf{M}} := (\mathbf{I}_d - \mathbf{U}_1 \mathbf{U}_1^\top)\mathbf{M}(\mathbf{I}_d - \mathbf{U}_1 \mathbf{U}_1^\top)$ with $\mathbf{U}_2 \perp \mathbf{U}_1$, to form an approximate $(k_1 + k_2)$-PCA of $\mathbf{M}$. We describe how we analyze compositions under both types of approximation we consider.

Energy k-PCA. In the case of ePCA, we analyze this composition using Lemma 16, an application of Cauchy's interlacing theorem. Interestingly, we show that the composition of an ϵ-k_1-ePCA and an ϵ-k_2-ePCA in the sense described previously results in an ϵ-$(k_1 + k_2)$-ePCA of $\mathbf{M}$ (Corollary 17), i.e. we show that ePCA composition suffers no approximation parameter degradation. We defer the details of our proof to Section 3, owing to the conciseness of the argument.

Correlation k-PCA. In the case of cPCA, our proof relies on Lemma 19, which bounds the quality of a composition of two cPCAs. To describe this result, we introduce some notation. Let $\mathbf{U}_1$ be a (δ, γ)-k_1-cPCA of $\mathbf{M}$ and $\mathbf{U}_2$ be a (δ, γ)-k_2-cPCA of $\widetilde{\mathbf{M}}$, and let $\mathbf{U} := (\mathbf{U}_1 \quad \mathbf{U}_2)$ and $k := k_1 + k_2$. Moreover, for parameters (Δ, Γ) to be set, let $\mathbf{L} := \mathbf{V}^{\geq (1-\Gamma)\lambda_k(\mathbf{M})}$ be the large eigenspace of $\mathbf{M}$ that (Δ, Γ)-k-cPCAs are required to be correlated with, let $\mathbf{\Lambda}$ be diagonal containing eigenvalues of $\mathbf{M}$ corresponding to $\mathbf{L}$, and let $\widetilde{\mathbf{L}} := \mathbf{V}^{\geq (1-\gamma_2)\lambda_{k_2}(\widetilde{\mathbf{M}})}(\widetilde{\mathbf{M}})$ be the analogous large eigenspace of the

residual matrix $\widetilde{\mathbf{M}}$. Then, Lemma 19 states that $\mathbf{U}$ is a (Δ, Γ)-k-cPCA of $\mathbf{M}$, for

$$\Delta := 3\delta + \frac{4\delta}{\gamma^2} \cdot \frac{\left\| \mathbf{U}_1^\top \mathbf{L}\mathbf{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}} \right\|_{\mathrm{op}}^2}{\lambda_k \left(\mathbf{M} \right)^2}, \ \Gamma := 2\gamma. \tag{5}$$

In other words, as long as we can appropriately control the quantity $\|\mathbf{U}_1^\top \mathbf{L}\mathbf{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}}\|_{\mathrm{op}}$, we can bound the parameter degradation (from (δ, γ) to (Δ, Γ)) of composing two cPCAs. For instance, if this quantity is $O(\gamma\lambda_k(\mathbf{M}))$, then cPCA composition blows up both the δ and γ parameters by constants.

Our proof of (5) applies the cPCA guarantees of $\mathbf{U}_1$ and $\mathbf{U}_2$ in turn, and uses a decomposition inspired by the gap-free Wedin theorem of Allen-Zhu and Li (2016) to bound a cross term arising in the analysis. In light of (5), we focus on bounding $\|\mathbf{U}_1^\top \mathbf{L}\mathbf{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}}\|_{\mathrm{op}}$, i.e. for arbitrary unit-norm $u \in \mathrm{span}(\mathbf{U}_1)$ and $w \in \mathrm{span}(\widetilde{\mathbf{L}})$, our goal is to bound $u^\top \mathbf{L}\mathbf{\Lambda}\mathbf{L}^\top w$, where by definition of our algorithm, $u \perp w$. Note that by our cPCA assumption, u is highly-uncorrelated with $\mathbf{S} := \mathbf{L}_\perp$, the small eigenspace of $\mathbf{M}$.

Our first observation is that when the spectrum of $\mathbf{M}$ is gapped, in that $(1-\gamma)\lambda_{k_1}(\mathbf{M}) > \lambda_{k_1+1}(\mathbf{M})$, the cPCA guarantee of $\mathbf{U}_1$ affords an additional *head guarantee* (Definition 15) by taking complements. Concretely, $\mathbf{L}$ is simply the top-k_1 eigenspace of $\mathbf{M}$ in the gapped setting, and because $\mathbf{U}_1$ captures almost all of the spectral mass of $\mathbf{L}$, w (which is orthogonal to $\mathbf{U}_1$) is highly-uncorrelated with $\mathbf{L}$. In other words, we have shown that in the gapped setting, u and w lie almost entirely in complementary eigenspaces of $\mathbf{M}$. By exploiting this fact along with $u \perp w$, it is straightforward to bound $u^\top \mathbf{L}\mathbf{\Lambda}\mathbf{L}^\top w = O(\gamma\lambda_k(\mathbf{M}))$,[7] as we show in Lemma 24, so that $\Delta = O(\delta)$ in (5).

Our second observation is that whenever the spectrum of $\mathbf{M}$ is not gapped, but $\mathbf{M}$ has a well-conditioned top-k eigenspace, we can nonetheless use our head guarantee strategy to bound (5) with some amount of approximation parameter degradation, shown in Lemma 28. To do this, we show that in this setting, we can artificially induce a head guarantee with a loss of $\approx k$ in the gap parameter by carefully inducting on intermediate cPCA guarantees (see Lemma 29). This lets us conclude that while u, w do not lie in entirely-complementary eigenspaces, the spectrum of $\mathbf{M}$ corresponding to the overlapping eigenspace lies in a very narrow range. By plugging in our bounds into (5), we show that $\Delta = \delta \cdot \mathrm{poly}(k)$ and $\Gamma = 2\gamma$ suffices in the well-conditioned regime.

After establishing complementary results in Lemma 24 (handling the gapped setting) and Lemma 28 (handling the well-conditioned setting), we show how to bucket the top-k spectrum of $\mathbf{M}$ so that one of these two analyses always applies. We then recursively apply our composition results via dyadic merging strategies, leading to Theorem 9, where the quasipolynomial $k^{\Theta(\log(k))}$ loss arises due to the number of merges. We consider the possibility of $\mathrm{poly}(k)$-overhead black-box cPCA reductions to be an interesting open direction suggested by our work, which likely requires new ideas.

1.3. Related work

We are aware of few papers considering deflation-based approaches to k-PCA without strong gap assumptions on the target matrix (see e.g. Li and Zhang (2015) for a gap-based analysis). The most direct comparison is Allen-Zhu and Li (2016), who analyze Algorithm 1 when $\mathcal{O}_{1\mathrm{PCA}}$ is a 1-cPCA

---

7. In this step, we require that (Δ, Γ) do not lie in the infeasible parameter regime ruled out by Proposition 32.

oracle, giving guarantees as in (1). Our Theorem 4 shows the lossiness in the Allen-Zhu and Li (2016) reduction is unnecessary if approximations are measured through Definition 2 (ePCA), and yields a competitive guarantee (3) for cPCA. Our Theorem 9 further removes all overhead in this cPCA reduction for $k = \Theta(1)$, in the valid parameter regimes not ruled out by Proposition 32.

We also draw inspiration from the extensive body of work focusing on computationally efficient and statistically optimal algorithms for PCA, e.g. Musco and Musco (2015); Garber et al. (2016); Shamir (2016) which give runtime bounds in the white-box setting, and e.g. Hardt and Price (2014); Sa et al. (2015); Balcan et al. (2016); Shamir (2016); Jain et al. (2016); Li et al. (2018) which focus on offline and online variants of statistical PCA. In particular, the gap-free notion of approximation in Definition 5 is patterned off analogous definitions in this latter line. Further, several other notions of PCA approximation have been proposed in the white-box setting, e.g. Eqs. (1)-(3) of Musco and Musco (2015); however, all such results we are aware of use matrix-vector products. Our motivation is analyzing Algorithm 1 as a reduction, so we focused on error metrics where black-box 1-PCA guarantees already exist, though it is interesting to prove similar reductions for other metrics.

Additionally, we mention contemporaneous work of Meyer et al. (2024), which gives analyses of variants of k-to-1 reductions for block Krylov methods targeting low-rank approximation, very close in spirit to our work. However, Meyer et al. (2024) again requires access to matrix-vector products for their reductions, which makes their results not directly applicable in our black-box setting.

Finally, iterative deflation has been popular in the context of sparse PCA d'Aspremont et al. (2004); Moghaddam et al. (2006); Mackey (2008); Yuan and Zhang (2013), as designing direct k-PCA methods is challenging in this setting. However, we are not aware of any general theoretical guarantees on this technique.

2. Preliminaries

For $n \in \mathbb{N}$ we let $[n] := \{i \in \mathbb{N} \mid i \le n\}$. We denote the ith canonical basis vector in $\mathbb{R}^d$ by e_i.

We denote matrices in capital boldface letters. For matrices $\mathbf{A}, \mathbf{B}$ with the same number of rows, we let $\begin{pmatrix} \mathbf{A} & \mathbf{B} \end{pmatrix}$ denote the horizontal concatenation of the matrices. We let $\mathbf{I}_d$ be the $d \times d$ identity matrix and $\mathbf{0}_d$ be the all-zeroes matrix. We say $\mathbf{V} \in \mathbb{R}^{d \times r}$ is orthonormal if its columns are orthonormal, i.e. $\mathbf{V}^\top \mathbf{V} = \mathbf{I}_r$; note that $d \ge r$ necessarily in this case. We let $\mathbb{S}^{d \times d}$ be the set of real symmetric $d \times d$ matrices, which we equip with the Loewner partial ordering $\preceq$ and the Frobenius inner product $\langle \mathbf{M}, \mathbf{N} \rangle = \text{Tr}(\mathbf{M}\mathbf{N})$. We let $\mathbb{S}^{d \times d}_{\succeq \mathbf{0}}$ and $\mathbb{S}^{d \times d}_{\succ \mathbf{0}}$ respectively denote the positive semidefinite and positive definite subsets of $\mathbb{S}^{d \times d}$. We say $\mathbf{A} \in \mathbb{R}^{m \times n}$ has singular value decomposition (SVD) $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ if $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ has rank-r, $\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$ is diagonal, and $\mathbf{U} \in \mathbb{R}^{m \times r}$, $\mathbf{V} \in \mathbb{R}^{n \times r}$ are orthonormal. We refer to the ith largest eigenvalue of $\mathbf{M} \in \mathbb{S}^{d \times d}$ by $\lambda_i(\mathbf{M})$, and the ith largest singular value of $\mathbf{M} \in \mathbb{R}^{m \times n}$ by $\sigma_i(\mathbf{M})$. For $\mathbf{M} \in \mathbb{S}^{d \times d}_{\succeq \mathbf{0}}$ and $k \in [d]$ we use $\|\mathbf{M}\|_k := \sum_{i \in [k]} \lambda_i(\mathbf{M})$ to refer to its k-Ky Fan norm, and $\|\mathbf{M}\|_{\text{op}} := \sigma_1(\mathbf{M})$ is the ℓ_2 operator norm of $\mathbf{M} \in \mathbb{R}^{m \times n}$. For $\mathbf{A} \in \mathbb{R}^{m \times n}$, the span of $\mathbf{A}$'s columns, i.e. $\{\mathbf{A}v \mid v \in \mathbb{R}^n\}$, is denoted $\text{span}(\mathbf{A})$. For $\mathbf{M} \in \mathbb{S}^{d \times d}_{\succ \mathbf{0}}$ and $k \in [d]$, we define

$$\kappa_k(\mathbf{M}) := \frac{\lambda_1(\mathbf{M})}{\lambda_k(\mathbf{M})},$$

to be the k-condition number of $\mathbf{M}$, i.e. the condition number of the top-k eigenspace of $\mathbf{M}$, and we abbreviate $\kappa(\mathbf{M}) := \kappa_d(\mathbf{M})$ for the (standard) condition number of $\mathbf{M}$.

For a subspace $V \subseteq \mathbb{R}^d$, we let $\mathbf{P}_V \in \mathbb{R}^{d \times d}$ be the orthogonal projection matrix onto V. We often identify a subspace $V \subseteq \mathbb{R}^d$ with a generating matrix $\mathbf{V} \in \mathbb{R}^{d \times r}$, where $r := \dim(V)$, such that $\mathbf{V}$ has orthonormal columns and $\mathrm{span}(\mathbf{V}) = V$. In this case, note that $\mathbf{P}_V = \mathbf{V}\mathbf{V}^\top$ for any valid $\mathbf{V}$. We write $U \perp V$ for subspaces $U, V \subset \mathbb{R}^d$ if every element of U is orthogonal to every element of V. Similarly, for two orthonormal matrices $\mathbf{U}, \mathbf{V}$ with the same number of rows, we write $\mathbf{U} \perp \mathbf{V}$ if $\mathrm{span}(\mathbf{U}) \perp \mathrm{span}(\mathbf{V})$. For orthonormal $\mathbf{U} \in \mathbb{R}^{d \times r}$, we write $\mathbf{U}_\perp$ to mean any orthonormal $d \times (d - r)$ matrix so that columns of $\mathbf{U}$ and $\mathbf{U}_\perp$ form a basis of $\mathbb{R}^d$, and define $U_\perp$ to be the orthogonal complement of a subspace $U \subseteq \mathbb{R}^d$. For $\mathbf{M} \in \mathbb{S}^{d \times d}$ and $\lambda > 0$, we denote the subspace spanned by eigenvectors of $\mathbf{M}$ corresponding to eigenvalues at least λ by $V^{\geq \lambda}(\mathbf{M})$, and the corresponding orthonormal generating matrix by $\mathbf{V}^{\geq \lambda}(\mathbf{M})$. We similarly define $V^{> \lambda}(\mathbf{M})$, $V^{\leq \lambda}(\mathbf{M})$, etc.

Linear algebra preliminaries. We use the following well-known linear algebra results.

Fact 1 (Cauchy's interlacing theorem) *Let $d \in \mathbb{N}$ and $r \in [d]$. For any $\mathbf{M} \in \mathbb{S}^{d \times d}$ and orthonormal $\mathbf{V} \in \mathbb{R}^{d \times r}$, $\lambda_{d-r+i}(\mathbf{M}) \leq \lambda_i(\mathbf{V}^\top \mathbf{M} \mathbf{V}) \leq \lambda_i(\mathbf{M})$ for all $i \in [r]$.*

Fact 2 (Weyl's perturbation inequality) *Let $\mathbf{M}, \mathbf{D} \in \mathbb{S}^{d \times d}$ have $\|\mathbf{D}\|_{\mathrm{op}} \leq \delta$. Then,*

$$\lambda_j(\mathbf{M}) - \delta \leq \lambda_j(\mathbf{M} + \mathbf{D}) \leq \lambda_j(\mathbf{M}) + \delta, \text{ for all } j \in [d].$$

Fact 3 *For any $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$, $2\|\mathbf{A}\|_\mathrm{F}^2 + 2\|\mathbf{B}\|_\mathrm{F}^2 \geq \|\mathbf{A} + \mathbf{B}\|_\mathrm{F}^2$.*

Fact 4 *For any $\{x_i\}_{i \in [n]} \subset \mathbb{R}_{>0}$ and $\{w_i\}_{i \in [n]} \subset \mathbb{R}_{\geq 0}$ with $\sum_{i \in [n]} w_i = 1$, we have for all $p \geq 1$,*

$$\left(\sum_{i \in [n]} w_i x_i^p \right)^{\frac{1}{p}} \geq \langle w, x \rangle.$$

Fact 5 *For $\mathbf{A}, \mathbf{B} \in \mathbb{S}^{d \times d}$ with $\mathbf{A} \preceq \mathbf{B}$ and $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$, $\mathrm{Tr}(\mathbf{A}\mathbf{M}) \leq \mathrm{Tr}(\mathbf{B}\mathbf{M})$.*

Finally, we prove the following basic observations relating Definitions 2 and 5, stated earlier, which allow us to convert an ePCA guarantee to a cPCA one and vice-versa.

Lemma 12 *If $\mathbf{U} \in \mathbb{R}^{d \times k}$ is an ϵ-k-ePCA of $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$, it is a $(\frac{\epsilon \|\mathbf{M}\|_k}{\Gamma \lambda_k(\mathbf{M})}, \Gamma)$-$k$-cPCA of $\mathbf{M}$ for any $\Gamma \in (0, 1)$.*

Proof Define $\mathbf{V} := \mathbf{V}^{\geq (1-\Gamma)\lambda_k(\mathbf{M})}(\mathbf{M})$, $\mathbf{W} := \mathbf{V}^{< (1-\Gamma)\lambda_k(\mathbf{M})}(\mathbf{M})$, and $\Delta := \|\mathbf{W}^\top \mathbf{U}\|_\mathrm{F}^2$. Note $\mathrm{Tr}(\mathbf{V}^\top \mathbf{U}\mathbf{U}^\top \mathbf{V}) = k - \Delta$, and $\mathbf{V}^\top \mathbf{U}\mathbf{U}^\top \mathbf{V}$ has $\leq k$ nonzero eigenvalues, all in $[0, 1]$. We bound

$$\begin{aligned}
(1 - \epsilon) \|\mathbf{M}\|_k &\leq \left\langle \mathbf{U}\mathbf{U}^\top, \mathbf{M} \right\rangle = \left\langle \mathbf{U}\mathbf{U}^\top, \mathbf{V}\mathbf{V}^\top \mathbf{M}\mathbf{V}\mathbf{V}^\top \right\rangle + \left\langle \mathbf{U}\mathbf{U}^\top, \mathbf{W}\mathbf{W}^\top \mathbf{M}\mathbf{W}\mathbf{W}^\top \right\rangle \\
&\leq \left\langle \mathbf{V}^\top \mathbf{U}\mathbf{U}^\top \mathbf{V}, \mathbf{V}^\top \mathbf{M}\mathbf{V} \right\rangle + \Delta \left\| \mathbf{W}^\top \mathbf{M}\mathbf{W} \right\|_{\mathrm{op}} \\
&\leq \|\mathbf{M}\|_k - \Delta \lambda_k(\mathbf{M}) + \Delta(1 - \Gamma)\lambda_k(\mathbf{M}) = \|\mathbf{M}\|_k - \Delta \Gamma \lambda_k(\mathbf{M}). \quad (6)
\end{aligned}$$

The second line used Fact 5 by noting that $\left\langle \mathbf{U}\mathbf{U}^\top, \mathbf{W}\mathbf{W}^\top \mathbf{M}\mathbf{W}\mathbf{W}^\top \right\rangle = \mathrm{Tr}(\mathbf{W}^\top \mathbf{M}\mathbf{W}\mathbf{W}^\top \mathbf{U}\mathbf{U}^\top \mathbf{W})$ and $\mathrm{Tr}(\mathbf{W}^\top \mathbf{U}\mathbf{U}^\top \mathbf{W}) \leq \Delta$, and the third used the von Neumann trace inequality, which shows that if $\{\lambda_i\}_{i \in [\ell]}$ are the nonincreasing eigenvalues of $\mathbf{V}^\top \mathbf{M}\mathbf{V}$ and $\{\sigma_i\}_{i \in [\ell]}$ are the nonincreasing eigenvalues of $\mathbf{V}^\top \mathbf{U}\mathbf{U}^\top \mathbf{V}$, $\left\langle \mathbf{V}^\top \mathbf{U}\mathbf{U}^\top \mathbf{V}, \mathbf{V}^\top \mathbf{M}\mathbf{V} \right\rangle \leq \sum_{i \in [\ell]} \sigma_i \lambda_i$. Since the σ_i sum to $k - \Delta$ and are

each at most 1, it is clear that the right-hand side above is bounded by $\|\mathbf{M}\|_k - \Delta\lambda_k(\mathbf{M})$ as claimed. The conclusion follows via rearranging (6). ∎

Lemma 13 *If $u \in \mathbb{R}^d$ is a (Δ, Γ)-1-cPCA of $\mathbf{M} \in \mathbb{S}_{\succeq\mathbf{0}}^{d\times d}$, it is a $(\Gamma + \Delta)$-1-ePCA of $\mathbf{M}$.*

Proof Define $\mathbf{V} := \mathbf{V}^{\geq(1-\Gamma)\lambda_1(\mathbf{M})}$ and $\mathbf{W} := \mathbf{V}^{<(1-\Gamma)\lambda_1(\mathbf{M})}$. The conclusion follows from

$$\left\langle uu^\top, \mathbf{M}\right\rangle = \left\langle uu^\top, \mathbf{VV}^\top\mathbf{MVV}^\top\right\rangle + \left\langle uu^\top, \mathbf{WW}^\top\mathbf{MWW}^\top\right\rangle$$
$$\geq \left\langle \mathbf{V}^\top uu^\top\mathbf{V}, \mathbf{V}^\top\mathbf{MV}\right\rangle \geq \left\langle \mathbf{V}^\top uu^\top\mathbf{V}, (1-\Gamma)\lambda_1(\mathbf{M})\mathbf{I}_{\dim(\mathrm{span}(\mathbf{V}))}\right\rangle$$
$$\geq (1-\Delta)(1-\Gamma)\lambda_1(\mathbf{M}) \geq (1-\Delta-\Gamma)\lambda_1(\mathbf{M}),$$

where $\mathbf{VV}^\top$, $\mathbf{WW}^\top$, and $\mathbf{M}$ all commute since $\mathbf{V}$ and $\mathbf{W}$ are eigenvectors of $\mathbf{M} \in \mathbb{S}_{\succeq\mathbf{0}}^{d\times d}$. ∎

Remark 14 *One interesting asymmetry between Lemma 8 and its counterpart, Lemma 7, is that there does not appear to be a natural k-PCA conversion analog of Lemma 8. For example, consider the case where $\mathbf{U} \in \mathbb{R}^{d\times k}$ is a (Δ, Γ)-k-cPCA of $\mathbf{M} \in \mathbb{S}_{\succeq\mathbf{0}}^{d\times d}$, but $\mathbf{M}$ has $\geq k$ eigenvalues in the range $(\lambda_k(\mathbf{M}), (1-\frac{\Gamma}{2})\lambda_k(\mathbf{M}))$, and $\mathrm{span}(\mathbf{U})$ is entirely contained in the corresponding subspace. Then, $\left\langle\mathbf{UU}^\top, \mathbf{M}\right\rangle < k\lambda_k(\mathbf{M})$, which can be arbitrarily smaller than $\|\mathbf{M}\|_k$ if $\kappa_k(\mathbf{M})$ is large.*

It is important to note that naïvely applying such conversions with our lossless ePCA result in Theorem 4 do not provide us with optimal cPCA overheads, as in Theorem 9, even when $k = 2$ and $\kappa_k(\mathbf{M}) = 1$. Our Theorem 9 does not suffer from such lossiness, up to an overhead only depending on k, except in a parameter regime ruled out by Proposition 32.

Before we begin describing our techniques, we define another desirable property of the output of the cPCA algorithms, which is complementary to the guarantee of Definition 5.

Definition 15 (Head guarantee for k-cPCA) *For $k \in [d]$, $\mathbf{M} \in \mathbb{S}_{\succeq\mathbf{0}}^{d\times d}$, let $\mathbf{U} \in \mathbb{R}^{d\times k}$ be orthonormal. We say $\mathbf{U}$ satisfies a (h, ω, Δ)-head guarantee with respect to $\mathbf{M}$ for $\omega, \Delta \in \mathbb{R}_{\geq 0}$, $0 \leq h \leq d$, if*

$$\lambda_h(\mathbf{M}) > \lambda_{h+1}(\mathbf{M}), \quad \left\|\left(\mathbf{V}^{\geq\lambda_h(\mathbf{M})}(\mathbf{M})\right)^\top \mathbf{U}_\perp\right\|_\mathrm{F}^2 \leq \Delta, \quad \text{and} \quad \frac{\lambda_{h+1}(\mathbf{M})}{\lambda_k(\mathbf{M})} \leq 1 + \omega. \quad (7)$$

If $h = 0$, we only require the third condition above.

