Optimal Rates for Adaptive Private k-PCA

Johanna Düngler
Department of Computer Science
University of Copenhagen
jodu@di.ku.dk

Amartya Sanyal
Department of Computer Science
University of Copenhagen
amsa@di.ku.dk

Abstract

Given n i.i.d. random matrices $A_i \in \mathbb{R}^{d\times d}$ with common expectation Σ, the goal of *Differentially Private Stochastic PCA* is to identify a k-dimensional subspace capturing the leading variance directions of Σ, while preserving differential privacy (DP) for each individual sample A_i. Düngler and Sanyal [2025] introduced k-DP-PCA, the first algorithm to simultaneously (1) achieve sample complexity $n = \widetilde{O}(d)$ for sub-Gaussian data, (2) adapt its privacy noise to the intrinsic randomness of the data, and (3) extend seamlessly to any target dimension $k \leq d$. However, its sample complexity has a suboptimal dependence on k. We propose the first algorithm that achieves optimal sample complexity in both d and k, while retaining properties (2) and (3) of k-DP-PCA. Additionally, prior utility guarantees require a minimum sample size with exponential dependence on the eigengap. Our method removes this dependence. It also improves the dependence on spectral parameters, matching known lower bounds for the spiked covariance model. Unlike deflation-based approaches like k-DP-PCA, our algorithm updates the full $d \times k$ subspace jointly rather than one eigenvector at a time. This non-deflation structure simplifies the algorithm, reduces the number of hyper-parameters, and improves computational efficiency, particularly when using block linear algebra libraries.

1 Introduction

Principal Component Analysis (PCA) is a basic technique in statistics and machine learning. Given data vectors $x_1, \ldots, x_n \in \mathbb{R}^d$, classical PCA estimates the leading eigenspace of the empirical second-moment matrix $\frac{1}{n}\sum_i x_i x_i^\top$. The task of recovering the top k eigenvectors is typically referred to as k-PCA. In this paper, we study a stochastic variant of this problem. The input is a sequence of independent symmetric matrices $A_1, \ldots, A_n \in \mathbb{R}^{d\times d}$ with common expectation Σ, and the goal is to estimate the top-k eigenspace of Σ. In addition, we want to protect the privacy of the individual samples A_i in the sense of differential privacy [Dwork et al., 2006].

Differential privacy offers a formal framework for protecting the privacy of individuals represented in a dataset, and has been deployed in settings involving sensitive information, including census data analysis and large-scale industrial analytics [Abowd et al., 2020, Apple, 2017]. Private PCA has commonly been studied in the fixed-dataset setting for nearly two decades [Blum et al., 2005, Chaudhuri et al., 2013, Hardt and Roth, 2013, Dwork et al., 2014], where algorithms typically rely on deterministic norm bounds. When such methods are applied to stochastic PCA, these bounds must be enforced through clipping, but this transfers a worst-case fixed-data requirement into a setting where the observations are random. The resulting privacy-utility guarantees are therefore often suboptimal: they either require sample sizes that grow super-linearly with the ambient dimension d, or add privacy noise at a level that does not exploit the randomness already present in the observations. Consequently, applying fixed-dataset private PCA methods directly in the stochastic setting typically yields error bounds of order $O(\sqrt{dk/n} + d^{3/2}k/(\varepsilon n))$, where ε denotes the privacy parameter, while the known optimal rate is of order $O(\sqrt{dk/n} + dk/(\varepsilon n))$ (Theorems 6 and 7).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The first adaptive private algorithm for stochastic PCA was given by Liu et al. [2022] for the case $k = 1$. Their key idea was to privately estimate the fluctuation scale of the stochastic update and to clip around a private center rather than around zero. This makes the privacy noise scale with the randomness in the data. Düngler and Sanyal [2025] extended this adaptive principle to $k > 1$. Their method, however, is deflation-based: it estimates one direction, projects it out, and repeats. This sequential structure introduces an additional dependence on k: their error bounds are of order $O(\sqrt{dk^{3/2}/n} + dk^{3/2}/(\varepsilon n))$. Additionally, their algorithm requires separate parameter tuning for each eigenvector it estimates, making it both more cumbersome to tune and computationally ill-suited for block linear-algebra implementations that are efficient in practice. Cai et al. [2024] obtained optimal rates for Gaussian data when the unknown population covariance Σ satisfies the flat-tail condition $\lambda_{k+1} = \cdots = \lambda_d$. However, because Σ is itself unknown, it is unclear if this condition can be privately certified from the input data in general. As a result, their privacy guarantees rely on an unverified distributional assumption on the data-generating process.

Our Contributions Our first contribution is ADADPO (Algorithm 2), a non-deflation-based adaptive private variant of Oja's algorithm (Algorithm 6) that achieves the optimal dependence on k in sample complexity. Unlike deflation-based methods such as k-DP-PCA [Düngler and Sanyal, 2025], which reduce k-PCA to k private one-dimensional problems, our algorithm estimates the target k-dimensional subspace jointly. Moreover, its adaptive privacy mechanism estimates and uses one single scale parameter and one clipping center for all k columns instead of using k separate range estimators. It then adds calibrated Gaussian noise that preserves symmetry. This joint treatment is what avoids the extra cost of handling the k directions separately. We show that ADADPO achieves the optimal dependence on d and k (Theorem 2), and performs well on simulated data.

Our second contribution is TADADPO (Algorithm 3), a refinement of ADADPO that improves the dependence of the private error bound on the spectrum of Σ. The main idea is that, after obtaining a sufficiently accurate initial estimate of the top-k eigenspace, the remaining estimation problem has a more favorable local geometry. TADADPO exploits this local structure in a second phase to reduce the amount of privacy noise needed for the final refinement. We prove a general utility guarantee for this procedure in Theorem 5. In the spiked covariance model, this yields the optimal spectral dependence up to logarithmic factors, matching the lower bound of Cai et al. [2024].

Corollary 1 (Informal version of Corollary 3)**.** *Assume $A_i = x_i x_i^\top$ with $x_i \sim \mathcal{N}(0, \Sigma)$, where Σ is a spiked covariance with a flat tail, meaning that for some $k \in [d]$, $\lambda_{k+1} = \cdots = \lambda_d$. Let $V_k \in \mathbb{R}^{d\times k}$ be the true top-k eigenvectors of Σ, and let $\lambda_k - \lambda_{k+1} > 0$. Then the output Q_T of* TADADPO *satisfies*

$$\left\|Q_T Q_T^\top - V_k V_k^\top\right\|_F \leq \tilde{O}\left(\frac{\sqrt{\lambda_1\lambda_{k+1}}}{\lambda_k - \lambda_{k+1}}\left[\sqrt{\frac{dk}{n}} + \frac{dk}{\varepsilon n}\right]\right).$$

The rest of the paper is organized as follows: Section 2 defines the stochastic PCA problem, the privacy model, the main assumptions, and a theoretical warm up for the following sections. Section 3 introduces ADADPO, its guarantees, and experiments. Section 4 presents TADADPO and its guarantees. Finally, Section 5 gives an overview of related work and presents the conclusion.

2 Problem Formulation

Let $A_1, \ldots, A_n \in \mathbb{R}^{d\times d}$ be independent symmetric random matrices with common expectation $\Sigma = \mathbb{E}[A_i]$. We assume that Σ is positive semidefinite with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$, and we denote by $V_k \in \mathbb{R}^{d\times k}$ the matrix containing the top k eigenvectors of Σ. The goal of *Stochastic k-PCA* is to output $Q \in \mathbb{R}^{d\times k}$ with orthonormal columns such that $QQ^\top$ is close to $V_k V_k^\top$, with error measured by $\left\|QQ^\top - V_k V_k^\top\right\|_F$. Throughout the paper, we assume a nonzero eigengap $\Delta_k := \lambda_k - \lambda_{k+1} > 0$, and we write $\kappa := \lambda_1/\Delta_k$ and $\kappa_\perp := (\lambda_1 - \lambda_d)/\Delta_k \leq \kappa$.

Let $\mathbb{S}_{d,k} := \left\{Q \in \mathbb{R}^{d\times k} : Q^\top Q = I_k\right\}$. For $Q \in \mathbb{S}_{d,k}$, let $\Pi_Q := I - QQ^\top$ and

$$\mathcal{S}_Q := \left\{Y \in \mathbb{R}^{d\times k} : Q^\top Y \text{ is symmetric}\right\}.$$

For a square matrix B, write $\mathrm{sym}(B) := \frac{B+B^\top}{2}$. The Frobenius-orthogonal projection onto $\mathcal{S}_Q$ is

$$\mathsf{P}_{\mathcal{S}_Q}(Y) := Q\,\mathrm{sym}(Q^\top Y) + \Pi_Q Y.$$

We use $\|\cdot\|_2$ for operator norm, $\|\cdot\|_F$ for Frobenius norm, $\langle A, B\rangle = \mathrm{Tr}(A^\top B)$ for the Frobenius inner product, and $\|\cdot\|_{(2,k)}$ for the Schatten-$(2,k)$ norm, and $\|\cdot\|_{S_p}$ for the Schatten-p norm $(\sum_i \sigma_i(\cdot)^p)^{1/p}$, so that $\|\cdot\|_{S_2} = \|\cdot\|_F$ and $\|\cdot\|_{(2,k)}$ is the rank-k truncation of $\|\cdot\|_{S_2}$. The notation $\tilde{O}(\cdot)$ hides polylogarithmic factors.

Definition 1 (Differential privacy)**.** Two datasets S and S' of the same size are neighboring ($S \sim S'$) if they differ by replacing one entry. A randomized algorithm M is (ε, δ)-differentially private if, for every neighboring pair S, S' and every measurable event $\mathcal{E}$,

$$\Pr(M(S) \in \mathcal{E}) \leq e^{\varepsilon} \Pr(M(S') \in \mathcal{E}) + \delta.$$

A standard way to ensure differential privacy is the Gaussian mechanism.

Definition 2 (Gaussian Mechanism)**.** For a query f with sensitivity $\Delta(f) := \sup_{S \sim S'} \|f(S) - f(S')\|_2$, one releases

$$M(S) = f(S) + g, \qquad g \sim \mathcal{N}(0, \sigma^2 I).$$

For $\sigma \gtrsim \Delta(f)\sqrt{\log(1/\delta)}/\varepsilon$, this mechanism is (ε, δ)-differentially private.

ADADPO is analyzed under the following global model, which is the same type of model used in prior stochastic private PCA analyses. We emphasize that our privacy guarantees are distribution-free; the assumptions below are used only for utility. In Assumption B, we introduce a local tangent-space refinement of this model, which applies after a sufficiently accurate warm-up phase and leads to sharper bounds for TADADPO.

Assumption A ($(\Sigma, \{\lambda_i\}_{i=1}^d, M, V, K, a, \gamma)$-model)**.** *The matrices $A_1, \ldots, A_n \in \mathbb{R}^{d\times d}$ are independent, symmetric, and satisfy:*

A.1 $\mathbb{E}[A_i] = \Sigma$*, where $\Sigma \succeq 0$ has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$, corresponding eigenvectors $v_1, \ldots, v_d$, and $\lambda_k - \lambda_{k+1} > 0$.*

A.2 $\|A_i - \Sigma\|_{(2,k)} = \sup_{P \in \mathbb{S}_{d,k}} \left\|P^\top (A_i - \Sigma)\right\|_F \leq \lambda_1 M$ *a.s.*

A.3 $\left\|\mathbb{E}\left[(A_i - \Sigma)^2\right]\right\|_2 \leq \lambda_1^2 V$.

A.4 $\max_{\|u\|_2=1, \|v\|_2=1} \mathbb{E}\left[\exp\left(\left(\frac{|u^\top (A_i - \Sigma) v|^2}{K^2 \lambda_1^2 \|H_u\|_2}\right)^{1/2a}\right)\right] \leq 2$*, where* $H_u = \frac{1}{\lambda_1^2}\mathbb{E}[(A_i - \Sigma) u\, u^\top (A_i - \Sigma)^\top]$ *and we denote* $\gamma^2 = \max_{\|u\|=1} \|H_u\|_2$.

Assumptions A.1 to A.3 are standard for matrix concentration (e.g., under the matrix Bernstein inequality [Tropp, 2012]) and thus also required for the utility guarantees of Oja's algorithm even in the non-private setting. Assumption A.4 controls the size of the bilinear form $u^\top (A_i - \Sigma) v$ and can be seen as a Gaussian-like tail bound, which tells us that the magnitude of the projection of the fluctuation $A_i - \Sigma$ is bounded with high probability. It is a standard assumption in related private stochastic PCA analyses [Liu et al., 2022, Düngler and Sanyal, 2025].

2.1 Theoretical Warmup

One of the most widely used algorithms for Principal Component Analysis (PCA) is Oja's algorithm [Oja, 1982, Oja and Karhunen, 1985]. It is a streaming algorithm that receives a sequence of matrices A_i and at each time step updates a k-dimensional subspace Q_t by

$$Q_t = \mathrm{QR}\left(Q_{t-1} + \eta_t A_t Q_{t-1}\right).$$

where η_t is the learning rate at time t. The standard utility theorem for Oja's method [Huang et al., 2021] assumes that the matrices A_t are independent, symmetric, and have common expectation Σ. We recall the full statement and pseudocode in Theorem 8 and Algorithm 6. These two requirements are exactly what make a private adaptive block update delicate.

If one could release $(A_t + G_t)Q_{t-1}$, where G_t is an independent symmetric Gaussian matrix, then the update would simply be Oja's algorithm run on perturbed symmetric matrices $A_t + G_t$. The challenge, however, is not merely privacy, but adaptivity: the algorithm must privately estimate the

range of the matrix being privatized, choose a clipping threshold, and add Gaussian noise calibrated to the resulting sensitivity.

Adding noise to the full matrix A_t would preserve the standard analysis of Oja's algorithm, but would require clipping at the range of A_t, which can be much larger than the range of the projected update A_tQ_{t-1}. Since privacy only needs to protect $A_tQ_{t-1} \in \mathbb{R}^{d\times k}$, this would be unnecessarily conservative and would destroy the adaptive gain. Instead, we estimate the range and add Gaussian noise directly in $\mathbb{R}^{d\times k}$, at the scale of A_tQ_{t-1}. This is the right scale for privacy, but the noise is generally not of the form G_tQ_{t-1} for any symmetric G_t. Hence the effective matrices are no longer symmetric, and the standard theorem for Oja's algorithm no longer applies. This is one reason prior adaptive k-PCA algorithms use deflation. Düngler and Sanyal [2025] reduce the problem to a sequence of private one-dimensional PCA calls. For $k = 1$, the Oja update becomes

$$q_t = \mathrm{QR}\left(q_{t-1} + \eta_t A_t q_{t-1}\right),$$

with $q_{t-1} \in \mathbb{R}^d$. In this $k = 1$ setting, the available utility result does not require the input matrices in the update to be symmetric. Deflation therefore avoids the symmetry obstruction that arises for the full $d \times k$ block update. However, it requires fresh batches of data to maintain independence across directions, which loses a factor in the dependence on k.

We retain the block update through two new observations. First, for fixed Q, the action of any symmetric matrix on Q lies in the linear space

$$\mathcal{S}_Q = \left\{Y \in \mathbb{R}^{d\times k} : Q^\top Y \text{ is symmetric}\right\}.$$

Thus the privacy noise only needs to be added inside $\mathcal{S}_Q$, rather than in all of $\mathbb{R}^{d\times k}$. This already restores a symmetric representation: every element of $\mathcal{S}_Q$ can be written as G_QQ for some symmetric matrix G_Q. However, with the natural isotropic noise on $\mathcal{S}_Q$, the representing matrix G_Q depends on Q. Along the algorithm, $Q = Q_{t-1}$ is random and adaptive, so the resulting effective matrices are not an i.i.d. sequence of symmetric perturbations, and the standard theorem for Oja's algorithm still does not apply.

The second observation is that adding slightly more noise gives a Q-independent symmetric representation, while preserving the same privacy guarantee. To state this precisely, we use the following notation.

Definition 3 (Gaussian Orthogonal Ensemble)**.** A symmetric matrix $G \in \mathbb{R}^{d\times d}$ is sampled from $\mathrm{GOE}_d(\sigma^2)$ if $G_{ii} \sim \mathcal{N}(0, 2\sigma^2)$ and $G_{ij} \sim \mathcal{N}(0, \sigma^2)$ independently for $i < j$.

Lemma 1. *Fix $Q \in \mathbb{R}^{d\times k}$ with $Q^\top Q = I_k$. Let $Z \in \mathbb{R}^{d\times k}$ have i.i.d. $\mathcal{N}(0, \sigma^2)$ entries, let $N_{\mathrm{sym}} \sim \mathrm{GOE}_k(\sigma^2)$, and let $G \sim \mathrm{GOE}_d(\sigma^2)$. Then*

$$QN_{\mathrm{sym}} + (I - QQ^\top)Z \overset{d}{=} GQ.$$

To privatize the projected update AQ, we need a uniform sensitivity bound. Since AQ is not deterministically bounded under the stochastic model, we enforce such a bound by clipping in Frobenius norm. For $R > 0$, define

$$\mathrm{clip}_R(Y) = \begin{cases} Y, & \|Y\|_F \le R, \\ RY/\|Y\|_F, & \|Y\|_F > R. \end{cases}$$

Thus, at an iteration with current iterate Q_{t-1}, we use the private clipped update

$$Q_t = \mathrm{QR}\left(Q_{t-1} + \eta_t\left(\mathrm{clip}_R(A_tQ_{t-1}) + Q_{t-1}N_t + (I - Q_{t-1}Q_{t-1}^\top)Z_t\right)\right),$$

where $N_t \sim \mathrm{GOE}_k(\sigma^2)$ and $Z_t \in \mathbb{R}^{d\times k}$ has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries.

For any R for which clipping is rare, the update agrees on the no-clipping event with non-private Oja's algorithm applied to $A_t + G_{t-1}$, where $G_{t-1} \sim \mathrm{GOE}_d(\sigma^2)$. Thus we can privatize the full $d \times k$ block update while preserving the structure required by Oja's analysis. For a rigorous privacy proof of Algorithm 1 see Theorem 3 in Section C. The next theorem gives a fixed-radius guarantee, choosing R under Assumption A so that clipping is unlikely and the dependence on k is correct.

Theorem 1 (Utility of Algorithm 1)**.** *Assume $A_1, \ldots, A_n$ satisfy Assumption A and write $\kappa = \frac{\lambda_1}{\lambda_k - \lambda_{k+1}}$. Set*

$$R = C\lambda_1\sqrt{dk}\big(K\gamma\log^a(ndk/\zeta) + 1\big), \qquad T = \lceil C_T\kappa\log(n)\rceil,$$

Algorithm 1 Private Oja's

Input: $S = \{A_i\}_{i=1}^n$, k, batch count T, privacy parameters (ε, δ), learning rates $\{\eta_i\}_{i=1}^T$, clipping radius R.

Set $B \leftarrow \lfloor n/T \rfloor$, $\sigma \leftarrow 2R\sqrt{2\log(1.25/\delta)}/(B\varepsilon)$.
Choose $Q_0' \in \mathbb{R}^{d\times k}$ uniformly at random and set $Q_0 \leftarrow \mathrm{QR}(Q_0')$.
for $t = 1, \dots, T$ **do**
    Set $\Pi_{Q_{t-1}} \leftarrow I - Q_{t-1}Q_{t-1}^\top$.
    Draw $Z_t \in \mathbb{R}^{d\times k}$ with i.i.d. entries $\mathcal{N}(0,\sigma^2)$, and draw $N_{\mathrm{sym},t} \sim \mathrm{GOE}_k(\sigma^2)$.
    $Q_t \leftarrow \mathrm{QR}\left(Q_{t-1} + \eta_t\left[\frac{1}{B}\sum_{i=B(t-1)+1}^{tB} \mathrm{clip}_R(A_iQ_{t-1}) + Q_{t-1}N_{\mathrm{sym},t} + \Pi_{Q_{t-1}}Z_t\right]\right)$.
end for
return Q_T.

where C_T is a sufficiently large absolute constant. If

$$n \gtrsim \max\left\{\frac{k^2\lambda_1^2 V}{\zeta^2(\lambda_k-\lambda_{k+1})^2}, \kappa^{3/2}\frac{kM}{\zeta}, \kappa^{3/2}\frac{(K\gamma+1)k^{3/2}d}{\varepsilon\zeta}\right\},$$

then there exist learning rates $\{\eta_t\}$ so that with probability at least $1-\zeta$, Algorithm 1 outputs Q_T satisfying

$$\left\|Q_TQ_T^\top - V_kV_k^\top\right\|_F \le \tilde{O}\left(\kappa\frac{\sqrt{kV}}{\sqrt{n}} + \kappa^{3/2}\frac{\sqrt{k}M}{n} + \kappa^{3/2}\frac{(K\gamma+1)kd}{\varepsilon n}\right).$$

The first two terms are the stochastic error of Oja's algorithm, and the third is the privacy cost. In particular, the privacy term scales as $dk/(\varepsilon n)$, reflecting the block update over all k directions.

Example 1 (Gaussian covariance and the k-dependence)**.** *Let $A_i = x_ix_i^\top$ with $x_i \sim \mathcal{N}(0,\Sigma)$. Then, with high probability, Assumption A holds with $M = \tilde{O}(d)$, $V = O(d)$, $K = O(1)$, $a = 1$, and $\gamma = O(1)$. Thus, applying Theorem 1 on this high-probability event gives*

$$\left\|Q_TQ_T^\top - V_kV_k^\top\right\|_F \le \tilde{O}\left(\kappa\left(\frac{\sqrt{kd}}{\sqrt{n}} + \sqrt{\kappa}\frac{dk}{\varepsilon n}\right)\right).$$

However, Algorithm 1 is not adaptive. Since R is fixed before the algorithm runs and clipping is around zero, the radius must cover both $(A_i-\Sigma)Q$ and ΣQ. Consequently, the privacy noise need not vanish even when the stochastic fluctuation does. Thus the block update has the right k-dependence, but its privacy cost is not adaptive to the randomness in the data, as illustrated in the example below.