Definition 15 asks that $\mathbf{U}_\perp$ is highly-uncorrelated with the top-h eigenspace of $\mathbf{M}$, so that $\mathbf{U}$ has picked up most of the spectral mass, for an h such that $\lambda_{h+1}(\mathbf{M})$ is not very separated from $\lambda_k(\mathbf{M})$. Note that Definition 5 does not explicitly guarantee a head guarantee in the absence of a gap.

3. k-to-1-ePCA reduction

In this section, we provide a guarantee on Algorithm 1's performance when $\mathcal{O}_{1\mathrm{PCA}}$ in Line 1 is an ϵ-1-ePCA oracle. We begin by stating one helper lemma used in the analysis.

Lemma 16 *Let $\mathbf{M} \in \mathbb{S}_{\succeq\mathbf{0}}^{d\times d}$ and $r \in [d]$. If $\mathbf{P} \in \mathbb{R}^{d\times d}$ is a rank-$(d-r)$ orthogonal projection matrix, then $\|\mathbf{PMP}\|_\mathrm{op} \geq \lambda_{r+1}(\mathbf{M})$.*

Proof This is a consequence of Cauchy's interlacing theorem (Fact 1). ∎

Theorem 4 (**k-to-1-ePCA reduction**) *Let $\epsilon \in (0,1)$, let $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$, and let $\mathcal{O}_{1\text{PCA}}$ be an ϵ-1-ePCA oracle (Definition 3). Then, Algorithm 1, when run on $\mathbf{M}$, returns $\mathbf{U} \in \mathbb{R}^{d \times k}$, an ϵ-k-ePCA of $\mathbf{M}$.*

Proof We proceed by induction on $i \in [k]$; for disambiguation let $\mathbf{U}_i \in \mathbb{R}^{d \times i}$ denote the horizontal concatenation of the first i calls to $\mathcal{O}_{1\text{PCA}}$, so that $\mathbf{P}_i = \mathbf{I}_d - \mathbf{U}_i \mathbf{U}_i^\top$. Observe that

$$
\begin{aligned}
\text{Tr}\left(\mathbf{U}_{i+1}^\top \mathbf{M} \mathbf{U}_{i+1}\right) &= \text{Tr}\left(\mathbf{U}_i^\top \mathbf{M} \mathbf{U}_i\right) + u_{i+1}^\top \mathbf{M} u_{i+1} \\
&\geq (1-\epsilon)\left\|\mathbf{M}\right\|_i + (1-\epsilon)\left\|\mathbf{P}_i \mathbf{M} \mathbf{P}_i\right\|_{\text{op}} \\
&\geq (1-\epsilon)\left\|\mathbf{M}\right\|_i + (1-\epsilon)\sigma_{i+1}(\mathbf{M}) = (1-\epsilon)\left\|\mathbf{M}\right\|_{i+1}.
\end{aligned}
$$

The first inequality used the inductive hypothesis (the base case is $i = 0$) and the guarantee on u_{i+1} from Definition 3, and the second used Lemma 16. The conclusion follows by taking $i = k$. ∎

We mention one elegant generalization of the proof strategy of Theorem 4, which shows that arbitrary compositions of two ϵ-ePCAs remain an ϵ-ePCA, regardless of the block size.

Corollary 17 *Let $\epsilon \in (0,1)$, let $\mathbf{U}_1 \in \mathbb{R}^{d \times k_1}$ be an ϵ-k_1-ePCA of $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$, and define $\widetilde{\mathbf{M}} := (\mathbf{I}_d - \mathbf{U}_1 \mathbf{U}_1^\top)\mathbf{M}(\mathbf{I}_d - \mathbf{U}_1 \mathbf{U}_1^\top)$. Let $\mathbf{U}_2 \in \mathbb{R}^{d \times k_2}$ be an ϵ-k_2-ePCA of $\widetilde{\mathbf{M}}$ with $\mathbf{U}_2 \perp \mathbf{U}_1$, and let $k := k_1 + k_2$. Then, $\mathbf{U} := \begin{pmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{pmatrix} \in \mathbb{R}^{d \times k}$ is an ϵ-k-ePCA of $\mathbf{M}$.*

Proof The proof is identical to Theorem 4, where we use Fact 1 to conclude

$$
\text{Tr}\left(\mathbf{U}_2^\top \mathbf{M} \mathbf{U}_2\right) \geq (1-\epsilon)\left\|\left(\mathbf{I}_d - \mathbf{U}_1 \mathbf{U}_1^\top\right)\mathbf{M}\left(\mathbf{I}_d - \mathbf{U}_1 \mathbf{U}_1^\top\right)\right\|_{k_2} \geq (1-\epsilon)\sum_{i \in [k_2]} \lambda_{k_1 + i}(\mathbf{M}).
$$

∎

4. k-to-1-cPCA reduction

In this section, we give our second black-box k-to-1-PCA reduction, which applies when $\mathcal{O}_{1\text{PCA}}$ in Line 1 of Algorithm 1 is a (δ, γ)-1-cPCA oracle. We first give two helper tools which are used throughout to bound the spectra of matrices under projections formed via perturbations of its eigenvectors. We defer a proof of our impossibility result characterizing invalid black-box cPCA regimes to Appendix B. In the complement of this parameter regime, we analyze the performance of Algorithm 1 and provide our main result, Theorem 9, deferring proofs to Appendix A.

Our first helper tool is a matrix decomposition inspired by the gap-free Wedin theorem (Lemma B.3 of Allen-Zhu and Li (2016)). At a high level, Lemma 18 applies noncommuting projection matrices compatible with only one of two matrices $\mathbf{M}, \widetilde{\mathbf{M}}$ on different sides of their difference, and rearranges terms. Its conclusion allows us to develop tools to bound the correlation between the small eigenspace of $\mathbf{M}$ and the large eigenspace of $\widetilde{\mathbf{M}}$.

Lemma 18 *Let* $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$, *let* $\mathbf{U} \in \mathbb{R}^{d \times r}$ *be orthonormal for* $r \in [d]$, *and let* $\widetilde{\mathbf{M}} := (\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top)\mathbf{M}(\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top)$. *Suppose* $\mathbf{M}, \widetilde{\mathbf{M}}$ *have eigendecompositions*

$$\mathbf{M} = \mathbf{L}\mathbf{\Lambda}\mathbf{L}^\top + \mathbf{S}\mathbf{\Sigma}\mathbf{S}^\top, \ \widetilde{\mathbf{M}} = \widetilde{\mathbf{L}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{L}}^\top + \widetilde{\mathbf{S}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{S}}^\top, \ for \ \mathbf{U} \perp \widetilde{\mathbf{L}}, \ \mathbf{U} \perp \widetilde{\mathbf{S}},$$

i.e. $\mathbf{L}, \mathbf{S}, \widetilde{\mathbf{L}}, \widetilde{\mathbf{S}}$ *are all orthonormal,* $\mathbf{L}, \mathbf{S}$ *have columns forming the eigenvectors of* $\mathbf{M}$, *and* $\widetilde{\mathbf{L}}, \widetilde{\mathbf{S}}, \mathbf{U}$ *have columns forming the eigenvectors of* $\widetilde{\mathbf{M}}$. *Then, assuming* $\widetilde{\mathbf{\Lambda}} \in \mathbb{S}_{\succ \mathbf{0}}^{d \times d}$,

$$\mathbf{S}^\top \widetilde{\mathbf{L}} = \left(\mathbf{I}_d - \mathbf{S}^\top \mathbf{U}\mathbf{U}^\top \mathbf{S} \right) \mathbf{\Sigma}\mathbf{S}^\top \widetilde{\mathbf{L}}\widetilde{\mathbf{\Lambda}}^{-1} - \mathbf{S}^\top \mathbf{U}\mathbf{U}^\top \mathbf{L}\mathbf{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}}\widetilde{\mathbf{\Lambda}}^{-1}.$$

We give a useful consequence of Lemma 18, repeatedly used to analyze compositions of cPCAs. The proof first applies two cPCA guarantees, and then uses Lemma 18 to bound a cross term.

Lemma 19 *Let* $\delta_1, \delta_2, \gamma_1, \gamma_2 \in [0, \frac{1}{10}]$. *Let* $\mathbf{U}_1 \in \mathbb{R}^{d \times k_1}$ *be a* (δ_1, γ_1)-k_1-*cPCA of* $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$, *and define* $\widetilde{\mathbf{M}} := (\mathbf{I}_d - \mathbf{U}_1\mathbf{U}_1^\top)\mathbf{M}(\mathbf{I}_d - \mathbf{U}_1\mathbf{U}_1^\top)$. *Let* $\mathbf{U}_2 \in \mathbb{R}^{d \times k_2}$ *be a* (δ_2, γ_2)-k_2-*cPCA of* $\widetilde{\mathbf{M}}$ *with* $\mathbf{U}_2 \perp \mathbf{U}_1$, *and let* $\gamma := \max(\gamma_1, 2\gamma_2)$, $k := k_1 + k_2$. *Let the eigendecompositions of* $\mathbf{M}, \widetilde{\mathbf{M}}$ *be*

$$\mathbf{M} = \mathbf{L}\mathbf{\Lambda}\mathbf{L}^\top + \mathbf{S}\mathbf{\Sigma}\mathbf{S}^\top, \ where \ \mathbf{L} := \mathbf{V}^{\geq (1-\gamma)\lambda_k(\mathbf{M})}(\mathbf{M}),$$
$$\widetilde{\mathbf{M}} = \widetilde{\mathbf{L}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{L}}^\top + \widetilde{\mathbf{S}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{S}}^\top, \ where \ \widetilde{\mathbf{L}} := \mathbf{V}^{\geq (1-\gamma_2)\lambda_{k_2}(\widetilde{\mathbf{M}})}(\widetilde{\mathbf{M}}), \ and \ \mathbf{U} \perp \widetilde{\mathbf{L}}, \ \mathbf{U} \perp \widetilde{\mathbf{S}},$$

i.e. orthonormal $\mathbf{L}, \mathbf{S}$ *have columns forming eigenvectors of* $\mathbf{M}$, *and orthonormal* $\widetilde{\mathbf{L}}, \widetilde{\mathbf{S}}, \mathbf{U}_1$ *have columns forming eigenvectors of* $\widetilde{\mathbf{M}}$. *Then,* $\mathbf{U} := \begin{pmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{pmatrix} \in \mathbb{R}^{d \times k}$ *is a* (Δ, γ)-k-*cPCA of* $\mathbf{M}$, *with* $\Delta := \delta_1 + 2\delta_2 + \frac{4\delta_1}{\gamma_2^2} \cdot \frac{\|\mathbf{U}_1^\top \mathbf{L}\mathbf{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}}\|_{\mathrm{op}}^2}{\lambda_k(\mathbf{M})^2}$.

We now overview our proof of Theorem 9, deferring details to Appendix A. Our high-level strategy is to provide two complementary composition results used to bound the performance of Algorithm 1 in the regime $\Gamma \geq \sqrt{\Delta} \cdot \kappa_k(\mathbf{M})$, which is the asymptotically slowest decay on Γ not ruled out by Proposition 32. Our analysis of the composition of black-box cPCAs proceeds in two steps.

1. In Lemma 20, we first consider the case where we perform a sequence of recursive cPCAs on $\mathbf{M}$, and there exist gaps in the spectrum of $\mathbf{M}$. We show how in this gapped case, cPCAs are composable with only a $\mathrm{poly}(k)$ blowup in cPCA parameters by applying Lemma 19.

2. In Lemma 21, we then consider the gap-free case, where we use a careful argument on head guarantees (see Definition 15) that we show are recursively afforded by approximate cPCAs, to bound the parameter blowup by a quasipolynomial factor in k.

Our first helper result is based on a simple strategy for bounding the term $\|\mathbf{U}_1^\top \mathbf{L}\mathbf{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}}\|_{\mathrm{op}}$ arising in Lemma 19, where we observe that when the spectrum is gapped $\mathbf{U}_1$ and $\widetilde{\mathbf{L}}$ span almost-orthogonal subspaces, by using an auxiliary head guarantee afforded by cPCA. We give our analysis of a single such composition in Lemma 24 in Appendix A. The following Lemma 20 then applies a careful double induction argument based on dyadic bucketing, via recursive calls to Lemma 24.

Lemma 20 *Let* $k \in [d]$ *and* $r \in [k]$, *such that* $\sum_{j \in [r]} k_j = k$ *for* $\{k_j\}_{j \in [r]} \subset \mathbb{N}$. *Suppose for all* $j \in [r]$, *the following recursively hold for some* $\delta, \gamma \in [0, \frac{1}{10}]$, *where we initialize* $\mathbf{P}_0 \leftarrow \mathbf{I}_d$.

1. $\mathbf{U}_j$ *is a* (δ, γ)-*approximate* k_j-*cPCA of* $\mathbf{P}_{j-1}\mathbf{M}\mathbf{P}_{j-1}$ *with* $\mathrm{span}(\mathbf{U}_j) \subseteq \mathrm{span}(\mathbf{P}_{j-1})$.

2. $\mathbf{P}_j = \mathbf{P}_{j-1} - \mathbf{U}_j \mathbf{U}_j^\top$.

Suppose for $\Delta := 4r^2\delta$, $\Gamma := 2r\gamma$, $\max(\Delta, \Gamma) \le \frac{1}{10}$, and letting $K_j := \sum_{j' \in [j]} k_{j'}$ for all $j \in [r]$,

$$\lambda_{K_j+1}(\mathbf{M}) < (1 - \Gamma)\lambda_{K_j}(\mathbf{M}) \text{ for all } j \in [r-1], \text{ and } \Delta \le \frac{\Gamma^2}{64\kappa_k(\mathbf{M})^2}. \tag{8}$$

Then $\mathbf{U} = \begin{pmatrix} \mathbf{U}_1 & \ldots & \mathbf{U}_r \end{pmatrix}$ is a (Δ, Γ)-approximate k-cPCA of $\mathbf{M}$.

We also mention that as a consequence of our proof in Lemma 20, the condition number of these blocked eigenspaces does not significantly degrade under deflation, deferring a formal statement to Lemma 27. Following a similar strategy, we provide a different analysis of cPCA composition in the following, assuming an entire eigenspace of the target is well-conditioned. To obtain this result, we induce head guarantees in recursive cPCA calls even in the absence of a gap, losing a k factor in the composition parameters, which propagates into a quasipolynomial dependence upon recursing.

Lemma 21 *Let $k \in [d]$, $\kappa_k(\mathbf{M}) \le 2$. Let $\mathbf{U} \in \mathbb{R}^{d \times k}$ be the output of $\mathsf{BlackBoxPCA}(\mathbf{M}, k, \mathcal{O}_{\text{cPCA}})$ (Algorithm 1) where $\mathcal{O}_{\text{cPCA}}$ is a (δ, γ)-1-cPCA oracle. Let $\Delta := (132k^2)^{\lceil \log_2(k) \rceil} \delta$, $\Gamma := 2^{\lceil \log_2(k) \rceil} \gamma$. Assume $\Delta \le \frac{\Gamma^2}{288}$. Then $\mathbf{U}$ is a (Δ, Γ)-k-cPCA of $\mathbf{M}$.*

We finally combine our helper results, Lemmas 20 and 21, and provide a generic analysis on the parameter degradation of $\mathsf{BlackBoxPCA}$'s guarantees when using 1-cPCA oracles.

Theorem 9 (k-to-1-cPCA reduction) *Let $(\Delta, \Gamma) \in (0, 1)$ and $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$ satisfy $\Delta \cdot \kappa_k(\mathbf{M})^2 \le \Gamma^2$, and let $\mathcal{O}_{\text{1PCA}}$ be a (δ, γ)-1-cPCA oracle (Definition 6), where $\delta := \frac{1}{k^{\Theta(\log k)}} \cdot \Delta$, $\gamma := \frac{1}{\Theta(k^3)} \cdot \Gamma$, for appropriate constants. Then, Algorithm 1 returns $\mathbf{U} \in \mathbb{R}^{d \times k}$, a (Δ, Γ)-k-cPCA of $\mathbf{M}$.*

Proof For convenience in this proof, let

$$\overline{\Delta} := \frac{\Delta}{640k^2}, \ \Delta' := \frac{\overline{\Delta}}{18k^2}, \ \delta := \frac{\Delta'}{(132k^2)^{\lceil \log_2(k) \rceil}}, \ \overline{\Gamma} := \frac{\Gamma}{10k}, \ \Gamma' := \frac{\overline{\Gamma}}{2k}, \ \gamma := \frac{\Gamma'}{2^{\lceil \log_2(k) \rceil}}.$$

Also, let $\mathcal{G}$ be the set of indices $i \in [\min(k, d-1)]$ such that $\lambda_{i+1}(\mathbf{M}) < (1 - \overline{\Gamma})\lambda_i(\mathbf{M})$, and let the elements of $\mathcal{G}$ be denoted $\{K_j\}_{j \in [r-1]}$. Let $K_0 := 0$ and $K_r := k$, and define $k_j := K_j - K_{j-1}$ for all $j \in [r]$. Following the notation of $\mathcal{O}_{\text{1PCA}}$, we let $\mathbf{U}_j := \{u_i\}_{i \in [K_{j-1}+1, K_j]} \in \mathbb{R}^{d \times k_i}$, and we iteratively define $\mathbf{P}^{(0)} := \mathbf{I}_d$, $\mathbf{P}^{(j)} := \mathbf{P}^{(j-1)} - \mathbf{U}_j \mathbf{U}_j^\top$ for all $j \in [r-1]$. We claim inductively that for all $j \in [r]$, $\mathbf{U}_j$ is a (Δ', Γ')-approximate k_j-cPCA of $\mathbf{P}^{(j-1)} \mathbf{M} \mathbf{P}^{(j-1)}$. Because $\overline{\Gamma} \le \frac{1}{10k}$,

$$\frac{\lambda_{K_j}(\mathbf{M})}{\lambda_{K_{j-1}+1}(\mathbf{M})} = \prod_{i=K_{j-1}+1}^{K_j-1} \frac{\lambda_{i+1}(\mathbf{M})}{\lambda_i(\mathbf{M})} > (1 - \overline{\Gamma})^k \ge \frac{2}{3}.$$

It is also straightforward to check that $\overline{\Delta}, \overline{\Gamma}$ satisfy the requirements of Lemma 20. Therefore, the inductive assumption and Lemma 27 show that $\kappa_{k_j}(\mathbf{P}^{(j-1)} \mathbf{M} \mathbf{P}^{(j-1)}) \le 2$. By applying Lemma 21 and $\Delta' \le \frac{1}{288} \cdot (\Gamma')^2$, we hence have completed our inductive argument. Finally, applying Lemma 20 shows that $\mathbf{U}$ is a $(\overline{\Delta}, \overline{\Gamma})$-$k$-cPCA, so it is clearly also a (Δ, Γ)-cPCA as claimed. ∎

References

Zeyuan Allen-Zhu and Yuanzhi Li. Lazysvd: Even faster svd decomposition yet without agonizing pain. *Advances in neural information processing systems*, 29, 2016.

Zeyuan Allen-Zhu and Yuanzhi 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. IEEE, 2017.

Maria-Florina Balcan, Simon Shaolei Du, Yining Wang, and Adams Wei Yu. An improved gap-dependency analysis of the noisy power method. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 284–309. JMLR.org, 2016.

Alexandre d'Aspremont, Laurent Ghaoui, Michael Jordan, and Gert Lanckriet. A direct formulation for sparse pca using semidefinite programming. *Advances in neural information processing systems*, 17, 2004.

I. Diakonikolas, D. M. Kane, and A. Pensia. Outlier Robust Mean Estimation with Subgaussian Rates via Stability. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

I. Diakonikolas, D. M. Kane, A. Pensia, and T. Pittas. Streaming Algorithms for High-Dimensional Robust Statistics. In *International Conference on Machine Learning*, 2022.

Ilias Diakonikolas, Daniel Kane, Ankit Pensia, and Thanasis Pittas. Nearly-linear time and streaming algorithms for outlier-robust pca. In *International Conference on Machine Learning*, pages 7886–7921. PMLR, 2023.

Dan Garber, Elad Hazan, Chi Jin, Cameron Musco, Praneeth Netrapalli, Aaron Sidford, et al. Faster eigenvector computation via shift-and-invert preconditioning. In *International Conference on Machine Learning*, pages 2626–2634. PMLR, 2016.

Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pages 2861–2869, 2014.

Prateek Jain, Chi Jin, Sham Kakade, Praneeth Netrapalli, and Aaron Sidford. Streaming pca: Matching matrix bernstein and near-optimal finite sample guarantees for oja's algorithm. In *Proceedings of The 29th Conference on Learning Theory (COLT)*, June 2016. URL https://arxiv.org/abs/1602.06929.

A. Jambulapati, J. Li, and K. Tian. Robust sub-gaussian principal component analysis and width-independent schatten packing. *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020. arxiv preprint at https://arxiv.org/abs/2006.06980.

Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc. A.*, 2016.

W. Kong, R. Somani, S. Kakade, and S. Oh. Robust meta-learning for mixed linear regression with small batches. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

P. K Kothari and D. Steurer. Outlier-robust moment-estimation via sum-of-squares. *arXiv preprint arXiv:1711.11581*, 2017.

Syamantak Kumar and Purnamrita Sarkar. Streaming PCA for markovian data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=FQGRkwmRzm.

Junghyun Lee, Hanseul Cho, Se-Young Yun, and Chulhee Yun. Fair streaming principal component analysis: Statistical and algorithmic viewpoint. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=TW3ipYdDQG.

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

Ren-Cang Li and Lei-Hong Zhang. Convergence of the block lanczos method for eigenvalue clusters. *Numerische Mathematik*, 131(1):83–113, 2015.

Xin Liang. On the optimality of the oja's algorithm for online pca. *Statistics and Computing*, 33 (3):62, 2023.

X. Liu, W. Kong, and S. Oh. Differential privacy and robust statistics in high dimensions. In *Proc. 35th Annual Conference on Learning Theory (COLT)*, 2022a.

Xiyang Liu, Weihao Kong, Prateek Jain, and Sewoong Oh. Dp-pca: Statistically optimal and differentially private pca. *arXiv preprint arXiv:2205.13709*, 2022b.

Lester Mackey. Deflation methods for sparse pca. *Advances in neural information processing systems*, 21, 2008.

Shahar Mendelson and Nikita Zhivotovskiy. Robust covariance estimation under ℓ_4-ℓ_2 norm equivalence. *The Annals of Statistics*, 2018. URL https://api.semanticscholar.org/CorpusID:60410055.

Raphael Meyer, Cameron Musco, and Christopher Musco. On the unreasonable effectiveness of single vector krylov methods for low-rank approximation. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 811–845. SIAM, 2024.

Stanislav Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 2015.

Baback Moghaddam, Yair Weiss, and Shai Avidan. Generalized spectral bounds for sparse lda. In *Proceedings of the 23rd international conference on Machine learning*, pages 641–648, 2006.

Cameron Musco and Christopher Musco. Randomized block krylov methods for stronger and faster approximate singular value decomposition. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, pages 1396–1404, 2015.

Joe Neeman, Bobby Shi, and Rachel Ward. Concentration inequalities for sums of markov dependent random matrices, 2023.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15:267–273, 1982.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720. URL https://doi.org/10.1080/14786440109462720.

Christopher De Sa, Christopher Ré, and Kunle Olukotun. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2332–2341. JMLR.org, 2015.

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, page 257–265. JMLR.org, 2016.

Xiaohan Wei and Stanislav Minsker. Estimation of the covariance structure of heavy-tailed distributions. *Advances in neural information processing systems*, 30, 2017.

Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14(4), 2013.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

Appendix A. Deferred proofs from Section 4

We first give proofs of Lemmas 18 and 19.

Lemma 22 *Let $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$, let $\mathbf{U} \in \mathbb{R}^{d \times r}$ be orthonormal for $r \in [d]$, and let $\widetilde{\mathbf{M}} := (\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top)\mathbf{M}(\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top)$. Suppose $\mathbf{M}, \widetilde{\mathbf{M}}$ have eigendecompositions*

$$\mathbf{M} = \mathbf{L}\boldsymbol{\Lambda}\mathbf{L}^\top + \mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^\top,\ \widetilde{\mathbf{M}} = \widetilde{\mathbf{L}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\mathbf{L}}^\top + \widetilde{\mathbf{S}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\mathbf{S}}^\top,\ \text{for } \mathbf{U} \perp \widetilde{\mathbf{L}},\ \mathbf{U} \perp \widetilde{\mathbf{S}},$$

i.e. $\mathbf{L}, \mathbf{S}, \widetilde{\mathbf{L}}, \widetilde{\mathbf{S}}$ are all orthonormal, $\mathbf{L}, \mathbf{S}$ have columns forming the eigenvectors of $\mathbf{M}$, and $\widetilde{\mathbf{L}}, \widetilde{\mathbf{S}}, \mathbf{U}$ have columns forming the eigenvectors of $\widetilde{\mathbf{M}}$. Then, assuming $\widetilde{\boldsymbol{\Lambda}} \in \mathbb{S}_{\succ \mathbf{0}}^{d \times d}$,

$$\mathbf{S}^\top \widetilde{\mathbf{L}} = \left(\mathbf{I}_d - \mathbf{S}^\top \mathbf{U}\mathbf{U}^\top \mathbf{S}\right)\boldsymbol{\Sigma}\mathbf{S}^\top \widetilde{\mathbf{L}}\widetilde{\boldsymbol{\Lambda}}^{-1} - \mathbf{S}^\top \mathbf{U}\mathbf{U}^\top \mathbf{L}\boldsymbol{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}}\widetilde{\boldsymbol{\Lambda}}^{-1}.$$

Proof Let $\mathbf{D} := \widetilde{\mathbf{M}} - \mathbf{M} = (\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top)\mathbf{M}(\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top) - \mathbf{M}$. Then,

$$\begin{aligned}
\mathbf{S}^\top \mathbf{D}\widetilde{\mathbf{L}} &= \mathbf{S}^\top \left(\left(\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top\right)\mathbf{M}\left(\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top\right)\widetilde{\mathbf{L}} - \mathbf{M}\widetilde{\mathbf{L}}\right) \\
&= \mathbf{S}^\top \left(\left(\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top\right)\mathbf{M}\widetilde{\mathbf{L}} - \mathbf{M}\widetilde{\mathbf{L}}\right) = -\mathbf{S}^\top \mathbf{U}\mathbf{U}^\top \mathbf{M}\widetilde{\mathbf{L}} \\
&= -\mathbf{S}^\top \mathbf{U}\mathbf{U}^\top \left(\mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^\top + \mathbf{L}\boldsymbol{\Lambda}\mathbf{L}^\top\right)\widetilde{\mathbf{L}} \\
&= -\mathbf{S}^\top \mathbf{U}\mathbf{U}^\top \mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^\top \widetilde{\mathbf{L}} - \mathbf{S}^\top \mathbf{U}\mathbf{U}^\top \mathbf{L}\boldsymbol{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}}, &(9)
\end{aligned}$$

where the second equality used $\widetilde{\mathbf{L}} \perp \mathbf{U}$. We also have,

$$\mathbf{S}^\top \mathbf{D}\widetilde{\mathbf{L}} = \mathbf{S}^\top \left(\widetilde{\mathbf{M}}\widetilde{\mathbf{L}} - \mathbf{M}\widetilde{\mathbf{L}}\right) = \mathbf{S}^\top \left(\widetilde{\mathbf{L}}\widetilde{\boldsymbol{\Lambda}} - \mathbf{M}\widetilde{\mathbf{L}}\right) = \mathbf{S}^\top \widetilde{\mathbf{L}}\widetilde{\boldsymbol{\Lambda}} - \mathbf{S}^\top \mathbf{M}\widetilde{\mathbf{L}} = \mathbf{S}^\top \widetilde{\mathbf{L}}\widetilde{\boldsymbol{\Lambda}} - \boldsymbol{\Sigma}\mathbf{S}^\top \widetilde{\mathbf{L}} \quad (10)$$

where the second equality used $\widetilde{\mathbf{M}} = \widetilde{\mathbf{L}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\mathbf{L}}^\top + \widetilde{\mathbf{S}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\mathbf{S}}^\top$ and $\widetilde{\mathbf{L}} \perp \widetilde{\mathbf{S}}$, and the fourth equality used $\mathbf{M} = \mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^\top + \mathbf{L}\boldsymbol{\Lambda}\mathbf{L}^\top$. Combining (9) and (10) and right-multiplying by $\widetilde{\boldsymbol{\Lambda}}^{-1}$, we have the claim. ∎

Lemma 23 *Let $\delta_1, \delta_2, \gamma_1, \gamma_2 \in [0, \frac{1}{10}]$. Let $\mathbf{U}_1 \in \mathbb{R}^{d \times k_1}$ be a (δ_1, γ_1)-k_1-cPCA of $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$, and define $\widetilde{\mathbf{M}} := (\mathbf{I}_d - \mathbf{U}_1\mathbf{U}_1^\top)\mathbf{M}(\mathbf{I}_d - \mathbf{U}_1\mathbf{U}_1^\top)$. Let $\mathbf{U}_2 \in \mathbb{R}^{d \times k_2}$ be a (δ_2, γ_2)-k_2-cPCA of $\widetilde{\mathbf{M}}$ with $\mathbf{U}_2 \perp \mathbf{U}_1$, and let $\gamma := \max(\gamma_1, 2\gamma_2)$, $k := k_1 + k_2$. Let the eigendecompositions of $\mathbf{M}, \widetilde{\mathbf{M}}$ be*

$$\mathbf{M} = \mathbf{L}\boldsymbol{\Lambda}\mathbf{L}^\top + \mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^\top,\ \text{where } \mathbf{L} := \mathbf{V}^{\geq (1-\gamma)\lambda_k(\mathbf{M})}(\mathbf{M}),$$

$$\widetilde{\mathbf{M}} = \widetilde{\mathbf{L}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\mathbf{L}}^\top + \widetilde{\mathbf{S}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\mathbf{S}}^\top,\ \text{where } \widetilde{\mathbf{L}} := \mathbf{V}^{\geq (1-\gamma_2)\lambda_{k_2}(\widetilde{\mathbf{M}})}(\widetilde{\mathbf{M}}),\ \text{and } \mathbf{U} \perp \widetilde{\mathbf{L}},\ \mathbf{U} \perp \widetilde{\mathbf{S}},$$

i.e. orthonormal $\mathbf{L}, \mathbf{S}$ have columns forming eigenvectors of $\mathbf{M}$, and orthonormal $\widetilde{\mathbf{L}}, \widetilde{\mathbf{S}}, \mathbf{U}_1$ have columns forming eigenvectors of $\widetilde{\mathbf{M}}$. Then, $\mathbf{U} := \begin{pmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{pmatrix} \in \mathbb{R}^{d \times k}$ is a (Δ, γ)-k-cPCA of $\mathbf{M}$, with $\Delta := \delta_1 + 2\delta_2 + \frac{4\delta_1}{\gamma_2^2} \cdot \frac{\|\mathbf{U}_1^\top \mathbf{L}\boldsymbol{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}}\|_{\mathrm{op}}^2}{\lambda_k(\mathbf{M})^2}$.

Proof We first observe

$$\left\|\mathbf{S}^\top \mathbf{U}\right\|_{\mathbf{F}}^2 = \mathrm{Tr}\left(\mathbf{U}^\top \mathbf{S}\mathbf{S}^\top \mathbf{U}\right) = \mathrm{Tr}\left(\mathbf{U}_1^\top \mathbf{S}\mathbf{S}^\top \mathbf{U}_1\right) + \mathrm{Tr}\left(\mathbf{U}_2^\top \mathbf{S}\mathbf{S}^\top \mathbf{U}_2\right). \tag{11}$$

Since $\mathbf{U}_1$ is a (δ_1, γ_1)-k_1-cPCA of $\mathbf{M}$, it is also a (δ_1, γ)-k-cPCA (since $\gamma \geq \gamma_1$), so we have $\text{Tr}\left(\mathbf{U}_1^\top \mathbf{S}\mathbf{S}^\top \mathbf{U}_1\right) \leq \delta_1$. Next, we bound $\text{Tr}\left(\mathbf{S}\mathbf{S}^\top \mathbf{U}_2 \mathbf{U}_2^\top\right)$. Then, since $\mathbf{U}_2 \perp \mathbf{U}_1$,

$$
\begin{aligned}
\text{Tr}\left(\mathbf{U}_2^\top \mathbf{S}\mathbf{S}^\top \mathbf{U}_2\right) &= \text{Tr}\left(\mathbf{U}_2^\top \left(\widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\top + \widetilde{\mathbf{S}}\widetilde{\mathbf{S}}^\top\right) \mathbf{S}\mathbf{S}^\top \left(\widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\top + \widetilde{\mathbf{S}}\widetilde{\mathbf{S}}^\top\right) \mathbf{U}_2\right) \\
&= \text{Tr}\left(\left(\mathbf{U}_2^\top \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\top \mathbf{S} + \mathbf{U}_2^\top \widetilde{\mathbf{S}}\widetilde{\mathbf{S}}^\top \mathbf{S}\right)\left(\mathbf{S}^\top \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\top \mathbf{U}_2 + \mathbf{S}^\top \widetilde{\mathbf{S}}\widetilde{\mathbf{S}}^\top \mathbf{U}_2\right)\right) \\
&\leq 2\text{Tr}\left(\mathbf{U}_2^\top \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\top \mathbf{S}\mathbf{S}^\top \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\top \mathbf{U}_2\right) + 2\text{Tr}\left(\mathbf{U}_2^\top \widetilde{\mathbf{S}}\widetilde{\mathbf{S}}^\top \mathbf{S}\mathbf{S}^\top \widetilde{\mathbf{S}}\widetilde{\mathbf{S}}^\top \mathbf{U}_2\right), \text{ using Fact 3} \\
&= 2\text{Tr}\left(\widetilde{\mathbf{L}}^\top \mathbf{U}_2 \mathbf{U}_2^\top \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\top \mathbf{S}\mathbf{S}^\top \widetilde{\mathbf{L}}\right) + 2\text{Tr}\left(\widetilde{\mathbf{S}}^\top \mathbf{S}\mathbf{S}^\top \widetilde{\mathbf{S}}\widetilde{\mathbf{S}}^\top \mathbf{U}_2 \mathbf{U}_2^\top \widetilde{\mathbf{S}}\right) \\
&\leq 2\text{Tr}\left(\widetilde{\mathbf{L}}^\top \mathbf{S}\mathbf{S}^\top \widetilde{\mathbf{L}}\right) + 2\text{Tr}\left(\widetilde{\mathbf{S}}^\top \mathbf{U}_2 \mathbf{U}_2^\top \widetilde{\mathbf{S}}\right), \text{ using Fact 5,} \\
&\qquad\qquad\qquad\qquad \left\|\widetilde{\mathbf{L}}^\top \mathbf{U}_2 \mathbf{U}_2^\top \widetilde{\mathbf{L}}\right\|_{\text{op}} \leq 1, \text{ and } \left\|\widetilde{\mathbf{S}}^\top \mathbf{S}\mathbf{S}^\top \widetilde{\mathbf{S}}\right\|_{\text{op}} \leq 1 \\
&\leq 2\left\|\mathbf{S}^\top \widetilde{\mathbf{L}}\right\|_{\text{F}}^2 + 2\delta_2 \qquad\qquad\qquad\qquad\qquad\qquad (12)
\end{aligned}
$$

where the last line used the cPCA guarantee. Now we bound $\|\mathbf{S}^\top \widetilde{\mathbf{L}}\|_{\text{F}}$. By Lemma 18,

$$
\mathbf{S}^\top \widetilde{\mathbf{L}} = \left(\mathbf{I}_d - \mathbf{S}^\top \mathbf{U}_1 \mathbf{U}_1^\top \mathbf{S}\right) \mathbf{\Sigma}\mathbf{S}^\top \widetilde{\mathbf{L}}\widetilde{\mathbf{\Lambda}}^{-1} - \mathbf{S}^\top \mathbf{U}_1 \mathbf{U}_1^\top \mathbf{L}\mathbf{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}}\widetilde{\mathbf{\Lambda}}^{-1}. \qquad (13)
$$

We next focus on the first term. Since $\|\mathbf{\Sigma}\|_{\text{op}} < (1 - \gamma)\lambda_k(\mathbf{M})$, and $\lambda_1(\mathbf{S}^\top \mathbf{U}_1 \mathbf{U}_1^\top \mathbf{S}) \leq 1$, we have

$$
\widetilde{\mathbf{\Lambda}}^{-1}\widetilde{\mathbf{L}}^\top \mathbf{S}\mathbf{\Sigma}\left(\mathbf{I}_d - \mathbf{S}^\top \mathbf{U}_1 \mathbf{U}_1^\top \mathbf{S}\right)^2 \mathbf{\Sigma}\mathbf{S}^\top \widetilde{\mathbf{L}}\widetilde{\mathbf{\Lambda}}^{-1} \preceq (1 - \gamma)^2 \lambda_k(\mathbf{M})^2 \widetilde{\mathbf{\Lambda}}^{-1}\widetilde{\mathbf{L}}^\top \mathbf{S}\mathbf{S}^\top \widetilde{\mathbf{L}}\widetilde{\mathbf{\Lambda}}^{-1}.
$$

Further, since Fact 1 implies $\lambda_k(\mathbf{M}) \leq \lambda_{k_2}(\widetilde{\mathbf{M}})$, Fact 5 implies that

$$
\begin{aligned}
\text{Tr}\left(\widetilde{\mathbf{\Lambda}}^{-1}\widetilde{\mathbf{L}}^\top \mathbf{S}\mathbf{S}^\top \widetilde{\mathbf{L}}\widetilde{\mathbf{\Lambda}}^{-1}\right) &= \text{Tr}\left(\widetilde{\mathbf{\Lambda}}^{-2}\widetilde{\mathbf{L}}^\top \mathbf{S}\mathbf{S}^\top \widetilde{\mathbf{L}}\right) \\
&\leq \left\|\widetilde{\mathbf{\Lambda}}^{-1}\right\|_{\text{op}}^2 \left\|\mathbf{S}^\top \widetilde{\mathbf{L}}\right\|_{\text{F}}^2 \leq \frac{1}{(1 - \gamma_2)^2 \lambda_k(\mathbf{M})^2} \left\|\mathbf{S}^\top \widetilde{\mathbf{L}}\right\|_{\text{F}}^2.
\end{aligned}
$$

Combining the above two displays proves $\|(\mathbf{I}_d - \mathbf{S}^\top \mathbf{U}_1 \mathbf{U}_1^\top \mathbf{S})\mathbf{\Sigma}\mathbf{S}^\top \widetilde{\mathbf{L}}\widetilde{\mathbf{\Lambda}}^{-1}\|_{\text{F}} \leq (1 - \gamma_2)\|\mathbf{S}^\top \widetilde{\mathbf{L}}\|_{\text{F}}$, since $\frac{(1-\gamma)^2}{(1-\gamma_2)^2} \leq (1 - \gamma_2)^2$ for $\gamma \geq 2\gamma_2$. Plugging this into (13), along with a triangle inequality, yields

$$
\left\|\mathbf{S}^\top \widetilde{\mathbf{L}}\right\|_{\text{F}} \leq (1 - \gamma_2)\left\|\mathbf{S}^\top \widetilde{\mathbf{L}}\right\|_{\text{F}} + \left\|\mathbf{S}^\top \mathbf{U}_1 \mathbf{U}_1^\top \mathbf{L}\mathbf{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}}\widetilde{\mathbf{\Lambda}}^{-1}\right\|_{\text{F}},
$$

which implies

$$
\left\|\mathbf{S}^\top \widetilde{\mathbf{L}}\right\|_{\text{F}} \leq \frac{\left\|\mathbf{S}^\top \mathbf{U}_1 \mathbf{U}_1^\top \mathbf{L}\mathbf{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}}\widetilde{\mathbf{\Lambda}}^{-1}\right\|_{\text{F}}}{\gamma_2}. \qquad (14)
$$

Finally, the conclusion follows by combining (11), (12), (14) and $\|\mathbf{AB}\|_{\mathrm{F}} \leq \|\mathbf{A}\|_{\mathrm{op}} \|\mathbf{B}\|_{\mathrm{F}}$:

$$\left\|\mathbf{S}^\top \mathbf{U}_1 \mathbf{U}_1^\top \mathbf{L}\boldsymbol{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}}\widetilde{\boldsymbol{\Lambda}}^{-1}\right\|_{\mathrm{F}}^2 \leq \left\|\mathbf{U}_1^\top \mathbf{S}\right\|_{\mathrm{F}}^2 \left\|\mathbf{U}_1^\top \mathbf{L}\boldsymbol{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}}\widetilde{\boldsymbol{\Lambda}}^{-1}\right\|_{\mathrm{op}}^2$$

$$\leq \left\|\mathbf{U}_1^\top \mathbf{S}\right\|_{\mathrm{F}}^2 \left\|\widetilde{\boldsymbol{\Lambda}}^{-1}\right\|_{\mathrm{op}}^2 \left\|\mathbf{U}_1^\top \mathbf{L}\boldsymbol{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}}\right\|_{\mathrm{op}}^2$$

$$\leq \delta_1 \frac{1}{(1-\gamma_2)^2 \lambda_{k_2}\left(\widetilde{\mathbf{M}}\right)^2} \left\|\mathbf{U}_1^\top \mathbf{L}\boldsymbol{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}}\right\|_{\mathrm{op}}^2$$

$$\leq \frac{\delta_1(1+3\gamma_2)}{\lambda_k(\mathbf{M})^2} \left\|\mathbf{U}_1^\top \mathbf{L}\boldsymbol{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}}\right\|_{\mathrm{op}}^2$$

$$\leq \frac{2\delta_1}{\lambda_k(\mathbf{M})^2} \left\|\mathbf{U}_1^\top \mathbf{L}\boldsymbol{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}}\right\|_{\mathrm{op}}^2 .$$

∎

In the remainder of the section, we give formal proofs of Lemma 20 and Lemma 21, our main composition results in gapped and well-conditioned settings respectively, which are combined to prove Theorem 9.

A.1. cPCA composition: the gapped case

In this section, we prove Lemma 20, our composition lemma which applies under gap assumptions in the spectrum. Before providing our full gapped composition result, we state two helper lemmas.

Lemma 24 *In the setting of Lemma 19, suppose that $(1-\gamma_1)\lambda_{k_1}(\mathbf{M}) > \lambda_{k_1+1}(\mathbf{M})$ and $\delta_1 \leq \frac{\gamma_2^2}{16\kappa_k(\mathbf{M})^2}$. Then, $\mathbf{U}$ is a $(2(\delta_1+\delta_2),\gamma)$-$k$-cPCA of $\mathbf{M}$, for $\gamma = \max(\gamma_1, 2\gamma_2)$.*

Proof By Lemma 19, it suffices to provide a bound on $\|\mathbf{U}_1^\top \mathbf{L}\boldsymbol{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}}\|_{\mathrm{op}}$.

Throughout the proof, let $w \in \mathrm{span}(\widetilde{\mathbf{L}}) \subseteq \mathrm{span}([\mathbf{U}_1]_\perp)$, $u \in \mathrm{span}(\mathbf{U}_1)$ be arbitrary unit vectors, and note that $u \perp w$ by definition. Let $\mathrm{rank}(\mathbf{L}) = k' \geq k$. Then,

$$u^\top \mathbf{L}\boldsymbol{\Lambda}\mathbf{L}^\top w = \sum_{i\in[k']} \lambda_i(\mathbf{M})[\mathbf{L}^\top u]_i[\mathbf{L}^\top w]_i$$

$$= \underbrace{\sum_{i=1}^{k_1} \lambda_i(\mathbf{M})[\mathbf{L}^\top u]_i[\mathbf{L}^\top w]_i}_{:=R_1} + \underbrace{\sum_{i=k_1+1}^{k'} \lambda_i(\mathbf{M})[\mathbf{L}^\top u]_i[\mathbf{L}^\top w]_i}_{:=R_2} . \qquad (15)$$

Let orthonormal $\mathbf{L}_1 \in \mathbb{R}^{d\times k_1}$ span the subspace of eigenvectors corresponding to the largest k_1 eigenvalues of $\mathbf{M}$, and let $\mathbf{S}_1$ be the complement subspace. We assumed that $(1-\gamma_1)\lambda_{k_1}(\mathbf{M}) > \lambda_{k_1+1}(\mathbf{M})$, so that $\|\mathbf{S}_1^\top \mathbf{U}_1\|_{\mathrm{F}}^2 \leq \delta_1$ by the cPCA guarantee. This also lets us conclude

$$\left\|\mathbf{L}_1^\top \mathbf{U}_1\right\|_{\mathrm{F}}^2 = \|\mathbf{U}_1\|_{\mathrm{F}}^2 - \left\|\mathbf{S}_1^\top \mathbf{U}_1\right\|_{\mathrm{F}}^2 \geq k_1 - \delta_1$$

$$\implies \left\|\mathbf{L}_1^\top w\right\|_2^2 \leq \|\mathbf{L}_1\|_{\mathrm{F}}^2 - \left\|\mathbf{L}_1^\top \mathbf{U}_1\right\|_{\mathrm{F}}^2 \leq \delta_1 \text{ for all } w \in \mathbb{R}^d, \ w \perp \mathbf{U}_1. \qquad (16)$$

20

Let $\mathbf{\Lambda}_1 \in \mathbb{R}^{k_1 \times k_1}$ and $\mathbf{\Lambda}_2 \in \mathbb{R}^{k_2 \times k_2}$ be diagonal matrices, so that $\mathbf{\Lambda}_1$ has diagonal entries given by $\{\lambda_i(\mathbf{M})\}_{i \in [k_1]}$ and $\mathbf{\Lambda}_2$ has diagonal entries $\{\lambda_i(\mathbf{M})\}_{i \in [k] \setminus [k_1]}$. We now bound the terms in (15): first,

$$|R_1| = \left| \left(\mathbf{L}_1^\top u\right)^\top \mathbf{\Lambda}_1 \left(\mathbf{L}_1^\top w\right) \right| \le \|\mathbf{\Lambda}_1\|_{\mathrm{op}} \left\|\mathbf{L}_1^\top w\right\|_2 \left\|\mathbf{L}_1^\top u\right\|_2 \le \lambda_1(\mathbf{M})\sqrt{\delta_1}.$$

In the last inequality we used (16). Next, we similarly have

$$|R_2| \le \|\mathbf{\Lambda}_2\|_{\mathrm{op}} \left\|\mathbf{S}_1^\top w\right\|_2 \left\|\mathbf{S}_1^\top u\right\|_2 \le \lambda_1(\mathbf{M})\sqrt{\delta_1},$$

since $\|\mathbf{S}_1^\top u\|_2 \le \|\mathbf{S}_1^\top \mathbf{U}_1\|_{\mathrm{F}} \le \sqrt{\delta_1}$. Plugging in the above two displays into (15), and applying Lemma 19, then yields the result due to the assumed bound on $\frac{\delta_1}{\gamma_2}$:

$$\left\|\mathbf{S}^\top \mathbf{U}\right\|_{\mathrm{F}}^2 \le \delta_1 + 2\delta_2 + \frac{4\delta_1}{\gamma_2^2 \lambda_k(\mathbf{M})^2} \cdot \left(2\sqrt{\delta_1}\lambda_1(\mathbf{M})\right)^2 = \delta_1 + 2\delta_2 + \frac{16\delta_1^2 \kappa_k(\mathbf{M})^2}{\gamma_2^2} \le 2(\delta_1 + \delta_2).$$

∎

Lemma 25 *Let $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$ have $\lambda_{k+1}(\mathbf{M}) < (1 - \Gamma)\lambda_k(\mathbf{M})$ for $k \in [d]$ and $\Gamma \in (0,1)$, let $\mathbf{V} \in \mathbb{R}^{d \times k}$ be an exact k-PCA of $\mathbf{M}$, and let $\mathbf{U} \in \mathbb{R}^{d \times k}$ be a (Δ, Γ)-k-cPCA of $\mathbf{M}$. Then,*

$$\left\|\left(\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top\right) \mathbf{M} \left(\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top\right) - \left(\mathbf{I}_d - \mathbf{V}\mathbf{V}^\top\right) \mathbf{M} \left(\mathbf{I}_d - \mathbf{V}\mathbf{V}^\top\right)\right\|_{\mathrm{op}} \le 4\sqrt{\Delta}\lambda_1(\mathbf{M}).$$

Proof By our gap assumption, the cPCA guarantees imply $\|(\mathbf{I}_d - \mathbf{V}\mathbf{V}^\top)\mathbf{U}\mathbf{U}^\top\|_{\mathrm{F}}^2 \le \Delta$, so that

$$\left\|\left(\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top\right)\mathbf{V}\mathbf{V}^\top\right\|_{\mathrm{F}}^2 = k - \left\langle\mathbf{U}\mathbf{U}^\top, \mathbf{V}\mathbf{V}^\top\right\rangle = \left\|\left(\mathbf{I}_d - \mathbf{V}\mathbf{V}^\top\right)\mathbf{U}\mathbf{U}^\top\right\|_{\mathrm{F}}^2 \le \Delta.$$

Next, observe that, since $\mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top = \mathbf{U}\mathbf{U}^\top(\mathbf{I}_d - \mathbf{V}\mathbf{V}^\top) - (\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top)\mathbf{V}\mathbf{V}^\top$,

$$\left\|\mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top\right\|_{\mathrm{op}} \le \left\|\mathbf{U}\mathbf{U}^\top(\mathbf{I}_d - \mathbf{V}\mathbf{V}^\top)\right\|_{\mathrm{op}} + \left\|(\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top)\mathbf{V}\mathbf{V}^\top\right\|_{\mathrm{op}} \le 2\sqrt{\Delta},$$

where we used our earlier bounds. The conclusion follows from

$$\left(\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top\right) \mathbf{M} \left(\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top\right) - \left(\mathbf{I}_d - \mathbf{V}\mathbf{V}^\top\right) \mathbf{M} \left(\mathbf{I}_d - \mathbf{V}\mathbf{V}^\top\right)$$
$$= \left(\mathbf{I}_d - \mathbf{V}\mathbf{V}^\top\right) \mathbf{M} \left(\mathbf{V}\mathbf{V}^\top - \mathbf{U}\mathbf{U}^\top\right) - \left(\mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top\right) \mathbf{M} \left(\mathbf{I}_d - \mathbf{U}\mathbf{U}^\top\right),$$

after using the triangle inequality and our earlier bound on $\|\mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top\|_{\mathrm{op}}$. ∎

We now apply Lemma 24 recursively to obtain our gapped composition result, aided by Lemma 25.