Example 2 (Spiked covariance and adaptivity)**.** *Consider the rank-one spiked model $x_i = s_i + n_i$, where $s_i \sim \mathrm{Unif}\{v,-v\}$, $n_i \sim \mathcal{N}(0,\sigma^2I_d)$, and $v \in \mathbb{R}^d$ is unit norm. Let $A_i = x_ix_i^\top$, so $\Sigma = \mathbb{E}[A_i] = vv^\top + \sigma^2I_d$. As $\sigma \to 0$, the samples become nearly deterministic: $A_i \to vv^\top = \Sigma$, and the stochastic fluctuation scale γ tends to zero. However, the radius used in Theorem 1 is $R = C\lambda_1\sqrt{dk}\big(K\gamma\log^a(ndk/\zeta)+1\big)$. Thus R remains of order $\lambda_1\sqrt{dk}$ even when $\gamma \to 0$, and the privacy noise does not vanish in the nearly deterministic regime. The adaptive algorithm in Section 3 removes this obstruction by privately estimating a center for A_iQ and clipping around that center.*

3 Adaptive Private Oja's

The warmup in the previous section gives a private block version of Oja's algorithm with the correct dependence on k, but its clipping radius is fixed in advance and centered at zero. We now make the block update adaptive. At each iteration t, Algorithm 2 estimates a center for the projected samples $A_iQ_{t-1} \in \mathbb{R}^{d\times k}$ and a range parameter controlling their fluctuations around that center (PRIVRANGE, Algorithm 4). It then clips around the estimated center and computes a truncated batch mean (TRUNCATEDMEAN, Algorithm 5) and finally privatizes the mean by adding Gaussian noise. The sensitivity, and hence the Gaussian noise scale, is calibrated to the empirical fluctuation

radius rather than to a worst-case radius fixed before the algorithm begins. Conditional on Q_{t-1}, the population center of the projected samples is ΣQ_{t-1}, so the relevant fluctuation is

$$A_i Q_{t-1} - \Sigma Q_{t-1} = (A_i - \Sigma) Q_{t-1}.$$

This is the quantity controlled by the fluctuation parameter γ. Consequently, when the stochastic fluctuations are small, the clipping radius and privacy noise become small as well. This is the adaptivity captured in Theorem 2.

Algorithm 2 Adaptive Private Oja's (ADADPO)

Input: $S = \{A_1, \dots, A_n\}$, $k \in [d]$, batch count T, warm-up length T_0, base clipping radius R, privacy parameters (ε, δ), learning rates $\{\eta_t\}_{t=1}^T$, and failure probability ζ

Set $B \leftarrow \lfloor n/T \rfloor$, $\sigma \leftarrow 2R\sqrt{2\log(1.25/\delta)}/(B\varepsilon)$
Choose $Q_0' \in \mathbb{R}^{d\times k}$ uniformly at random, $Q_0 \leftarrow \mathrm{QR}[Q_0']$
for $t = 1, 2, \dots, T_0$ **do** ▷ Phase I: non-adaptive warm-up
  $\bar{G}_t \leftarrow \frac{1}{B}\sum_{i=B(t-1)+1}^{tB} \mathrm{clip}_R(A_i Q_{t-1})$ **and** $\sigma_t \leftarrow \sigma$
  $Z_t \in \mathbb{R}^{d\times k}$ with coordinates $\overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma_t^2)$, $N_{\mathrm{sym},t} \sim GOE_k(\sigma_t^2)$
  $Q_t \leftarrow \mathrm{QR}\left[Q_{t-1} + \eta_t\left(\mathsf{P}_{\mathcal{S}_{Q_{t-1}}}(\bar{G}_t) + Q_{t-1}N_{\mathrm{sym},t} + (I - Q_{t-1}Q_{t-1}^\top)Z_t\right)\right]$
end for
for $t = T_0 + 1, \dots, T$ **do** ▷ Phase II: adaptive
  Estimate Range $\widehat{\Lambda}_t \leftarrow \textsc{PrivRange}\left(\{A_i Q_{t-1}\}_{i=B\cdot(t-1)+1}^{B\cdot(t-1)+B/2}, (\varepsilon, \delta), \zeta/2T\right)$
  Set Truncation Threshold $R_t \leftarrow 3K\sqrt{\widehat{\Lambda}_t}\,\log^a(Bdk/(2\zeta))$
  $\bar{G}_t \leftarrow \textsc{TruncatedMean}\left(\{A_i Q_{t-1}\}_{i=B\cdot(t-1)+B/2+1}^{tB}, R_t, \widehat{\Lambda}_t, (\varepsilon/2, \delta/2), \zeta/2T\right)$
  Set privacy noise multiplier $\sigma_t \leftarrow \frac{8R_t\sqrt{dk}}{B\varepsilon}\sqrt{2\log(2.5/\delta)}$
  $Z_t \in \mathbb{R}^{d\times k}$ with coordinates $\overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma_t^2)$, $N_{\mathrm{sym},t} \sim GOE_k(\sigma_t^2)$
  $Q_t \leftarrow \mathrm{QR}\left[Q_{t-1} + \eta_t\left(\mathsf{P}_{\mathcal{S}_{Q_{t-1}}}(\bar{G}_t) + Q_{t-1}N_{\mathrm{sym},t} + (I - Q_{t-1}Q_{t-1}^\top)Z_t\right)\right]$
end for
return Q_T

Lemma 2 (Privacy of Algorithm 2)**.** *If $0 < \varepsilon \le 1$ then Algorithm 2 is (ε, δ)-DP.*

Remark. The restriction on ε can be removed by replacing the standard Gaussian calibration with the analytic Gaussian mechanism. This calibration is less convenient to state in closed form, but is readily implemented in our experiments.

Theorem 2 (Utility of Algorithm 2)**.** *Given matrices $A_1, \dots, A_n$ fulfilling Assumption A write $\kappa = \frac{\lambda_1}{\lambda_k - \lambda_{k+1}}$ and set the base clipping radius, the batch count and the warm-up length to*

$$R = C\lambda_1\sqrt{dk}\big(K\gamma\log^a(ndk/\zeta) + 1\big), \qquad T = \lceil C_T\kappa\log(n)\rceil, \qquad T_0 = \lceil T/2\rceil,$$

where C_T is a sufficiently large absolute constant. If

$$n \gtrsim \max\left\{\frac{k^2\lambda_1^2 V}{\zeta^2(\lambda_k - \lambda_{k+1})^2}, \kappa^{3/2}\cdot\frac{kM}{\zeta}, \kappa^{3/2}\cdot\frac{(K\gamma+1)k^{3/2}d}{\varepsilon\zeta}, \frac{dk}{\varepsilon}, \frac{K^2 d}{\varepsilon}\right\},$$

then there exist learning rates $\{\eta_t\}$ such that, with probability at least $1 - \zeta$, Algorithm 2 outputs Q_T with

$$\|Q_T Q_T^\top - V_k V_k^\top\|_F \le \tilde{O}\left(\kappa\frac{\sqrt{kV}}{\sqrt{n}} + \kappa^{3/2}\frac{\sqrt{k}M}{n} + \kappa^{3/2}\frac{Kdk\gamma}{\varepsilon n}\right).$$

Remark. The choice $T_0 = \lceil T/2\rceil$ is sufficient but far from necessary: We fix it at $T/2$ because that choice requires no knowledge of V, M or γ.

Corollary 2 (Rank-k spiked covariance)**.** *Let $V_k \in \mathbb{R}^{d\times k}$ have orthonormal columns, let $\Theta = \mathrm{diag}(\theta_1, \dots, \theta_k)$ with $\theta_1 \ge \cdots \ge \theta_k > 0$, and let $X_i = V_k\Theta^{1/2} + \sigma Z_i$, $(Z_i)_{ab} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Set*

$A_i = X_i X_i^\top$, *then* $\Sigma = \mathbb{E}[A_i] = V_k \Theta V_k^\top + k\sigma^2 I_d$. *Running Algorithm 2 with input* $A_1, \ldots, A_n$ *outputs* Q_T *with high probability satisfying*

$$\|Q_T Q_T^\top - V_k V_k^\top\|_F \leq \tilde{O}\left(\frac{\sigma\sqrt{(\theta_1 + k\sigma^2)kd}}{\theta_k\sqrt{n}} + \sigma \cdot \sqrt{\frac{\theta_1 + k\sigma^2}{\theta_k}}\,\frac{Kdk\,\sqrt{\theta_1 + \sigma^2 k}}{\varepsilon\theta_k n}\right)$$

Remark. The privacy term in Corollary 2 scales with the fluctuation level σ and vanishes as $\sigma \to 0$.

To estimate fluctuations around the mean (Algorithm 4), we build on the private range-estimation idea of Liu et al. [2022] (Algorithm 7). The new difficulty is that our updates are matrix-valued: $A_i Q_{t-1} \in \mathbb{R}^{d\times k}$, rather than vector-valued as in the top-eigenvector setting. A naive extension would estimate a range separately for each column and compose privacy over k runs. Instead, Algorithm 4 estimates one range parameter for the whole batch by taking blockwise maxima over the k column fluctuations and running the private histogram learner only once per update step. This lets us update the entire k-dimensional subspace at once, rather than using a deflation-based sequence of private one-dimensional problems as in Düngler and Sanyal [2025].

3.1 Experiments

We compare AdaDPO against the stochastic DP-PCA methods: k-DP-PCA and k-DP-Ojas [Düngler and Sanyal, 2025], and against stochastic adaptations of the DP-Gauss algorithms of Dwork et al. [2014] and the noisy power method of Hardt and Price [2014]. The latter methods were designed for deterministic bounded-data settings, so we adapt them to our stochastic model by clipping.

The algorithms of Dwork et al. [2014] assume a data matrix $X \in \mathbb{R}^{n\times d}$ with rows bounded in Euclidean norm by one and estimate the top eigenspace of $X^\top X$. The noisy power method was originally analyzed under single-entry matrix changes of size at most one. Later analyses [Florina Balcan et al., 2016, Nicolas et al., 2024] showed privacy under perturbations $A' = A + C$ satisfying $\sqrt{\sum_i \|C_{i,:}\|_1^2} \leq 1$. In contrast, our setting observes independent matrices A_i, with no deterministic norm bound, and the goal is to estimate the top eigenspace of $\mathbb{E}[A_i] = \Sigma$.

To use deterministic baselines, we clip the samples. A global normalization $\tilde{x}_i = x_i / \max_j \|x_j\|_2$ is not private, since one sample can change the normalization of all samples [Hu et al., 2024]. Normalizing each sample individually, $\tilde{x}_i = x_i/\|x_i\|_2$, changes the target because generally $\mathbb{E}[xx^\top/\|x\|_2^2] \neq \Sigma$. Instead, as in [Düngler and Sanyal, 2025], we clip at a deterministic threshold R chosen so that clipping is inactive with high probability. For the spiked covariance model, we take $R = C\sqrt{\lambda_1} + \sigma\sqrt{d\log(n/\vartheta)}$, which ensures $\|x_i\|_2 \leq R$ for all i with probability at least $1 - \vartheta$, up to constants. The DP-Gauss privacy noise is then scaled to this clipping radius.

With this clipping, we adapt Algorithms 1 and 2 of Dwork et al. [2014], denoted `DP-Gauss-1` and `DP-Gauss-2`. `DP-Gauss-1` clips each sample, adds Gaussian noise to the empirical second-moment matrix, and runs non-private PCA on the noisy matrix. `DP-Gauss-2` privately estimates the eigengap of the clipped covariance matrix, computes the non-private top-k eigenspace of the clipped data, and perturbs the eigenvectors using noise calibrated to the private eigengap estimate.

We also adapt the noisy power method. Replacing one sample by a adds the rank-one perturbation $C = aa^\top$, for which $\sqrt{\sum_i \|C_{i,:}\|_1^2} = \|a\|_2\|a\|_1$. Thus, to satisfy the adjacency condition of Florina Balcan et al. [2016], Nicolas et al. [2024], we clip so that $\|a\|_2 \leq R$ (same R as for `DP-Gauss`) and $\|a\|_1 \leq R'$, and scale the privacy noise accordingly. In the spiked covariance model, we use $R' = \sigma d + \sqrt{\lambda_1 d} + \sigma\sqrt{d\log(n/\vartheta)}$, so that $\|x_i\|_1 \leq R'$ with probability at least $1 - \vartheta$, up to constants. We call this baseline `DP-Power-Method`.

Experimental Results on Spiked Covariance Data In Figure 1a we vary the ambient dimension, eigengap, and noise level while keeping the other parameters fixed. The plots show that AdaDPO consistently attains lower subspace error. In Figure 1b we vary the sample size across several choices of k and d, showing that the advantage persists across different ranks and dimensions and that AdaDPO benefits from the predicted block k-dependence. Across all plots, the deterministic bounded-data baselines are less competitive, reflecting the cost of clipping and calibrating noise to worst-case bounds rather than to the stochastic fluctuations of the data. Additional experiments and implementation details are provided in Section G.

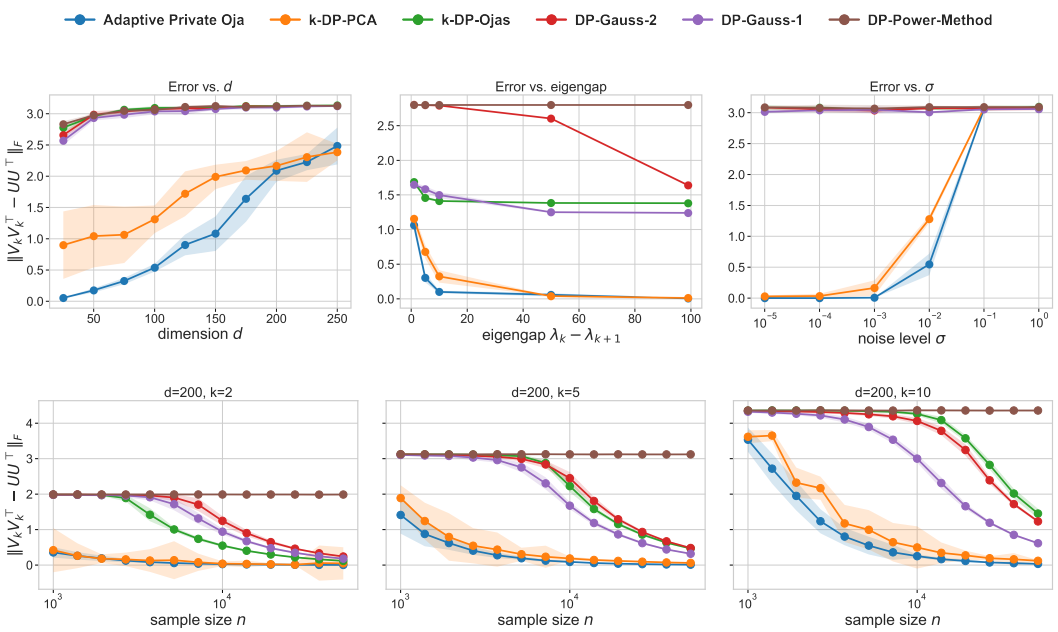


Figure 1: Comparison of AdaDPO with baselines on the spiked covariance model. We plot the mean over 10 trials, with shaded regions representing 95% confidence intervals. We set $\varepsilon = 1$, $\delta = 0.01$. For the top row figures we set $k = 5$.

4 Fine-tuning in tangent space for optimal spectral dependence

AdaDPO has optimal dependence on d, k, n, and ε for general Gaussian inputs, because its privacy noise adapts to the stochastic fluctuation of the data. In the Gaussian spiked covariance model of Cai et al. [2024], the data are Gaussian with covariance $\Sigma = U\Theta U^\top + \sigma^2 I_d$, where $U \in \mathbb{R}^{d\times k}$ has orthonormal columns. Equivalently, all directions outside the signal subspace have the same eigenvalue, $\lambda_{k+1} = \cdots = \lambda_d = \sigma^2$. This additional structure allows a sharper spectral dependence: the minimax lower bound (Theorem 7) scales as $\frac{\sqrt{\lambda_1\lambda_{k+1}}}{\lambda_k - \lambda_{k+1}}\left(\sqrt{\frac{dk}{n}} + \frac{dk}{\varepsilon n}\right)$, whereas applying Theorem 2 to $A_i = x_i x_i^\top$ with $x_i \sim \mathcal{N}(0, , \Sigma)$ defined as above, gives

$$\left\|Q_T Q_T^\top - V_k V_k^\top\right\|_F \le \widetilde{O}\left(\frac{\lambda_1\sqrt{kd}}{(\lambda_k - \lambda_{k+1})\sqrt{n}} + \left(\frac{\lambda_1}{\lambda_k - \lambda_{k+1}}\right)^{3/2}\frac{Kdk}{\varepsilon n}\right).$$

Thus AdaDPO has the correct dependence on the statistical and privacy parameters, but not the sharper spectral dependence available in this special Gaussian family. To close this gap, we use AdaDPO only as a warm-up phase. After it returns a private estimate Q_0 within constant subspace error of the true top-k eigenspace, we run a second-stage refinement on fresh samples. The key observation is that the component of $A_i Q_{t-1}$ inside $\mathrm{span}(Q_{t-1})$ changes the basis of Q_{t-1}, not the represented subspace. The subspace only moves through the tangent component Π_{t-1}. Hence, instead of updating with $A_i Q_{t-1}$, we update with

$$Y_i = \Pi_{t-1} A_i Q_{t-1}, \qquad \Pi_{t-1} = I - Q_{t-1}Q_{t-1}^\top.$$

This removes the component along the current estimate and only estimates the residual correction. In the flat-tail spiked model, this residual fluctuation is governed by the interaction between signal and noise directions, giving the scale $\sqrt{\lambda_1\lambda_{k+1}}$ rather than λ_1. We present the full result of Algorithm 3 in Theorem 5 and Lemma 12 and a short corollary below.

Corollary 3 (Gaussian spiked covariance, Algorithm 3)**.** *Let $A_i = x_i x_i^\top$, where $x_i \sim \mathcal{N}(0, \Sigma)$ and $\Sigma = V_k \Lambda V_k^\top + \sigma^2 I_d$. Suppose the starting point Q_0 is within a constant error of V_k, Algorithm 3 returns Q_T satisfying*

$$\left\|Q_T Q_T^\top - V_k V_k^\top\right\|_F \le \mathrm{Opt}_T + \tilde{O}\left(\frac{\sqrt{\lambda_1\lambda_{k+1}}}{\lambda_k - \lambda_{k+1}}\left[\sqrt{\frac{dk}{n}} + \frac{Kdk}{\varepsilon n}\right] + \frac{\sqrt{k}M_\perp}{(\lambda_k - \lambda_{k+1})n}\right).$$

Algorithm 3 Tangent Adaptive Private Ojas (TADADPO)

Input: samples $S = \{A_i\}_{i=1}^n$, target rank k, batch size B, privacy parameters (ε, δ), failure probability ζ, learning rates $\{\eta_t\}_{t=1}^n$, and a private starting point $Q_0 \in \mathbb{R}^{d\times k}$.

for $t = 1, \dots, T = \lfloor n/B \rfloor$ **do**
  Set $\Pi_{t-1} = I - Q_{t-1}Q_{t-1}^\top$
  Project updates to tangent space $Y_i \leftarrow \Pi_{t-1}A_iQ_{t-1}$ for $i \in [B(t-1)+1, Bt]$
  Estimate Range $\widehat{\Lambda}_t \leftarrow \textsc{PrivRange}(\{Y_i\}_{i=B\cdot(t-1)+1}^{B\cdot(t-1)+B/2}, \varepsilon, \delta, \zeta/(8T))$.
  Set Truncation Threshold $r_t = C_R K\sqrt{\widehat{\Lambda}_t}\log^a\left(\frac{BdkT}{\zeta}\right)$
  $\bar{Y}_t \leftarrow \textsc{TruncatedMean}\Big(\{Y_i\}_{i=B(t-1)+B/2+1}^{Bt}, r_t, \widehat{\Lambda}_t, (\varepsilon/2, \delta/2), \zeta/(8T)\Big)$
  Set privacy noise multiplier $\sigma_t \leftarrow \frac{8r_t\sqrt{dk}}{B\varepsilon}\sqrt{2\log(1.25/\delta)}$
  Draw $Z_t \in \mathbb{R}^{d\times k}$ with i.i.d. entries $\mathcal{N}(0, \sigma_t^2)$
  $Q_t \leftarrow \mathrm{QR}\left(Q_{t-1} + \eta_t\Pi_{t-1}(\bar{Y}_t + Z_t)\right)$
end for
return Q_T

where $\mathrm{Opt}_T := \exp\Big(-c\Delta\sum_{t=1}^T \eta_t\Big)\left\|Q_0Q_0^\top - V_kV_k^\top\right\|_F$, *$\eta_t$ are learning rates, and* $\Delta = \lambda_k - \lambda_{k+1}$. *When the linear Bernstein term is lower order* $\left(M_\perp \lesssim \sqrt{\lambda_1\lambda_{k+1}}\sqrt{dn}\right)$ *and* $K = O(1)$*, this matches the known flat-tail spiked-covariance lower-bound scaling up to logarithmic factors.*

To analyze the refinement phase, we replace the global fluctuation parameters from Assumption A by local tangent parameters. These only need to hold inside the constant-accuracy basin reached by the initializer, because the refinement algorithm is never intended to be run from an arbitrary point.

For $Q \in \mathbb{S}_{d,k}$, define $e(Q) := \left\|QQ^\top - V_kV_k^\top\right\|_F$, and $Z_Q(A) := \Pi_Q(A-\Sigma)Q$. The matrix Bernstein variance of the full tangent update is

$$v_Q^2 := \max\left\{\left\|\mathbb{E}\left[Z_Q(A)Z_Q(A)^\top\right]\right\|_2, \left\|\mathbb{E}\left[Z_Q(A)^\top Z_Q(A)\right]\right\|_2\right\}. \tag{1}$$

Writing the columns of Q as $q_1, \dots, q_k$, define the largest one-column tangent scale by

$$g_Q^2 := \max_{r\in[k]}\left\|\Pi_Q\mathbb{E}\left[(A-\Sigma)q_rq_r^\top(A-\Sigma)^\top\right]\Pi_Q\right\|_2. \tag{2}$$

Here v_Q controls concentration of the tangent minibatch mean, while g_Q is the scale used by the private range-estimation and truncation steps.