Lemma 26 *Let $k \in [d]$ and $r \in [k]$, such that $\sum_{j \in [r]} k_j = k$ for $\{k_j\}_{j \in [r]} \subset \mathbb{N}$. Suppose for all $j \in [r]$, the following recursively hold for some $\delta, \gamma \in [0, \frac{1}{10}]$, where we initialize $\mathbf{P}_0 \leftarrow \mathbf{I}_d$.*

1. $\mathbf{U}_j$ is a (δ, γ)-approximate k_j-cPCA of $\mathbf{P}_{j-1}\mathbf{M}\mathbf{P}_{j-1}$ with $\mathrm{span}(\mathbf{U}_j) \subseteq \mathrm{span}(\mathbf{P}_{j-1})$.

2. $\mathbf{P}_j = \mathbf{P}_{j-1} - \mathbf{U}_j \mathbf{U}_j^\top$.

Suppose for $\Delta := 4r^2\delta$, $\Gamma := 2r\gamma$, $\max(\Delta, \Gamma) \leq \frac{1}{10}$, and letting $K_j := \sum_{j' \in [j]} k_{j'}$ for all $j \in [r]$,

$$\lambda_{K_j+1}(\mathbf{M}) < (1 - \Gamma)\lambda_{K_j}(\mathbf{M}) \text{ for all } j \in [r-1], \text{ and } \Delta \leq \frac{\Gamma^2}{64\kappa_k(\mathbf{M})^2}. \tag{8}$$

Then $\mathbf{U} = \begin{pmatrix} \mathbf{U}_1 & \dots & \mathbf{U}_r \end{pmatrix}$ is a (Δ, Γ)-approximate k-cPCA of $\mathbf{M}$.

Proof For all $j \in [r]$, let $\mathbf{W}_j := \begin{pmatrix} \mathbf{U}_1 & \dots & \mathbf{U}_j \end{pmatrix} \in \mathbb{R}^{d \times K_j}$, so that $\mathbf{U} = \mathbf{W}_k$. Our proof will be by double induction. Throughout the proof, refer to the following assumption by Prefix$_{r-1}$: for all $j \in [r-1]$, $\mathbf{W}_j$ is a (Δ, Γ)-K_j-cPCA of $\mathbf{M}$. We will prove the stated claim assuming Prefix$_{r-1}$, which means that Prefix$_{r-1}$ implies Prefix$_r$. Because all the assumptions continue to apply if the lemma statement took smaller values of r and k, this means we can induct on the base of Prefix, justifying our assumption of Prefix$_{r-1}$. For the remainder of the proof, suppose Prefix$_{r-1}$ is true.

We first state a consequence of Prefix$_{r-1}$. Let $1 \leq j' \leq j \leq r - 1$, so $\mathbf{W}_{j'}$ is a (Δ, Γ)-$K_{j'}$-cPCA of $\mathbf{M}$ by assumption, and $\mathbf{W}_{j'}\mathbf{W}_{j'}^\top = \mathbf{I}_d - \mathbf{P}_{j'}$. Moreover, let $\mathbf{V}_{j'} \in \mathbb{R}^{d \times K_{j'}}$ be an exact $K_{j'}$-cPCA of $\mathbf{M}$, which is unique by the gap assumption (8), and let $\mathbf{Q}_{j'} := \mathbf{V}_{j'}\mathbf{V}_{j'}^\top$ for convenience. We have:

$$\begin{aligned}
\lambda_{K_j - K_{j'}+1}\left((\mathbf{I}_d - \mathbf{P}_{j'})\mathbf{M}(\mathbf{I}_d - \mathbf{P}_{j'})\right) &\leq \lambda_{K_j - K_{j'}+1}\left((\mathbf{I}_d - \mathbf{Q}_{j'})\mathbf{M}(\mathbf{I}_d - \mathbf{Q}_{j'})\right) \\
&\quad + \left\|(\mathbf{I}_d - \mathbf{P}_{j'})\mathbf{M}(\mathbf{I}_d - \mathbf{P}_{j'}) - (\mathbf{I}_d - \mathbf{Q}_{j'})\mathbf{M}(\mathbf{I}_d - \mathbf{Q}_{j'})\right\|_{\mathrm{op}} \\
&\leq \lambda_{K_j+1}(\mathbf{M}) + 4\sqrt{\Delta}\lambda_1(\mathbf{M}) \\
&< (1 - \Gamma)\lambda_{K_j}(\mathbf{M}) + \frac{\Gamma}{2}\lambda_k(\mathbf{M}) \leq \left(1 - \frac{\Gamma}{2}\right)\lambda_{K_j}(\mathbf{M}) \\
&\leq \left(1 - \frac{\Gamma}{2}\right)\lambda_{K_j - K_{j'}}\left((\mathbf{I}_d - \mathbf{P}_{j'})\mathbf{M}(\mathbf{I}_d - \mathbf{P}_{j'})\right).
\end{aligned} \tag{17}$$

Here, the first inequality was by Fact 2, the second used Lemma 25 and the gap assumption (8) with $j \leftarrow j'$, the third used (8) with $j \leftarrow j$, the fourth used $\lambda_k(\mathbf{M}) \leq \lambda_{K_j}(\mathbf{M})$, and the last used Fact 1. In other words, (17) shows that after any j' steps of our recursive cPCA procedure, all remaining gaps (ensured for our original matrix by (8)) continue to hold in the residual matrix $(\mathbf{I}_d - \mathbf{P}_{j'})\mathbf{M}(\mathbf{I}_d - \mathbf{P}_{j'})$, up to a small multiplicative loss in the gap parameter.

Next, we divide $[r]$ into dyadic intervals, using $L + 1$ layers for $L := \lceil \log_2(r) \rceil$, labeled $0 \leq \ell \leq L$. In particular, the 0th layer is $S_{0,1} := [r]$, the first layer consists of the intervals $S_{1,0} := [2^{L-1}]$ and $S_{1,1} := [r] \setminus S_{1,0}$, and so on. More generally, for each $0 \leq \ell \leq L$ and $0 \leq i < 2^\ell$, we let

$$S_{\ell,i} := \left\{ j \in [r] \mid i2^{L-\ell} + 1 \leq j \leq (i+1)2^{L-\ell} \right\},$$

$$a_{\ell,i} := \sum_{j \in [i2^{L-\ell}]} k_j = K_{i2^{L-\ell}}, \; b_{\ell,i} := \sum_{j \in [\min(r,(i+1)2^{L-\ell})]} k_j = K_{\min(r,(i+1)2^{L-\ell})}.$$

In other words, this dyadic splitting induces a binary tree, where the (L^{th}-layer) leaves correspond to single elements in $[r]$, and the (0th-layer) root corresponds to the set $[r]$. Moreover, the node in the tree associated with $S_{\ell,i} \subseteq [r]$ captures the interval $[a_{\ell,i} + 1, b_{\ell,i}] \subseteq [k]$. We refer to this node as the $(\ell, i)^{\text{th}}$ node, and we associate it with $\mathbf{W}_{\ell,i}$, an approximate $(b_{\ell,i} - a_{\ell,i})$-cPCA of the matrix

$\mathbf{P}_{i2^{L-\ell}}\mathbf{M}\mathbf{P}_{i2^{L-\ell}}$. In particular, $\mathbf{W}_{\ell,i}$ is just the horizontal concatenation of $\mathbf{U}_j$ for all $j \in S_{\ell,i}$. We inductively analyze the approximation quality of the intermediate $\mathbf{W}_{\ell,i}$, to show that for all $0 \le \ell \le L$, $\mathbf{W}_{\ell,i}$ is a $(\delta_\ell, \gamma_\ell)$-approximate $(b_{\ell,i} - a_{\ell,i})$-cPCA of $\mathbf{P}_{i2^{L-\ell}}\mathbf{M}\mathbf{P}_{i2^{L-\ell}}$ for all $0 \le i < 2^\ell$ and

$$\delta_\ell := 4^{L-\ell}\delta, \ \gamma_\ell := 2^{L-\ell}\gamma.$$

The base case, $\ell = L$, follows since all $\mathbf{W}_{L,i} = \mathbf{U}_{i+1}$ are (δ, γ)-approximate cPCAs of their corresponding $\mathbf{P}_i \mathbf{M} \mathbf{P}_i$. Next, suppose inductively that all the $\mathbf{W}_{\ell,i}$ are $(\delta_\ell, \gamma_\ell)$-approximate cPCAs for some $0 \le \ell < L$, and all $0 \le i < 2^\ell$. Then, consider some $\mathbf{W}_{\ell-1,i}$, which is the composition of $\mathbf{W}_{\ell,2i}$ and $\mathbf{W}_{\ell,2i+1}$. To analyze this composition, we apply Lemma 24, which requires that

$$\lambda_{b_{\ell,2i}-a_{\ell,2i}}\left(\mathbf{P}_{i2^{L-\ell+1}}\mathbf{M}\mathbf{P}_{i2^{L-\ell+1}}\right) \ge (1+\gamma_\ell)\lambda_{b_{\ell-1,i}-a_{\ell-1,i}}\left(\mathbf{P}_{i2^{L-\ell+1}}\mathbf{M}\mathbf{P}_{i2^{L-\ell+1}}\right). \tag{18}$$

Because $\gamma_\ell \le \gamma_0 \le \frac{\Gamma}{2}$ for all $0 \le \ell < L$, (18) is implied by (17). Lemma 24 also requires $\delta_\ell \le \frac{\gamma_\ell^2}{16\kappa_k(\mathbf{M})^2}$, which is invariant to the choice of ℓ since δ_ℓ and γ_ℓ^2 grow at the same rate, so this is implied by our bounds in (8). Therefore, we can apply Lemma 24 and $\mathbf{W}_{\ell-1,i}$ is indeed a $(4\delta_\ell, 2\gamma_\ell) = (\delta_{\ell-1}, \gamma_{\ell-1})$-cPCA as claimed. The conclusion follows by taking $\ell = 0$, as $4^L \le 4r^2$, $2^L \le 2r$. ∎

The following byproduct of our proof of Lemma 20 is useful in our development.

Lemma 27 *In the setting of Lemma 20, suppose that $\lambda_{K_j}(\mathbf{M}) \ge \frac{2}{3}\lambda_{K_{j-1}+1}(\mathbf{M})$ for all $j \in [r]$, where we let $K_0 := 0$. Then, for all $j \in [r]$, we have $\kappa_{k_j}\left(\mathbf{P}_{j-1}\mathbf{M}\mathbf{P}_{j-1}\right) \le 2$.*

Proof First of all, we have $\lambda_{k_j}(\mathbf{P}_{j-1}\mathbf{M}\mathbf{P}_{j-1}) \ge \lambda_{K_j}(\mathbf{M})$ by Fact 1. Moreover, an analogous argument to (17) shows that the largest eigenvalue of $\mathbf{P}_{j-1}\mathbf{M}\mathbf{P}_{j-1}$ is perturbed by at most $\frac{\Gamma}{2}\lambda_k(\mathbf{M}) \le \frac{1}{20}\lambda_{K_{j-1}+1}(\mathbf{M})$, when compared to $\lambda_{K_{j-1}+1}(\mathbf{M})$, so $\lambda_1(\mathbf{P}_{j-1}\mathbf{M}\mathbf{P}_{j-1}) \le 1.05\lambda_{K_{j-1}+1}(\mathbf{M})$. The conclusion follows from combining these inequalities, because $1.05 \cdot \frac{3}{2} \le 2$. ∎

A.2. cPCA composition: the well-conditioned case

We now analyze Algorithm 1 under the promise that $\kappa_k(\mathbf{M}) \le 2$, and prove Lemma 21. We begin with a basic helper lemma on composition under a head guarantee (Definition 15), patterned off of Lemma 24.

Lemma 28 *In the setting of Lemma 19, let $\delta_1 \le \frac{\gamma_2^2}{288}$, $\kappa_k(\mathbf{M}) \le 2$, and $\gamma_1 \le \gamma_2$. For $0 \le h \le d$, suppose that $\mathbf{U}_1$ satisfies a $(h, 2k_1\gamma_1, \delta_1)$-head guarantee (Definition 15) with respect to $\mathbf{M}$. Then, $\mathbf{U}$ is a (δ, γ)-k-cPCA of $\mathbf{M}$, for $\delta := 130k_1^2\delta_1 + 2\delta_2$ and $\gamma := 2\gamma_2$.*

Proof Throughout the proof, for convenience we denote

$$\mathbf{L}_h := \mathbf{V}^{\ge \lambda_h(\mathbf{M})}(\mathbf{M}) \implies \left\|\mathbf{L}_h^\top[\mathbf{U}_1]_\perp\right\|_F^2 \le \delta_1 \text{ and } \frac{\lambda_{h+1}(\mathbf{M})}{\lambda_{k_1}(\mathbf{M})} \le 1 + 2k_1\gamma_1.$$

By Lemma 19, it suffices to bound $\|\mathbf{U}_1^\top\mathbf{L}\mathbf{\Lambda}\mathbf{L}^\top\widetilde{\mathbf{L}}\|_{\mathrm{op}}$. Throughout the proof, let $w \in \mathrm{span}(\widetilde{\mathbf{L}}) \subseteq \mathrm{span}([\mathbf{U}_1]_\perp)$, $u \in \mathrm{span}(\mathbf{U}_1)$ be arbitrary unit vectors, so $u \perp w$. Next, let $m \in [d]$ be the largest index such that $\lambda_m(\mathbf{M}) \ge (1 - \gamma_1)\lambda_{k_1}(\mathbf{M})$. Finally, let $k' := \mathrm{rank}(\mathbf{L}) \ge k$. Note that

$h \leq k_1 \leq m \leq k'$, where $h \leq k_1$ follows because otherwise $\|\mathbf{L}_h^\top [\mathbf{U}_1]_\perp\|_{\mathrm{op}} < 1$ is impossible, as this would mean

$$\dim\left(\mathrm{span}\left(\mathbf{L}_h\right)\right) + \dim\left(\mathrm{span}\left([\mathbf{U}_1]_\perp\right)\right) = h + d - k_1 > d \implies \mathrm{span}(\mathbf{L}_h) \cap \mathrm{span}([\mathbf{U}_1]_\perp) \neq \emptyset.$$

Next, letting $\mathbf{V} := \begin{pmatrix} \mathbf{L} & \mathbf{S} \end{pmatrix}$ be a full set of orthonormal eigenvectors for $\mathbf{M}$, we have

$$
\begin{aligned}
u^\top \mathbf{L}\mathbf{\Lambda}\mathbf{L}^\top w &= \sum_{i \in [k']} \lambda_i\left(\mathbf{M}\right) [\mathbf{L}^\top u]_i [\mathbf{L}^\top w]_i \\
&= \sum_{i \in [m]} \lambda_i\left(\mathbf{M}\right) [\mathbf{L}^\top u]_i [\mathbf{L}^\top w]_i + \sum_{i \in [k']\setminus[m]} \lambda_i\left(\mathbf{M}\right) [\mathbf{L}^\top u]_i [\mathbf{L}^\top w]_i \\
&\quad - \lambda_{k_1}\left(\mathbf{M}\right) \left\langle \mathbf{V}^\top u, \mathbf{V}^\top w \right\rangle \\
&= \underbrace{\sum_{i \in [h]} \left(\lambda_i\left(\mathbf{M}\right) - \lambda_{k_1}\left(\mathbf{M}\right)\right) [\mathbf{L}^\top u]_i [\mathbf{L}^\top w]_i}_{:=R_1} \\
&\quad + \underbrace{\sum_{i \in [m]\setminus[h]} \left(\lambda_i\left(\mathbf{M}\right) - \lambda_{k_1}\left(\mathbf{M}\right)\right) [\mathbf{L}^\top u]_i [\mathbf{L}^\top w]_i}_{:=R_2} \\
&\quad + \underbrace{\sum_{i \in [k']\setminus[m]} \lambda_i\left(\mathbf{M}\right) [\mathbf{L}^\top u]_i [\mathbf{L}^\top w]_i}_{:=R_3} - \underbrace{\sum_{i \in [d]\setminus[m]} \lambda_{k_1}\left(\mathbf{M}\right) [\mathbf{V}^\top u]_i [\mathbf{V}^\top w]_i}_{:=R_4}.
\end{aligned}
\tag{19}
$$

In the second equality, we used $\langle u, w \rangle = 0$ and $\mathbf{V}\mathbf{V}^\top = \mathbf{I}_d$, and in the third, we used that the first m rows of $\mathbf{V}^\top$ agree with $\mathbf{L}^\top$. For convenience in the following, let $\mathbf{\Lambda}_1 \in \mathbb{R}^{(h-1)\times(h-1)}$, $\mathbf{\Lambda}_2 \in \mathbb{R}^{(m-h+1)\times(m-h+1)}$, $\mathbf{\Lambda}_3 \in \mathbb{R}^{(k'-m)\times(k'-m)}$ and $\mathbf{\Lambda}_4 \in \mathbb{R}^{(d-m)\times(d-m)}$ be diagonal matrices such that:

1. $\mathbf{\Lambda}_1$ has diagonal entries $\{\lambda_i\left(\mathbf{M}\right) - \lambda_{k_1}\left(\mathbf{M}\right)\}_{i \in [h-1]}$.

2. $\mathbf{\Lambda}_2$ has diagonal entries $\{\lambda_i\left(\mathbf{M}\right) - \lambda_{k_1}\left(\mathbf{M}\right)\}_{i \in [m]\setminus[h-1]}$.

3. $\mathbf{\Lambda}_3$ has diagonal entries $\{\lambda_i\left(\mathbf{M}\right)\}_{i \in [k']\setminus[m]}$.

4. $\mathbf{\Lambda}_4$ has diagonal entries $\{\lambda_{k_1}\left(\mathbf{M}\right)\}_{i \in [d]\setminus[m]}$.

We now bound the terms in (19). For R_1,

$$|R_1| = \left| \left(\mathbf{L}_h^\top u\right)^\top \mathbf{\Lambda}_1 \left(\mathbf{L}_h^\top w\right) \right| \leq \|\mathbf{\Lambda}_1\|_{\mathrm{op}} \left\|\mathbf{L}_h^\top w\right\|_2 \left\|\mathbf{L}_h^\top u\right\|_2 \leq \lambda_1(\mathbf{M})\sqrt{\delta_1}.$$

The last inequality used $\|\mathbf{L}_h^\top w\|_2 \leq \left\|\mathbf{L}_h^\top [\mathbf{U}_1]_\perp\right\|_{\mathrm{F}} \leq \sqrt{\delta_1}$. For R_2, since $\|u\|_2 = \|w\|_2 = 1$,

$$|R_2| \leq \|\mathbf{\Lambda}_2\|_{\mathrm{op}} \leq \max\left\{ \frac{\lambda_{h+1}\left(\mathbf{M}\right)}{\lambda_{k_1}\left(\mathbf{M}\right)} - 1, \gamma_1 \right\} \lambda_{k_1}(\mathbf{M}).$$

For R_3, R_4, since $\left\|\mathbf{S}^\top u\right\|_2 \le \left\|\mathbf{S}^\top \mathbf{U}_1\right\|_{\mathrm{F}} \le \sqrt{\delta_1}$ by the cPCA guarantee on $\mathbf{U}_1$, and the jth rows of $\mathbf{V}^\top$ and $\mathbf{S}^\top$ agree for any $j > m$ by definition,

$$|R_3| \le \|\mathbf{\Lambda}_3\|_{\mathrm{op}} \left\|\mathbf{S}^\top w\right\|_2 \left\|\mathbf{S}^\top u\right\|_2 \le \lambda_1(\mathbf{M})\sqrt{\delta_1},$$

$$|R_4| \le \|\mathbf{\Lambda}_4\|_{\mathrm{op}} \left\|\mathbf{S}^\top w\right\|_2 \left\|\mathbf{S}^\top u\right\|_2 \le \lambda_1(\mathbf{M})\sqrt{\delta_1}.$$

Plugging in the above displays into (19), we have

$$\|\mathbf{U}_1^\top \mathbf{L}\mathbf{\Lambda}\mathbf{L}^\top \widetilde{\mathbf{L}}\|_{\mathrm{op}}^2 \le \left(3\lambda_1(\mathbf{M})\sqrt{\delta_1} + \max\left\{\frac{\lambda_{h+1}(\mathbf{M})}{\lambda_{k_1}(\mathbf{M})} - 1, \gamma_1\right\}\lambda_{k_1}(\mathbf{M})\right)^2$$

$$\le 18\lambda_1(\mathbf{M})^2 \delta_1 + 2\max\left\{(2k_1\gamma_1)^2, \gamma_1^2\right\}\lambda_{k_1}(\mathbf{M})^2$$

$$= 18\lambda_1(\mathbf{M})^2 \delta_1 + 8k_1^2\gamma_1^2\lambda_1(\mathbf{M})^2.$$

Applying Lemma 19 and using the assumed bounds in the lemma statement then yields the result:

$$\left\|\mathbf{S}^\top \mathbf{U}\right\|_{\mathrm{F}}^2 \le \delta_1 + 2\delta_2 + \frac{4\delta_1}{\gamma_2^2\lambda_k(\mathbf{M})^2} \cdot \left(18\lambda_1(\mathbf{M})^2 \delta_1 + 8k_1^2\gamma_1^2\lambda_1(\mathbf{M})^2\right)$$

$$= \delta_1 + 2\delta_2 + \frac{8\delta_1\kappa_k(\mathbf{M})^2}{\gamma_2^2} \cdot \left(9\delta_1 + 4k_1^2\gamma_1^2\right) \le 130k_1^2\delta_1 + 2\delta_2.$$

∎

In the absence of an explicit gap in the spectrum of $\mathbf{M}$, Lemma 28 shows how to nonetheless apply Lemma 19, assuming a head guarantee. The following lemma shows how to inductively use cPCA bounds to enforce such a head guarantee. Informally, the requirement (20) states that after i steps of Algorithm 1, the next h steps are a cPCA of the residual matrix, for all $h \in [m]$. We show that (20) either implies a gap in the spectrum of the residual matrix, or that the entire residual matrix is well-conditioned in the top-m subspace. Either case yields a guarantee compatible with Lemma 28.

Lemma 29 *Let $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d\times d}$ and let $\mathbf{U} \in \mathbb{R}^{d\times d}$ be orthonormal, with columns $\{u_\ell\}_{\ell\in[d]}$. For $1 \le i \le j \le d$, let $\mathbf{U}_{[i,j]} \in \mathbb{R}^{d\times(j-i+1)}$ have columns $\{u_\ell\}_{\ell\in[i,j]}$. Further, for $i \in [d]$, suppose for all $h \in [m]$,*

$$\mathbf{U}_{[i+1,i+h]} \text{ is a } (\delta, \gamma)\text{-}h\text{-cPCA of } \widetilde{\mathbf{M}}_i := \left(\mathbf{I}_d - \mathbf{U}_{[1,i]}\mathbf{U}_{[1,i]}^\top\right)\mathbf{M}\left(\mathbf{I}_d - \mathbf{U}_{[1,i]}\mathbf{U}_{[1,i]}^\top\right). \tag{20}$$

There is $h \in [0, m-1]$ so $\mathbf{U}_{[i+1,i+m]}$ satisfies a $(h, 2m\gamma, \delta)$-head guarantee with respect to $\widetilde{\mathbf{M}}_i$.

Proof Let $h \in [1, m-1]$ be maximal such that $\lambda_{h+1}(\widetilde{\mathbf{M}}_i) \le (1-\gamma)\lambda_h(\widetilde{\mathbf{M}}_i)$; if no such index exists, then we set $h = 0$. Then, by definition the first condition in (7) is satisfied for this value h if $h \ne 0$, and otherwise it is irrelevant. Further, since h is the largest such index, $\lambda_{j+1}(\widetilde{\mathbf{M}}_i) > (1-\gamma)\lambda_j(\widetilde{\mathbf{M}}_i)$ for all $j \in [h+1, m-1]$, so telescoping across all such j, we have the third claim in (7):

$$\frac{\lambda_{h+1}(\widetilde{\mathbf{M}}_i)}{\lambda_m(\widetilde{\mathbf{M}}_i)} \le \left(\frac{1}{1-\gamma}\right)^{m-h-1} \le 1 + 2m\gamma.$$

This holds even if $h = 0$, and the second claim in (7) is irrelevant in this case, so we assume $h \geq 1$ henceforth. For notational convenience, let $\widetilde{\mathbf{L}}_{i,h} := \mathbf{V}^{\geq \lambda_h(\widetilde{\mathbf{M}}_i)}(\widetilde{\mathbf{M}}_i) \in \mathbb{R}^{d \times h}$. The definition of h implies $\widetilde{\mathbf{L}}_{i,h} = [\mathbf{V}^{\leq (1-\gamma)\lambda_h(\widetilde{\mathbf{M}}_i)}(\widetilde{\mathbf{M}}_i)]_\perp$, so (20) with $m \leftarrow h$ yields

$$\left\| \mathbf{U}_{[i+1,i+h]}^\top \mathbf{V}^{\leq (1-\gamma)\lambda_h(\widetilde{\mathbf{M}}_i)}(\widetilde{\mathbf{M}}_i) \right\|_{\mathrm{F}}^2 \leq \delta \implies \left\| \mathbf{U}_{[i+1,i+h]}^\top \widetilde{\mathbf{L}}_{i,h} \right\|_{\mathrm{F}}^2 \geq h - \delta$$

$$\implies \left\| \widetilde{\mathbf{L}}_{i,h}^\top \left[\mathbf{U}_{[i+1,i+h]} \right]_\perp \right\|_{\mathrm{F}}^2 \leq \delta$$

$$\implies \left\| \widetilde{\mathbf{L}}_{i,h}^\top \left[\mathbf{U}_{[i+1,i+m]} \right]_\perp \right\|_{\mathrm{F}}^2 \leq \delta.$$

In the last claim, we used that $\mathrm{span}(\mathbf{U}_{[i+1,i+h]}) \supset \mathrm{span}(\mathbf{U}_{[i+1,i+m]})$. ∎

Finally, we apply Lemmas 28 and 29 to complete our analysis of Algorithm 1.

Lemma 30 *Let $k \in [d]$, $\kappa_k(\mathbf{M}) \leq 2$. Let $\mathbf{U} \in \mathbb{R}^{d \times k}$ be the output of $\mathsf{BlackBoxPCA}(\mathbf{M}, k, \mathcal{O}_{\mathrm{cPCA}})$ (Algorithm 1) where $\mathcal{O}_{\mathrm{cPCA}}$ is a (δ, γ)-1-cPCA oracle. Let $\Delta := (132k^2)^{\lceil \log_2(k) \rceil} \delta$, $\Gamma := 2^{\lceil \log_2(k) \rceil} \gamma$. Assume $\Delta \leq \frac{\Gamma^2}{288}$. Then $\mathbf{U}$ is a (Δ, Γ)-k-cPCA of $\mathbf{M}$.*

Proof We divide $[k]$ into dyadic intervals as in Lemma 20, using $L + 1$ layers for $L := \lceil \log_2(k) \rceil$, labeled $0 \leq \ell \leq L$. In particular, the 0th layer is $S_{0,1} := [k]$, the first layer consists of the intervals $S_{1,0} := [2^{L-1}]$ and $S_{1,1} := [k] \setminus S_{1,0}$, and so on. For each $0 \leq \ell \leq L$ and $0 \leq i < 2^\ell$, we let

$$S_{\ell,i} := \left\{ j \in [k] \mid i2^{L-\ell} + 1 \leq j \leq (i+1)2^{L-\ell} \right\}.$$

This dyadic splitting again induces a binary tree, where singletons in $[k]$ are leaves and the root is $[k]$. The node in the tree associated with $S_{\ell,i} \subseteq [k]$ captures the interval $[i2^{L-\ell} + 1, \min((i + 1)2^{L-\ell}, k)] \subseteq [k]$, and (following notation in Lemma 29) we associate it with

$$\mathbf{W}_{i,\ell} := \mathbf{U}_{[i2^{L-\ell}+1, \min((i+1)2^{L-\ell}, k)]}.$$

Moreover, we define $m_{i,\ell} := \dim(\mathrm{span}(\mathbf{W}_{i,\ell}))$, and for all $h \in [m_{i,\ell}]$, we let

$$\mathbf{W}_{i,\ell}^{(h)} := \mathbf{U}_{[i2^{L-\ell}+1, i2^{L-\ell}+h]},$$

i.e. $\mathbf{W}_{i,\ell}^{(h)}$ is the first h columns of $\mathbf{W}_{i,\ell}$. We inductively claim that for all $0 \leq \ell \leq L$, $\mathbf{W}_{\ell,i}^{(h)}$ is a $(\delta_\ell, \gamma_\ell)$-approximate h-cPCA of $\mathbf{P}_{i2^{L-\ell}} \mathbf{M} \mathbf{P}_{i2^{L-\ell}}$, for all $0 \leq i < 2^\ell$, and all $h \in [m_{i,\ell}]$, where

$$\delta_\ell := (132k^2)^{L-\ell} \delta, \ \gamma_\ell := 2^{L-\ell} \gamma.$$

The base case $\ell = L$ follows by assumption. Next, suppose the claim above holds for some $0 \leq \ell < L$, and all $0 \leq i < 2^\ell$. Consider some $\mathbf{W}_{\ell-1,i}$, the composition of $\mathbf{W}_{\ell,2i}$ and $\mathbf{W}_{\ell,2i+1}$. If $\mathbf{W}_{\ell,2i+1}$ is empty, then all $\mathbf{W}_{\ell-1,i}^{(h)}$ are $(\delta_{\ell-1}, \gamma_{\ell-1})$-$h$-cPCAs of $\mathbf{P}_{i2^{L-\ell+1}} \mathbf{M} \mathbf{P}_{i2^{L-\ell+1}}$ by the inductive assumption.

Otherwise, by induction, (20) holds with $i \leftarrow i2^{L-\ell+1}$ and $m \leftarrow 2^{L-\ell}$, for $\delta \leftarrow \delta_\ell, \gamma \leftarrow \gamma_\ell$. So, we can apply Lemma 29 to obtain $h \in [0, 2^{L-\ell} - 1]$ satisfying the head guarantee needed by Lemma 28. We then use Lemma 28 with $\mathbf{M} \leftarrow \mathbf{P}_{i2^{L-\ell+1}} \mathbf{M} \mathbf{P}_{i2^{L-\ell+1}}$ to analyze our composition. In particular,

$$\kappa_{k'}(\mathbf{P}_{i2^{L-\ell+1}} \mathbf{M} \mathbf{P}_{i2^{L-\ell+1}}) \leq 2$$

26

where $k' := \min(2^{L-\ell+1}, k - i2^{L-\ell+1})$, by Fact 1 and $\kappa_k(\mathbf{M}) \le 2$. Moreover, δ_ℓ grows faster than γ_ℓ^2, so the remaining requirement in Lemma 28 is satisfied by our choices of Δ, Γ. Finally, Lemma 28 proves that $\mathbf{W}_{i,\ell-1}$ is a $(\delta_{\ell-1}, \gamma_{\ell-1})$-$m_{i,\ell-1}$-cPCA of $\mathbf{P}_{i2^{L-\ell+1}}\mathbf{M}\mathbf{P}_{i2^{L-\ell+1}}$. It is straightforward to check the same argument we used also shows that $\mathbf{W}_{i,\ell-1}^{(h)}$ is a $(\delta_{\ell-1}, \gamma_{\ell-1})$-$h$-cPCA of $\mathbf{P}_{i2^{L-\ell+1}}\mathbf{M}\mathbf{P}_{i2^{L-\ell+1}}$ for all $h \in [m_{i,\ell-1}]$, since we can truncate the composition early. This completes the induction, and the conclusion follows by taking $\ell \leftarrow 0$. ∎

Appendix B. Invalid regimes for black-box cPCA

To complement our main result on cPCA (Theorem 9), in this section, we point out that not all values of (δ, γ) pairs are feasible inputs for Algorithm 1 to serve as a lossless black-box reduction even for constant k, unlike in the case of ePCA. Formally, we characterize two types of parameter regimes for black-box cPCA using the following definition.

Definition 31 (Black-box cPCA regimes) *Let $g : [0,1] \times [1, \infty) \to [0,1]$. We say g induces an* invalid black-box cPCA regime *if, for every $f : \mathbb{N} \to \mathbb{R}_{>0}$, there is a value $\Delta \in [0,1]$, a $k \in \mathbb{N}$, and $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$ with $d \ge k$ such that, letting $\Gamma := g(\Delta, \kappa_k(\mathbf{M}))$,[8] and defining*

$$\delta := \frac{\Delta}{f(k)}, \; \gamma := \frac{\Gamma}{f(k)}, \tag{21}$$

BlackBoxPCA *does not always return a (Δ, Γ)-k-cPCA if $\mathcal{O}_{\text{1PCA}}$ is a (δ, γ)-1-cPCA oracle.*

Conversely, we say g induces a valid black-box cPCA regime *if there exists $f : \mathbb{N} \to \mathbb{R}_{>0}$, such that for all $\mathbf{M} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$, $\Delta \in [0,1]$, $\Gamma := g(\Delta, \kappa_k(\mathbf{M}))$, and (δ, γ) in (21), BlackBoxPCA always returns a (Δ, Γ)-k-cPCA if $\mathcal{O}_{\text{1PCA}}$ is a (δ, γ)-1-cPCA oracle.*

For example, suppose we establish g which induces an invalid black-box cPCA regime. This means that for BlackBoxPCA to serve generically as a $(\Delta, \Gamma := g(\Delta, \kappa))$-$k$-cPCA algorithm for all $\Delta \in [0,1]$ and $\kappa > 1$, we must ask for $\mathcal{O}_{\text{1PCA}}$ to be a (δ, γ)-1-cPCA oracle where either $\delta = o_k(\Delta)$ or $\gamma = o_k(\Gamma)$, i.e. we necessarily suffer some loss in either the Δ or Γ parameter by more than a function of k. The following impossibility result establishes functions g which induce invalid black-box cPCA regimes.

Our choice to parameterize the regimes g in Definition 31 by $\kappa_k(\mathbf{M})$ is motivated by a result of Liang (2023) (see Eq. (1.3) in that paper), which proves that some sample complexity overhead in $\kappa_k(\mathbf{M})$ is necessary in general. On the other hand, the requirement that $\Gamma \gtrsim \sqrt{\Delta}$ appears to be an artifact of the specific deflation strategy in Algorithm 1. We now state our main impossibility result.

Proposition 32 (Invalid black-box cPCA regimes) *Let $\alpha, \beta, \nu \in \mathbb{R}_{>0}$, and $g : [0,1] \times (1, \infty) \to [0,1]$ be defined as $g(\Delta, \kappa) := \nu\Delta^\alpha\kappa^\beta$. If $\alpha > \frac{1}{2}$ or $\beta < 1$, g induces an invalid black-box cPCA regime.*

Moreover, Theorem 9 shows that the complement of the regimes ruled out by Proposition 32 are all valid regimes, where we suffer parameter degradation only in k. Hence, Proposition 32 and Theorem 9 completely characterize valid and invalid black-box cPCA regimes, under Definition 31.

---

8. We let $\kappa_k(\mathbf{M})$ denote the k-condition number of $\mathbf{M}$, i.e. the ratio of $\lambda_1(\mathbf{M})$ and $\lambda_k(\mathbf{M})$ (see Section 2).

We establish Proposition 32 by giving 3-dimensional examples where Algorithm 1 fails to return a (Δ, Γ)-2-cPCA if $\Gamma \lesssim \kappa_2(\mathbf{M})\sqrt{\Delta}$ asymptotically, assuming $\mathcal{O}_{1\text{PCA}}$ is a (δ, γ)-1-cPCA oracle, for any δ, γ smaller than Δ, Γ by a constant. First, we show that sublinear functions induce invalid regimes.

Lemma 33 *Let $\alpha, \beta, \nu \in \mathbb{R}_{>0}$, and define $g : [0,1] \times (1, \infty) \to [0,1]$ as $g(\Delta, \kappa) := \nu \Delta^\alpha \kappa^\beta$. If $\alpha > 1$ or $\beta < 1$, then g induces an invalid black-box cPCA regime.*

Proof Following the notation in Theorem 31, fix $f : \mathbb{N} \to \mathbb{R}_{>0}$. Without loss of generality, we assume that f is nondecreasing. Let $C := f(2)$. We divide the analysis into two cases:

1. $\boldsymbol{\alpha > 1}$: For any $\kappa > 1$, $\frac{g(\Delta, \kappa)}{\Delta(\kappa-1)} \leq \min\left(\frac{1}{2C}, \frac{1}{4}\right)$ for all sufficiently small Δ.

2. $\boldsymbol{\beta < 1}$: For any $\Delta > 0$, $\frac{g(\Delta, \kappa)}{\Delta(\kappa-1)} \leq \min\left(\frac{1}{2C}, \frac{1}{4}\right)$ for all sufficiently large κ.

Therefore, we can always find a pair (Δ, κ) in both cases, such that $\frac{g(\Delta, \kappa)}{\Delta(\kappa-1)} \leq \min\left(\frac{1}{2C}, \frac{1}{4}\right)$. We will show an example with $k = 2$ and a matrix $\mathbf{M}$ with k-condition number $\kappa_k(\mathbf{M}) = \kappa$. If we can show that for $\delta = \frac{\Delta}{C}$, two $(\delta, 0)$-1-cPCA oracles fail to produce a (Δ, Γ)-1-cPCA, where $\Gamma := g(\Delta, \kappa)$, the same will be true for calling two $(\frac{\Delta}{f(k)}, \frac{\Gamma}{f(k)})$-1-cPCA oracles, since $\delta = \frac{\Delta}{f(k)}$ and $0 \leq \frac{\Gamma}{f(k)}$. This gives us our desired contradiction, since it shows g induces an invalid black-box cPCA regime.

To this end, we now show that there is $\mathbf{M} \in \mathbb{S}_{\succeq 0}^{3 \times 3}$ with $\kappa_k(\mathbf{M}) = \kappa$ such that calling two $(\delta, 0)$-1-cPCA oracles as in Algorithm 1 fails to return a (Δ, Γ)-2-cPCA. Define the 3×3 diagonal matrix

$$\mathbf{M} := \begin{pmatrix} \kappa & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 - 2\Gamma \end{pmatrix}.$$

We let $u_1^\top = \begin{pmatrix} \sqrt{1-\delta} & 0 & \sqrt{\delta} \end{pmatrix}$ and observe u_1 is a $(\delta, 0)$-1-cPCA of $\mathbf{M}$. We also let u_2 be the top eigenvector of $\widetilde{\mathbf{M}} := (\mathbf{I}_3 - u_1 u_1^\top)\mathbf{M}(\mathbf{I}_3 - u_1 u_1^\top)$, which is a $(\delta, 0)$-1-cPCA of $\widetilde{\mathbf{M}}$. To compute u_2, note

$$\widetilde{\mathbf{M}} = \begin{pmatrix} \delta((1-2\Gamma)(1-\delta) + \kappa\delta) & 0 & -\sqrt{\delta(1-\delta)}((1-2\Gamma)(1-\delta) + \kappa\delta) \\ 0 & 1 & 0 \\ -\sqrt{\delta(1-\delta)}((1-2\Gamma)(1-\delta) + \kappa\delta) & 0 & (1-\delta)((1-2\Gamma)(1-\delta) + \kappa\delta) \end{pmatrix},$$

so that the eigenvector-eigenvalue pairs of $\widetilde{\mathbf{M}}$ are

$$(e_2, 1), \quad \left(\begin{pmatrix} -\sqrt{\delta} \\ 0 \\ \sqrt{1-\delta} \end{pmatrix}, (1-2\Gamma)(1-\delta) + \kappa\delta \right).$$

Since $\delta = \frac{\Delta}{C} \geq \frac{2\Gamma}{\kappa-1}$, we have

$$(1-2\Gamma)(1-\delta) + \kappa\delta = 1 + 2\Gamma\delta + (\kappa-1)\delta - 2\Gamma \geq 1,$$

so $u_2^\top = \begin{pmatrix} -\sqrt{\delta} & 0 & \sqrt{1-\delta} \end{pmatrix}$ is an exact 1-cPCA of $\widetilde{\mathbf{M}}$. In this case, $\mathbf{U} = \begin{pmatrix} u_1 & u_2 \end{pmatrix}$ fails to be a (Δ, Γ)-2-cPCA of $\mathbf{M}$ as claimed, since

$$\left\| \mathbf{U}^\top \mathbf{V}^{<(1-\Gamma)\lambda_2(\mathbf{M})} \right\|_F^2 = \left\| \mathbf{U}^\top e_3 \right\|_F^2 = 1 \geq \Delta.$$

The second equality can also be seen by noting that $e_3 \in \mathrm{span}(\mathbf{U})$. ∎

We next improve Lemma 33 to show that we can rule out g asymptotically smaller than $\sqrt{\cdot}$ in Δ. Namely for a fixed κ, while Lemma 33 rules out black-box cPCA reductions without parameter losses for $\Gamma \ll \Delta$, when Δ is small enough, Proposition 32 further rules out such reductions for $\Gamma \ll \sqrt{\Delta}$.

Proposition 34 (Invalid black-box cPCA regimes) *Let $\alpha, \beta, \nu \in \mathbb{R}_{>0}$, and $g : [0, 1] \times (1, \infty) \to [0, 1]$ be defined as $g(\Delta, \kappa) := \nu \Delta^\alpha \kappa^\beta$. If $\alpha > \frac{1}{2}$ or $\beta < 1$, g induces an invalid black-box cPCA regime.*

Proof If $\alpha > 1$ or $\beta < 1$, then applying Lemma 33 yields the claim. Thus, in the remainder of the proof suppose $\frac{1}{2} < \alpha \leq 1$ and $\beta \geq 1$. Fix $f : \mathbb{N} \to \mathbb{R}_{>0}$, and define $C := f(2)$.

Since $\alpha \leq 1$, there exists a $c > 0$ such that $g(\Delta, 2) \geq c\Delta$ for all sufficiently small Δ. Let

$$K := \min\left(\frac{c}{10}, \frac{1}{10C}, \frac{1}{100}\right).$$

Since $\alpha > \frac{1}{2}$, $g(\Delta, 2) \leq K\sqrt{\Delta}$ for all sufficiently small Δ.

Assume $\Delta < 1$ is small enough that $\Gamma := g(\Delta)$ satisfies

$$\Gamma \in \left[c\Delta, K\sqrt{\Delta}\right].$$

Let $\delta := 10K\Delta \leq \frac{\Delta}{C}$. Note that $\delta \leq c\Delta \leq \Gamma$. Let

$$\mathbf{M} := \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 - 2\Gamma \end{pmatrix}, \quad \widetilde{\mathbf{M}} := \left(\mathbf{I}_3 - u_1 u_1^\top\right) \mathbf{M} \left(\mathbf{I}_3 - u_1 u_1^\top\right).$$

We prove g induces an invalid black-box cPCA regime by showing that calling two $\left(\frac{\Delta}{C}, 0\right)$-1-cPCA oracles on $\mathbf{M}$ as in Algorithm 1 fails to return a (Δ, Γ)-2-cPCA.

Define $u_1 = \left(\sqrt{1 - \delta} \quad \sqrt{\delta/2} \quad \sqrt{\delta/2}\right)^\top$, which is a $\left(\frac{\Delta}{C}, 0\right)$-1-cPCA for $\mathbf{M}$.