Assumption B (Tangent refinement model)**.** *There exist constants $c_0 \in (0,1)$, $M_\perp \geq 0$, $v_0, v_1 \geq 0$, $g_0, g_1 \geq 0$, $K \geq 1$, and $a \geq 1/2$ such that the following hold for every $Q \in \mathbb{S}_{d,k}$ satisfying $\left\|QQ^\top - V_kV_k^\top\right\|_2 \leq c_0$.*

B.1 $\|\Pi_Q(A_i-\Sigma)Q\|_2 \leq M_\perp$, $v_Q \leq v_0 + v_1e(Q)$, $g_Q \leq g_0 + g_1e(Q)$

B.2 *For every column q_r of Q, every unit vector $u \perp \mathrm{span}(Q)$, and every $\vartheta \in (0,1)$,*

$$\Pr\left(\left|u^\top(A_i-\Sigma)q_r\right| > Kg_Q\log^a(1/\vartheta)\right) \leq \vartheta.$$

Remark. Assumption B is a localized version of the global model used for ADADPO. The global assumptions imply the boundedness and variance parts of this model with $M_\perp \leq M$, $v_0 = \lambda_1\sqrt{V}$, and $v_1 = 0$, and give the crude tail scale $\lambda_1\gamma$. In structured settings, such as the flat-tail spiked covariance model, the sharper g_Q-scale tail condition can be verified directly.

The experiment in Figure 2 illustrates the effect of the tangent refinement phase. We first run ADADPO to obtain a private constant-accuracy warm start, and then run TADADPO on fresh samples. Once the warm start is accurate enough, the tangent phase exploits the smaller residual fluctuations and achieves lower error than continuing with ADADPO alone. For the general privacy and utility guarantees of TADADPO, see Section C.5.

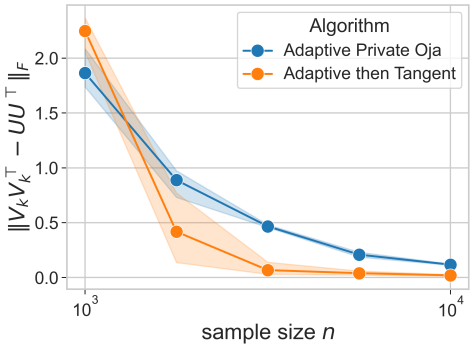


Figure 2: Comparison of AdaDPO with a two-stage variant that warm-starts with AdaDPO and then runs TAdaDPO.

5 Related Work and Conclusion

Related Work. Private PCA has a long history in the fixed-dataset setting [Blum et al., 2005, Chaudhuri et al., 2013, Hardt and Roth, 2013, Dwork et al., 2014, Florina Balcan et al., 2016, Nicolas et al., 2024]. These methods provide input-wise privacy guarantees, but when applied directly to stochastic PCA they typically rely on worst-case sensitivity bounds or clipping, leading to suboptimal stochastic rates.

Adaptive private algorithms for stochastic PCA were introduced by Liu et al. [2022] for $k = 1$, using privacy noise calibrated to stochastic fluctuations. The extension of this idea to $k > 1$ by Düngler and Sanyal [2025] is deflation-based and therefore incurs extra dependence on k. In contrast, our method treats the top-k subspace directly through block updates. Cai et al. [2024] obtain optimal guarantees for Gaussian data under a flat-tail condition on the unknown covariance Σ. In contrast our privacy guarantee instead holds input-wise for every dataset.

Dimension-independent sample complexity bounds have also been obtained under strong multiplicative eigengap assumptions [Singhal and Steinke, 2021, Tsfadia, 2024]. Recent work on private spectral perturbation and low-rank approximation [Tran et al., 2025] analyzes a fixed noisy matrix, whereas our algorithm releases only adaptive projected block actions. Other results study coherent matrix models [d'Orsi and Novikov, 2026], or decentralized privacy models [Campbell et al., 2025, Nicolas et al., 2024]. However, they use different adjacency notions which when transformed to our more general adjacency notion yield suboptimal error-guarantees. A more detailed comparison with these results is provided in Section B.

Open Problems and Conclusion. Several directions remain open. Since AdaDPO is not deflation-based and updates the full $d \times k$ iterate directly, it is naturally suited to streaming implementations. An interesting next step is to develop a fully online version with adaptive privacy calibration and comparable utility guarantees. A second direction is to weaken the tail assumptions behind the private range estimator. Our current analysis relies on concentration of the projected fluctuations, and extending the range-estimation and truncation steps to heavy-tailed data would substantially broaden the applicability of the method.

To the best of our knowledge, our results give the first differentially private algorithm for stochastic k-PCA that is sample-complexity optimal in both d and k as well as spectral parameters in some settings, while calibrating its privacy noise to the inherent stochasticity of the data.

6 Acknowledgement

JD acknowledges support from the Danish Data Science Academy, which is funded by the Novo Nordisk Foundation (NNF21SA0069429) and VILLUM FONDEN (40516). AS acknowledges the Novo Nordisk Foundation for support via the Startup grant (NNF24OC0087820) and VILLUM FONDEN via the Young Investigator program (72069). The authors would also like to thank Rasmus Pagh for very insightful discussions.

References

John M Abowd, Gary L Benedetto, Simson L Garfinkel, Scot A Dahl, Aref N Dajani, Matthew Graham, Michael B Hawes, Vishesh Karwa, Daniel Kifer, Hang Kim, et al. The modernization of statistical disclosure limitation at the us census bureau. *URL: bit. ly/DPcensus20*, 2020.

Differential Privacy Team Apple. Learning with privacy at scale, 2017. `https://machinelearning.apple.com/research/learning-with-privacy-at-scale`.

Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the sulq framework. In *Principles of Database Systems (PODS)*, 2005.

T Tony Cai, Dong Xia, and Mengyue Zha. Optimal differentially private pca and estimation for spiked covariance matrices. *arXiv:2401.03820*, 2024.

Andrew Campbell, Anna Scaglione, and Sean Peisert. Decentralized differentially private power method. *arXiv preprint arXiv:2507.22849*, 2025.

Kamalika Chaudhuri, Anand D Sarwate, and Kaushik Sinha. A near-optimal algorithm for differentially-private principal components. *Journal of Machine Learning Research*, 2013.

Johanna Düngler and Amartya Sanyal. An iterative algorithm for differentially private k-pca with adaptive noise. In *Neural Information Processing Systems (NeurIPS)*, 2025.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography (TCC)*, 2006.

Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Symposium on Theory of Computing (STOC)*, 2014.

Tommaso d'Orsi and Gleb Novikov. Tight differentially private pca via matrix coherence. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 10–51. SIAM, 2026.

Maria Florina Balcan, Simon S Du, Yining Wang, and Adams Wei Yu. An improved gap-dependency analysis of the noisy power method. In *Conference on Algorithmic Learning Theory (ALT)*, 2016.

Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. *Neural Information Processing Systems (NeurIPS)*, 2014.

Moritz Hardt and Aaron Roth. Beyond worst-case analysis in private singular vector computation. In *Symposium on Theory of Computing (STOC)*, 2013.

Yaxi Hu, Amartya Sanyal, and Bernhard Schölkopf. Provable privacy with non-private pre-processing. In *International Conference on Learning Representations (ICLR)*, 2024.

De Huang, Jonathan Niles-Weed, and Rachel Ward. Streaming k-pca: Efficient guarantees for oja's algorithm, beyond rank-one updates. In *Conference on Algorithmic Learning Theory (ALT)*, 2021.

Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. In *Innovations in Theoretical Computer Science Conference*, 2017.

Xiyang Liu, Weihao Kong, Prateek Jain, and Sewoong Oh. Dp-pca: Statistically optimal and differentially private pca. In *Neural Information Processing Systems (NeurIPS)*, 2022.

Julien Nicolas, César Sabater, Mohamed Maouche, Sonia Ben Mokhtar, and Mark Coates. Differentially private and decentralized randomized power method. *arXiv:2411.01931*, 2024.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 1982.

Erkki Oja and Juha Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of mathematical analysis and applications*, 106(1): 69–84, 1985.

Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

Vikrant Singhal and Thomas Steinke. Privately learning subspaces. *Neural Information Processing Systems (NeurIPS)*, 2021.

Phuc Tran, Nisheeth K Vishnoi, and Van H Vu. Spectral perturbation bounds for low-rank approximation with applications to privacy. *arXiv preprint arXiv:2510.25670*, 2025.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics (FoCM)*, 2012.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and trends® in machine learning*, 8(1-2):1–230, 2015.

Eliad Tsfadia. On differentially private subspace estimation in a distribution-free setting. In *Neural Information Processing Systems (NeurIPS)*, 2024.

A Math Preliminaries

Lemma 3 (Variational characterization of the Schatten-$(2,k)$ norm)**.** *Let $B \in \mathbb{R}^{d\times d}$, and let*

$$\mathbb{S}_{d,k} := \{P \in \mathbb{R}^{d\times k} : P^\top P = I_k\}.$$

Then

$$\sup_{P\in\mathbb{S}_{d,k}} \|P^\top B\|_F = \|B\|_{(2,k)} := \left(\sum_{j=1}^{k} \sigma_j(B)^2\right)^{1/2},$$

where $\sigma_1(B) \geq \cdots \geq \sigma_d(B) \geq 0$ are the singular values of B.

Proof. For any $P \in \mathbb{S}_{d,k}$,

$$\|P^\top B\|_F^2 = \text{Tr}(B^\top PP^\top B) = \text{Tr}(P^\top BB^\top P).$$

Since $P^\top P = I_k$, the matrix $PP^\top$ is the orthogonal projector onto a k-dimensional subspace of $\mathbb{R}^d$. Hence

$$\|P^\top B\|_F^2 = \text{Tr}(P^\top BB^\top P)$$

is the sum of the Rayleigh quotients of $BB^\top$ over an orthonormal k-frame. By Ky Fan's maximum principle,

$$\sup_{P\in\mathbb{S}_{d,k}} \text{Tr}(P^\top BB^\top P) = \sum_{j=1}^{k} \lambda_j(BB^\top),$$

where $\lambda_1(BB^\top) \geq \cdots \geq \lambda_d(BB^\top) \geq 0$ are the eigenvalues of $BB^\top$. Since

$$\lambda_j(BB^\top) = \sigma_j(B)^2,$$

it follows that

$$\sup_{P\in\mathbb{S}_{d,k}} \|P^\top B\|_F^2 = \sum_{j=1}^{k} \sigma_j(B)^2.$$

Taking square roots yields the claim. □

Lemma 4 (Minibatch concentration in top-k Frobenius norm)**.** *Let $X_i := A_i - \Sigma$, and suppose Assumption A holds. Let $I_t \subset [n]$ be disjoint batches of size m, and define*

$$\bar{X}_t := \frac{1}{m}\sum_{i\in I_t} X_i.$$

Then there exists a universal constant $C > 0$ such that, with probability at least $1-\zeta$, simultaneously for all $t \in [T]$,

$$\|\bar{X}_t\|_{(2,k)} \leq C\sqrt{k}\left(\lambda_1\sqrt{\frac{V\log(2dT/\zeta)}{m}} + \frac{M\log(2dT/\zeta)}{m}\right),$$

Proof. Let $X_i = A_i - \Sigma$. Then X_i are independent, symmetric, and mean-zero.

First, by Assumption A.2,

$$\|X_i\|_{\mathrm{op}} \le \|X_i\|_{(2,k)} = \sup_{P \in \mathbb{S}_{d,k}} \|P^\top X_i\|_F \le \lambda_1 M \qquad \text{a.s.}$$

Hence each summand is almost surely bounded in operator norm by $\lambda_1 M$.

By the Liu-style variance assumption Assumption A.3,

$$\left\|\mathbb{E}[X_i^2]\right\|_{\mathrm{op}} \le \lambda_1^2 V.$$

Now fix a batch I_t of size m. Matrix Bernstein applied to

$$\sum_{i \in I_t} X_i$$

with almost-sure bound $\lambda_1 M$ and variance proxy

$$\left\|\sum_{i \in I_t} \mathbb{E}[X_i^2]\right\|_{\mathrm{op}} \le m\lambda_1^2 V$$

implies that, with probability at least $1 - \alpha$,

$$\left\|\frac{1}{m}\sum_{i \in I_t} X_i\right\|_{\mathrm{op}} \le C\lambda_1 \left(\sqrt{\frac{V \log(2d/\alpha)}{m}} + \frac{M \log(2d/\alpha)}{m}\right).$$

Set $\alpha = \zeta/T$ and union bound over $t \in [T]$. Then, with probability at least $1 - \zeta$, the above operator-norm bound holds simultaneously for every batch.

Finally, for every matrix B,

$$\|B\|_{(2,k)} = \left(\sum_{j=1}^{k} \sigma_j(B)^2\right)^{1/2} \le \sqrt{k}\,\|B\|_{\mathrm{op}}.$$

Applying this inequality to $B = \bar{X}_t$ gives

$$\|\bar{X}_t\|_{(2,k)} \le C\sqrt{k}\lambda_1 \left(\sqrt{\frac{V \log(2dT/\zeta)}{m}} + \frac{M \log(2dT/\zeta)}{m}\right)$$

simultaneously for all $t \in [T]$, as claimed. □

Definition 4 (ζ-approximate Utility)**.** We say $U \in \mathbb{R}^{d \times k}$ is ζ-approximate if U has orthonormal columns and

$$\langle UU^\top, \Sigma \rangle \ge (1 - \zeta^2)\langle V_k V_k^\top, \Sigma \rangle.$$

Lemma 5. *If* $\langle UU^\top, \Sigma \rangle \ge (1 - \zeta^2)\langle V_k V_k^\top, \Sigma \rangle$ *then*

$$\|UU^\top - V_k V_k^\top\|_F \le \zeta \sqrt{\frac{2\sum_{i=1}^{k} \lambda_i}{\lambda_k - \lambda_{k+1}}}$$

Proof. Let

$$P := UU^\top, \qquad P_\star := V_k V_k^\top.$$

Since $P_\star$ projects onto the top-k eigenspace of Σ, we have

$$\langle P_\star, \Sigma \rangle = \sum_{i=1}^{k} \lambda_i.$$

The assumption implies

$$\langle P_\star, \Sigma\rangle - \langle P, \Sigma\rangle \le \zeta^2 \langle P_\star, \Sigma\rangle = \zeta^2 \sum_{i=1}^{k} \lambda_i.$$

We next lower bound the same quantity in terms of the distance between projectors. Let $v_1, \dots, v_d$ be an orthonormal eigenbasis of Σ, ordered so that

$$\Sigma = \sum_{i=1}^{d} \lambda_i v_i v_i^\top.$$

Define

$$a_i := v_i^\top P v_i.$$

Because P is a rank-k orthogonal projector, $0 \le a_i \le 1$ and

$$\sum_{i=1}^{d} a_i = \mathrm{tr}(P) = k.$$

Therefore,

$$\begin{aligned}\langle P_\star, \Sigma\rangle - \langle P, \Sigma\rangle &= \sum_{i=1}^{k} \lambda_i (1 - a_i) - \sum_{i=k+1}^{d} \lambda_i a_i \\ &\ge \lambda_k \sum_{i=1}^{k} (1 - a_i) - \lambda_{k+1} \sum_{i=k+1}^{d} a_i.\end{aligned}$$

Since

$$\sum_{i=1}^{k} (1 - a_i) = k - \sum_{i=1}^{k} a_i = \sum_{i=k+1}^{d} a_i,$$

we get

$$\langle P_\star, \Sigma\rangle - \langle P, \Sigma\rangle \ge (\lambda_k - \lambda_{k+1}) \left(k - \sum_{i=1}^{k} a_i \right).$$

Moreover,

$$\sum_{i=1}^{k} a_i = \mathrm{tr}(P_\star P),$$

so

$$k - \sum_{i=1}^{k} a_i = k - \mathrm{tr}(P_\star P).$$

Using the fact that P and $P_\star$ are rank-k orthogonal projectors,

$$\|P - P_\star\|_F^2 = \mathrm{tr}(P) + \mathrm{tr}(P_\star) - 2\,\mathrm{tr}(P P_\star) = 2\left(k - \mathrm{tr}(P P_\star)\right).$$

Hence

$$\langle P_\star, \Sigma\rangle - \langle P, \Sigma\rangle \ge \frac{\lambda_k - \lambda_{k+1}}{2} \|P - P_\star\|_F^2.$$

Combining this with the upper bound from the assumption gives

$$\frac{\lambda_k - \lambda_{k+1}}{2} \|P - P_\star\|_F^2 \le \zeta^2 \sum_{i=1}^{k} \lambda_i.$$

Therefore,

$$\|P - P_\star\|_F^2 \le \zeta^2 \frac{2 \sum_{i=1}^{k} \lambda_i}{\lambda_k - \lambda_{k+1}}.$$

Taking square roots and substituting back $P = UU^\top$ and $P_\star = V_k V_k^\top$ yields

$$\|UU^\top - V_k V_k^\top\|_F \le \zeta \sqrt{\frac{2 \sum_{i=1}^{k} \lambda_i}{\lambda_k - \lambda_{k+1}}}.$$

□

Lemma 6 (High-probability bound for GOE)**.** *Let $G \sim \mathrm{GOE}_d(\sigma^2)$. Then for every $1 \le k \le d$ and every $t \ge 0$,*

$$\Pr\Big(\|G\|_{(2,k)} \;\le\; \sqrt{k}\,\big(2\sigma\sqrt{d}+t\big)\Big) \ge 1 - 2\exp\!\left(-\frac{t^2}{4\sigma^2}\right).$$

Equivalently, for every $\delta \in (0,1)$,

$$\Pr\left(\|G\|_{(2,k)} \;\le\; 2\sigma\sqrt{kd} + 2\sigma\sqrt{k\log\!\left(\frac{2}{\delta}\right)}\right) \ge 1-\delta.$$

Proof. By Lemma 3,

$$\|G\|_{(2,k)} = \sup_{P\in\mathbb{S}_{d,k}} \|P^\top G\|_F = \left(\sum_{i=1}^{k}\sigma_i(G)^2\right)^{1/2} \le \sqrt{k}\,\sigma_1(G) = \sqrt{k}\,\|G\|_{\mathrm{op}}.$$

For $G \sim \mathrm{GOE}_d(\sigma^2)$, the spectral norm satisfies

$$\Pr\Big(\|G\|_{\mathrm{op}} \ge 2\sigma\sqrt{d}+t\Big) \le 2\exp\!\left(-\frac{t^2}{4\sigma^2}\right), \qquad t \ge 0,$$

which proves the first claim. Setting $t = 2\sigma\sqrt{\log(2/\delta)}$ yields the second. □

Lemma 7 (Neumann series)**.** *Let A be a random square matrix, let G be an event, and suppose that $\|A\mathbf{1}_G\|_2 \le \varepsilon$ almost surely for a constant $\varepsilon < 1$. Then on G the matrices $I - A$, $I + A$ and $I - A^2$ are invertible, and*

$$\big\|(I-A^2)^{-1}\mathbf{1}_G\big\|_2 \le \frac{1}{1-\varepsilon^2}, \tag{3}$$

$$\big\|\big[(I-A^2)^{-1} - I\big]\mathbf{1}_G\big\|_2 \le \frac{\varepsilon^2}{1-\varepsilon^2}. \tag{4}$$

Consequently, for every random matrix X with $X = X\mathbf{1}_G$,

$$\|X\|_F \le \frac{1}{1-\varepsilon^2}\,\big\|X(I-A^2)\big\|_F \qquad \textit{pointwise.} \tag{5}$$

Outside G nothing is asserted: the hypothesis constrains A only where $\mathbf{1}_G = 1$, so $I - A^2$ need not be invertible there, and $(I-A^2)^{-1}$ is to be read as an arbitrary measurable choice, say I. The choice is immaterial, since $(I-A^2)^{-1}$ occurs only against $\mathbf{1}_G$.

Proof. Fix $\omega \in G$ and put $B := A(\omega)^2$. Submultiplicativity of $\|\cdot\|_2$ and the hypothesis give $\|B\|_2 \le \|A(\omega)\|_2^2 \le \varepsilon^2 < 1$. Hence $\|B^j\|_2 \le \|B\|_2^j \le \varepsilon^{2j}$ and the series $\sum_{j\ge0} B^j$ converges absolutely in operator norm. Its partial sums telescope against $I - B$,

$$(I-B)\sum_{j=0}^{m} B^j = I - B^{m+1} \xrightarrow[m\to\infty]{} I,$$

because $\|B^{m+1}\|_2 \le \varepsilon^{2(m+1)} \to 0$; the same computation on the other side gives $\sum_{j=0}^{m} B^j(I-B) \to I$. Therefore $I - B = I - A(\omega)^2$ is invertible with $(I-A(\omega)^2)^{-1} = \sum_{j\ge0} B^j$ and

$$\big\|(I-A(\omega)^2)^{-1}\big\|_2 \le \sum_{j\ge0}\|B\|_2^j \le \sum_{j\ge0}\varepsilon^{2j} = \frac{1}{1-\varepsilon^2},$$

which is (3). Since A commutes with itself, $I - A^2 = (I-A)(I+A)$, and a product of square matrices is invertible only if both factors are; this gives the invertibility of $I \pm A$. Subtracting the term $j = 0$ from the series,

$$(I-A(\omega)^2)^{-1} - I = \sum_{j\ge1} B^j = B\,(I-A(\omega)^2)^{-1},$$

so its norm is at most $\varepsilon^2(1-\varepsilon^2)^{-1}$, which is (4).

For (5): on G we may write $X = X(I-A^2)(I-A^2)^{-1}$, while off G both sides of the asserted inequality are unaffected because $X = X\mathbf{1}_G$ vanishes there. Hence, pointwise, $\|X\|_F \le \|X(I-A^2)\|_F\,\|(I-A^2)^{-1}\mathbf{1}_G\|_2 \le (1-\varepsilon^2)^{-1}\|X(I-A^2)\|_F$ by (3), which is the claim. □

B Comparisons to Prior Work

We provide additional details on the comparisons summarized in the related-work section (Section 5).

Comparison to deterministic bounded-input methods. Several works address k-PCA in the standard setting, while assuming an additive eigengap [Blum et al., 2005, Chaudhuri et al., 2013, Hardt and Roth, 2013, Dwork et al., 2014, Nicolas et al., 2024]. These works operate in a deterministic setting where each sample is assumed to be bounded ($\|x_i\| \leq \beta$). When applied to the stochastic setting, these works generally yield suboptimal error rates. This is partially due to the fact that all of these works assume a data independent bound ($\beta = 1$), which we cannot easily enforce in the stochastic setting. Considering Gaussian data with $x_i \sim \mathcal{N}(0, \Sigma)$, we know $\|x_i\| \leq \beta = O(\sqrt{\lambda_1 d \log(n/\zeta)}$ for all i with probability $1 - \zeta$. [Blum et al., 2005, Dwork et al., 2014, Nicolas et al., 2024] use the Gaussian mechanism, so when scaling the privacy noise with a factor β we ensure (ε, δ)-DP in the stochastic setting. The tightest of the previous discussed result then achieves

$$O\left(\sqrt{dk/n} + d^{3/2}k/(\varepsilon n)\right).$$

Comparison to Düngler and Sanyal [2025]. Under assumptions comparable to Assumption A, and for inputs of the form $A_i = x_i x_i^\top$ with $x_i \sim \mathcal{N}(0, \Sigma)$, the main result of Düngler and Sanyal [2025] implies that, with high probability, the output subspace U is ζ-approximate in the sense that

$$\langle UU^\top, \Sigma \rangle \geq (1 - \zeta^2)\langle V_k V_k^\top, \Sigma \rangle,$$

where

$$\zeta = \tilde{O}\left(\kappa'\left(\sqrt{\frac{dk}{n}} + \frac{dk\sqrt{\log(1/\delta)}}{\varepsilon n}\right)\right).$$

Here $\tilde{O}(\cdot)$ suppresses polylogarithmic factors in $n, d, 1/\varepsilon$, and $\log(1/\delta)$. By Lemma 5, this energy guarantee implies the Frobenius-norm bound

$$\|UU^\top - V_k V_k^\top\|_F \leq \zeta\sqrt{\frac{2\sum_{i=1}^k \lambda_i}{\lambda_k - \lambda_{k+1}}}.$$

Comparing this rate with the lower bound Corollary 5 in Section D.1 shows that the guarantee of Düngler and Sanyal [2025] is loose by a factor $\sqrt{k}$ in Frobenius norm. The same loss appears when their energy-based utility is compared to the corresponding lower bound. In addition, the error scaling in their main theorem holds only under a sample-size condition containing an exponential eigengap-dependent term. Our result removes this condition and attains the optimal dependence on k.

A further difference is algorithmic. The method of Düngler and Sanyal [2025] recovers the leading eigenspace one direction at a time, requiring stage-wise choices of parameters. Our method instead updates the entire k-dimensional subspace simultaneously. This avoids retuning across stages.

Comparison to fixed-matrix perturbation analyses. The results of Tran et al. [2025] give refined perturbation bounds for private low-rank approximation by analyzing a fixed noisy matrix

$$\widetilde{M} = M + E$$

and controlling the error between the best rank-p approximations of M and $\widetilde{M}$. These results are not directly applicable to our algorithm because we do not release a noisy full covariance matrix. Instead, at step t, our method releases only the projected block action

$$\frac{1}{n}\sum_{i=1}^n A_i Q_{t-1} + Z_t,$$

where Q_{t-1} depends on previous private iterates. The privacy noise is therefore calibrated to the range of the projected matrices $A_i Q_{t-1}$, rather than to a global $d \times d$ perturbation of the covariance matrix. This distinction is essential: the projected stochastic fluctuations can be substantially smaller than the worst-case full-matrix sensitivity.

Comparison to coherent matrix models. The work of d'Orsi and Novikov [2026] studies DP-PCA under matrix coherence assumptions in a non-stochastic setting where the input is a single matrix. Their bounds are stated for

$$\|(I - UU^\top)V_k\|_2,$$

which is related to the Frobenius subspace error by

$$\|V_k V_k^\top - UU^\top\|_F \le \sqrt{2k}\,\|(I - UU^\top)V_k\|_2.$$

To apply their result to the empirical covariance

$$\widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^\top,$$

one must translate their matrix-level adjacency notion into sample-level privacy. In their notation, two matrices M, M' are Δ_{mat}-adjacent if, for $E = M - M'$,

$$\sqrt{\|EE^\top\|_1} \le \Delta_{\mathrm{mat}}.$$

Suppose the Gaussian samples are clipped so that $\|x_i\|_2^2 \le R_\eta^2$. Replacing one sample x by another sample y changes the empirical covariance by

$$E = \frac{1}{N}(xx^\top - yy^\top).$$

Hence

$$\Delta_{\mathrm{mat}} \le \frac{1}{N}\sqrt{\|(xx^\top - yy^\top)^2\|_1} \le \frac{2\sqrt{d}\,R_\eta^2}{N}.$$

Plugging this into their rank-k private eigenspace bound, and using the worst-case coherence bound $k\mu_k(\widehat{\Sigma}) \le d$, gives a privacy error of order

$$\widetilde{O}\left(\frac{R_\eta^2 d}{N\varepsilon\gamma_k}\right), \qquad \gamma_k = \lambda_k(\Sigma) - \lambda_{k+1}(\Sigma).$$

For $x \sim \mathcal{N}(0, \Sigma)$, Gaussian quadratic-form concentration gives

$$R_\eta^2 \lesssim \mathrm{tr}(\Sigma) + \sqrt{\lambda_1(\Sigma)\,\mathrm{tr}(\Sigma)\log(N/\eta)} + \lambda_1(\Sigma)\log(N/\eta) = \widetilde{O}(\mathrm{tr}(\Sigma)).$$

Combining this privacy term with the usual non-private Davis–Kahan statistical error yields the population subspace bound

$$\|(I - \widehat{P})V_k\|_2 \lesssim_{\log} \frac{\lambda_1(\Sigma)}{\gamma_k}\sqrt{\frac{d}{N}} + \frac{\mathrm{tr}(\Sigma)d}{N\varepsilon\gamma_k}.$$

The privacy term arises from a worst-case matrix-adjacency conversion. It therefore does not adapt to the stochastic spread of the projected gradients, which is the source of the sharper dependence in our adaptive mechanism.

Decentralized and alternative adjacency models. There are also private PCA algorithms for decentralized settings [Campbell et al., 2025, Nicolas et al., 2024]. These works use privacy and partition models different from the sample-level stochastic setting considered here. For example, Campbell et al. [2025] use a power-method-based approach under a neighboring relation based on changing one row of the data matrix, with $\|X - X'\| \le 1$. This does not directly correspond to our model, where each sample contributes a matrix $A_i = x_i x_i^\top$. In particular, changing one sample changes one rank-one matrix contribution to the empirical covariance, rather than a single bounded row in the data matrix under their adjacency model.

Similarly, the original noisy power method was analyzed for matrix inputs under single-entry changes of magnitude at most one. Later analyses [Florina Balcan et al., 2016, Nicolas et al., 2024] extended the privacy guarantee to perturbations of the form $A' = A + C$ satisfying

$$\sqrt{\sum_i \|C_{i,:}\|_1^2} \le 1.$$

These matrix-adjacency guarantees are useful in fixed-dataset PCA, but they do not directly yield the stochastic, sample-level rates considered in this paper.

C Proofs

C.1 Privacy of Algorithm 1

Theorem 3 (Privacy of Algorithm 1)**.** *If $0 < \varepsilon \leq 1$, then Algorithm 1 is (ε, δ)-DP under replacement neighboring datasets.*

Proof. At update step t, the matrices A_i are accessed only through

$$\mu_t = \frac{1}{B} \sum_{i=B(t-1)+1}^{tB} \mathrm{clip}_R(A_i Q_{t-1}).$$

Although Q_{t-1} is data-dependent, it is determined entirely by the outputs of the previous update steps. Hence, conditional on all previous outputs, Q_{t-1} is fixed, and we may analyze the privacy of the t-th update step with Q_{t-1} treated as a constant.

If one matrix A_i in the t-th batch is replaced by A_i', producing μ_t', then

$$\|\mu_t - \mu_t'\|_F = \frac{1}{B} \left\|\mathrm{clip}_R(A_i Q_{t-1}) - \mathrm{clip}_R(A_i' Q_{t-1})\right\|_F \leq \frac{2R}{B}.$$

Therefore, by Lemma 8, the t-th update step is (ε, δ)-DP, even conditional on the previous outputs. Since the algorithm is one-pass and each datum A_i is used in exactly one batch, the overall algorithm is (ε, δ)-DP by parallel composition. □

Lemma 8 (Privacy of Single Update step of Algorithm 1)**.** *Fix $R > 0$ and $Q \in \mathbb{R}^{d\times k}$ with $Q^\top Q = I_k$, and define*

$$f_Q(A_1, \ldots, A_B) := \frac{1}{B} \sum_{i=1}^{B} \mathrm{clip}_R(A_i Q),$$

where $A_1, \ldots, A_B$ are symmetric. Consider the mechanism

$$M(A_1, \ldots, A_B; Q) = f_Q(A_1, \ldots, A_B) + W_Q,$$

where

$$W_Q := QN + N_\perp, \qquad N_\perp = (I - QQ^\top)Z,$$

with $N \sim GOE_k(\sigma^2)$ and $Z \in \mathbb{R}^{d\times k}$ having i.i.d. entries in $\mathcal{N}(0, \sigma^2)$. Then $M(\cdot; Q)$ is (ε, δ)-differentially private whenever

$$\sigma \geq \frac{\Delta_Q}{\varepsilon}\sqrt{2\log(1.25/\delta)}, \qquad \Delta_Q := \sup_{(A_1,\ldots,A_B)\sim(A_1',\ldots,A_B')} \|f_Q(A_1, \ldots, A_B) - f_Q(A_1', \ldots, A_B')\|_F.$$

In particular, if neighboring batches differ in one entry, then

$$\Delta_Q \leq \frac{2R}{B},$$

so it suffices to choose

$$\sigma \geq \frac{2R}{B\varepsilon}\sqrt{2\log(1.25/\delta)}.$$

Proof. First, each $\mathrm{clip}_R(A_i Q)$ lies in

$$\mathcal{S}_Q = \{M \in \mathbb{R}^{d\times k} : Q^\top M \text{ is symmetric}\}.$$

Indeed, $A_i Q \in \mathcal{S}_Q$ because $Q^\top A_i Q$ is symmetric whenever $A_i = A_i^\top$, and clipping only rescales by a nonnegative scalar. Hence the average $f_Q(A_1, \ldots, A_B)$ also lies in $\mathcal{S}_Q$.

Next, since $N \sim GOE_k(\sigma^2)$, we may write

$$N \overset{d}{=} N_1 + N_2,$$

where N_1, N_2 are independent draws from $\mathrm{GOE}_k(\sigma^2/2)$. Therefore

$$W_Q = QN + N_\perp \overset{d}{=} QN_1 + N_\perp + QN_2.$$

Define

$$W_Q^{(1)} := QN_1 + N_\perp, \qquad W_Q^{(2)} := QN_2.$$

Then

$$M(A_1, \dots, A_B; Q) \stackrel{d}{=} f_Q(A_1, \dots, A_B) + W_Q^{(1)} + W_Q^{(2)}.$$

We now verify that $W_Q^{(1)}$ is an isotropic Gaussian on $\mathcal{S}_Q$. Every element of $\mathcal{S}_Q$ can be written uniquely as

$$M = QS + C, \qquad S = S^\top, \qquad Q^\top C = 0.$$

The two components QS and C are orthogonal in Frobenius inner product. The noise $W_Q^{(1)}$ has exactly this form:

$$W_Q^{(1)} = QN_1 + N_\perp, \qquad N_1 = N_1^\top, \qquad Q^\top N_\perp = 0.$$

By construction, N_1 is the Frobenius-isotropic Gaussian on the space of symmetric $k \times k$ matrices, and $N_\perp = (I - QQ^\top)Z$ is the Frobenius-isotropic Gaussian on the subspace $\{C : Q^\top C = 0\}$. Since these two parts are independent and orthogonal, $W_Q^{(1)}$ is a centered Gaussian with covariance $\sigma^2 I_{\mathcal{S}_Q}$ on $\mathcal{S}_Q$.

Therefore $f_Q + W_Q^{(1)}$ is the Gaussian mechanism in the Hilbert space $(\mathcal{S}_Q, \langle \cdot, \cdot \rangle_F)$, and it is (ε, δ)-DP whenever

$$\sigma \geq \frac{\Delta_Q}{\varepsilon} \sqrt{2 \log(1.25/\delta)}.$$

Since $W_Q^{(2)}$ is independent of the input data, adding it is post-processing. Hence $M(\cdot; Q)$ is also (ε, δ)-DP under the same condition.

If neighboring batches differ only in the j-th entry, then

$$f_Q(A_1, \dots, A_B) - f_Q(A_1', \dots, A_B') = \frac{1}{B}\Big(\mathrm{clip}_R(A_j Q) - \mathrm{clip}_R(A_j' Q)\Big),$$

and so

$$\|f_Q(A_1, \dots, A_B) - f_Q(A_1', \dots, A_B')\|_F \leq \frac{1}{B}\Big(\|\,\mathrm{clip}_R(A_j Q)\|_F + \|\,\mathrm{clip}_R(A_j' Q)\|_F\Big) \leq \frac{2R}{B}.$$

Thus $\Delta_Q \leq 2R/B$, which yields the claim. □

C.2 Utility of Algorithm 1

Theorem 1 (Utility of Algorithm 1)**.** *Assume $A_1, \dots, A_n$ satisfy Assumption A and write $\kappa = \frac{\lambda_1}{\lambda_k - \lambda_{k+1}}$. Set*

$$R = C\lambda_1 \sqrt{dk}\big(K\gamma \log^a(ndk/\zeta) + 1\big), \qquad T = \lceil C_T \kappa \log(n) \rceil,$$

where C_T is a sufficiently large absolute constant. If

$$n \gtrsim \max\left\{ \frac{k^2 \lambda_1^2 V}{\zeta^2 (\lambda_k - \lambda_{k+1})^2}, \kappa^{3/2} \frac{kM}{\zeta}, \kappa^{3/2} \frac{(K\gamma + 1) k^{3/2} d}{\varepsilon \zeta} \right\},$$

then there exist learning rates $\{\eta_t\}$ so that with probability at least $1 - \zeta$, Algorithm 1 outputs Q_T satisfying

$$\big\| Q_T Q_T^\top - V_k V_k^\top \big\|_F \leq \tilde{O}\left(\kappa \frac{\sqrt{kV}}{\sqrt{n}} + \kappa^{3/2} \frac{\sqrt{k} M}{n} + \kappa^{3/2} \frac{(K\gamma + 1) kd}{\varepsilon n} \right).$$

Proof. Let $B = \lfloor n/T \rfloor$. Lemma 9 implies that for our choice of R with probability $1 - \zeta$, Algorithm 1 does not have any clipping. Therefore the update step becomes

$$Q_t' = Q_{t-1} + \eta_t \left(\frac{1}{B} \sum_{i = B \cdot (t-1) + 1}^{tB} A_i Q_{t-1} + Q_{t-1} N_{sym,t} + (I - Q_{t-1} Q_{t-1}^\top) N_t \right)$$

By Lemma 1 in distribution this is equal to

$$Q'_t = Q_{t-1} + \eta_t \left(\frac{1}{B} \sum_{i=B\cdot(t-1)+1}^{tB} A_i Q_{t-1} + G_t Q_{t-1} \right)$$
$$= Q_{t-1} + \eta_t \left(\frac{1}{B} \sum_{i=B\cdot(t-1)+1}^{tB} A_i + G_t \right) Q_{t-1}$$

where G_t is independently sampled from $G_t \sim GOE_d(\sigma^2)$.

Let

$$C_t = \frac{1}{B} \sum_{i=B\cdot(t-1)+1}^{tB} A_i + G_t.$$

We will show that C_t satisfies Assumption C, so that Theorem 13 applies. The updates are i.i.d., as the noise multiplier is fixed before the run, so the conditional law of C_t does not depend on the past. Since the blocks are disjoint and the Gaussian noise at step t is fresh, we see straight away that

$$\mathbb{E}\big[C_t \mid \mathcal{F}_{t-1}\big] = \frac{1}{B} \sum_{i=B(t-1)+1}^{tB} \mathbb{E}\big[A_i \mid \mathcal{F}_{t-1}\big] + \mathbb{E}\big[G_t \mid \mathcal{F}_{t-1}\big] = \frac{1}{B} \sum_{i=B(t-1)+1}^{tB} \mathbb{E}[A_i] = \Sigma,$$

as every entry of G_t is a centered Gaussian.

Next we show the upper bound on

$$\|C_t - \Sigma\|_{(2,k)} = \sup_{P \in \mathbb{S}_{d,k}} \|P^\top (C_t - \Sigma)\|_F.$$

By the triangle inequality,

$$\|C_t - \Sigma\|_{(2,k)} \le \left\| \frac{1}{B} \sum_{i=B\cdot(t-1)+1}^{tB} (A_i - \Sigma) \right\|_{(2,k)} + \|G_t\|_{(2,k)}.$$

By Lemma 4, with probability at least $1-\zeta$, simultaneously for all $t \in [T]$,

$$\left\| \frac{1}{B} \sum_{i=B\cdot(t-1)+1}^{tB} (A_i - \Sigma) \right\|_{(2,k)} \le C\sqrt{k}\,\lambda_1 \left(\sqrt{\frac{V \log(2dT/\zeta)}{B}} + \frac{M \log(2dT/\zeta)}{B} \right) =: \mathsf{E}_B.$$

We know by Lemma 6 that with probability $1-\zeta$,

$$\|G_t\|_{(2,k)} \le C\sigma\sqrt{k(d + \log(T/\zeta))}$$

for all $t \in [T]$. Since

$$\sigma = \frac{2R\sqrt{2\ln(1.25/\delta)}}{B\varepsilon},$$

and

$$R = C\lambda_1 \sqrt{dk} \big(K\gamma \log^a(ndk/\zeta) + 1 \big),$$

we have

$$\|G_t\|_{(2,k)} \le C \frac{\lambda_1 (K\gamma + 1) k \sqrt{d(d + \log(T/\zeta))}}{\varepsilon B} \log^a(ndk/\zeta) \sqrt{\log(1/\delta)} =: \mathsf{N}_B.$$

Set

$$M' := \mathsf{E}_B + \mathsf{N}_B.$$

Then, on the above high-probability event,

$$\|C_t - \Sigma\|_{(2,k)} \le M' \qquad \text{for all } t \in [T].$$

We apply Theorem 13 to the sequence $C_1, \dots, C_T$ with M replaced by M'. For every $\delta \in (0,1)$ define learning rates

$$n_0 = \tilde{\Theta}\left(\frac{kM'^2}{\delta^2 \Delta_k^2}\right), \qquad \beta = \tilde{\Theta}\left(\max\left\{\frac{\lambda_1 \log(n)}{\Delta_k}, \frac{M'^2 \log^2(n)}{\Delta_k^2}\right\}\right),$$

and

$$\eta_t = \begin{cases} \tilde{\Theta}\left(\frac{1}{\Delta_k n_0}\right) & t \le n_0, \\ \tilde{\Theta}\left(\frac{1}{\Delta_k(\beta + t - n_0)}\right) & t > n_0. \end{cases}$$

Then for $V \in \mathbb{R}^{d\times k}$ the orthogonal matrix whose columns are the k leading eigenvectors of Σ, the algorithm satisfies

$$\left\|Q_T Q_T^\top - V_k V_k^\top\right\|_F \le \sqrt{2}e\left(\frac{\beta+1}{\beta+T-n_0}\right)^{\alpha} + C'\frac{M'}{\Delta_k}\sqrt{\frac{\Lambda}{T-n_0}}, \qquad \Lambda \asymp \log\frac{1}{\delta} + \log\left(e + \frac{M'^2}{\Delta_k^2 \delta}\right),$$

with probability at least $1-\delta$. With $c_\eta = \lceil \log n \rceil$ one has $\alpha = \frac{1}{2}\lceil \log n \rceil$ and $\beta \asymp \kappa \log n$, so $T = \lceil C_T \kappa \log n \rceil$ gives $S := \beta + T - n_0 \ge 2e^2\beta$ for C_T large enough. Since $\beta \ge 1$ we have $(\beta+1)/S \le e^{-2}$, so the first term of Theorem 13 is at most $\sqrt{2}\, e\, e^{-2\alpha} = \sqrt{2}\, e^{\,1-\lceil \log n \rceil} \le \sqrt{2}\, e/n \le n^{-1/2}$ for $n \ge 2e^2$, and is dominated by the second.

Taking δ to be a constant multiple of ζ, using $T \ge 2n_0$, and absorbing logarithmic factors into $\tilde{O}(\cdot)$,

$$\|Q_T Q_T^\top - V_k V_k^\top\|_F \le \tilde{O}\left(\frac{M'}{\Delta_k \sqrt{T}}\right).$$

Substituting $M' = \mathsf{E}_B + \mathsf{N}_B$, we get

$$\|Q_T Q_T^\top - V_k V_k^\top\|_F \le \tilde{O}\left(\frac{\mathsf{E}_B}{\Delta_k \sqrt{T}} + \frac{\mathsf{N}_B}{\Delta_k \sqrt{T}}\right).$$

Since $T = n/B$, the minibatch term satisfies

$$\frac{\mathsf{E}_B}{\sqrt{T}} = \tilde{O}\left(\lambda_1 \sqrt{\frac{kV}{n}} + \sqrt{\kappa}\frac{\sqrt{k}M}{n}\right),$$

while, using $B = n/T$ with $T = \lceil C_T \kappa \log n \rceil$,

$$\frac{\mathsf{N}_B}{\sqrt{T}} = \tilde{O}\left(\sqrt{\kappa}\frac{\lambda_1 (K\gamma + 1) kd}{\varepsilon n}\right).$$

Therefore,

$$\|Q_T Q_T^\top - V_k V_k^\top\|_F \le \tilde{O}\left(\frac{\lambda_1\sqrt{kV}}{\Delta_k\sqrt{n}} + \sqrt{\kappa}\frac{\sqrt{k}M}{\Delta_k n} + \sqrt{\kappa}\frac{\lambda_1(K\gamma+1)kd}{\varepsilon \Delta_k n}\right).$$

Finally, the stated lower bound on n is exactly the condition $T \gtrsim kM'^2/(\zeta^2\Delta_k^2)$, after substituting $B = n/T$, $T = \lceil C_T \kappa \log n \rceil$ and suppressing logarithmic factors. □

Lemma 9 (No clipping with high probability)**.** *Consider Algorithm 1 under Assumption A, and choose*

$$R \leftarrow C\lambda_1\sqrt{dk}\left(K\gamma \log^a(ndk/\zeta) + 1\right)$$

for a sufficiently large universal constant $C > 0$. Then, with probability at least $1-\zeta$, no clipping occurs throughout the algorithm, i.e.

$$\|A_i Q_{t-1}\|_F \le R \qquad \textit{for all } t \in [T],\ i \in \{B(t-1)+1, \dots, tB\}.$$

Equivalently, on this event,

$$\mathrm{clip}_R(A_i Q_{t-1}) = A_i Q_{t-1} \qquad \textit{for all such } i, t.$$

Proof. Fix $t \in [T]$ and $i \in \{B(t-1)+1, \dots, tB\}$. Since the algorithm is one-pass, Q_{t-1} depends only on the matrices $A_1, \dots, A_{B(t-1)}$, and is therefore independent of A_i. Hence, conditional on the past, Q_{t-1} is fixed.