Let $u_2 = \left(u_{21} \quad u_{22} \quad u_{23}\right)^\top$ be the top eigenvector of $\widetilde{\mathbf{M}}$ and λ be the corresponding eigenvalue, so u_2 is a $\left(\frac{\Delta}{C}, 0\right)$-1-cPCA for $\widetilde{\mathbf{M}}$. Since $\mathbf{V}^{<(1-\Gamma)\lambda_2(\mathbf{M})}(\mathbf{M}) = e_3$, to prove $\mathbf{U} = \left(u_1 \quad u_2\right)$ is not a (Δ, Γ)-cPCA of $\mathbf{M}$ we must show

$$u_{23}^2 > \Delta. \tag{22}$$

By a direct calculation,[9] we get

$$\lambda = 1 + \frac{\delta}{2} + \frac{\delta\Gamma}{2} + \left(\frac{\sqrt{4\Gamma^2 + \delta^2 + \delta\Gamma(2\delta + \delta\Gamma - 4\Gamma)}}{2} - \Gamma\right). \tag{23}$$

---

9. We provide a calculation of the eigenvalue here.

Next, we provide upper and lower bounds on λ, which will be useful for the rest of the analysis.

$$\sqrt{4\Gamma^2 + \delta^2 + \delta\Gamma \left(2\delta + \delta\Gamma - 4\Gamma\right)} \leq \sqrt{4\Gamma^2 + \delta^2 + 4\delta\Gamma} = 2\Gamma + \delta,$$

and

$$\sqrt{4\Gamma^2 + \delta^2 + \delta\Gamma \left(2\delta + \delta\Gamma - 4\Gamma\right)} \geq \sqrt{4\Gamma^2 + \delta^2 - 4\Gamma^2\delta} \geq 2\Gamma,$$

where the last inequality used $\delta = 10K\Delta \geq 4K^2\Delta \geq 4\Gamma^2$. Using these bounds in (23), we have

$$1 + \frac{\delta}{2} < \lambda < 1 + \frac{3\delta}{2}. \tag{24}$$

Since u_1 and u_2 are distinct eigenvectors of $\widetilde{\mathbf{M}} \succeq 0$, they are orthogonal to each other. Therefore,

$$\lambda u_2 = \widetilde{\mathbf{M}} u_2 = \left(\mathbf{I}_3 - u_1 u_1^\top\right) \mathbf{M} \left(\mathbf{I}_3 - u_1 u_1^\top\right) u_2 = \mathbf{M} u_2 - \left(u_1^\top \mathbf{M} u_2\right) u_1.$$

Since $\mathbf{M}$ is diagonal, rearranging the above equality and writing u_1, u_2 in the canonical basis,

$$\begin{pmatrix} (2 - \lambda)\, u_{21} \\ (1 - \lambda)\, u_{22} \\ (1 - 2\Gamma - \lambda)\, u_{23} \end{pmatrix} = \left(u_1^\top \mathbf{M} u_2\right) \begin{pmatrix} \sqrt{1 - \delta} \\ \sqrt{\frac{\delta}{2}} \\ \sqrt{\frac{\delta}{2}} \end{pmatrix}.$$

Using the bound on λ from (24), the above equation implies that $u_1^\top \mathbf{M} u_2, u_{21}, u_{22}$, and u_{23} are all nonzero, and $\lambda \notin \{2, 1, 1 - 2\Gamma\}$. Therefore,

$$\frac{u_{21}}{u_{22}} = \sqrt{\frac{2\left(1 - \delta\right)}{\delta}} \left(\frac{1 - \lambda}{2 - \lambda}\right), \quad \frac{u_{23}}{u_{22}} = \left(\frac{\lambda - 1}{\lambda - 1 + 2\Gamma}\right). \tag{25}$$

Taking absolute values and using the bounds on λ from (24),

$$\left|\frac{u_{21}}{u_{22}}\right| \leq \sqrt{\frac{2\left(1 - \delta\right)}{\delta}} \left(\frac{\left(1 + \frac{3\delta}{2}\right) - 1}{2 - \left(1 + \frac{3\delta}{2}\right)}\right) = \sqrt{\frac{2\left(1 - \delta\right)}{\delta}} \left(\frac{3\delta}{2 - 3\delta}\right) \leq 1, \tag{26}$$

and

$$\left|\frac{u_{23}}{u_{22}}\right| = \frac{\lambda - 1}{\lambda - 1 + 2\Gamma} \leq 1, \tag{27}$$

where (26) used that $\delta = 10K\Delta \leq \frac{1}{10}$. Then, (26) and (27) imply $|u_{22}| = \max\{|u_{21}|, |u_{22}|, |u_{23}|\}$. Since u_2 is a unit vector, $u_{22}^2 \geq \frac{1}{3}$. Using (24) and (25), our claim (22) follows as shown below:

$$u_{23}^2 = \left(\frac{\lambda - 1}{\lambda - 1 + 2\Gamma}\right)^2 u_{22}^2 \geq \left(\frac{\left(1 + \frac{\delta}{2}\right) - 1}{\left(1 + \frac{\delta}{2}\right) - 1 + 2\Gamma}\right)^2 \cdot \frac{1}{3}$$

$$= \frac{\delta^2}{3\left(\delta + 4\Gamma\right)^2} \geq \frac{\delta^2}{75\Gamma^2} \geq \frac{100K^2\Delta^2}{75\left(K\sqrt{\Delta}\right)^2} > \Delta.$$

$\blacksquare$

Appendix C. Applications

In this section, we give applications of our main results, Theorems 4 and 9, to the design of k-PCA algorithms for statistical problems. In Section C.1, we begin by proving several preliminary results which will be used throughout our applications. We provide our results on robust k-PCA under various distributional assumptions, Theorems 10 and 11, in Section C.2. Finally, in Section C.3, we develop a simple k-cPCA algorithm in a heavy-tailed online setting in Theorem 52.

C.1. Preliminaries

In this section, we establish some results about the clipping of sampled data points that will be used to handle heavy tails in subsequent applications. We first define the standard notion of (p, C_p)-hypercontractivity (see, for example, Mendelson and Zhivotovskiy (2018) where it was used in a similar context).

Definition 35 ((p, C_p)-hypercontractivity) *Let $p \geq 4$ be a positive even integer. A distribution $\mathcal{D}$ over $\mathbb{R}^d$ is said to be (p, C_p)-hypercontractive if for all $v \in \mathbb{R}^d$,*

$$\mathbb{E}_{X \sim \mathcal{D}} \left[\langle X - \mathbb{E}[X], v\rangle^p\right]^{\frac{1}{p}} \leq C_p \mathbb{E}_{X \sim \mathcal{D}} \left[\langle X - \mathbb{E}[X], v\rangle^2\right]^{\frac{1}{2}}.$$

We observe that Definition 35 is compatible with assuming $\mathcal{D}$ is centered and symmetric, with no loss in parameters. Namely, let $\{X_i\}_{i \in [n]} \sim \mathcal{D}$ with covariance $\mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^\top] = \Sigma$. If $\mathcal{D}$ is not symmetric or centered, we can consider the $\frac{n}{2}$ independent variables

$$Y_i := \frac{X_{2i-1} - X_{2i}}{\sqrt{2}}, \ i \in \left[1, \left\lfloor \frac{n}{2} \right\rfloor\right].$$

Indeed, if X and X' are independent with probability distribution $\mathcal{D}$ and $Y := (X - X')/\sqrt{2}$, then $\mathbb{E}Y = \mathbb{0}_d, \mathbb{E}[YY^\top] = \Sigma$, Y is symmetric about $\mathbb{0}_d$, and

$$\mathbb{E}_{X,X' \sim \mathcal{D}} \left[\langle Y, v\rangle^p\right]^{\frac{1}{p}} = \frac{1}{\sqrt{2}} \mathbb{E}_{X,X' \sim \mathcal{D}} \left[\left(\langle X - \mathbb{E}X, v\rangle - \langle X' - \mathbb{E}X', v\rangle\right)^p\right]^{\frac{1}{p}}$$

$$\leq \frac{1}{\sqrt{2}} \mathbb{E}_{X,X' \sim \mathcal{D}} \left[2^{p-1}\left(\langle X - \mathbb{E}X, v\rangle^p + \langle X' - \mathbb{E}X', v\rangle^p\right)\right]^{\frac{1}{p}}$$

$$= \sqrt{2} \mathbb{E}_{X \sim \mathcal{D}} \left[\langle X - \mathbb{E}X, v\rangle^p\right]^{\frac{1}{p}} \leq C_p \mathbb{E}_{X \sim \mathcal{D}} \left[2\langle X - \mathbb{E}X, v\rangle^2\right]^{\frac{1}{2}}$$

$$= C_p \mathbb{E}_{X,X' \sim \mathcal{D}} \left[\langle X - X', v\rangle^2\right]^{\frac{1}{2}},$$

where the inequality follows from Fact 4. So, Y also has a (p, C_p)-hypercontractive distribution. Therefore, in applications for hypercontractive $\mathcal{D}$, we assume without loss that $\mathcal{D}$ is centered and symmetric. We next define a clipping operation relevant to our heavy-tailed applications.

Definition 36 (Clipping) *Let $R \in \mathbb{R}_{>0}$. The R-clipping function $\mathcal{T}_R : \mathbb{R}^d \to \mathbb{R}^d$ is defined as*

$$\mathcal{T}_R(x) := \min\left(1, \frac{\sqrt{R}}{\|x\|_2}\right) v$$

for all $x \in \mathbb{R}^d$. Given distribution $\mathcal{D}$ over $\mathbb{R}^d$, the R-clipped distribution of $\mathcal{D}$ is defined as $\mathcal{T}_R(\mathcal{D})$.

We will often use $\mathcal{T}$ for the R-clipping function if the clipping radius R is obvious from context. In the next result, we analyze the bias due to the clipping function.

Lemma 37 (Bias of clipping) *Let $\mathcal{D}$ be (p, C_p)-hypercontractive over $\mathbb{R}^d$ with covariance Σ.*

1. *For all $u \in \mathbb{R}^d$, $u^\top (\Sigma - \mathbb{E}_{x \sim \mathcal{D}}[\mathcal{T}_R(x)\mathcal{T}_R(x)^\top])u \leq C_p^p \left(u^\top \Sigma u\right) \left(\frac{\mathrm{Tr}(\Sigma)}{R}\right)^{\frac{p}{2}-1}$.*

2. *For $x \sim \mathcal{D}$, $\Pr(\|x\|_2 \geq \sqrt{R}) \leq (C_p(\frac{\mathrm{Tr}(\Sigma)}{R})^{\frac{1}{2}})^{(p-2)}$.*

Proof For $x \sim \mathcal{D}$ and any unit vector $u \in \mathbb{R}^d$ we have,

$$
\begin{aligned}
\mathbb{E}\left[\langle u, x\rangle^2 - \langle u, \mathcal{T}_R(x)\rangle^2\right] &= \mathbb{E}\left[\langle u, x\rangle^2 \left(1 - \frac{R}{\|x\|_2^2}\right) \mathbb{1}\left(\|x\|_2 \geq \sqrt{R}\right)\right] \\
&\leq \mathbb{E}[\langle u, x\rangle^p]^{\frac{2}{p}} \mathbb{E}\left[\left(\mathbb{1}\left(\|x\|_2 \geq \sqrt{R}\right)\right)^{\frac{p}{p-2}}\right]^{\frac{p-2}{p}} \\
&\leq C_p^2 \mathbb{E}\left[\langle u, x\rangle^2\right] \mathbb{E}\left[\mathbb{1}\left(\|x\|_2 \geq \sqrt{R}\right)\right]^{\frac{p-2}{p}} \\
&\leq C_p^2 \mathbb{E}\left[\langle u, x\rangle^2\right] \Pr\left(\|x\|_2 \geq \sqrt{R}\right)^{\frac{p-2}{p}} \\
&\leq C_p^2 \mathbb{E}\left[\langle u, x\rangle^2\right] \left(\frac{\mathbb{E}\left[\|x\|_2^p\right]}{R^{\frac{p}{2}}}\right)^{\frac{p-2}{p}},
\end{aligned}
\tag{28}
$$

where the second, third, and fifth lines used Hölder's inequality, (p, C_p)-hypercontractivity, and Markov's inequality respectively. Denote $r := \frac{p}{2}$ for convenience of notation. Then,

$$
\begin{aligned}
\mathbb{E}\left[\|x\|_2^p\right] = \mathbb{E}\left[\left(x^\top x\right)^r\right] &= \mathbb{E}\left[\left(\sum_{i \in [d]} x_i^2\right)^r\right] \\
&= \sum_{j_1, j_2, \ldots j_r \in [d]} \mathbb{E}\left[x_{j_1}^2 x_{j_2}^2 \ldots x_{j_r}^2\right] \\
&\leq \sum_{j_1, j_2, \ldots j_r \in [d]} \mathbb{E}\left[x_{j_1}^{2r}\right]^{\frac{1}{r}} \mathbb{E}\left[x_{j_2}^{2r}\right]^{\frac{1}{r}} \ldots \mathbb{E}\left[x_{j_r}^{2r}\right]^{\frac{1}{r}}, \text{ using Holder's inequality} \\
&= \sum_{j_1, j_2, \ldots j_r \in [d]} \mathbb{E}\left[\langle x, e_{j_1}\rangle^{2r}\right]^{\frac{1}{r}} \mathbb{E}\left[\langle x, e_{j_2}\rangle^{2r}\right]^{\frac{1}{r}} \ldots \mathbb{E}\left[\langle x, e_{j_r}\rangle^{2r}\right]^{\frac{1}{r}} \\
&\leq C_p^{2r} \sum_{j_1, j_2, \ldots j_r \in [d]} \mathbb{E}\left[\langle x, e_{j_1}\rangle^2\right] \mathbb{E}\left[\langle x, e_{j_2}\rangle^2\right] \ldots \mathbb{E}\left[\langle x, e_{j_r}\rangle^2\right], \text{ by hypercontractivity} \\
&= C_p^{2r} \sum_{j_1, j_2, \ldots j_r \in [d]} \left(e_{j_1}^\top \Sigma e_{j_1}\right) \left(e_{j_2}^\top \Sigma e_{j_2}\right) \ldots \left(e_{j_r}^\top \Sigma e_{j_r}\right) \\
&= C_p^{2r} \left(\sum_{i \in [d]} e_i^\top \Sigma e_i\right)^r = C_p^p \mathrm{Tr}(\Sigma)^{\frac{p}{2}}.
\end{aligned}
$$

Substituting this into (28) we have proven the second claim, and

$$\mathbb{E}\left[\langle u, x\rangle^2 - \langle u, \mathcal{T}_R(x)\rangle^2\right] \leq C_p^2 \mathbb{E}\left[\langle u, x\rangle^2\right]\left(\frac{C_p^p \mathrm{Tr}(\boldsymbol{\Sigma})^{\frac{p}{2}}}{R^{\frac{p}{2}}}\right)^{\frac{p-2}{p}} = C_p^p \mathbb{E}\left[\langle u, x\rangle^2\right]\left(\frac{\mathrm{Tr}(\boldsymbol{\Sigma})}{R}\right)^{\frac{p}{2}-1}$$

which completes our proof of the first claim as well. ∎

As an immediate consequence of Lemma 37 we have the following corollary.

Corollary 38 *In the setting of Lemma 37, for any $\rho \in (0, \frac{1}{2})$ and $R \geq (\frac{C_p^p}{\rho})^{\frac{2}{p-2}} \mathrm{Tr}(\boldsymbol{\Sigma})$, we have*

1. $\|\mathbb{E}_{x\sim\mathcal{D}}[\mathcal{T}_R(x)\mathcal{T}_R(x)^\top] - \boldsymbol{\Sigma}\|_{\mathrm{op}} \leq \rho\|\boldsymbol{\Sigma}\|_{\mathrm{op}}$.

2. $\mathrm{Pr}_{x\sim\mathcal{D}}(\|x\|_2 \geq \sqrt{R}) \leq (\frac{\rho}{C_p^2})^{\frac{p}{p-2}}$.

3. *For all $v \in \mathbb{R}^d$, $\mathbb{E}_{x\sim\mathcal{D}}[\langle\mathcal{T}_R(x), v\rangle^p]^{\frac{1}{p}} \leq 2C_p \mathbb{E}_{x\sim\mathcal{D}}[\langle\mathcal{T}_R(x), v\rangle^2]^{\frac{1}{2}}$.*

Proof *Items 1 and 2 follow directly from substituting the value of R in Lemma 37. We next prove Item 3. For $x \sim \mathcal{D}$ we upper bound the pth moment of one-dimensional projections:*

$$\begin{aligned}
\mathbb{E}\left[\langle\mathcal{T}_R(x), v\rangle^p\right] &= \mathbb{E}\left[\langle x, v\rangle^p\left(\mathbb{1}\left(\|X\|_2 \leq \sqrt{R}\right) + \left(\frac{\sqrt{R}}{\|x\|_2}\right)^p \mathbb{1}\left(\|X\|_2 \geq \sqrt{R}\right)\right)\right]\\
&\leq 2\mathbb{E}\left[\langle x, v\rangle^p\right]\\
&\leq 2C_p^p \mathbb{E}\left[\langle x, v\rangle^2\right]^{\frac{p}{2}}, \ \textit{using } (p, C_p)\textit{-hypercontractivity of } X
\end{aligned} \tag{29}$$

We now obtain lower bounds on the second moments of one-dimensional projections:

$$\begin{aligned}
\mathbb{E}\left[\langle x, v\rangle^2 - \langle\mathcal{T}_R(x), v\rangle^2\right] &\leq C_p^p\left(u^\top \boldsymbol{\Sigma} u\right)\left(\frac{\mathrm{Tr}(\boldsymbol{\Sigma})}{R}\right)^{\frac{p}{2}-1} \ \textit{using Lemma 37,}\\
&\leq C_p^2 \mathbb{E}\left[\langle X, v\rangle^2\right]\left(\left(\frac{\rho}{C_p^2}\right)^{\frac{p}{p-2}}\right)^{\frac{p-2}{p}} = \rho\mathbb{E}\left[\langle X, v\rangle^2\right].
\end{aligned}$$

Substituting this into (29) completes our proof:

$$\mathbb{E}\left[\langle\mathcal{T}_R(x), v\rangle^p\right] \leq 2C_p^p\left(\frac{\mathbb{E}\left[\langle x, v\rangle^2\right]}{1-\rho}\right)^{\frac{p}{2}} \leq (2C_p)^p \mathbb{E}\left[\langle x, v\rangle^2\right]^{\frac{p}{2}}.$$

∎

C.2. Robust PCA

In this section, we develop algorithms for robust PCA. The robust PCA problem asks us to output an approximate PCA of a covariance matrix from independent draws from the inducing distribution, even after a fraction of draws are corrupted. We define our contamination model formally below.

Definition 39 (Strong contamination model) *Given a* corruption parameter $\epsilon \in (0, \frac{1}{2})$ *and a distribution* $\mathcal{P}$, *an algorithm obtains samples from* $\mathcal{P}$ *with* ϵ-contamination *as follows.*

1. *The algorithm specifies the number* n *of samples it requires.*

2. *A set* S *of* n *i.i.d. samples from* $\mathcal{P}$ *is drawn but not yet shown to the algorithm.*

3. *An arbitrarily powerful adversary then inspects* S, *before deciding to replace any subset of* $\lceil \epsilon n \rceil$ *samples with arbitrarily corrupted points ("outliers") to obtain the contaminated set* T, *which is then returned to the algorithm.*

We say T *is an* ϵ-corrupted version *of* S *and a set of* ϵ-corrupted samples *from* $\mathcal{P}$.

Robust estimation typically requires that the inliers satisfy structural properties (that hold with high probability) that distinguishes them from harmful outliers. In the context of PCA, we use the following stability condition from Jambulapati et al. (2020); Diakonikolas et al. (2023) to design our algorithms.

Definition 40 (Stability) *For* $\epsilon \in (0, \frac{1}{2})$ *and* $\gamma \geq \epsilon$, *we say a distribution* G *on* $\mathbb{R}^d$ *is* (ϵ, γ)-stable *with respect to* $\mathbf{\Sigma} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$ *if for all functions* $w : \mathbb{R}^d \to [0, 1]$ *such that* $\mathbb{E}_{X \sim G}[w(X)] \geq 1 - \epsilon$, *the weighted second moment matrix* $\mathbf{\Sigma}_{G_w} := \frac{\mathbb{E}_{X \sim G}[w(X)XX^\top]}{\mathbb{E}_{X \sim G}[w(X)]}$ *satisfies* $(1 - \gamma)\mathbf{\Sigma} \preceq \mathbf{\Sigma}_{G_w} \preceq (1 + \gamma)\mathbf{\Sigma}$.

In Jambulapati et al. (2020); Diakonikolas et al. (2023), the distribution G in Definition 40 corresponds to the uniform distribution over the remaining inliers, i.e. $S \cap T$ in Theorem 39. In this notation, the uniform distribution on the corrupted set T can be written as $(1 - \epsilon)G + \epsilon B$, where B corresponds to the uniform distribution over the outliers $T \setminus S$. Under the stability condition stated in Definition 40, the following result from Diakonikolas et al. (2023) gives a generic 1-ePCA algorithm that runs in nearly-linear time.

Proposition 41 ((Diakonikolas et al., 2023, Theorem 3.1)) *Let* ϵ_0, γ_0 *be sufficiently small absolute constants. Let* T *be a set of* n *data points in* $\mathbb{R}^d$. *For* $\epsilon \in (0, \epsilon_0)$ *and* $\gamma \in (0, \gamma_0)$, *suppose the uniform distribution on* T *can be written as* $(1 - \epsilon)G + \epsilon B$, *where* G *is* (ϵ, γ)-stable *with respect to* $\mathbf{\Sigma}$. *There is an algorithm* $\mathcal{A}_1$ *taking* T, ϵ, γ, *and* $\delta \in (0, 1)$ *as inputs.* $\mathcal{A}_1$ *outputs a unit vector* $v \in \mathbb{R}^d$ *such that, with probability* $\geq 1 - \delta$, v *is an* $O(\gamma)$ *1-ePCA of* $\mathbf{\Sigma}$, *and* v *lies in the span of* T, *in time*

$$O\left(\frac{nd}{\gamma^2}\text{polylog}\left(\frac{d}{\epsilon\delta}\right)\right).$$

In the context of our reduction in Theorem 4, we will repeatedly invoke the above algorithm $\mathcal{A}$ on deflated data, i.e. after projecting out all reported principal components so far. Conveniently, stability is preserved under arbitrary deflations, as shown below.

Lemma 42 *Let* $S \subset \mathbb{R}^d$ *be such that the uniform distribution over* S *is* (ϵ, γ)-stable *with respect to* $\mathbf{\Sigma}$. *For any* $\mathbf{P} \in \mathbb{R}^{d \times d}$, *the uniform distribution over* $\{\mathbf{P}x\}_{x \in S}$ *is* (ϵ, γ)-stable *with respect to* $\mathbf{P}\mathbf{\Sigma}\mathbf{P}$.

Proof For $\mathbf{A}, \mathbf{B} \in \mathbb{S}^{d \times d}$, $\mathbf{A} \preceq \mathbf{B}$ implies $\mathbf{P}\mathbf{A}\mathbf{P} \preceq \mathbf{P}\mathbf{B}\mathbf{P}$ for all $\mathbf{P} \in \mathbb{R}^{d \times d}$, giving the claim. ∎

Thus, Theorem 41 is compatible with our framework in Theorem 4 and can be used as a 1-ePCA oracle for $\mathbf{\Sigma}$. Combining this observation with Theorem 4, we obtain nearly-linear time algorithms

for robust k-ePCA. Importantly, the sample complexity of our robust k-ePCA algorithm does not increase with k, as we can reuse the same samples in each call to Theorem 41 due to Lemma 42.

Corollary 43 (k-ePCA in nearly-linear time under stability) *In the setting of Proposition 41, there is an algorithm $\mathcal{A}_k$ taking T, ϵ, γ, $\delta \in (0, 1)$, and $k \in [d]$ as inputs. $\mathcal{A}_k$ outputs orthonormal $\mathbf{U} \in \mathbb{R}^{d \times k}$, such that, with probability $\geq 1 - \delta$, $\mathbf{U}$ is an $O(\gamma)$-k-ePCA of $\boldsymbol{\Sigma}$, in time*

$$O\left(\frac{ndk}{\gamma^2}\text{polylog}\left(\frac{d}{\epsilon\delta}\right)\right).$$

Theorem 43 also leads to k-cPCA guarantees using Theorem 7 and Theorem 9 in some parameter regimes, but we focus on ePCA for compatibility with Theorem 41. In Sections C.2.1 and C.2.2, we apply Corollary 43 to two fundamental distribution families: sub-Gaussian distributions and hypercontractive distributions. Our results follow via establishing appropriate stability conditions.

C.2.1. SUB-GAUSSIAN DISTRIBUTIONS

In this section, we show that by leveraging known stability conditions for sub-Gaussian distributions, we can directly apply Corollary 43 to design a robust k-PCA algorithm. We first give a definition of the sub-Gaussian distribution family under consideration.

Definition 44 (Sub-Gaussian distribution) *We say a distribution $\mathcal{D}$ on $\mathbb{R}^d$ is r-sub-Gaussian for a parameter $r \geq 1$ if it has mean $\mathbb{0}_d$ and covariance $\boldsymbol{\Sigma}$,[10] and for all unit vectors $v \in \mathbb{R}^d$ and $t \in \mathbb{R}$,*

$$\mathbb{E}_{X \sim D}\left[\exp(tv^\top X)\right] \leq \exp\left(\frac{t^2 r}{2}v^\top \boldsymbol{\Sigma} v\right).$$

Next, we present a result from Jambulapati et al. (2020) bounding the sample complexity required for the uniform distribution over a set of corrupted samples to be stable with respect to a covariance matrix.