Write the columns of Q_{t-1} as $q_1^{(t-1)}, \dots, q_k^{(t-1)}$. For any $r \in [d]$ and $s \in [k]$, Assumption A.4 applied with $P = I$, $u = e_r$, and $v = q_s^{(t-1)}$ yields

$$\mathbb{E}\left[\exp\left(\left(\frac{|e_r^\top (A_i - \Sigma) q_s^{(t-1)}|^2}{K^2 \lambda_1^2 \gamma^2}\right)^{1/(2a)}\right) \,\middle|\, Q_{t-1}\right] \le 1.$$

By a standard sub-Weibull tail bound, this implies that for a sufficiently large universal constant $C > 0$,

$$|e_r^\top (A_i - \Sigma) q_s^{(t-1)}| \le C \lambda_1 K \gamma \log^a(ndk/\zeta)$$

simultaneously for all $i \in [n]$, $r \in [d]$, and $s \in [k]$ with probability at least $1 - \zeta$, by a union bound over the ndk such quantities.

On this event,

$$\|(A_i - \Sigma) Q_{t-1}\|_F \le \sqrt{dk} \max_{r \in [d],\, s \in [k]} |e_r^\top (A_i - \Sigma) q_s^{(t-1)}| \le C \lambda_1 \sqrt{dk}\, K \gamma \log^a(ndk/\zeta).$$

Moreover,

$$\|\Sigma Q_{t-1}\|_F \le \|\Sigma\|_2 \|Q_{t-1}\|_F = \lambda_1 \sqrt{k} \le \lambda_1 \sqrt{dk}.$$

Therefore,

$$\|A_i Q_{t-1}\|_F \le \|\Sigma Q_{t-1}\|_F + \|(A_i - \Sigma) Q_{t-1}\|_F \le C \lambda_1 \sqrt{dk}\, \big(K \gamma \log^a(ndk/\zeta) + 1\big).$$

By the choice of R, it follows that

$$\|A_i Q_{t-1}\|_F \le R$$

for all $t \in [T]$ and all $i \in \{B(t-1)+1, \dots, tB\}$. Thus no clipping occurs throughout the algorithm. □

Lemma 1. *Fix $Q \in \mathbb{R}^{d \times k}$ with $Q^\top Q = I_k$. Let $Z \in \mathbb{R}^{d \times k}$ have i.i.d. $\mathcal{N}(0, \sigma^2)$ entries, let $N_{\mathrm{sym}} \sim \mathrm{GOE}_k(\sigma^2)$, and let $G \sim \mathrm{GOE}_d(\sigma^2)$. Then*

$$Q N_{\mathrm{sym}} + (I - QQ^\top) Z \overset{d}{=} GQ.$$

Proof of Lemma 1. Let $U = [Q, Q_\perp] \in O(d)$. Since Z has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries and $Q_\perp$ has orthonormal columns, $Q_\perp^\top Z$ has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries and is independent of N_{sym}.

Now

$$U^\top (Q N_{\mathrm{sym}} + (I - P) Z) = \begin{pmatrix} N_{\mathrm{sym}} \\ Q_\perp^\top Z \end{pmatrix}.$$

On the other hand, by orthogonal invariance of GOE,

$$U^\top G U \sim \mathrm{GOE}_d(\sigma^2).$$

Writing

$$U^\top G U = \begin{pmatrix} A_0 & B^\top \\ B & H \end{pmatrix},$$

we have

$$A_0 \sim \mathrm{GOE}_k(\sigma^2), \qquad B \text{ has i.i.d. } \mathcal{N}(0, \sigma^2) \text{ entries},$$

and A_0, B are independent. Therefore

$$U^\top G Q = \begin{pmatrix} A_0 \\ B \end{pmatrix} \overset{d}{=} \begin{pmatrix} N_{\mathrm{sym}} \\ Q_\perp^\top Z \end{pmatrix} = U^\top (Q N_{\mathrm{sym}} + (I - P) Z).$$

Multiplying by U proves the claim. □

C.3 Privacy of Algorithm 2

Lemma 2 (Privacy of Algorithm 2)**.** *If $0 < \varepsilon \leq 1$ then Algorithm 2 is (ε, δ)-DP.*

Proof. Fix an iteration t, and condition on all previous outputs. Then $Q = Q_{t-1}$ is fixed. The range is estimated PRIVRANGE, which is (ε, δ)-DP by Theorem 4, and uses samples disjoint from TRUNCATEDMEAN and therefore parallel DP composition applies.

Now consider TRUNCATEDMEAN: Let $C = (\bar{g}_{jr})$ be the coordinatewise center computed by TRUNCATEDMEAN. The dk coordinate histograms each use privacy budget $(\varepsilon/(2dk), \delta/(2dk))$, so by basic composition C is $(\varepsilon/2, \delta/2)$-DP.

For any fixed value of C, let $\widetilde{G}_i(C)$ be A_iQ after coordinatewise truncation to $[C_{jr} - R_t, C_{jr} + R_t]$. Since each coordinate of both truncated summands lies in an interval of length $2R_t$,

$$\|\widetilde{G}_i(C) - \widetilde{G}'_i(C)\|_F \leq 2R_t\sqrt{dk}.$$

and therefore for neighboring mean batches S, S', coordinatewise truncation gives

$$\|\frac{1}{m}\sum_{i=1}^{m}\widetilde{G}_i(C) - \frac{1}{m}\sum_{i=1}^{m}\widetilde{G}'_i(C)\| \leq \frac{2R_t\sqrt{dk}}{m}.$$

The algorithm then applies a projection $\mathsf{P}_{\mathcal{S}_Q}$ to the output of TRUNCATEDMEAN. Define

$$h_C(S) = \mathsf{P}_{\mathcal{S}_Q}\left(\frac{1}{m}\sum_{i=1}^{m}\widetilde{G}_i(C)\right).$$

Then $h_C(S) \in \mathcal{S}_Q$ so by Lemma 8 adding the noise matrix

$$W_Q = QN_{\mathrm{sym}} + (I - QQ^\top)Z.$$

calibrated to twice the sensitivity of $h_C(S)$ is private. Since $\mathsf{P}_{\mathcal{S}_Q}$ is a projection and therefore cannot increase the norm the sensitivity is

$$\|h_C(S) - h_C(S')\|_F \leq \frac{2R_t\sqrt{dk}}{m}.$$

For batch size $m = B/2$, and privacy parameters $(\varepsilon/2, \delta/2)$ this gives

$$\sigma_t \geq \frac{8R_t\sqrt{dk}}{B\varepsilon}\sqrt{2\log(2.5/\delta)}.$$

which is exactly the calibration in Algorithm 2.

By adaptive composition,

$$\big(C, h_C(S) + W_Q\big)$$

is (ε, δ)-DP. Since the algorithm releases only $h_C(S) + W_Q$, discarding C is post-processing. By the budget choice in Algorithm 2, the mean-half mechanism is therefore (ε, δ)-DP.

For $t > T_0$ the PRIVRANGE and TRUNCATEDMEAN act on a disjoint set of inputs, so by parallel composition the adaptive update is (ε, δ)-DP. For $t \leq T_0$ the update is that of Algorithm 1: the batch mean has Frobenius-norm sensitivity $2R/B$ in a single A_i, and $\sigma = 2R\sqrt{2\log(1.25/\delta)}/(B\varepsilon)$ is the Gaussian mechanism calibrated to it, so the warm-up update is (ε, δ)-DP by Theorem 3. Since the algorithm is one-pass and the update batches are disjoint across t, the final output is also (ε, δ)-DP. □

C.4 Utility of Algorithm 2

Theorem 2 (Utility of Algorithm 2)**.** *Given matrices $A_1, \ldots, A_n$ fulfilling Assumption A write $\kappa = \frac{\lambda_1}{\lambda_k - \lambda_{k+1}}$ and set the base clipping radius, the batch count and the warm-up length to*

$$R = C\lambda_1\sqrt{dk}\big(K\gamma\log^a(ndk/\zeta) + 1\big), \qquad T = \lceil C_T\kappa\log(n)\rceil, \qquad T_0 = \lceil T/2\rceil,$$

where C_T is a sufficiently large absolute constant. If

$$n \gtrsim \max\left\{ \frac{k^2\lambda_1^2 V}{\zeta^2(\lambda_k - \lambda_{k+1})^2}, \kappa^{3/2}\cdot\frac{kM}{\zeta}, \kappa^{3/2}\cdot\frac{(K\gamma+1)k^{3/2}d}{\varepsilon\zeta}, \frac{dk}{\varepsilon}, \frac{K^2 d}{\varepsilon} \right\},$$

then there exist learning rates $\{\eta_t\}$ such that, with probability at least $1-\zeta$, Algorithm 2 outputs Q_T with

$$\|Q_T Q_T^\top - V_k V_k^\top\|_F \le \tilde{O}\left(\kappa\frac{\sqrt{kV}}{\sqrt{n}} + \kappa^{3/2}\frac{\sqrt{k}M}{n} + \kappa^{3/2}\frac{Kdk\gamma}{\varepsilon n} \right).$$

Proof. Let $B = \lfloor n/T \rfloor$ and let

$$I_t := \{B(t-1) + B/2 + 1, \dots, tB\}.$$

Lemma 10 together with Theorem 4 imply that, under the stated sample-size conditions and after a union bound over $t \in [T]$, with probability $1-\zeta$ Algorithm 5 does not clip. Hence TRUNCATEDMEAN returns

$$\frac{1}{m}\sum_{i\in I_t} A_i Q_{t-1}, \qquad m = B/2.$$

Since each A_i is symmetric, $A_i Q_{t-1} \in \mathcal{S}_{Q_{t-1}}$. Therefore the raw mean is also in $\mathcal{S}_{Q_{t-1}}$, and

$$\mathsf{P}_{\mathcal{S}_{Q_{t-1}}}\left(\frac{1}{m}\sum_{i\in I_t} A_i Q_{t-1}\right) = \frac{1}{m}\sum_{i\in I_t} A_i Q_{t-1}.$$

Therefore the update step becomes

$$Q_t' = Q_{t-1} + \eta_t\left(\frac{2}{B}\sum_{i=B\cdot(t-1)+B/2+1}^{tB} A_i Q_{t-1} + Q_{t-1}N_{\mathrm{sym},t} + (I - Q_{t-1}Q_{t-1}^\top)Z_t\right).$$

By Lemma 1 in distribution this is equal to

$$\begin{aligned} Q_t' &= Q_{t-1} + \eta_t\left(\frac{2}{B}\sum_{i=B\cdot(t-1)+B/2+1}^{tB} A_i Q_{t-1} + G_t Q_{t-1}\right) \\ &= Q_{t-1} + \eta_t\left(\frac{2}{B}\sum_{i=B\cdot(t-1)+B/2+1}^{tB} A_i + G_t\right) Q_{t-1}, \end{aligned}$$

where G_t is independently sampled from $GOE_d(\sigma_t^2)$.

Let $\mathcal{F}_{t-1}$ denote the sigma-field generated by all randomness used before iteration t, and let $\mathcal{R}_t$ denote the randomness in the range-estimation half of the t-th batch. On the no-clipping event, Lemma 1 lets us write the private update in the form

$$Q_t = \mathrm{QR}\big(Q_{t-1} + \eta_t C_t Q_{t-1}\big), \qquad C_t = \frac{2}{B}\sum_{i=B(t-1)+B/2+1}^{tB} A_i + G_t,$$

where $G_t \sim \mathrm{GOE}_d(\sigma_t^2)$, and σ_t is $\sigma(\mathcal{F}_{t-1}, \mathcal{R}_t)$-measurable.

We will show that C_t satisfies Assumption C, conditioned on a high-probability event, and that the warm-up updates are i.i.d., so that Theorem 13 applies. The samples in the mean half of the batch are disjoint from the range half, hence independent of $(\mathcal{F}_{t-1}, \mathcal{R}_t, \sigma_t)$. Therefore

$$\mathbb{E}[C_t \mid \mathcal{F}_{t-1}, \mathcal{R}_t] = \mathbb{E}\left[\frac{2}{B}\sum_{i=B(t-1)+B/2+1}^{tB} A_i\right] + \mathbb{E}[G_t \mid \mathcal{F}_{t-1}, \mathcal{R}_t] = \Sigma.$$

Moreover, conditional on $(\mathcal{F}_{t-1}, \mathcal{R}_t)$, the matrix C_t is symmetric and uses fresh randomness independent of the current iterate Q_{t-1}.

Next we show the upper bound on

$$\|C_t - \Sigma\|_{(2,k)} = \sup_{P \in \mathbb{S}_{d,k}} \|P^\top (C_t - \Sigma)\|_F.$$

By the triangle inequality,

$$\|C_t - \Sigma\|_{(2,k)} \le \left\| \frac{2}{B} \sum_{i=B\cdot(t-1)+B/2+1}^{tB} (A_i - \Sigma) \right\|_{(2,k)} + \|G_t\|_{(2,k)}.$$

By Lemma 4, with probability at least $1-\zeta$, simultaneously for all $t \in [T]$,

$$\left\| \frac{2}{B} \sum_{i=B\cdot(t-1)+B/2+1}^{tB} (A_i - \Sigma) \right\|_{(2,k)} \le C\sqrt{k}\lambda_1 \left(\sqrt{\frac{V \log(2dT/\zeta)}{B}} + \frac{M \log(2dT/\zeta)}{B} \right) =: \mathsf{E}_B.$$

Moreover, by Theorem 4,

$$\widehat{\Lambda}_t \le C\lambda_1^2 \gamma^2$$

simultaneously for all $t \in [T]$. Hence

$$R_t = 3K\sqrt{\widehat{\Lambda}_t} \log^a(Bdk/\zeta) \le CK\lambda_1 \gamma \log^a(BdkT/\zeta) =: R_\star.$$

Since the Gaussian noise variance in iteration t is calibrated using R_t,

$$\sigma_t \le \frac{8R_\star \sqrt{2dk \log(2.5/\delta)}}{\varepsilon B} =: \sigma_\star.$$

By Lemma 6 and a union bound over $t \in [T]$, with probability at least $1-\zeta$,

$$\|G_t\|_{(2,k)} \le C\sigma_\star \sqrt{k(d + \log(T/\zeta))}$$

for all $t \in [T]$. Therefore,

$$\|G_t\|_{(2,k)} \le C \frac{\lambda_1 K \gamma\, k\sqrt{d(d+\log(T/\zeta))}}{\varepsilon B} \log^a(BdkT/\zeta) \sqrt{\log(1/\delta)} =: \mathsf{N}_B.$$

Set $M' := E_B + N_B$, and let $\mathcal{E}$ be the intersection of the range-estimation, no-clipping, minibatch-concentration, and Gaussian-noise events above. On $\mathcal{E}$,

$$\|C_t - \Sigma\|_{(2,k)} \le M' \qquad \text{for all } t \in [T].$$

Thus, on $\mathcal{E}$, ADADPO coincides with Oja's method run on the symmetric effective matrices $C_1, \dots, C_T$. Since the range half is disjoint from the mean half, the adaptive choice of σ_t is measurable with respect to $\widetilde{\mathcal{F}}_{t-1} = \sigma(\mathcal{F}_{t-1}, \mathcal{R}_t)$, while the mean-half samples and the GOE base noise at time t are fresh. Assumption C therefore holds on $\mathcal{E}$ with respect to the pre-update filtration $(\widetilde{\mathcal{F}}_t)$ of (24) and with constant M': each update matrix is symmetric, satisfies $\mathbb{E}[C_t \mid \widetilde{\mathcal{F}}_{t-1}] = \Sigma$, and obeys $\|C_t - \Sigma\|_{(2,k)} \le M'$. This is Lemma 27, with $M = M'$. It remains to check the hypothesis that Theorem 13 places on Phase I. On the warm-up batches $t \le T_0$ the multiplier $\sigma_t \equiv \sigma$ is fixed before the run and clipping is inactive on $\mathcal{E}$, so

$$C_t = \frac{1}{B} \sum_{i=B(t-1)+1}^{tB} A_i + G_t$$

does not depend on Q_{t-1}: those updates are i.i.d. and independent of Z_0, with the constant M_{I}' of Theorem 1 in place of M'. Phase II, on $(T_0, T]$, uses the adaptive constant M' and nothing beyond Assumption C. Applying Theorem 13 and reducing the final success probability by $\mathbb{P}(\mathcal{E}^c)$ gives the following.

For every $\delta \in (0,1)$ define learning rates

$$n_0 = \tilde{\Theta}\left(\frac{kM'^2}{\delta^2 \Delta_k^2} \right), \qquad \beta = \tilde{\Theta}\left(\frac{M'^2}{\Delta_k^2} \right),$$

and

$$\eta_t = \begin{cases} \tilde{\Theta}\left(\frac{1}{\Delta_k n_0}\right), & t \le n_0, \\ \tilde{\Theta}\left(\frac{1}{\Delta_k(\beta+t-n_0)}\right), & t > n_0. \end{cases}$$

Then for $V \in \mathbb{R}^{d\times k}$ the orthogonal matrix whose columns are the k leading eigenvectors of Σ, the algorithm satisfies

$$\|Q_T Q_T^\top - V_k V_k^\top\|_F \le \sqrt{2}\, e \left(\frac{\beta+1}{\beta+T-T_0}\right)^{\alpha} + C' \frac{M'}{\Delta_k}\sqrt{\frac{\Lambda}{T-T_0}}, \qquad \Lambda \asymp \log\frac{1}{\delta} + \log\left(e + \frac{M'^2}{\Delta_k^2 \delta}\right),$$

with probability at least $1-\delta$.

With $c_\eta = \lceil \log n \rceil$ one has $\alpha = \frac{1}{2}\lceil \log n \rceil$ and $\beta \asymp \kappa \log n$, so $T = \lceil C_T \kappa \log n \rceil$ and $T_0 = \lceil T/2 \rceil$ give $S := \beta + T - T_0 \ge 2e^2\beta$ for C_T large enough. Since $\beta \ge 1$ we have $(\beta+1)/S \le e^{-2}$, so the first term is at most $\sqrt{2}\, e\, e^{-2\alpha} \le \sqrt{2}\, e/n \le n^{-1/2}$ for $n \ge 2e^2$ and is dominated by the second. Taking δ to be a constant multiple of ζ and absorbing logarithmic factors into $\tilde{O}(\cdot)$,

$$\|Q_T Q_T^\top - V_k V_k^\top\|_F \le \tilde{O}\left(\frac{M'}{\Delta_k\sqrt{T}}\right).$$

Substituting $M' = \mathsf{E}_B + \mathsf{N}_B$ gives

$$\|Q_T Q_T^\top - V_k V_k^\top\|_F \le \tilde{O}\left(\frac{\mathsf{E}_B}{\Delta_k\sqrt{T}} + \frac{\mathsf{N}_B}{\Delta_k\sqrt{T}}\right).$$

Since $BT = n$ up to rounding, the minibatch term satisfies

$$\frac{\mathsf{E}_B}{\sqrt{T}} = \tilde{O}\left(\lambda_1\sqrt{\frac{kV}{n}} + \sqrt{\kappa}\,\frac{\sqrt{k}M}{n}\right),$$

while, using $B = \lfloor n/T \rfloor$ with $T = \lceil C_T \kappa \log n \rceil$,

$$\frac{\mathsf{N}_B}{\sqrt{T}} = \tilde{O}\left(\sqrt{\kappa}\,\frac{\lambda_1 K\gamma\, kd}{\varepsilon n}\right).$$

Therefore,

$$\|Q_T Q_T^\top - V_k V_k^\top\|_F \le \tilde{O}\left(\frac{\lambda_1\sqrt{kV}}{\Delta_k\sqrt{n}} + \sqrt{\kappa}\,\frac{\sqrt{k}M}{\Delta_k n} + \sqrt{\kappa}\,\frac{\lambda_1 K\gamma\, kd}{\varepsilon\Delta_k n}\right).$$

□

Lemma 10 (No clipping for matrix-valued truncated mean)**.** *Fix an iteration t, and let $Q = Q_{t-1} \in \mathbb{R}^{d\times k}$ have orthonormal columns $q_1, \dots, q_k$, where Q is measurable with respect to the previous batches and hence independent of the current mean batch $\{A_i\}_{i=1}^m$. Let*

$$G_i = A_i Q \in \mathbb{R}^{d\times k}, \qquad i = 1, \dots, m.$$

Assume the matrices A_i satisfy Assumption A and assume that $\widehat{\Lambda}$ returned by PrivRange *(Algorithm 4) satisfies*

$$\widehat{\Lambda} \;\ge\; \max_{r\in[k]} \lambda_1^2 \|H q_r\|_2.$$

Let TruncatedMean *(Algorithm 5) be run with*

$$R = 3K\sqrt{\widehat{\Lambda}}\log^a(mdk/\zeta) \qquad \text{and privacy parameters } (\varepsilon_C, \delta_C).$$

If the number of samples m passed to TruncatedMean *satisfies*

$$m \ge C\frac{dk}{\varepsilon_C}\log\left(\frac{dk}{\zeta\delta_C}\right)$$

for a sufficiently large universal constant $C > 0$, then with probability at least $1-\zeta$, for every $i \in [m]$, $j \in [d]$, and $r \in [k]$,

$$(G_i)_{jr} \in \Big[\bar{g}_{jr} - R,\ \bar{g}_{jr} + R\Big].$$

Proof. Fix a column $r \in [k]$. Since q_r is a unit vector and is independent of the current mean batch, the vector-valued gradients

$$g_i^{(r)} := A_i q_r \in \mathbb{R}^d$$

satisfy the same one-vector assumptions as in the paper with $u = q_r$. In particular, if we write

$$\mu^{(r)} := \mathbb{E}[g_i^{(r)}] = \Sigma q_r,$$

then for every coordinate $j \in [d]$, the scalar random variable $(g_i^{(r)})_j - \mu_j^{(r)}$ obeys the same sub-Weibull tail bound, namely

$$\Pr\Big(|(g_i^{(r)})_j - \mu_j^{(r)}| > K\lambda_1 \sqrt{\|Hq_r\|_2}\log^a(mdk/\zeta)\Big) \le \frac{\zeta}{4mdk}.$$

Since

$$\lambda_1^2 \|Hq_r\|_2 \le \widehat{\Lambda},$$

we obtain

$$\Pr\Big(|(g_i^{(r)})_j - \mu_j^{(r)}| > K\sqrt{\widehat{\Lambda}}\log^a(mdk/\zeta)\Big) \le \frac{\zeta}{4mdk}.$$

A union bound over all triples $(i, j, r) \in [m] \times [d] \times [k]$ gives that with probability at least $1 - \zeta/4$,

$$|(g_i^{(r)})_j - \mu_j^{(r)}| \le K\sqrt{\widehat{\Lambda}}\log^a(mdk/\zeta) \qquad \forall\, i \in [m],\, j \in [d],\, r \in [k]. \tag{6}$$

Next fix (j, r), and write $L := K\sqrt{\widehat{\Lambda}}\log^a(mdk/\zeta)$, so that on the event (6) every sample lies in

$$I_{jr} := \Big[\mu_j^{(r)} - L,\ \mu_j^{(r)} + L\Big].$$

Consider the histogram used by TRUNCATEDMEAN on the samples $\{(G_i)_{jr}\}_{i=1}^m = \{(g_i^{(r)})_j\}_{i=1}^m$, with bin width $\sqrt{\widehat{\Lambda}}$, run with privacy budget $(\varepsilon_C/(dk), \delta_C/(dk))$. Because *all* m samples lie in I_{jr}, every non-empty bin intersects I_{jr} and every bin disjoint from I_{jr} is empty; how many bins I_{jr} spans is immaterial.

It remains to check that the release is not identically zero. Conditionally on the past the variables $(G_i)_{jr}$, $i \in [m]$, are i.i.d. with mean $\mu_j^{(r)}$ and variance at most $\lambda_1^2\|Hq_r\|_2 \le \widehat{\Lambda}$, so Chebyshev's inequality gives $\Pr\left(|(G_i)_{jr} - \mu_j^{(r)}| > 2\sqrt{\widehat{\Lambda}}\right) \le \frac{1}{4}$. The interval $[\mu_j^{(r)} - 2\sqrt{\widehat{\Lambda}},\ \mu_j^{(r)} + 2\sqrt{\widehat{\Lambda}}]$ has length $4\sqrt{\widehat{\Lambda}}$ and therefore meets at most five bins, one of which carries expected mass at least $3/20$; by a Chernoff bound and a union bound over the dk coordinates its empirical mass is at least $\frac{1}{10}$ for every (j, r) with probability at least $1 - \zeta/8$. By Lemma 11 at accuracy $\beta < \frac{1}{20}$, applied with per-coordinate privacy budget $(\varepsilon_C/(dk), \delta_C/(dk))$, which the stated lower bound on m supplies, that bin is released with positive count; since every empty bin is released as 0, the selected bin is non-empty and therefore intersects I_{jr}, for every coordinate, with probability at least $1 - \zeta/4$.

Since $\bar{g}_{jr}$ is the left endpoint of the selected bin and that bin meets I_{jr} and has width $\sqrt{\widehat{\Lambda}}$, it follows that $|\bar{g}_{jr} - \mu_j^{(r)}| \le L + \sqrt{\widehat{\Lambda}}$, and hence, using $K \ge 1$ and $\log^a(mdk/\zeta) \ge 1$,

$$\left|\bar{g}_{jr} - \mu_j^{(r)}\right| \le 2K\sqrt{\widehat{\Lambda}}\log^a(mdk/\zeta) \tag{7}$$

for all (j, r) simultaneously with probability at least $1 - \zeta/4$. By basic composition over the dk coordinate histograms, the released coordinate center $(\bar{g}_{jr})_{j\in[d],r\in[k]}$ is $(\varepsilon_C, \delta_C)$-DP. This accounts only for the private center selection; the unnoised truncated average returned by TRUNCATEDMEAN is not a standalone DP mean release.

Finally, on the intersection of the events (6) and (7), for every i, j, r,

$$|(G_i)_{jr} - \bar{g}_{jr}| \le |(G_i)_{jr} - \mu_j^{(r)}| + |\mu_j^{(r)} - \bar{g}_{jr}| \le 3K\sqrt{\widehat{\Lambda}}\log^a(mdk/\zeta) = R.$$

Thus

$$(G_i)_{jr} \in [\bar{g}_{jr} - R,\ \bar{g}_{jr} + R]$$

for all i, j, r simultaneously. Hence no coordinate is clipped. Let $\widetilde{G}_i$ be the coordinatewise truncated version of G_i. The preceding display implies $\widetilde{G}_i = G_i = A_i Q$ for all i. Since each A_i is symmetric, $Q^\top A_i Q$ is symmetric, and hence $A_i Q \in \mathcal{S}_Q$. Therefore the raw mean also belongs to $\mathcal{S}_Q$, so

$$\mathsf{P}_{\mathcal{S}_Q}\left(\frac{1}{m}\sum_{i=1}^{m}\widetilde{G}_i\right) = \frac{1}{m}\sum_{i=1}^{m} A_i Q.$$

□

Algorithm 4 PRIVRANGE

Input: $S = \{G_i\}_{i=1}^{B} \subset \mathbb{R}^{d\times k}$, privacy parameters (ε, δ), failure probability ζ

for $i = 1, 2, \ldots, \lfloor B/2 \rfloor$ **do**
  Let $\widetilde{G}_i \leftarrow G_{2i} - G_{2i-1} \in \mathbb{R}^{d\times k}$
end for
Let $\widetilde{S} = \{\widetilde{G}_i\}_{i=1}^{\lfloor B/2 \rfloor}$
Let $m \leftarrow C_1 \log(1/(\delta\zeta))/\varepsilon$
Partition $\widetilde{S}$ into m subsets and denote each subset by $\mathcal{G}_j$, where each subset has size $b = \lfloor B/(2m) \rfloor$
for $j = 1, \ldots, m$ **do**
  for $r = 1, \ldots, k$ **do**
    Let $G_j^{(r)} \in \mathbb{R}^{d\times b}$ be the matrix whose columns are the rth columns of the matrices in $\mathcal{G}_j$
    Let $\lambda_{j,r} \leftarrow \lambda_1\left(\frac{1}{2b} G_j^{(r)} \big(G_j^{(r)}\big)^\top\right)$
  end for
  Let $m_j \leftarrow \max_{r\in[k]} \lambda_{j,r}$
end for
Partition $[0,\infty)$ into $\Omega \leftarrow \{\ldots, [2^{-2/4}, 2^{-1/4}), [2^{-1/4}, 1), [1, 2^{1/4}), [2^{1/4}, 2^{2/4}), \ldots\} \cup \{[0,0]\}$
Run (ε, δ)-DP histogram learner of Lemma 11 on $\{m_j\}_{j=1}^{m}$ over Ω
if all the bins are empty **then**
  return $\perp$
end if
Let $[l, r]$ be a non-empty bin that contains the maximum number of points in the DP histogram
return $\widehat{\Lambda} = 2l$

Algorithm 5 TRUNCATEDMEAN

Input: $S = \{G_1, \ldots, G_B\} \subset \mathbb{R}^{d\times k}$, truncation threshold R, range estimate $\widehat{\Lambda}$, center privacy parameters $(\varepsilon_C, \delta_C)$

for $j = 1, \ldots, d$ **do**
  for $r = 1, \ldots, k$ **do**
    Run a private histogram learner on $\{(G_\ell)_{jr}\}_{\ell=1}^{B}$ over bins of width $\sqrt{\widehat{\Lambda}}$ using privacy budget $(\varepsilon_C/(dk), \delta_C/(dk))$
    Let $[\ell_{jr}, h_{jr}]$ be the bin with the largest private count
    Set $\bar{g}_{jr} \leftarrow \ell_{jr}$
  end for
end for
for $\ell = 1, \ldots, B$ **do**
  for $j = 1, \ldots, d$ **do**
    for $r = 1, \ldots, k$ **do**
      Truncate $(G_\ell)_{jr}$ to $[\bar{g}_{jr} - R,\ \bar{g}_{jr} + R]$.
    end for
  end for
  Let $\widetilde{G}_\ell$ be the truncated matrix
end for
return $\widetilde{\mu} \leftarrow \frac{1}{B}\sum_{\ell=1}^{B} \widetilde{G}_\ell$

Theorem 4 (Privacy and accuracy of PRIVRANGE)**.** *Algorithm 4 is (ε, δ)-DP. Let $Q \in \mathbb{R}^{d\times k}$ have orthonormal columns $q_1, \dots, q_k$, and suppose Q is fixed independently of the current batch $\{A_\ell\}_{\ell=1}^B$. Define $G_\ell = A_\ell Q \in \mathbb{R}^{d\times k}$, $\ell = 1, \dots, B$, and let $\Lambda(Q) := \max_{r\in[k]} \lambda_1^2 \|Hq_r\|_2$.*

Assume that the matrices A_ℓ satisfy Assumptions A.1 and A.4. If the block size $b = \lfloor \frac{B}{2m} \rfloor$ satisfies $b \geq \tilde{O}\big(K^2 d \log(km/\zeta)\big)$, equivalently, $B \geq \tilde{O}\Big(\frac{K^2 d \log(k/(\delta\zeta)) \log(1/(\delta\zeta))}{\varepsilon}\Big)$, up to the same logarithmic factors, then with probability at least $1-\zeta$, Algorithm 4 outputs $\widehat{\Lambda}$ satisfying $\widehat{\Lambda} \in [\Lambda(Q), C_\Lambda \Lambda(Q)]$.

Proof. Privacy is immediate. The algorithm first transforms the batch $\{G_\ell\}_{\ell=1}^B$ deterministically into the block maxima $\{m_j\}_{j=1}^m$, and then applies the (ε, δ)-DP histogram learner (Lemma 11). Hence Algorithm 4 is (ε, δ)-DP.

We now prove the accuracy guarantee. For each column $r \in [k]$, define

$$g_\ell^{(r)} := G_\ell e_r = A_\ell q_r \in \mathbb{R}^d.$$

Since Q is fixed independently of the current batch, each q_r is fixed and independent of the samples $\{A_\ell\}_{\ell=1}^B$. Therefore the vectors $\{g_\ell^{(r)}\}$ are in the setting of Theorem 6.1 of Liu et al. [2022].

For each block $j \in [m]$ and column $r \in [k]$, let $G_j^{(r)} \in \mathbb{R}^{d\times b}$ be the matrix whose columns are the r-th columns of the paired differences in $\mathcal{G}_j$, and define

$$\lambda_{j,r} = \lambda_1\left(\frac{1}{2b} G_j^{(r)} (G_j^{(r)})^\top\right).$$

The columns of $G_j^{(r)}$ are pair differences, whose covariance is twice the target covariance. Thus the factor $1/(2b)$ normalizes the block covariance to the original target scale. By the concentration result underlying Theorem 6.1 of Liu et al. [2022], provided

$$b \geq \tilde{O}\big(K^2 d \log(km/\zeta)\big),$$

we have with probability at least $1 - \zeta/(10m)$ that, simultaneously for all $r \in [k]$,

$$\lambda_{j,r} \in \left[\frac{1}{\sqrt{2}} \lambda_1^2 \|Hq_r\|_2,\ \sqrt{2}\, \lambda_1^2 \|Hq_r\|_2\right].$$

On this event,

$$m_j := \max_{r\in[k]} \lambda_{j,r} \in \left[\frac{1}{\sqrt{2}} \Lambda(Q),\ \sqrt{2}\, \Lambda(Q)\right].$$

Taking a union bound over the m blocks, with probability at least $1 - \zeta/10$ *every* m_j lies in

$$I := \Big[\frac{1}{\sqrt{2}}\Lambda(Q),\ \sqrt{2}\, \Lambda(Q)\Big].$$

The endpoints of I have ratio 2 while the bins of Ω have ratio $2^{1/4}$, so I spans $\log_{2^{1/4}} 2 = 4$ bin widths and intersects at most five consecutive bins; and because every m_j lies in I, every non-empty bin intersects I, and every bin disjoint from I is empty. At least one of the intersecting bins therefore holds at least a $1/5$-fraction of the m_j. By Lemma 11 at accuracy $\beta < \frac{1}{10}$, which the choice $m = C_1 \log(1/(\delta\zeta))/\varepsilon$ supplies, the DP histogram learner returns, with probability at least $1 - \zeta/2$, a non-empty bin $[l, 2^{1/4} l)$ intersecting I.

Consequently $l < \sqrt{2}\, \Lambda(Q)$, while $2^{1/4} l > \frac{1}{\sqrt{2}} \Lambda(Q)$ gives $l > 2^{-3/4} \Lambda(Q) > \frac{1}{2} \Lambda(Q)$, so the output

$$\widehat{\Lambda} = 2l$$

satisfies

$$\widehat{\Lambda} \in \big[\Lambda(Q),\, C_\Lambda \Lambda(Q)\big], \qquad C_\Lambda = 2\sqrt{2}.$$

Combining the above events proves the theorem. □

Lemma 11 (Stability-based histogram [Karwa and Vadhan, 2017, Lemma 2.3])**.** *For every $K \in \mathbb{N} \cup \{\infty\}$, domain Ω, every collection of disjoint bins $B_1, \ldots, B_K$ defined on Ω, $n \in \mathbb{N}$, $\varepsilon \geq 0$, $\delta \in (0, 1/n)$, $\beta > 0$, and $\alpha \in (0, 1)$, under the replacement notion of neighboring datasets there exists an (ε, δ)-differentially private algorithm*

$$\mathcal{M} : \Omega^n \to \mathbb{R}^K$$

such that for any dataset $X_1, \ldots, X_n \in \Omega^n$,

1. $$\hat{p}_k = \frac{1}{n} \sum_{X_i \in B_k} 1,$$

2. $$(\tilde{p}_1, \ldots, \tilde{p}_K) \leftarrow \mathcal{M}(X_1, \ldots, X_n),$$

   and

3. *if*
   $$n \geq \min\left\{ \frac{8}{\varepsilon\beta} \log\left(\frac{2K}{\alpha}\right), \frac{8}{\varepsilon\beta} \log\left(\frac{4}{\alpha\delta}\right) \right\},$$
   then
   $$\mathbb{P}\big(|\tilde{p}_k - \hat{p}_k| \leq \beta\big) \geq 1 - \alpha.$$

4. *and $\hat{p}_k = 0$ implies $\tilde{p}_k = 0$; that is, every bin released with a positive count contains at least one sample.*

Corollary 2 (Rank-k spiked covariance)**.** *Let $V_k \in \mathbb{R}^{d \times k}$ have orthonormal columns, let $\Theta = \mathrm{diag}(\theta_1, \ldots, \theta_k)$ with $\theta_1 \geq \cdots \geq \theta_k > 0$, and let $X_i = V_k \Theta^{1/2} + \sigma Z_i$, $(Z_i)_{ab} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. Set $A_i = X_i X_i^\top$, then $\Sigma = \mathbb{E}[A_i] = V_k \Theta V_k^\top + k\sigma^2 I_d$. Running Algorithm 2 with input $A_1, \ldots, A_n$ outputs Q_T with high probability satisfying*

$$\|Q_T Q_T^\top - V_k V_k^\top\|_F \leq \tilde{O}\left(\frac{\sigma\sqrt{(\theta_1 + k\sigma^2)kd}}{\theta_k\sqrt{n}} + \sigma \cdot \sqrt{\frac{\theta_1 + k\sigma^2}{\theta_k}} \frac{Kdk\sqrt{\theta_1 + \sigma^2 k}}{\varepsilon\theta_k n} \right)$$

Proof. Let

$$S_i \coloneqq A_i - \Sigma.$$

First,

$$A_i = (V\Lambda^{1/2} + \sigma Z_i)(V\Lambda^{1/2} + \sigma Z_i)^\top.$$

Expanding and taking expectations gives

$$\mathbb{E}[A_i] = V\Lambda V^\top + \sigma V \Lambda^{1/2} \mathbb{E}[Z_i^\top] + \sigma \mathbb{E}[Z_i] \Lambda^{1/2} V^\top + \sigma^2 \mathbb{E}[Z_i Z_i^\top].$$

The two cross terms vanish because $\mathbb{E}[Z_i] = 0$. Since $Z_i \in \mathbb{R}^{d \times k}$ has i.i.d. standard Gaussian entries,

$$\mathbb{E}[Z_i Z_i^\top] = kI_d.$$

Therefore

$$\Sigma = \mathbb{E}[A_i] = V\Lambda V^\top + k\sigma^2 I_d.$$

The eigenvalues of Σ are

$$\mu_j = \lambda_j + k\sigma^2, \qquad j \leq k,$$

and

$$\mu_{k+1} = \cdots = \mu_d = k\sigma^2.$$

Hence

$$\mu_1 = \lambda_1 + k\sigma^2, \qquad \mu_k - \mu_{k+1} = \lambda_k.$$

Next we control the variance parameter in Assumption A.3. Write

$$B \coloneqq V\Lambda V^\top.$$

Then

$$\begin{aligned} S_i &= A_i - \Sigma \\ &= \sigma V \Lambda^{1/2} Z_i^\top + \sigma Z_i \Lambda^{1/2} V^\top + \sigma^2 (Z_i Z_i^\top - k I_d). \end{aligned}$$

Define

$$C_i \coloneqq V \Lambda^{1/2} Z_i^\top + Z_i \Lambda^{1/2} V^\top, \qquad R_i \coloneqq Z_i Z_i^\top - k I_d.$$

Thus

$$S_i = \sigma C_i + \sigma^2 R_i.$$

Since $(A+B)^2 \preceq 2A^2 + 2B^2$ for symmetric matrices,

$$\mathbb{E}[S_i^2] \preceq 2\sigma^2 \mathbb{E}[C_i^2] + 2\sigma^4 \mathbb{E}[R_i^2].$$

We now compute the two terms. By Gaussian moment calculations,

$$\mathbb{E}[C_i^2] = (d+2)B + \operatorname{tr}(\Lambda) I_d.$$

Therefore

$$\|\mathbb{E}[C_i^2]\|_2 \le (d+2)\lambda_1 + \operatorname{tr}(\Lambda).$$

Also, since $Z_i Z_i^\top$ is a Wishart matrix with k degrees of freedom,

$$\mathbb{E}[(Z_i Z_i^\top - k I_d)^2] = k(d+1) I_d.$$

Hence

$$\|\mathbb{E}[R_i^2]\|_2 = k(d+1).$$

Combining the previous displays,

$$\begin{aligned} \|\mathbb{E}[S_i^2]\|_2 &\le 2\sigma^2 \left((d+2)\lambda_1 + \operatorname{tr}(\Lambda) \right) + 2\sigma^4 k(d+1) \\ &\lesssim d\sigma^2(\lambda_1 + k\sigma^2). \end{aligned}$$

Since $\mu_1 = \lambda_1 + k\sigma^2$, Assumption A.3 holds with

$$V_{\mathrm{thm}} \lesssim \frac{d\sigma^2(\lambda_1 + k\sigma^2)}{(\lambda_1 + k\sigma^2)^2} = \frac{d\sigma^2}{\lambda_1 + k\sigma^2}.$$

Next we bound γ. For any unit vector u,

$$S_i u = \sigma C_i u + \sigma^2 R_i u.$$

Again using $(x+y)(x+y)^\top \preceq 2xx^\top + 2yy^\top$, we have

$$\mathbb{E}[S_i u u^\top S_i] \preceq 2\sigma^2 \mathbb{E}[C_i u u^\top C_i] + 2\sigma^4 \mathbb{E}[R_i u u^\top R_i].$$

Now

$$C_i u = V \Lambda^{1/2} Z_i^\top u + Z_i \Lambda^{1/2} V^\top u.$$

Since $Z_i^\top u \sim \mathcal{N}(0, I_k)$,

$$\mathbb{E}[V \Lambda^{1/2} Z_i^\top u u^\top Z_i \Lambda^{1/2} V^\top] = V \Lambda V^\top = B.$$

Also, letting $a = \Lambda^{1/2} V^\top u$,

$$\mathbb{E}[Z_i a a^\top Z_i^\top] = \|a\|_2^2 I_d \le \lambda_1 I_d.$$

Therefore

$$\|\mathbb{E}[C_i u u^\top C_i]\|_2 \lesssim \lambda_1.$$

For the Wishart fluctuation term, another standard Gaussian moment calculation gives

$$\mathbb{E}[(Z_i Z_i^\top - k I_d) u u^\top (Z_i Z_i^\top - k I_d)] = k(I_d + u u^\top),$$

and hence

$$\|\mathbb{E}[R_i u u^\top R_i]\|_2 \le 2k.$$

Thus

$$\left\| \mathbb{E}[S_i u u^\top S_i] \right\|_2 \lesssim \sigma^2 \lambda_1 + k\sigma^4 = \sigma^2(\lambda_1 + k\sigma^2).$$

By the definition of γ,

$$\gamma^2 = \max_{\|u\|=1} \frac{\left\|\mathbb{E}[S_i u u^\top S_i]\right\|_2}{\mu_1^2},$$

so

$$\gamma^2 \lesssim \frac{\sigma^2(\lambda_1 + k\sigma^2)}{(\lambda_1 + k\sigma^2)^2} = \frac{\sigma^2}{\lambda_1 + k\sigma^2}.$$