Lemma 45 (Corollary 4, Jambulapati et al. (2020)) *Let $\mathcal{D}$ be an $O(1)$-sub-Gaussian distribution on $\mathbb{R}^d$ with covariance $\boldsymbol{\Sigma}$. Let $\epsilon \in (0, \epsilon_0)$ for an absolute constant ϵ_0, $\delta \in (0, 1)$, and for an absolute constant C, let $\gamma := C\epsilon\log(\frac{1}{\epsilon})$. If S is a set of n i.i.d. samples from $\mathcal{D}$ where, for an appropriate constant,*

$$n = \Theta\left(\frac{d + \log(\frac{1}{\delta})}{\gamma^2}\right),$$

then with probability $\geq 1 - \delta$, the uniform distribution over S is (ϵ, γ) stable with respect to $\boldsymbol{\Sigma}$.

Applying Theorem 41 with Lemma 45 in hand, (Diakonikolas et al., 2023, Theorem 1.2) obtained a nearly-linear time algorithm for robust 1-ePCA.[11] By simply replacing the use of Theorem 41 with Theorem 43, we further obtain the following robust k-PCA result for sub-Gaussian distributions.

---

10. The same symmetrization strategy as discussed in Section C.1 shows that the assumption that $\mathcal{D}$ is mean-zero is without loss of generality. We defer additional discussion to Jambulapati et al. (2020).

11. Theorem 41 is applicable because if the uniform distribution over a set S is (ϵ, γ)-stable, then the uniform distribution over $S' \subset S$ with $|S'| \geq (1 - \epsilon)|S|$ is also $(\epsilon, O(\gamma))$-stable; we apply this with $S' = S \cap T$.

Theorem 10 (Robust sub-Gaussian k-ePCA) *Let $\mathcal{D}$ be an $O(1)$-sub-Gaussian distribution on $\mathbb{R}^d$ with covariance Σ. Let $\epsilon \in (0, \epsilon_0)$ for an absolute constant ϵ_0, and $\delta \in (0, 1)$. Let T be an ϵ-corrupted set of samples from $\mathcal{D}$ with $|T| = \Theta(\frac{d + \log(1/\delta)}{\epsilon^2 \log^2(1/\epsilon)})$ for an appropriate constant. Algorithm $\mathcal{A}_k$ (Corollary 43) run on inputs T, ϵ, $\gamma = \Theta(\epsilon \log(\frac{1}{\epsilon}))$, δ, and $k \in [d]$ outputs orthonormal $\mathbf{U} \in \mathbb{R}^{d \times k}$ such that, with probability $\geq 1 - \delta$, $\mathbf{U}$ is an $O(\epsilon \log(\frac{1}{\epsilon}))$-$k$-ePCA of Σ, in time $O(\frac{ndk}{\epsilon^2} \mathrm{polylog}(\frac{d}{\epsilon\delta}))$.*

As discussed previously, the sample complexity of Theorem 10 notably does not grow as k increases.

C.2.2. HYPERCONTRACTIVE DISTRIBUTIONS

In this section, we relax the sub-Gaussianity assumption used in Jambulapati et al. (2020); Diakonikolas et al. (2023) and instead assume that the data is drawn from a (p, C_p)-hypercontractive distribution (Definition 35). Our main result in this section is the following algorithm for k-ePCA in the hypercontractive setting.

Theorem 11 (Robust hypercontractive k-ePCA) *For an even integer $p \geq 4$, let $\mathcal{D}$ be (p, C_p)-hypercontractive on $\mathbb{R}^d$ with mean $\mathbb{0}_d$ and covariance Σ. Let $\epsilon \in (0, \epsilon_0)$, $\delta \in (0, 1)$, and $\gamma = \Theta(C_p^2 \epsilon^{1 - \frac{2}{p}})$ such that $\gamma \in (0, \gamma_0)$ for absolute constants ϵ_0, γ_0. Let T be an ϵ-corrupted set of samples from $\mathcal{D}$ with $|T| = \Theta(\beta(\frac{d \log d + \log(1/\delta)}{\gamma^2}))$ for an appropriate constant, where $\beta := C_p^6 \epsilon^{-\frac{2}{p}}$. Algorithm $\mathcal{A}_k$ (Corollary 43) run on inputs T, ϵ, γ, δ, and $k \in [d]$ outputs orthonormal $\mathbf{U} \in \mathbb{R}^{d \times k}$ such that, with probability $\geq 1 - \delta$, $\mathbf{U}$ is an $O(\gamma)$-k-ePCA of Σ, in time $O(\frac{ndk}{\gamma^2} \mathrm{polylog}(\frac{d}{\epsilon\delta}))$.*

We mention that the approximation factor of $\gamma = \Theta(C_p^2 \epsilon^{1 - \frac{2}{p}})$ achieved by Theorem 11 is optimal under ϵ-corruption for a (p, C_p)-hypercontractive distribution (cf. Theorem 48). Further, the sample complexity of Theorem 11 differs from the information-theoretic sample complexity by a factor of $\beta = C_p^6 \epsilon^{-\frac{2}{p}}$ and a $\log d$ term (Liu et al., 2022a, Proposition 6.7). In particular, for constant C_p, ϵ, the sample complexity of Theorem 11 nearly-matches that of Gaussian data even for heavy-tailed distributions, while being robust to a constant fraction of corruption and running in nearly-linear time.

At the end of the section, we give a detailed comparison between Theorem 11 and the closest related work of Kong et al. (2020), which handled distributions with bounded fourth moments under an assumption related to (but not directly comparable to) our assumption of $(4, C_4)$-hypercontractivity.

Theorem 11 follows by establishing that a large enough set of samples from hypercontractive distributions satisfy the conditions of Theorem 43. To establish stability, one could hope that the samples drawn from $\mathcal{D}$ are stable as is (without removing any ϵ-fraction). A common approach towards this goal would be using matrix Chernoff bounds, which would lead to a multiplicative dependence on $\log(\frac{1}{\delta})$, where δ is the failure probability. Instead, we shall use the following result from Diakonikolas et al. (2020), phrased in our notation, that does a more white-box analysis and uses the flexibility permitted by removing $O(\epsilon)$-fraction of samples to establish a sharper dependence on $\log(\frac{1}{\delta})$.

Lemma 46 (Lemma 4.1 and 4.2, Diakonikolas et al. (2020)) *Let $\mathcal{D}'$ be a distribution on $\mathbb{R}^d$ satisfying uncentered hypercontractivity, i.e. for an even integer $p \geq 4$, for all $v \in \mathbb{R}^d$,*

$$(\mathbb{E}_{x \sim \mathcal{D}'}[\langle x, v \rangle^p])^{\frac{1}{p}} \leq \sigma_p(\mathbb{E}_{x \sim \mathcal{D}'}[\langle x, v \rangle^2])^{\frac{1}{2}}.$$

Let $\epsilon \in (0, \frac{1}{2})$ and $\delta \in (0, 1)$. Suppose $\mathcal{D}'$ is supported on a ball of radius $\sqrt{R} \geq 1$, and $\Sigma' := \mathbb{E}_{x \sim \mathcal{D}'}[xx^\top]$ satisfies $\|\Sigma'\|_{\mathrm{op}} = O(1)$. Let S be a set of n i.i.d. samples from $\mathcal{D}'$, for

$$n = \Theta\left(\frac{R \log d}{\alpha^2} + \frac{\sigma_p^4 \log(\frac{1}{\delta})}{\alpha^2} + \frac{\sigma_p^2 \log(\frac{1}{\delta})}{\alpha \epsilon^{\frac{2}{p}}} + \sqrt{\frac{d \log(\frac{1}{\delta})}{\epsilon^2 \alpha}} + \frac{\log(\frac{1}{\delta})}{\epsilon} \right),$$

for an appropriate constant, and $\alpha \in (0, 1)$. Then with probability $\geq 1 - \delta$, the uniform distribution on S can be written as $(1 - \epsilon)P + \epsilon B$, where P is (ϵ, γ)-stable with respect to $\mathbf{I}_d$, for

$$\gamma = O\left(\|\Sigma' - \mathbf{I}_d\|_{\mathrm{op}} + \sigma_p^2 \epsilon^{1 - \frac{2}{p}} + \alpha \right).$$

We are now ready to give our sample complexity bound for hypercontractive stability.

Proposition 47 *For an even integer $p \geq 4$, let $\mathcal{D}$ be a (p, C_p)-hypercontractive on $\mathbb{R}^d$ with mean $\mathbb{0}_d$ and covariance Σ. Let $\epsilon \in (0, \epsilon_0)$ for an absolute constant ϵ_0, $\delta \in (0, 1)$, and $\gamma = \Theta(\sigma_p^2 \epsilon^{1 - \frac{2}{p}})$ be such that $\gamma \leq \frac{1}{2}$. If S is a set of n i.i.d. samples from $\mathcal{D}$ where, for an appropriate constant,*

$$n = \Theta\left(C_p^2 \epsilon^{-\frac{2}{p}} \cdot \frac{d \log(d)}{\gamma^2} + \max(C_p^4, C_p^8 \epsilon^{1 - \frac{2}{p}}) \cdot \frac{\log(\frac{1}{\delta})}{\gamma^2} \right),$$

then with probability $\geq 1 - \delta$, the uniform distribution on S can be written as $(1 - \epsilon)P + \epsilon B$ for two distributions P and B such that P is (ϵ, γ)-stable with respect to Σ.

Proof We first observe that if the uniform distribution over a set S' satisfies stability with respect to $\mathbf{I}_d$, then $S := \{\Sigma^{\frac{1}{2}}x\}_{x \in S'}$ satisfies stability with respect to Σ. We let the samples in S' be drawn as $\Sigma^{-\frac{1}{2}}x$ for $x \sim \mathcal{D}$, and thus follow an isotropic distribution. Moreover, it is straightforward to see that this transformed distribution also satisfies (p, C_p)-hypercontractivity. Thus, we assume $\mathcal{D}$ is isotropic in the rest of the proof and establish stability with respect to $\mathbf{I}_d$.

Next, we consider a clipped variant of $\mathcal{D}$. Let $\mathcal{D}'$ be the distribution of $\mathcal{T}_R(x)$ for $R = (\frac{C_p^p}{\rho})^{\frac{2}{p-2}} d$ and $x \sim \mathcal{D}$, so $\mathcal{D}'$ is supported on a ball of radius $\sqrt{R} \geq 1$. Let $\Sigma' := \mathbb{E}_{x \sim \mathcal{D}'}[xx^\top]$. We shall choose $\rho = C' C_p^2 \epsilon^{1 - \frac{2}{p}} \leq \frac{\gamma}{2}$ for a large enough constant C', such that $\rho \leq 1$ by an upper bound on γ. By Theorem 38, $\|\Sigma' - \Sigma\|_{\mathrm{op}} = \|\Sigma' - \mathbf{I}_d\|_{\mathrm{op}} \leq \rho$ and the probability that a sample from $\mathcal{D}$ is clipped is $\leq \frac{\epsilon}{100}$. In particular, by a Chernoff bound, this second conclusion shows that with probability at least $1 - \exp(-\Omega(n\epsilon)) \geq 1 - \frac{\delta}{2}$, less than an $\frac{\epsilon}{2}$ fraction of samples will be clipped. We let the clipped samples form the distribution B. In the sequel, we prove stability of the unclipped samples.

Recall that $\mathcal{D}'$ has second moment matrix Σ' with $\|\Sigma'\|_{\mathrm{op}} \leq 2$, and $\mathcal{D}'$ satisfies uncentered hypercontractivity with $\sigma_p = 2C_p$ by Theorem 38. We hence may apply Lemma 46 with $\epsilon \leftarrow \frac{\epsilon}{2}$, $\delta \leftarrow \frac{\delta}{2}$,

and $\alpha \leftarrow \rho = \Theta(C_p^2 \epsilon^{1-\frac{2}{p}})$, which again lies in $(0,1)$. Since $\rho \geq \|\Sigma' - \mathbf{I}_d\|_{\mathrm{op}}$, with probability $1 - \frac{\delta}{2}$, P is $(\frac{\epsilon}{2}, \gamma)$-stable for $\gamma = O(C_p^2 \epsilon^{1-\frac{2}{p}})$. Finally, we bound the sample complexity:

$$
n = O\left(\frac{R \log d}{\alpha^2} + \frac{\sigma_p^4 \log(\frac{1}{\delta})}{\alpha^2} + \frac{\sigma_p^2 \log(\frac{1}{\delta})}{\alpha \epsilon^{\frac{2}{p}}} + \sqrt{\frac{d \log(\frac{1}{\delta})}{\epsilon^2 \alpha}} + \frac{\log(\frac{1}{\delta})}{\epsilon} \right)
$$

$$
= O\left(\left(\frac{C_p^p}{\rho}\right)^{\frac{2}{p-2}} \frac{d \log d}{\gamma^2} + \frac{C_p^4 \log(\frac{1}{\delta})}{\gamma^2} + \frac{\log(\frac{1}{\delta})}{\gamma \epsilon^{\frac{2}{p}}} + \sqrt{\frac{d}{\gamma^2} \frac{\log(\frac{1}{\delta})}{(\epsilon^2/\gamma)}} + \frac{\log(\frac{1}{\delta})}{\epsilon} \right)
$$

$$
\text{(using the value of } R)
$$

$$
= O\left(C_p^2 \epsilon^{-\frac{2}{p}} \frac{d \log d}{\gamma^2} + \frac{C_p^4 \log(\frac{1}{\delta})}{\gamma^2} + \frac{\log(\frac{1}{\delta})}{C_p^2 \epsilon} + \sqrt{\frac{d}{\gamma^2} \frac{\log(\frac{1}{\delta})}{(\epsilon^2/\gamma)}} + \frac{\log(\frac{1}{\delta})}{\epsilon} \right)
$$

$$
\text{(using the value of } \gamma, \rho)
$$

$$
= O\left(C_p^2 \epsilon^{-\frac{2}{p}} \frac{d \log d}{\gamma^2} + \frac{C_p^4 \log(\frac{1}{\delta})}{\min(\gamma^2, C_p^2 \epsilon)} + C_p^2 \epsilon^{-\frac{2}{p}} \frac{d \log d}{\gamma^2} + \frac{\log(\frac{1}{\delta})}{C_p^2 \epsilon^{-\frac{2}{p}}(\epsilon^2/\gamma)} \right)
$$

$$
\text{(using } \sqrt{ab} \leq 2a + 2b)
$$

$$
= O\left(C_p^2 \epsilon^{-\frac{2}{p}} \frac{d \log d}{\gamma^2} + \frac{C_p^4 \log(\frac{1}{\delta})}{\min(\gamma^2, C_p^2 \epsilon, C_p^2 \epsilon^{-\frac{2}{p}}(\epsilon^2/\gamma))} \right)
$$

$$
= O\left(C_p^2 \epsilon^{-\frac{2}{p}} \frac{d \log d}{\gamma^2} + C_p^4 \frac{\log(\frac{1}{\delta})}{\min(\gamma^2, \epsilon)} \right) \qquad \text{(using the value of } \gamma)
$$

$$
= O\left(C_p^2 \epsilon^{-\frac{2}{p}} \frac{d \log d}{\gamma^2} + \max(C_p^4, C_p^8 \epsilon^{1-\frac{2}{p}}) \frac{\log(\frac{1}{\delta})}{\gamma^2} \right). \qquad \text{(using the value of } \gamma)
$$

By a union bound, with probability $1 - \delta$, the uniform distribution on S can be written as $(1-\epsilon)P + \epsilon B$, where P is $(\frac{\epsilon}{2}, \gamma)$-stable. Finally, $(\frac{\epsilon}{2}, \gamma)$-stability of P implies $(\epsilon, O(\gamma))$ stability of P.[12] ∎

We now provide the proof of Theorem 11, by leveraging Proposition 47 and Corollary 43.

Proof [Proof of Theorem 11]

Let S be the set of i.i.d. samples from $\mathcal{D}$. Theorem 47 implies that, with probability $\geq 1 - \delta$, the uniform distribution on S can be written as $(1 - \epsilon)P + \epsilon B$, where P is (ϵ, γ) stable with respect to Σ. Since ϵ is small enough, the (ϵ, γ)-stability of P implies P is also $(10\epsilon, O(\gamma))$-stable. Let T be the corrupted set of samples and observe that the uniform distribution on T is $\leq 2\epsilon$-far from P in total variation. Thus, the uniform distribution on T can be written as $(1 - 2\epsilon)P' + 2\epsilon B'$, where P' conditions P on a probability $(1 - O(\epsilon))$ event and consequently inherits $(8\epsilon, \gamma)$-stability with respect to Σ from P; see, for example, (Diakonikolas et al., 2022, Lemma 2.12). Reparameterizing ϵ to 2ϵ, we see that Theorem 43 is applicable. ∎

---

12. To see this, note that stability implies every second moment matrix of a ϵ-weighted subset of the data is bounded in the Loewner order by $O(\gamma)\Sigma$, a bound which can be applied twice.

We now show that the error guarantee of Theorem 11 is optimal (up to constants) for small ϵ and $k = 1$.

Lemma 48 (Lower bound for robust hypercontractive ePCA) *Let $c > 0$ be a small enough constant and let $d \in \mathbb{N}$ be sufficiently large. Let $\mathcal{D}$ be a (p, C_p) hypercontractive distribution on $\mathbb{R}^d$ for an even integer $p \geq 4$ and $C_p > 2$, with mean $\mathbb{0}_d$ and covariance $\boldsymbol{\Sigma}$. Let $\epsilon \in (0, \epsilon_0)$ for $\epsilon_0 = (\frac{C_p^2}{8})^{\frac{p}{2}}$ and let $\gamma = C_p^2 \epsilon^{1-\frac{2}{p}}$. There is no algorithm that, with probability $\geq \frac{1}{2}$, outputs a unit vector u that is a $c\gamma$-1-ePCA of $\boldsymbol{\Sigma}$, given infinite samples and runtime.*

Proof We exhibit $d+1$ mean-$\mathbb{0}_d$, (p, C_p)-hypercontractive distributions $\mathcal{P} = \{\mathcal{D}_i\}_{0 \leq i \leq d}$. Moreover, all of the distributions $\{\mathcal{D}_i\}_{i \in [d]}$ will have total variation distance $\leq \epsilon$ from $\mathcal{D}_0$. Finally, letting $\{\boldsymbol{\Sigma}_i\}_{0 \leq i \leq d}$ be the corresponding covariances, our construction will have $\boldsymbol{\Sigma}_0 = \mathbf{I}_d$, and $\boldsymbol{\Sigma}_i = \mathbf{I}_d + s e_i e_i^\top$ for $s := \epsilon(0.25 C_p^2 \epsilon^{-\frac{2}{p}} - 1) = \Theta(C_p^2 \epsilon^{1-\frac{2}{p}}) = \Theta(\gamma)$, where we used $\epsilon \leq \epsilon_0$.

We now give our construction. Let P be the uniform distribution on $\{-1, 1\}$ and let P' be the distribution $(1 - \epsilon) + \epsilon B$, where B is the uniform distribution on $\{\pm 0.5 C_p \epsilon^{-\frac{1}{p}}\}$. Observe that both P and P' are mean-zero and (p, C_p)-hypercontractive; indeed, for $X \sim P'$, $\mathbb{E}[X^2] = 1 - \epsilon + 0.25 \epsilon C_p^2 \epsilon^{-\frac{2}{p}} = 1 + s$, while

$$\mathbb{E}[X^p] = (1 - \epsilon) + \epsilon 2^{-p} C_p^p \epsilon^{-1} \leq C_p^p (\mathbb{E}[X^2])^{\frac{p}{2}},$$

where we use that $C_p \geq 2$ and $\mathbb{E}[X^2] \geq 1$.

Define $\mathcal{D}_0$ to be the distribution on $\mathbb{R}^d$ such that coordinate is sampled independently from P. For each $i \in [d]$, define $\mathcal{D}_i$ to be the distribution on $\mathbb{R}^d$ such that all coordinates are sampled independently, the ith coordinate is $\sim P'$, and all other coordinates are $\sim P$. Since all $\mathcal{D}_i$ are product distributions, the arguments of (Kothari and Steurer, 2017, Lemma 5.9) imply that all $\mathcal{D}_i$ are also (p, C_p)-hypercontractive, and have the claimed covariances. Finally, the coupling formulation of total variation distance implies all $\{\mathcal{D}_i\}_{i \in [d]}$ are at most ϵ far from $\mathcal{D}_0$ in the total variation distance.

Thus, under these conditions, it is possible that the distribution of inliers is $\mathcal{D} = \mathcal{D}_{i^*}$ for some unknown $i^* \in [d]$, but the adversary corrupts the distribution to be $\mathcal{D}_0$. If the algorithm outputs a unit vector $\widehat{u}$ in this case, then it is an ρ-1-ePCA for $\mathcal{D}_{i^*}$ for $\rho = 1 - \frac{1 + s u_{i^*}^2}{1 + s} = \frac{s(1 - u_{i^*}^2)}{1 + s}$. Importantly, the error $\rho \geq \frac{\min(1, s)}{4}$ if $u_{i^*}^2 \leq 1/4$. Since $\mathcal{D}_0$ contains no information about the unknown index i^*, we conclude that no algorithm can output a unit vector $u \in \mathbb{R}^d$ with $u_{i^*}^2 \geq \frac{1}{2}$, for large enough d, and thus the ePCA approximation error is at least $c\gamma$ for a small constant c. ∎

Comparison to Kong et al. (2020). As mentioned previously, the most comparable result to Theorem 11 in the literature is due to Proposition 2.6 of Kong et al. (2020), which used a different set of distributional assumptions than hypercontractivity. The assumptions used in Proposition 2.6 of

Kong et al. (2020) are:[13]

$$\frac{1}{k}\mathbb{E}_{x\sim\mathcal{D}}\left[\left\langle \mathbf{U}\mathbf{U}^\top, xx^\top - \mathbf{\Sigma}\right\rangle^2\right] \le \nu(k)^2 \text{ for all orthonormal } \mathbf{U} \in \mathbb{R}^{d\times k},$$
$$\text{and } \left\|xx^\top - \mathbf{\Sigma}\right\|_{\text{op}} \le B \text{ with probability 1.} \tag{30}$$

For comparison, if $\mathcal{D}$ is a $(4, C_4)$-hypercontractive distribution,

$$\mathbb{E}_{x\sim\mathcal{D}}\left[\left\langle vv^\top, xx^\top\right\rangle^2\right] \le C_4^2 \left\langle vv^\top, \mathbf{\Sigma}\right\rangle^2 \text{ for all } v \in \mathbb{R}^d. \tag{31}$$

The assumption (30) differs in that applies to rank-k matrices (rather than rank-1), it enforces a uniform bound $\nu(k)$ (rather than a bound sensitive to v, which is tighter if the spectrum of $\mathbf{\Sigma}$ is uneven), and enforces an almost sure bound, which does not follow in general for hypercontractive distributions. We now bound the parameters in (30) under the assumption (31) for fair comparison.