Equivalently,

$$\mu_1 \gamma \lesssim \sigma\sqrt{\lambda_1 + k\sigma^2}.$$

It remains to verify the boundedness parameter on $\mathcal{E}_{\mathrm{bd}}$. The standard Gaussian operator norm bound gives

$$\mathbb{P}\left(\|Z_i\|_2 \le \sqrt{d} + \sqrt{k} + t\right) \ge 1 - e^{-t^2/2}.$$

Taking $t = 2\sqrt{\log n}$ and union bounding over $i \in [n]$ yields

$$\mathbb{P}(\mathcal{E}_{\mathrm{bd}}) \ge 1 - \frac{1}{n}.$$

On this event,

$$\|X_i\|_2 \le \sqrt{\lambda_1} + \sigma(\sqrt{d} + \sqrt{k} + 2\sqrt{\log n}).$$

Therefore

$$\begin{aligned}
\|A_i - \Sigma\|_2 &\le \|A_i\|_2 + \|\Sigma\|_2 \\
&= \|X_i\|_2^2 + \mu_1 \\
&\le \left(\sqrt{\lambda_1} + \sigma(\sqrt{d} + \sqrt{k} + 2\sqrt{\log n})\right)^2 + \lambda_1 + k\sigma^2.
\end{aligned}$$

Using

$$\|P^\top (A_i - \Sigma)\|_{(2,k)} \le \sqrt{k}\,\|A_i - \Sigma\|_2,$$

we may take

$$M = \frac{\sqrt{k}}{\lambda_1}\left[\left(\sqrt{\lambda_1} + \sigma(\sqrt{d} + \sqrt{k} + 2\sqrt{\log n})\right)^2 + \lambda_1 + k\sigma^2\right].$$

We now apply Theorem 2. Its first term is

$$\kappa\sqrt{\frac{kV_{\mathrm{thm}}}{n}}, \qquad \kappa = \frac{\mu_1}{\mu_k - \mu_{k+1}}.$$

Using

$$\mu_1 = \lambda_1 + k\sigma^2, \qquad \mu_k - \mu_{k+1} = \lambda_k,$$

and

$$V_{\mathrm{thm}} \lesssim \frac{d\sigma^2}{\lambda_1 + k\sigma^2},$$

we obtain

$$\begin{aligned}
\frac{\mu_1\sqrt{kV_{\mathrm{thm}}}}{(\mu_k - \mu_{k+1})\sqrt{n}} &\lesssim \frac{(\lambda_1 + k\sigma^2)\sqrt{k\frac{d\sigma^2}{\lambda_1 + k\sigma^2}}}{\lambda_k\sqrt{n}} \\
&= \frac{\sigma\sqrt{kd(\lambda_1 + k\sigma^2)}}{\lambda_k\sqrt{n}}.
\end{aligned}$$

The boundedness term becomes

$$\frac{\sqrt{k}M}{\lambda_k n} = \frac{k\left[\left(\sqrt{\lambda_1} + \sigma(\sqrt{d} + \sqrt{k} + 2\sqrt{\log n})\right)^2 + \lambda_1 + k\sigma^2\right]}{\lambda_k n}.$$

Finally, the privacy term is

$$\kappa^{3/2}\,\frac{Kdk\gamma}{\varepsilon n}.$$

Using $\mu_1\gamma \lesssim \sigma\sqrt{\lambda_1 + k\sigma^2}$ and $\mu_k - \mu_{k+1} = \lambda_k$, this is bounded by

$$\sqrt{\kappa}\,\frac{Kdk\,\sigma\sqrt{\lambda_1 + k\sigma^2}}{\varepsilon\lambda_k n}.$$

Combining these three bounds proves the claim. □

Corollary 4 (Gaussian spiked covariance, Algorithm 2)**.** *With input matrices $A_i = x_i x_i^\top$ with x_i defined as in Theorem 7 Algorithm 2 gives*

$$\|Q_T Q_T^\top - V_k V_k^\top\|_F \le \widetilde{O}\left(\kappa\sqrt{\frac{dk}{n}} + \kappa^{3/2}\,\frac{Kdk}{\varepsilon n}\right).$$

Proof. Since $x_i \sim \mathcal{N}(0, \Sigma)$, we have

$$\mathbb{E}[A_i] = \mathbb{E}[x_i x_i^\top] = \Sigma.$$

The eigenvectors corresponding to the top k eigenvalues of Σ are the columns of U. Moreover,

$$\lambda_j = \theta_j + \sigma^2, \qquad j \le k, \qquad \lambda_{k+1} = \cdots = \lambda_d = \sigma^2,$$

so the target eigenspace is $\mathrm{span}(U)$, and the eigengap is

$$\Delta_k = \lambda_k - \lambda_{k+1} = \theta_k.$$

Let

$$S_i \coloneqq A_i - \Sigma = x_i x_i^\top - \Sigma.$$

We first compute the variance parameter appearing in Assumption A.3. Since x_i is Gaussian,

$$\mathbb{E}\left[\|x_i\|_2^2 x_i x_i^\top\right] = \mathrm{tr}(\Sigma)\Sigma + 2\Sigma^2.$$

Therefore

$$\begin{aligned}\mathbb{E}[S_i^2] &= \mathbb{E}[(x_i x_i^\top - \Sigma)^2] \\ &= \mathbb{E}[\|x_i\|_2^2 x_i x_i^\top] - \Sigma^2 \\ &= \mathrm{tr}(\Sigma)\Sigma + \Sigma^2.\end{aligned}$$

Hence

$$\left\|\mathbb{E}[S_i^2]\right\|_2 = \lambda_1\,\mathrm{tr}(\Sigma) + \lambda_1^2 = \lambda_1\left(\lambda_1 + \sum_{j=1}^{d}\lambda_j\right).$$

Thus Assumption A.3 holds with

$$V_{\mathrm{thm}} = \frac{\lambda_1\left(\lambda_1 + \sum_{j=1}^{d}\lambda_j\right)}{\lambda_1^2} = \frac{\lambda_1 + \sum_{j=1}^{d}\lambda_j}{\lambda_1}.$$

Next we bound γ. For any unit vector u,

$$\begin{aligned}\mathbb{E}[S_i u u^\top S_i] &= \mathbb{E}[(x_i x_i^\top - \Sigma) u u^\top (x_i x_i^\top - \Sigma)] \\ &= \mathbb{E}[(u^\top x_i)^2 x_i x_i^\top] - \Sigma u u^\top \Sigma.\end{aligned}$$

Using the Gaussian fourth-moment identity,

$$\mathbb{E}[(u^\top x_i)^2 x_i x_i^\top] = (u^\top \Sigma u)\Sigma + 2\Sigma u u^\top \Sigma.$$

Therefore

$$\mathbb{E}[S_i u u^\top S_i] = (u^\top \Sigma u)\Sigma + \Sigma u u^\top \Sigma.$$

Taking operator norms gives

$$\begin{aligned}\left\|\mathbb{E}[S_i u u^\top S_i]\right\|_2 &\le (u^\top \Sigma u)\|\Sigma\|_2 + \|\Sigma u\|_2^2 \\ &\le \lambda_1^2 + \lambda_1^2 = 2\lambda_1^2.\end{aligned}$$

Hence

$$\gamma^2 = \max_{\|u\|=1} \frac{\left\|\mathbb{E}[S_i u u^\top S_i]\right\|_2}{\lambda_1^2} \le 2.$$

For Gaussian quadratic forms, Assumption A.4 holds with $a = 1$ and an absolute constant K. We absorb this absolute constant into the $\widetilde{O}(\cdot)$ notation.

It remains to verify the boundedness parameter on a high-probability event. Write $x_i = \Sigma^{1/2} g_i$, where $g_i \sim \mathcal{N}(0, I_d)$. Then

$$\|x_i\|_2 \le \sqrt{\lambda_1}\|g_i\|_2.$$

By the standard Gaussian norm bound,

$$\mathbb{P}\left(\|g_i\|_2 \le \sqrt{d} + t\right) \ge 1 - e^{-t^2/2}.$$

Taking $t = 2\sqrt{\log n}$ and applying a union bound over $i \in [n]$, we get

$$\mathbb{P}(\mathcal{E}_{\mathrm{bd}}) \ge 1 - \frac{1}{n}.$$

On $\mathcal{E}_{\mathrm{bd}}$,

$$\|A_i - \Sigma\|_2 \le \|x_i x_i^\top\|_2 + \|\Sigma\|_2 = \|x_i\|_2^2 + \lambda_1 \le \lambda_1 \left[\left(\sqrt{d} + 2\sqrt{\log n}\right)^2 + 1\right].$$

Since

$$\|P^\top (A_i - \Sigma)\|_{(2,k)} \le \sqrt{k}\,\|A_i - \Sigma\|_2,$$

Assumption A.2 holds on $\mathcal{E}_{\mathrm{bd}}$ with

$$M = \sqrt{k}\left[\left(\sqrt{d} + 2\sqrt{\log n}\right)^2 + 1\right].$$

We now apply Theorem 2. Its first term becomes

$$\begin{aligned}\kappa\sqrt{\frac{kV_{\mathrm{thm}}}{n}} &= \frac{\lambda_1\sqrt{k\frac{\lambda_1 + \sum_{j=1}^d \lambda_j}{\lambda_1}}}{\Delta_k\sqrt{n}} \\ &= \frac{\sqrt{k\lambda_1\left(\lambda_1 + \sum_{j=1}^d \lambda_j\right)}}{\Delta_k\sqrt{n}}.\end{aligned}$$

The boundedness term becomes

$$\frac{\sqrt{k}M}{\Delta_k n} = \frac{k\lambda_1\left[\left(\sqrt{d} + 2\sqrt{\log n}\right)^2 + 1\right]}{\Delta_k n}.$$

Finally, since $\gamma \le \sqrt{2}$ and K is an absolute constant for Gaussian quadratic forms, the privacy term satisfies

$$\frac{\lambda_1 K dk\gamma}{\varepsilon\Delta_k n} \lesssim \frac{dk\lambda_1}{\varepsilon\Delta_k n}.$$

Since $\varepsilon \le 1$ and $K = O(1)$, the boundedness term is dominated by the privacy term. Combining these three bounds proves the result. □

C.5 Proofs for the tangent refinement theorem

Theorem 5 (Utility of TADADPO)**.** *Suppose $A_1, \ldots, A_n$ satisfy Assumption A. Suppose further that, together with the starting point Q_0, they satisfy Assumption B, and that*

$$\left\|Q_0 Q_0^\top - V_k V_k^\top\right\|_F \le c_0/2$$

with probability at least $1 - \zeta/8$. Set

$$B = \lfloor n/T \rfloor, \qquad T = \left\lceil C_T \log\left(\frac{ndk}{\zeta\delta\varepsilon}\right)\right\rceil.$$

If

$$n \geq \tilde{\Omega}\left(\max\left\{\frac{K^2d + Kdk}{\varepsilon}, \frac{k(v_0^2 + v_1^2)}{\Delta^2}, \frac{\sqrt{k}M_\perp}{\Delta}, \frac{K(g_0 + g_1)dk}{\varepsilon\Delta}\right\}\right),$$

then there exists a choice of learning rates $\{\eta_t\}_{t=1}^T$ *such that, with probability at least* $1 - \zeta$*, Algorithm 3 outputs* Q_T *satisfying*

$$\left\|Q_T Q_T^\top - V_k V_k^\top\right\|_F \leq \mathrm{Opt}_T + \tilde{O}\left(\frac{\sqrt{k}v_0}{\Delta\sqrt{n}} + \frac{\sqrt{k}M_\perp}{\Delta n} + \frac{Kg_0dk}{\Delta\varepsilon n}\right),$$

where $\mathrm{Opt}_T := \exp\left(-c\Delta\sum_{t=1}^T \eta_t\right)\left\|Q_0Q_0^\top - V_kV_k^\top\right\|_F$.

Lemma 12 (Privacy of TADADPO)**.** *Suppose the warm-start samples and refinement samples are disjoint. If the warm-start algorithm is* (ε, δ)*-DP and each refinement iteration uses disjoint batches with the budget allocation in Algorithm 3, then the full two-stage algorithm is* (ε, δ)*-DP. If the warm-start and refinement phases reuse records, their privacy costs compose sequentially.*

Lemma 13 (Privacy of projected coordinatewise tangent mean)**.** *Fix* $Q \in \mathbb{S}_{d,k}$*,* $\Pi = I - QQ^\top$*, a coordinate center* $C \in \mathbb{R}^{d\times k}$*, and a coordinatewise radius* $r > 0$*. For a batch* $S = (A_1, \ldots, A_m)$*, let* $\widetilde{Y}_i(C)$ *be* $\Pi A_i Q$ *after coordinatewise truncation to* $[C_{jr} - r, C_{jr} + r]$*, and define*

$$h_C(S) := \Pi\left(\frac{1}{m}\sum_{i=1}^m \widetilde{Y}_i(C)\right).$$

Let $Z \in \mathbb{R}^{d\times k}$ *have i.i.d. entries* $\mathcal{N}(0, \sigma^2)$*, and release*

$$M_C(S) = h_C(S) + \Pi Z.$$

If

$$\sigma \geq \frac{2r\sqrt{dk}}{m\varepsilon_M}\sqrt{2\log(1.25/\delta_M)},$$

then M_C *is* $(\varepsilon_M, \delta_M)$*-DP conditional on* C*.*

Proof. By construction, $h_C(S)$ lies in the tangent subspace

$$\mathcal{T}_Q := \left\{Y \in \mathbb{R}^{d\times k} : Q^\top Y = 0\right\},$$

because $Q^\top\Pi = 0$. If neighboring batches differ in one entry, then

$$\|h_C(S) - h_C(S')\|_F \leq \frac{2r\sqrt{dk}}{m},$$

because each coordinatewise-truncated matrix lies in a coordinate box of radius r, and Π is non-expansive in Frobenius norm. The random matrix ΠZ is a centered isotropic Gaussian on $\mathcal{T}_Q$: if $Q_\perp$ is an orthonormal basis for $\mathrm{span}(Q)^\perp$, then $\Pi Z = Q_\perp(Q_\perp^\top Z)$, and $Q_\perp^\top Z$ has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. Thus this is the standard Gaussian mechanism in the Euclidean space $(\mathcal{T}_Q, \|\cdot\|_F)$ with sensitivity

$$\Delta_t \leq \frac{2r\sqrt{dk}}{m},$$

which proves the claim. □

Proof of Lemma 12. Fix an iteration t, and condition on all previous outputs. Then $Q = Q_{t-1}$ and $\Pi = I - QQ^\top$ are fixed.

The range half is handled by PRIVRANGE and is $(\varepsilon_R, \delta_R)$-DP. It is disjoint from the mean half.

Now consider the mean half. Let C be the coordinatewise center computed internally by TRUNCATEDMEAN. Since TRUNCATEDMEAN runs dk coordinate histograms, each with privacy budget $(\varepsilon_C/(dk), \delta_C/(dk))$, the center C is $(\varepsilon_C, \delta_C)$-DP by basic composition.

Conditional on any fixed value of C, the projected coordinatewise truncated mean plus ΠZ_t is $(\varepsilon_M, \delta_M)$-DP by Lemma 13, using the Gaussian calibration in Algorithm 3. By adaptive composition,

$$(C, \bar{Y}_t + W_t)$$

is $(\varepsilon_C + \varepsilon_M, \delta_C + \delta_M)$-DP on the mean half. The algorithm releases only $\bar{Y}_t + W_t$, so hiding C is post-processing. Since Algorithm 3 sets $(\varepsilon_C, \delta_C) = (\varepsilon_M, \delta_M) = (\varepsilon/2, \delta/2)$, the mean-half mechanism is (ε, δ)-DP.

The range and mean halves are disjoint, so one update is (ε, δ)-DP by parallel composition. The refinement batches are disjoint across t, so the full refinement phase is (ε, δ)-DP. Finally, because the warm-start and refinement samples are disjoint, the two-stage algorithm is (ε, δ)-DP by parallel composition. □

Proof of Theorem 5. Let $e_t := \left\|Q_t Q_t^\top - P_\star\right\|_F$. We first collect the high-probability events. By Lemma 14 and a union bound over $t \in [T]$,

$$g^2_{Q_{t-1}} \le \widehat{\Lambda}_t \le C_\Lambda g^2_{Q_{t-1}}$$

holds for all t with probability at least $1 - \zeta/8$. On this event, Lemma 15 implies that no coordinate in any mean batch is clipped. Thus, writing $Q = Q_{t-1}$ and $\Pi = I - QQ^\top$, on the no-clipping event,

$$\mu_t = \frac{1}{m} \sum_{i \in \mathcal{B}_t^{\text{mean}}} Y_i = \frac{1}{m} \sum_{i \in \mathcal{B}_t^{\text{mean}}} \Pi A_i Q.$$

Since each Y_i lies in the tangent space, $\Pi \mu_t = \mu_t$. Therefore the update has the form

$$Q_t = \mathrm{QR}\left(Q + \eta_t\{\Pi \Sigma Q + E_t\}\right),$$

where

$$E_t = \frac{1}{m} \sum_{i \in \mathcal{B}_t^{\text{mean}}} \Pi(A_i - \Sigma)Q + W_t.$$

This is exactly the form needed for the local tangent contraction lemma. By Lemma 16, again union bounded over t,

$$\|E_t\|_F \le \tilde{O}\left(\sqrt{k} v_{Q_{t-1}} m^{-1/2} + \sqrt{k} M_\perp m^{-1} + \frac{K g_{Q_{t-1}} dk}{m\varepsilon}\right)$$

simultaneously for all t, with probability at least $1 - \zeta/4$. Using Assumption B.1, this gives

$$\|E_t\|_F \le A + D e_{t-1},$$

where

$$A := \tilde{O}\left(\sqrt{k} v_0 m^{-1/2} + \sqrt{k} M_\perp m^{-1} + \frac{K g_0 dk}{m\varepsilon}\right), \qquad D := \tilde{O}\left(\sqrt{k} v_1 m^{-1/2} + \frac{K g_1 dk}{m\varepsilon}\right).$$

For every iterate that remains in the local basin, Lemma 17 gives

$$e_t \le (1 - c\eta_t \Delta) e_{t-1} + C\eta_t \|E_t\|_F.$$

The stated lower bound on n, together with $m \asymp n/T$ and $T \asymp \kappa_\perp \log(ndk/\zeta\delta\varepsilon)$, implies $D \le c\Delta$ for a small enough universal constant: the batch is a factor T smaller than n, so every branch of that lower bound controlling A or D picks up $\kappa_\perp$. Hence

$$e_t \le (1 - c'\eta_t \Delta) e_{t-1} + C\eta_t A.$$

Unrolling this recursion yields

$$e_T \le \exp\left(-c'\Delta \sum_{t=1}^{T} \eta_t\right) e_0 + \frac{CA}{\Delta}.$$

Since $m \asymp n/T$ with $T \asymp \kappa_\perp \log(ndk/\zeta\delta\varepsilon)$, and A carries $m^{-1/2}$ in its first term and m^{-1} in the other two,

$$\frac{A}{\Delta} = \tilde{O}\left(\sqrt{\kappa_\perp}\, \frac{\sqrt{k} v_0}{\Delta\sqrt{n}} + \kappa_\perp \frac{\sqrt{k} M_\perp}{\Delta n} + \kappa_\perp \frac{K g_0 dk}{\Delta\varepsilon n}\right).$$

This gives the claimed bound after adjusting constants in Opt_T.

It remains only to justify that the local condition is valid throughout the proof. The sample-size lower bound also makes $A/\Delta \le c_0$ after reducing the universal constants. Since the initializer satisfies $\left\|Q_0Q_0^\top - P_\star\right\|_F \le c_0/2$, induction gives $\left\|Q_tQ_t^\top - P_\star\right\|_F \le c_0$ for every $t \le T$. Combining the initializer event with the range, center, and perturbation events gives probability at least $1-\zeta$. $\square$

Lemma 14 (Accuracy of PRIVRANGE on tangent samples)**.** *Fix $Q \in \mathbb{S}_{d,k}$ independently of the current range batch, and let $Y_i = \Pi_Q A_i Q$. Suppose Assumption B.2 holds at Q. If the block size in Algorithm 4 satisfies*

$$b \ge \tilde{\Omega}\left(K^2 d \log(km_0/\alpha)\right),$$

then with probability at least $1-\alpha$, PRIVRANGE *outputs $\widehat{\Lambda}$ satisfying*

$$g_Q^2 \le \widehat{\Lambda} \le C_\Lambda g_Q^2.$$

Proof. The privacy argument is part of Lemma 12; here we prove accuracy. Here PRIVRANGE is applied to the tangent samples $Y_i = \Pi_Q A_i Q$. We now prove the accuracy guarantee. Fix the current iterate Q, and condition on all previous batches. Then Q is deterministic and independent of the samples used by PRIVRANGE. Let $P = QQ^\top, \Pi_Q = I - P$, and write the columns of Q as $q_1, \dots, q_k$. For each column $r \in [k]$, define the tangent covariance

$$C_r(Q) := \Pi_Q \mathbb{E}\left[(A-\Sigma) q_r q_r^\top (A-\Sigma)^\top\right] \Pi_Q,$$

and define the corresponding tangent scale

$$g_Q^2 := \max_{r\in[k]} \|C_r(Q)\|_2.$$

For a paired sample, define $\widetilde{Y}_s^{(r)} := \Pi_Q(A_{2s} - A_{2s-1})q_r$. Then,

$$\begin{aligned}\mathbb{E}\left[\widetilde{Y}_s^{(r)}\left(\widetilde{Y}_s^{(r)}\right)^\top\right] &= \mathbb{E}\left[\Pi_Q(A_{2s}-\Sigma) q_r q_r^\top (A_{2s}-\Sigma)^\top \Pi_Q\right] \\ &\quad + \mathbb{E}\left[\Pi_Q(A_{2s-1}-\Sigma) q_r q_r^\top (A_{2s-1}-\Sigma)^\top \Pi_Q\right] \\ &= 2C_r(Q).\end{aligned}$$

Thus the empirical matrix $\widehat{C}_{j,r} := \frac{1}{2b}\sum_{s\in\mathcal{G}_j} \widetilde{Y}_s^{(r)}\left(\widetilde{Y}_s^{(r)}\right)^\top$ is an unbiased estimator of $C_r(Q)$. We define $\lambda_{j,r} := \lambda_{\max}(\widehat{C}_{j,r}), \qquad m_j := \max_{r\in[k]} \lambda_{j,r}$.