Let $\mathbf{U} \in \mathbb{R}^{d\times k}$ with columns $\{u_i\}_{i\in[k]}$ be orthonormal, and suppose that (31) holds. Then,

$$\mathbb{E}_{x\sim\mathcal{D}}\left[\left\langle \mathbf{U}\mathbf{U}^\top, xx^\top - \mathbf{\Sigma}\right\rangle^2\right] = \mathbb{E}_{x\sim\mathcal{D}}\left[\left\langle \mathbf{U}\mathbf{U}^\top, xx^\top\right\rangle^2\right]$$
$$+ \mathbb{E}_{x\sim\mathcal{D}}\left[-2\left\langle \mathbf{U}\mathbf{U}^\top, xx^\top\right\rangle\left\langle \mathbf{U}\mathbf{U}^\top, \mathbf{\Sigma}\right\rangle + \left\langle \mathbf{U}\mathbf{U}^\top, \mathbf{\Sigma}\right\rangle^2\right]$$
$$= \mathbb{E}_{x\sim\mathcal{D}}\left[\left\langle \sum_{i\in[k]} u_i u_i^\top, xx^\top\right\rangle^2\right] - \left\langle \mathbf{U}\mathbf{U}^\top, \mathbf{\Sigma}\right\rangle^2$$
$$= \sum_{i\in[k]}\sum_{j\in[k]} \mathbb{E}_{x\sim\mathcal{D}}\left[\left\langle u_i, x_i\right\rangle^2 \left\langle u_j, x_j\right\rangle^2\right] - \left\langle \mathbf{U}\mathbf{U}^\top, \mathbf{\Sigma}\right\rangle^2$$
$$\le \sum_{i\in[k]}\sum_{j\in[k]} \sqrt{\mathbb{E}_{x\sim\mathcal{D}}\left[\left\langle u_i, x_i\right\rangle^4\right]\mathbb{E}_{x\sim\mathcal{D}}\left[\left\langle u_j, x_j\right\rangle^4\right]} - \left\langle \mathbf{U}\mathbf{U}^\top, \mathbf{\Sigma}\right\rangle^2$$
$$= \left(\sum_{i\in[k]} \sqrt{\mathbb{E}_{x\sim\mathcal{D}}\left[\left\langle u_i, x_i\right\rangle^4\right]}\right)^2 - \left\langle \mathbf{U}\mathbf{U}^\top, \mathbf{\Sigma}\right\rangle^2$$
$$\le \left(\sum_{i\in[k]} C_4 \left\langle u_i u_i^\top, \mathbf{\Sigma}\right\rangle\right)^2 - \left\langle \mathbf{U}\mathbf{U}^\top, \mathbf{\Sigma}\right\rangle^2$$
$$= \left(C_4^2 - 1\right)\left\langle \mathbf{U}\mathbf{U}^\top, \mathbf{\Sigma}\right\rangle^2 \le \left(C_4^2 - 1\right)\|\mathbf{\Sigma}\|_k^2.$$

The first inequality used the Cauchy-Schwarz inequality and the second inequality used (31). It follows that the value of $\nu(k)$ that can be derived in (30) satisfies

$$\nu(k) \le \sqrt{C_4^2 - 1} \cdot \frac{\|\mathbf{\Sigma}\|_k}{\sqrt{k}}.$$

---

13. Proposition 2.6 of Kong et al. (2020) assumes a stronger version of the first line in (30) where the inner product is taken against arbitrary rank-k matrices with Frobenius norm 1. However, inspection of their proof shows they only use the weaker (30), i.e. bounded inner products against matrices of the form $\frac{1}{\sqrt{k}}\mathbf{U}\mathbf{U}^\top$, which they term "semi-orthogonal."

This bound is tight up to constant factors, as witnessed by the hypercontractive mixture distribution $\frac{1}{2}\mathcal{N}(\mathbb{0}_d, \frac{1}{2}\mathbf{I}_d) + \frac{1}{2}\mathcal{N}(\mathbb{0}_d, \frac{3}{2}\mathbf{I}_d)$, whose covariance matrix is $\mathbf{\Sigma} = \mathbf{I}_d$ and whose hypercontractive parameter in (31) is a constant bounded away from 1, since the fourth moment of a standard Gaussian is 3. In this case, we claim that for any orthonormal matrix $\mathbf{U} \in \mathbb{R}^{d \times k}$, $\langle \mathbf{U}\mathbf{U}^\top, xx^\top \rangle$ deviates from its expectation by $\Omega(k)$ with constant probability, which is enough to lower bound $\nu(k)$ by $\Omega(\sqrt{k})$ by using the definition (30), since the left-hand side is $\frac{1}{k}$ times the variance of $\langle \mathbf{U}\mathbf{U}^\top, xx^\top \rangle$. The expectation of $\langle \mathbf{U}\mathbf{U}^\top, xx^\top \rangle$ is simply k. On the other hand, if $x \sim \mathcal{N}(\mathbb{0}_d, \frac{3}{2}\mathbf{I}_d)$ (which happens with probability $\frac{1}{2}$), the Hanson-Wright inequality implies that with high constant probability, we have $|x^\top \mathbf{U}\mathbf{U}^\top x - \frac{3}{2}k| \le \frac{1}{3}k$, so $x^\top \mathbf{U}\mathbf{U}^\top x \ge \frac{7}{6}k$. Therefore, the variance of $\langle \mathbf{U}\mathbf{U}^\top, xx^\top \rangle$ is indeed $\Omega(k^2)$, so $\nu(k) = \Omega(\sqrt{k})$, matching the above inequality up to a constant factor.

Next, by Proposition 2.6 of Kong et al. (2020), the output $\widetilde{\mathbf{U}}$ of their proposed algorithm satisfies

$$\|\mathbf{\Sigma}\|_k - \left\langle \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top, \mathbf{\Sigma} \right\rangle = O\left(\epsilon \|\mathbf{\Sigma}\|_k + \nu(k)\sqrt{k\epsilon} \right) = O(\gamma) \|\mathbf{\Sigma}\|_k,$$

where $\gamma = \sqrt{\epsilon}$, which is implied by the upper bound in Theorem 11 up to constant factors for any $C_4 = \Theta(1)$. Next, using the clipping argument in Lemma 37 for $B := C_4^2 \cdot \frac{\text{Tr}\mathbf{\Sigma}}{\epsilon}$, under (31), $\Pr(\|xx^\top\|_{\text{op}} \ge B) \le \epsilon$.[14] Therefore, the sample complexity used by Proposition 2.6 of Kong et al. (2020) is

$$n = O\left(\left(dk^2 + \frac{B}{\nu(k)}\sqrt{k\epsilon} \right) \frac{\log\left(\frac{d}{\delta\epsilon}\right)}{\epsilon} \right)$$

$$= O\left(\left(\frac{dk^2}{\epsilon} + \frac{\text{Tr}(\mathbf{\Sigma})}{\|\mathbf{\Sigma}\|_k} \frac{k}{\epsilon^{1.5}} \right) \log\left(\frac{d}{\delta\epsilon}\right) \right),$$

for $C_4 = \Theta(1)$, which is worse than the sample complexity required in Theorem 11 in the dependence on k. For example, in the reasonably well-conditioned regime where $\text{Tr}(\mathbf{\Sigma}) = \Theta(\frac{d}{k} \cdot \|\mathbf{\Sigma}\|_k)$, the respective sample complexities of Kong et al. (2020) and our algorithm for the same estimation rate are

$$O\left(\left(\frac{dk^2}{\epsilon} + \frac{d}{\epsilon^{1.5}} \right) \cdot \log\left(\frac{d}{\delta\epsilon}\right) \right), \quad O\left(\frac{d\log d + \log\frac{1}{\delta}}{\epsilon^{1.5}} \right).$$

C.3. Online heavy-tailed PCA

In this section, we provide a k-cPCA algorithm for heavy-tailed data without adversarial corruptions in an online setting, as a proof-of-concept application of our cPCA reduction. More precisely, we let $X_1, X_2, \ldots X_n \sim \mathcal{D}$ be i.i.d. draws from a (p, C_p)-hypercontractive distribution, $\mathcal{D}$, with covariance matrix $\mathbf{\Sigma}$. Our goal is to perform an approximate k-cPCA of $\mathbf{\Sigma}$, in the setting where the samples arrive online, i.e. we are limited to using $O(kd)$ space, where $k \ll d$ but n is potentially $\gg d$.

We first establish a helper result, Lemma 49, which shows that an approximate k-cPCA of $\mathbf{\Sigma} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$ is also an approximate k-cPCA of $\widehat{\mathbf{\Sigma}} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$ provided $\mathbf{\Sigma}$ and $\widehat{\mathbf{\Sigma}}$ are sufficiently close in operator norm. Our result follows from standard eigenvalue perturbation bounds from the literature.

Lemma 49 (cPCA perturbation) *Let $\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$ satisfy $\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\|_{\text{op}} \le \rho$. Let $\mathbf{U} \in \mathbb{R}^{d \times k}$ be a (δ, γ)-k-cPCA of $\widehat{\mathbf{\Sigma}}$ for $\max\{\delta, \gamma\} \le \frac{1}{10}$. Then for $\rho < \frac{\gamma \lambda_k(\mathbf{\Sigma})}{2}$, the following hold.*

---

14. This is the threshold at which point the clipped samples can be treated as outliers.

1. $\mathbf{U}$ *is a* (Δ, Γ)-k-*cPCA of* $\boldsymbol{\Sigma}$ *with* $\Delta := 8k(\frac{\rho}{\gamma\lambda_k(\boldsymbol{\Sigma})})^2 + 2\delta$, $\Gamma := 2\gamma$.

2. $\kappa_k(\widehat{\boldsymbol{\Sigma}}) \leq 2\kappa_k(\boldsymbol{\Sigma})$.

Proof Let $\widehat{\mathbf{S}} := \widehat{\mathbf{V}}^{\leq(1-\gamma)\lambda_k(\widehat{\boldsymbol{\Sigma}})}(\widehat{\boldsymbol{\Sigma}})$ denote the eigenspace of $\widehat{\boldsymbol{\Sigma}}$ with eigenvalues $\leq (1-\gamma)\lambda_k(\widehat{\boldsymbol{\Sigma}})$ and let $\widehat{\mathbf{L}}$ denote its complement subspace. Similarly, define $\mathbf{S} := \mathbf{V}^{\leq(1-\Gamma)\lambda_k(\boldsymbol{\Sigma})}(\boldsymbol{\Sigma})$ and let $\mathbf{L}$ denote its complement subspace. By definition of $\mathbf{U}$, we have $\|\widehat{\mathbf{S}}^\top\mathbf{U}\|_{\mathrm{F}}^2 = \mathrm{Tr}(\mathbf{U}^\top\widehat{\mathbf{S}}\widehat{\mathbf{S}}^\top\mathbf{U}) \leq \delta$. Hence,

$$
\begin{aligned}
\|\mathbf{S}^\top\mathbf{U}\|_{\mathrm{F}}^2 = \mathrm{Tr}\left(\mathbf{U}^\top\mathbf{S}\mathbf{S}^\top\mathbf{U}\right) &= \mathrm{Tr}\left(\mathbf{U}^\top\left(\widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top + \widehat{\mathbf{S}}\widehat{\mathbf{S}}^\top\right)\mathbf{S}\mathbf{S}^\top\left(\widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top + \widehat{\mathbf{S}}\widehat{\mathbf{S}}^\top\right)\mathbf{U}\right) \\
&= \mathrm{Tr}\left(\left(\mathbf{U}^\top\widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top\mathbf{S} + \mathbf{U}^\top\widehat{\mathbf{S}}\widehat{\mathbf{S}}^\top\mathbf{S}\right)\left(\mathbf{S}^\top\widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top\mathbf{U} + \mathbf{S}^\top\widehat{\mathbf{S}}\widehat{\mathbf{S}}^\top\mathbf{U}\right)\right) \\
&\leq 2\mathrm{Tr}\left(\mathbf{U}^\top\widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top\mathbf{S}\mathbf{S}^\top\widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top\mathbf{U}\right) \quad\quad (32) \\
&\quad + 2\mathrm{Tr}\left(\mathbf{U}^\top\widehat{\mathbf{S}}\widehat{\mathbf{S}}^\top\mathbf{S}\mathbf{S}^\top\widehat{\mathbf{S}}\widehat{\mathbf{S}}^\top\mathbf{U}\right), \text{(Fact 3)} \\
&= 2\mathrm{Tr}\left(\widehat{\mathbf{L}}^\top\mathbf{S}\mathbf{S}^\top\widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top\mathbf{U}\mathbf{U}^\top\widehat{\mathbf{L}}\right) + 2\mathrm{Tr}\left(\mathbf{S}\mathbf{S}^\top\widehat{\mathbf{S}}\widehat{\mathbf{S}}^\top\mathbf{U}\mathbf{U}^\top\widehat{\mathbf{S}}\widehat{\mathbf{S}}^\top\right) \\
&\leq 2\left\|\widehat{\mathbf{L}}^\top\mathbf{S}\mathbf{S}^\top\widehat{\mathbf{L}}\right\|_{\mathrm{op}}\mathrm{Tr}\left(\widehat{\mathbf{L}}^\top\mathbf{U}\mathbf{U}^\top\widehat{\mathbf{L}}\right) \quad\quad (33) \\
&\quad + 2\left\|\mathbf{S}\mathbf{S}^\top\right\|_{\mathrm{op}}\mathrm{Tr}\left(\widehat{\mathbf{S}}^\top\widehat{\mathbf{S}}\widehat{\mathbf{S}}^\top\mathbf{U}\mathbf{U}^\top\widehat{\mathbf{S}}\right) \\
&\leq 2\left\|\mathbf{S}^\top\widehat{\mathbf{L}}\right\|_{\mathrm{op}}^2\mathrm{Tr}\left(\widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top\mathbf{U}\mathbf{U}^\top\right) + 2\left\|\mathbf{S}\mathbf{S}^\top\right\|_{\mathrm{op}}\mathrm{Tr}\left(\widehat{\mathbf{S}}^\top\mathbf{U}\mathbf{U}^\top\widehat{\mathbf{S}}\right) \\
&\leq 2\left\|\mathbf{S}^\top\widehat{\mathbf{L}}\right\|_{\mathrm{op}}^2\left\|\widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top\right\|_{\mathrm{op}}\mathrm{Tr}\left(\mathbf{U}\mathbf{U}^\top\right) + 2\left\|\mathbf{S}\mathbf{S}^\top\right\|_{\mathrm{op}}\left\|\widehat{\mathbf{S}}^\top\mathbf{U}\right\|_{\mathrm{F}}^2 \\
&\leq 2k\left\|\mathbf{S}^\top\widehat{\mathbf{L}}\right\|_{\mathrm{op}}^2 + 2\delta, \text{ using the cPCA guarantee.} \quad\quad (34)
\end{aligned}
$$

Further, using Fact 2,

$$
\left|\lambda_k(\widehat{\boldsymbol{\Sigma}}) - \lambda_k(\boldsymbol{\Sigma})\right| \leq \left\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\right\|_{\mathrm{op}} \leq \rho. \quad\quad (35)
$$

Therefore for $\Gamma = 2\gamma$,

$$
\begin{aligned}
(1-\gamma)\lambda_k(\widehat{\boldsymbol{\Sigma}}) - (1-\Gamma)\lambda_k(\boldsymbol{\Sigma}) &= \gamma\lambda_k(\boldsymbol{\Sigma}) + (1-\gamma)\left(\lambda_k(\widehat{\boldsymbol{\Sigma}}) - \lambda_k(\boldsymbol{\Sigma})\right) \\
&\geq \gamma\lambda_k(\boldsymbol{\Sigma}) - \rho(1-\gamma), \text{ using (35)} \\
&\geq \frac{\gamma\lambda_k(\boldsymbol{\Sigma})}{2}, \text{ using } \rho \leq \frac{\gamma\lambda_k(\boldsymbol{\Sigma})}{2}.
\end{aligned}
$$

Hence, by applying the gap-free Wedin theorem (Lemma B.3, Allen-Zhu and Li (2016)), we have

$$
\left\|\mathbf{S}^\top\widehat{\mathbf{L}}\right\|_{\mathrm{op}} \leq \frac{\rho}{(1-\gamma)\lambda_k(\widehat{\boldsymbol{\Sigma}}) - (1-\Gamma)\lambda_k(\boldsymbol{\Sigma})} \leq \frac{2\rho}{\gamma\lambda_k(\boldsymbol{\Sigma})}. \quad\quad (36)
$$

Combining (34) and (36) yields the first claim, as desired:

$$
\|\mathbf{S}^\top\mathbf{U}\|_{\mathrm{F}}^2 \leq \frac{8k\rho^2}{\gamma^2\lambda_k(\boldsymbol{\Sigma})^2} + 2\delta.
$$

Finally, for the claim regarding $\kappa_k(\widehat{\mathbf{\Sigma}})$, we have by two applications of Fact 2:

$$\kappa_k(\widehat{\mathbf{\Sigma}}) = \frac{\lambda_1(\widehat{\mathbf{\Sigma}})}{\lambda_k(\widehat{\mathbf{\Sigma}})} \leq \frac{\lambda_1(\mathbf{\Sigma}) + \left|\lambda_1(\mathbf{\Sigma}) - \lambda_1(\widehat{\mathbf{\Sigma}})\right|}{\lambda_k(\mathbf{\Sigma}) - \left|\lambda_k(\mathbf{\Sigma}) - \lambda_k(\widehat{\mathbf{\Sigma}})\right|} \leq \frac{\lambda_1(\mathbf{\Sigma}) + \frac{\gamma\lambda_k(\mathbf{\Sigma})}{2}}{\lambda_k(\mathbf{\Sigma}) - \frac{\gamma\lambda_k(\mathbf{\Sigma})}{2}} \leq \frac{1+\frac{\gamma}{2}}{1-\frac{\gamma}{2}}\kappa_k(\mathbf{\Sigma}) \leq 2\kappa_k(\mathbf{\Sigma}).$$

∎

We specify a choice of ρ for convenient application of Lemma 49 in the following.

Corollary 50 *In the setting of Lemma 49, if $\rho := \sqrt{\frac{\delta}{8k}}\gamma\lambda_k(\mathbf{\Sigma})$, $\mathbf{U}$ is a $(3\delta, 2\gamma)$-k-cPCA of $\mathbf{\Sigma}$.*

Corollaries 38 and 50 make our roadmap clear for heavy-tailed PCA. We first use Corollary 38 to devise a clipping threshold for our data such that the clipped covariance matrix is close in operator norm to the original distribution's covariance matrix. We then perform approximate k-cPCA on the clipped data, analyzed using our reduction in Theorem 9, and finally use Corollary 50 to convert the resulting k-cPCA on the clipped covariance to a corresponding k-cPCA on the original covariance.

To demonstrate an instantiation of this strategy, we propose an online gap-free k-cPCA algorithm for heavy-tailed data using Oja's algorithm Oja (1982) as the 1-cPCA oracle. More specifically, we use the following guarantee on Oja's algorithm from Allen-Zhu and Li (2017).

Proposition 51 (Theorem 2, Allen-Zhu and Li (2017))) *There is an algorithm* Oja *with the following guarantee. Let $\delta, \gamma \geq 0$ such that $\max(\delta, \gamma) \leq \frac{1}{10}$ and $\delta \leq \frac{\gamma^2}{256}$. Let $d \in \mathbb{N}$, $R > 0$, and $\beta \in (0,1)$. Let $n = \Omega(\frac{d}{\delta\gamma^2}\log(\frac{d}{\beta}))$, and let $\{X_i\}_{i\in[n]} \in \mathbb{R}^d$ be drawn i.i.d. from $\mathcal{D}$, a distribution on $\mathbb{R}^d$ with mean $\mathbb{0}_d$ and covariance $\mathbf{\Sigma}$, where $\|X\|_2^2 \leq R$ almost surely for $X \sim \mathcal{D}$. Then, with probability at least $1 - \beta$, $\mathsf{Oja}(\delta, \gamma, \{X_i\}_{i\in[n]}, R)$ returns a (δ, γ)-cPCA of $\mathbf{M}$, using $O(d)$ space.*

As a consequence of Proposition 51 and our roadmap described earlier, we conclude with the following.

Theorem 52 *For an even integer $p \geq 4$, let $\mathcal{D}$ be (p, C_p)-hypercontractive on $\mathbb{R}^d$ with mean $\mathbb{0}_d$ and covariance $\mathbf{\Sigma}$. Let $(\Delta, \Gamma) \in (0,1)$, assume $\Delta \cdot \kappa_k(\mathbf{\Sigma})^2 \leq \Gamma^2$, and set $\delta = \frac{1}{k^{\Theta(\log k)}} \cdot \Delta$, $\gamma = \frac{1}{\Theta(k^3)} \cdot \Gamma$, for appropriate constants. Let*

$$\alpha := \left(\frac{C_p^2 \kappa_k(\mathbf{\Sigma})\sqrt{k}}{\Gamma\sqrt{\Delta}}\right)^{\frac{1}{p-2}}, \quad R := \Theta(\alpha \mathrm{Tr}(\mathbf{\Sigma})).$$

Let $\beta \in (0,1)$. If $n = \Theta(\alpha \frac{d\kappa_k(\mathbf{\Sigma})^2}{\delta\gamma^2}\log(\frac{d}{\beta}))$ for an appropriate constant, BlackBoxPCA *(Algorithm 1) using* Oja *as a (δ, γ)-1-cPCA oracle on n samples from $\mathcal{T}_R(\mathcal{D})$ returns a (Δ, Γ)-k-cPCA of $\mathbf{\Sigma}$ with probability $\geq 1 - \beta$, in $O(dk)$ space.*

Proof Let $\rho := (\frac{\Delta}{24k})^{1/2}\frac{\Gamma}{\kappa_k(\mathbf{\Sigma})}$. Using Corollary 50, for $R \geq \left(\frac{C_p^p}{\rho}\right)^{\frac{2}{p-2}}\mathrm{Tr}(\mathbf{\Sigma})$, the covariance of the clipped distribution $\mathcal{T}_R(\mathcal{D})$, $\mathbf{\Sigma}'$, satisfies

$$\left\|\mathbf{\Sigma}' - \mathbf{\Sigma}\right\|_{\mathrm{op}} \leq \rho\|\mathbf{\Sigma}\|_{\mathrm{op}}.$$

Next, we obtain a k-cPCA, $\mathbf{U}$, of $\boldsymbol{\Sigma}'$ by running Algorithm 1 on the clipped distribution. Using Theorem 9 and the guarantees of Proposition 51, we have that $\mathbf{U}$ is a $\left(\frac{\Delta}{3}, \frac{\Gamma}{2}\right)$-$k$-cPCA of $\boldsymbol{\Sigma}'$. Finally, we claim $\mathbf{U}$ is also a (Δ, Γ)-k-cPCA of $\boldsymbol{\Sigma}$ by Corollary 50, which requires

$$\left\|\boldsymbol{\Sigma}' - \boldsymbol{\Sigma}\right\|_{\text{op}} \leq \rho \left\|\boldsymbol{\Sigma}\right\|_{\text{op}} \leq 2\sqrt{\frac{\Delta}{24k}}\Gamma\lambda_k\left(\boldsymbol{\Sigma}\right). \tag{37}$$

The proof follows by substituting the definition of ρ in (37). The space complexity is immediate from the definition of Algorithm 1. ∎

We do note that the prior work Allen-Zhu and Li (2017), in addition to proving Proposition 51, gave a sophisticated analysis of a simultaneous variant of Oja's algorithm using a $d \times k$ block matrix, which applies to distributions with almost surely bounded supports (with rates parameterized by the covariance matrix and the almost sure bound). While this Allen-Zhu and Li (2017) result does not directly apply to hypercontractive distributions, following the same roadmap as in Theorem 52, i.e. first truncating the distribution (via Corollary 38) and then bounding the perturbation (via Corollary 50), but using the k-Oja result of Allen-Zhu and Li (2017) in place of Proposition 51 and Theorem 9, obtains an improvement over Theorem 52. For instance, Allen-Zhu and Li (2017)'s k-Oja analysis applies in all parameter regimes because it is not bottlenecked by the impossibility result in Proposition 32, and also incurs only a polynomial overhead in k.

We include Theorem 52 as a proof-of-concept of how to apply our reduction to give a more straightforward k-cPCA result by relying only on existence of the corresponding 1-cPCA algorithm, rather than designing a custom analysis as was done in Allen-Zhu and Li (2017). We are optimistic about the utility of the approach in Theorem 52 in providing tools for attacking future statistical PCA settings with various constraints (e.g. privacy, dependent data, and so forth as mentioned in Section 1) via reductions, as 1-cPCA algorithms are typically more straightforward to analyze than k-cPCA algorithms.