By the concentration result underlying Theorem 6.1 of Liu et al. [2022], applied conditionally on Q to the independent centered vectors $\widetilde{Y}_s^{(r)}/\sqrt{2}, s \in \mathcal{G}_j$, and then union bounded over $r \in [k]$, the batch-size condition

$$b \ge \widetilde{\Omega}\left(K^2\left(d + \log\frac{km_0}{\alpha}\right)\right)$$

implies that, for each fixed batch j, with probability at least $1 - \alpha/(10m_0)$,

$$\max_{r\in[k]} \left\|\widehat{C}_{j,r} - C_r(Q)\right\|_2 \le \eta_0 g_Q^2, \eta_0 := 1 - 2^{-1/2}.$$

A union bound over the m_0 batches gives that, with probability at least $1-\alpha/10$, this event holds simultaneously for all $j \in [m_0]$ and all $r \in [k]$. On this event, we prove that every batch statistic m_j lies in a constant factor interval around g_Q^2. Let $r_\star$ be a column attaining the maximum in the definition of g_Q^2, so that $\|C_{r_\star}(Q)\|_2 = g_Q^2$. By Weyl's inequality,

$$\lambda_{j,r_\star} = \lambda_{\max}(\widehat{C}_{j,r_\star}) \ge \lambda_{\max}(C_{r_\star}(Q)) - \|\widehat{C}_{j,r_\star} - C_{r_\star}(Q)\|_2 \ge (1-\eta_0) g_Q^2 = 2^{-1/2} g_Q^2.$$

Therefore

$$m_j = \max_{r\in[k]} \lambda_{j,r} \ge 2^{-1/2} g_Q^2.$$

For every $r \in [k]$,

$$\lambda_{j,r} \le \lambda_{\max}(C_r(Q)) + \left\|\widehat{C}_{j,r} - C_r(Q)\right\|_2 \le 2^{1/2} g_Q^2.$$

Taking the maximum over r yields simultaneously for all $j \in [m_0]$,

$$m_j \in \left[2^{-1/2} g_Q^2,\, 2^{1/2} g_Q^2\right].$$

It remains to pass from the non-private block statistics to the private range estimate. The geometric bins in the stable histogram algorithm have multiplicative width $2^{1/4}$. The interval

$$I_Q := \left[2^{-1/2} g_Q^2,\, 2^{1/2} g_Q^2\right]$$

has endpoint ratio 2, and therefore intersects at most a universal constant number C_Ω of consecutive geometric bins. Therefore, by Lemma 11, the bin selected by the private histogram intersects I_Q with probability at least $1 - \alpha/2$.

Let the selected bin have left endpoint ℓ. Then,

$$2^{-3/4} g_Q^2 \le \ell \le 2^{1/2} g_Q^2.$$

The algorithm outputs $\widehat{\Lambda} = 2\ell$, where the factor 2 is a universal safety constant. Hence

$$g_Q^2 \le \widehat{\Lambda} \le 2^{3/2} g_Q^2.$$

Absorbing $2^{3/2}$ into a universal constant C_Λ, we obtain

$$g_Q^2 \le \widehat{\Lambda} \le C_\Lambda g_Q^2.$$

Combining the concentration event and the histogram event gives the claim with probability at least $1 - \alpha$.

□

Lemma 15 (No clipping for projected tangent truncated mean)**.** *Fix an iteration and condition on the past, so that $Q \in \mathbb{S}_{d,k}$, $\Pi = I - QQ^\top$, and $\widehat{\Lambda}$ are fixed. Let*

$$Y_i = \Pi A_i Q, \qquad \mu_Q = \mathbb{E}\,[Y_i] = \Pi \Sigma Q, \qquad i = 1, \dots, m.$$

Suppose $g_Q^2 \le \widehat{\Lambda} \le C_\Lambda g_Q^2$. Run TRUNCATEDMEAN *on $Y_1, \dots, Y_m$, with coordinate radius*

$$r = C_R K \sqrt{\widehat{\Lambda}} \log^a\left(\frac{mdkT}{\alpha}\right)$$

and center privacy parameters $(\varepsilon_C, \delta_C)$. If

$$m \ge \tilde{\Omega}\left(\frac{dk}{\varepsilon_C}\right),$$

then with probability at least $1 - \alpha$,

$$(Y_i)_{jr} \in [\bar{g}_{jr} - r, \bar{g}_{jr} + r] \qquad \text{for all } i \in [m],\ j \in [d],\ r \in [k],$$

where $\bar{g}_{jr}$ is the coordinate center selected by TRUNCATEDMEAN*. Consequently, if $\widetilde{Y}_i$ denotes the truncated version of Y_i,*

$$\Pi\left(\frac{1}{m}\sum_{i=1}^m \widetilde{Y}_i\right) = \frac{1}{m}\sum_{i=1}^m Y_i.$$

Proof of Lemma 15. Condition on the past and on the successful event $g_Q^2 \le \widehat{\Lambda} \le C_\Lambda g_Q^2$. Then Q, Π, and $\widehat{\Lambda}$ are fixed, and the current mean batch is independent of them.

Fix $(j, r) \in [d] \times [k]$. Since

$$(Y_i - \mu_Q)_{jr} = e_j^\top \Pi (A_i - \Sigma) q_r = (\Pi e_j)^\top (A_i - \Sigma) q_r,$$

if $\Pi e_j = 0$, this coordinate is identically zero. Otherwise set $u_j = \Pi e_j / \|\Pi e_j\|_2$. Then $u_j \perp \operatorname{span}(Q)$, $\|u_j\|_2 = 1$, and

$$(Y_i - \mu_Q)_{jr} = \|\Pi e_j\|_2\, u_j^\top (A_i - \Sigma) q_r.$$

Thus Assumption B.2, together with $\widehat{\Lambda} \geq g_Q^2$, implies that for

$$L := C_{\text{tail}} K \sqrt{\widehat{\Lambda}} \log^a\left(\frac{mdkT}{\alpha}\right),$$

we have

$$\Pr[|(Y_i - \mu_Q)_{jr}| > L] \leq \frac{\alpha}{16mdkT}.$$

A union bound over all dk coordinates and all samples in the mean batch gives, with probability at least $1 - \alpha/4$,

$$|(Y_i - \mu_Q)_{jr}| \leq L$$

simultaneously for all $i \in [m]$ and all (j, r).

On this event, for each coordinate (j, r), all samples lie in

$$I_{jr} := [(\mu_Q)_{jr} - L, (\mu_Q)_{jr} + L].$$

Since the histogram bins have width $\sqrt{\widehat{\Lambda}}$, the interval I_{jr} intersects at most

$$C_{\text{bin}} \leq CK \log^a\left(\frac{mdkT}{\alpha}\right)$$

consecutive bins. By the stability-based histogram guarantee, equivalently by Lemma 11, applied with per-coordinate privacy budget $(\varepsilon_C/(dk), \delta_C/(dk))$ and union bounded over the dk coordinates, the selected private bin intersects I_{jr} for every coordinate with probability at least $1 - \alpha/8$, provided

$$m \geq \tilde{\Omega}\left(\frac{dk}{\varepsilon_C}\right),$$

with logarithmic dependence on dk, α, δ_C and the $K \log^a(mdkT/\alpha)$ factor hidden in $\tilde{\Omega}(\cdot)$.

Let $\bar{g}_{jr}$ be the left endpoint of the selected bin. Since the selected bin intersects I_{jr},

$$|\bar{g}_{jr} - (\mu_Q)_{jr}| \leq L + \sqrt{\widehat{\Lambda}} \leq CK\sqrt{\widehat{\Lambda}} \log^a\left(\frac{mdkT}{\alpha}\right).$$

Thus, on the intersection of the tail and histogram events, for all i, j, r,

$$|(Y_i)_{jr} - \bar{g}_{jr}| \leq |(Y_i - \mu_Q)_{jr}| + |(\mu_Q)_{jr} - \bar{g}_{jr}| \leq r$$

after choosing C_R sufficiently large. Hence no coordinate is clipped, so $\widetilde{Y}_i = Y_i$ for all i. Since $Y_i = \Pi A_i Q$, we have $\Pi Y_i = Y_i$, and therefore

$$\Pi\left(\frac{1}{m}\sum_{i=1}^m \widetilde{Y}_i\right) = \frac{1}{m}\sum_{i=1}^m Y_i.$$

Summing the failure probabilities gives the stated probability after adjusting constants. □

Lemma 16 (Tangent perturbation bound)**.** *Fix an iteration and condition on the past and the randomness that determines r_t, so $Q = Q_{t-1}$, $\Pi = I - QQ^\top$, and r_t are fixed. Let $m = |\mathcal{B}_t^{\text{mean}}|$, and define*

$$S_t := \frac{1}{m}\sum_{i \in \mathcal{B}_t^{\text{mean}}} \Pi(A_i - \Sigma)Q, \qquad E_t := S_t + W_t,$$

where $W_t = \Pi Z_t$, $(Z_t)_{jr} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_t^2)$, with $\sigma_t = \frac{2r_t\sqrt{dk}}{m\varepsilon_M}\sqrt{2\log(1.25/\delta_M)}$. Let

$$Y_i := \Pi(A_i - \Sigma)Q,$$

and define

$$v_Q^2 := \max\left\{\left\|\mathbb{E}[Y_i Y_i^\top]\right\|_2, \left\|\mathbb{E}[Y_i^\top Y_i]\right\|_2\right\}.$$

Assume

$$\|Y_i\|_2 \leq M_\perp \qquad \textit{a.s.}$$

Then, on the event $r_t \leq CKg_Q \operatorname{polylog}(BdkT/\zeta)$, with probability at least $1 - \xi$ over the mean batch and Gaussian noise, we have, suppressing logarithmic factors,

$$\|E_t\|_F \leq \tilde{O}\left(\sqrt{k}v_Q m^{-1/2} + \sqrt{k}M_\perp m^{-1} + \frac{Kg_Q dk}{m\varepsilon}\right).$$

Proof. Conditionally on the past and on r_t, the matrices $Y_i = \Pi(A_i - \Sigma)Q,\ i \in \mathcal{B}_t^{\mathrm{mean}}$, are independent and mean zero. We first bound the stochastic term S_t. For a rectangular matrix $Y \in \mathbb{R}^{d\times k}$, define its self-adjoint dilation

$$\mathcal{D}(Y) := \begin{pmatrix} 0 & Y \\ Y^\top & 0 \end{pmatrix} \in \mathbb{R}^{(d+k)\times(d+k)}.$$

Then $\|\mathcal{D}(Y_i)\|_2 = \|Y_i\|_2 \le M_\perp$, and by the definition of v_Q, $\left\|\sum_{i\in\mathcal{B}_t^{\mathrm{mean}}} \mathbb{E}[\mathcal{D}(Y_i)^2]\right\|_2 \le m v_Q^2$. Self-adjoint matrix Bernstein applied in dimension $d+k$ gives, with probability at least $1-\xi/2$,

$$\left\|\frac{1}{m}\sum_{i\in\mathcal{B}_t^{\mathrm{mean}}} Y_i\right\|_2 \le C\left(v_Q\sqrt{\frac{\log((d+k)/\xi)}{m}} + \frac{M_\perp \log((d+k)/\xi)}{m}\right).$$

Since $S_t \in \mathbb{R}^{d\times k}$, we have $\|S_t\|_F \le \sqrt{k}\|S_t\|_2$. Therefore, with probability at least $1-\xi/2$,

$$\|S_t\|_F \le C\sqrt{k}\left(v_Q\sqrt{\frac{\log((d+k)/\xi)}{m}} + \frac{M_\perp \log((d+k)/\xi)}{m}\right).$$

It remains to bound the Gaussian term. Let $Q_\perp$ be an orthonormal basis for $\mathrm{range}(\Pi)$, so that $\Pi = Q_\perp Q_\perp^\top$. Then $W_t = \Pi Z_t = Q_\perp(Q_\perp^\top Z_t)$, and $\|W_t\|_F = \|Q_\perp^\top Z_t\|_F$. By rotational invariance of the Gaussian distribution, $Q_\perp^\top Z_t \in \mathbb{R}^{(d-k)\times k}$ has i.i.d. $\mathcal{N}(0,\sigma_t^2)$ entries. Thus $\frac{\|W_t\|_F^2}{\sigma_t^2} \sim \chi^2_{(d-k)k}$. A standard chi-square tail bound implies that, with probability at least $1-\xi/2$,

$$\|W_t\|_F \le C\sigma_t\sqrt{dk + \log(1/\xi)}.$$

Substituting $\sigma_t = \frac{2r_t\sqrt{dk}}{m\varepsilon_M}\sqrt{2\log(1.25/\delta_M)}$ gives

$$\|W_t\|_F \le C\frac{r_t\sqrt{dk}}{m\varepsilon_M}\sqrt{(dk+\log(1/\xi))\log(1/\delta_M)}.$$

Combining the bounds for S_t and W_t by the triangle inequality and using the event,

$$r_t \le CKg_Q\,\mathrm{polylog}\left(BdkT/\zeta\right),$$

completes the proof. □

Lemma 17 (Local contraction of the noisy tangent step)**.** *There exist universal constants $c_0, c, c', C > 0$ such that the following holds. Let $P_\star = V_k V_k^\top$, let $\Delta = \lambda_k - \lambda_{k+1} > 0$, and let $Q \in \mathbb{S}_{d,k}$ satisfy*

$$\|QQ^\top - P_\star\|_F \le c_0.$$

Let $E \in \mathbb{R}^{d\times k}$ satisfy $Q^\top E = 0$. If $\eta \le c/(\lambda_1 - \lambda_d)$. and

$$Q^+ = \mathrm{QR}(Q + \eta(\Pi_Q \Sigma Q + E)), \qquad \Pi_Q = I - QQ^\top,$$

then

$$\|Q^+(Q^+)^\top - P_\star\|_F \le (1 - c'\eta\Delta)\|QQ^\top - P_\star\|_F + C\eta\|E\|_F.$$

Proof. Let $G_Q := \Pi_Q \Sigma Q$. We first prove contraction for the noiseless update

$$\bar{Q} = \mathrm{QR}(Q + \eta G_Q),$$

and then compare $\bar{Q}$ with the noisy update. Work in the eigenbasis $[V_k, V_\perp]$ of Σ, so that

$$\Sigma = \mathrm{diag}(\Lambda_1, \Lambda_2), \qquad \Lambda_1 = \mathrm{diag}(\lambda_1, \dots, \lambda_k), \qquad \|\Lambda_2\|_2 \le \lambda_{k+1}.$$

Set

$$A := V_k^\top Q, \qquad B := V_\perp^\top Q, \qquad H := Q^\top \Sigma Q, \qquad N := \Lambda_2 B - BH.$$

Since $Q^\top Q = I_k$, we have $A^\top A + B^\top B = I_k$. Moreover,

$$\|QQ^\top - P_\star\|_F = \sqrt{2}\|B\|_F,$$

Let $\bar{Y} := Q + \eta G_Q$. The bottom block of $\bar{Y}$ is $V_\perp^\top \bar{Y} = B + \eta N$. Also $Q^\top G_Q = 0$, and therefore

$$\bar{Y}^\top \bar{Y} = I_k + \eta^2 G_Q^\top G_Q \succeq I_k.$$

Thus QR normalization right-multiplies $\bar{Y}$ by a matrix of operator norm at most one, so

$$\|V_\perp^\top \bar{Q}\|_F^2 \le \|B + \eta N\|_F^2 = \|B\|_F^2 + 2\eta\langle B, N\rangle + \eta^2 \|N\|_F^2.$$

We now bound the two terms involving N.

$$-\langle B, N\rangle = \operatorname{tr}(B^\top B H) - \operatorname{tr}(B^\top \Lambda_2 B).$$

Equivalently, using $A^\top A + B^\top B = I_k$ and $\lambda_i - \lambda_j \ge \Delta$ for $i \le k < j$,

$$-\langle B, N\rangle = \sum_{i \le k < j} (\lambda_i - \lambda_j)|(AB^\top)_{ij}|^2 \ge \Delta \|AB^\top\|_F^2.$$

Finally,

$$\|AB^\top\|_F^2 = \operatorname{tr}(B^\top B A^\top A) = \operatorname{tr}\big(B^\top B(I_k - B^\top B)\big) \ge (1 - \|B\|_F^2)\|B\|_F^2.$$

Hence $-\langle B, N\rangle \ge \Delta(1 - \|B\|_F^2)\|B\|_F^2$. The same block calculation also gives, for c_0 sufficiently small,

$$\|N\|_F^2 \le C \sum_{i \le k < j} (\lambda_i - \lambda_j)^2 |(AB^\top)_{ij}|^2 \le C\lambda_1[-\langle B, N\rangle].$$

Therefore, if $\eta \le c/(\lambda_1 - \lambda_d)$,

$$\begin{aligned}
\|V_\perp^\top \bar{Q}\|_F^2 &\le \|B\|_F^2 - 2\eta[-\langle B, N\rangle] + C\eta^2 \lambda_1[-\langle B, N\rangle] \\
&\le \|B\|_F^2 - c\eta\Delta\|B\|_F^2 \\
&\le (1 - c\eta\Delta)\|B\|_F^2,
\end{aligned}$$

where in the second line we used $\|B\|_F \le c_0$ and chose c_0 and c small enough. Taking square roots and adjusting constants,

$$\|V_\perp^\top \bar{Q}\|_F \le (1 - c'\eta\Delta)\|B\|_F.$$

Since projector Frobenius distance is $\sqrt{2}\|V_\perp^\top Q\|_F$,

$$\|\bar{Q}\bar{Q}^\top - P_\star\|_F \le (1 - c'\eta\Delta)\|QQ^\top - P_\star\|_F.$$

Now consider the noisy update

$$Y := Q + \eta(G_Q + E).$$

Since $Q^\top G_Q = 0$ and $Q^\top E = 0$,

$$Q^\top Y = Q^\top \bar{Y} = I_k.$$

Thus both Y and $\bar{Y}$ have smallest singular value at least one. The orthogonal projectors onto their column spaces therefore satisfy

$$\|P_Y - P_{\bar{Y}}\|_F \le C\|Y - \bar{Y}\|_F = C\eta\|E\|_F,$$

where

$$P_Y = Q^+(Q^+)^\top, \qquad P_{\bar{Y}} = \bar{Q}\bar{Q}^\top.$$

The triangle inequality gives

$$\begin{aligned}
\|Q^+(Q^+)^\top - P_\star\|_F &\le \|\bar{Q}\bar{Q}^\top - P_\star\|_F + \|P_Y - P_{\bar{Y}}\|_F \\
&\le (1 - c'\eta\Delta)\|QQ^\top - P_\star\|_F + C\eta\|E\|_F.
\end{aligned}$$

This proves the lemma. □

Lemma 18 (Gaussian covariance)**.** *Let $A_i = x_i x_i^\top$ with $x_i \sim \mathcal{N}(0, \Sigma)$, and let*

$$\mathrm{tr}_{>k}(\Sigma) := \sum_{j>k} \lambda_j.$$

In a local basin around $P_\star$, Assumption B holds with

$$v_0 \asymp \sqrt{\lambda_1(\mathrm{tr}_{>k}(\Sigma) + k\lambda_{k+1})}, \qquad v_1 \asymp \lambda_1\sqrt{k},$$

and

$$g_0 \asymp \sqrt{\lambda_1\lambda_{k+1}}, \qquad g_1 \asymp \lambda_1,$$

up to logarithmic high-probability truncation factors in $M_\perp$. Consequently, under the absorption condition in Theorem 5,

$$\left\|Q_T Q_T^\top - P_\star\right\|_F \le \mathrm{Opt}_T + \tilde{O}\left(\frac{\sqrt{k\lambda_1(\mathrm{tr}_{>k}(\Sigma) + k\lambda_{k+1})}}{\Delta\sqrt{n}} + \frac{\sqrt{k}M_\perp}{\Delta n} + \frac{K\sqrt{\lambda_1\lambda_{k+1}}\,dk}{\Delta\varepsilon n}\right).$$

Proof. The Gaussian fourth-moment identity gives, for every fixed symmetric matrix B,

$$\mathbb{E}\left[(A-\Sigma)B(A-\Sigma)\right] = \mathrm{tr}(\Sigma B)\Sigma + \Sigma B\Sigma. \tag{8}$$

Fix $Q \in \mathbb{S}_{d,k}$, set $P = QQ^\top$, $\Pi = I - P$, and define $e_2 = \|P - P_\star\|_2$. For a column q of Q, (8) with $B = qq^\top$ gives

$$\Pi\mathbb{E}\left[(A-\Sigma)qq^\top(A-\Sigma)\right]\Pi = (q^\top\Sigma q)\Pi\Sigma\Pi + (\Pi\Sigma q)(\Pi\Sigma q)^\top.$$

If $z \in \mathrm{range}(\Pi)$ is unit norm, then $Pz = 0$, so $P_\star z = (P_\star - P)z$ and $\|P_\star z\| \le e_2$. Hence

$$z^\top\Sigma z \le \lambda_{k+1} + \lambda_1 e_2^2, \qquad \|\Pi\Sigma\Pi\|_2 \le \lambda_{k+1} + \lambda_1 e_2^2.$$

Moreover,

$$\|\Pi\Sigma q\|_2 \le \|\Pi\Sigma(P - P_\star)\|_2 + \|\Pi P_\star\Sigma P_\star\|_2 \le 2\lambda_1 e_2.$$

Since $q^\top\Sigma q \le \lambda_1$,

$$g_Q^2 \le C(\lambda_1\lambda_{k+1} + \lambda_1^2 e_2^2),$$

and therefore

$$g_Q \le C(\sqrt{\lambda_1\lambda_{k+1}} + \lambda_1 e(Q)).$$

This gives the stated choices of g_0, g_1.

For v_Q, write $Y = \Pi(A-\Sigma)Q$. Since $P = QQ^\top$,

$$YY^\top = \Pi(A-\Sigma)P(A-\Sigma)\Pi.$$

Applying (8) with $B = P$,

$$\mathbb{E}\left[YY^\top\right] = \Pi\{\mathrm{tr}(\Sigma P)\Sigma + \Sigma P\Sigma\}\Pi.$$

Using $\mathrm{tr}(\Sigma P) \le k\lambda_1$, $\|\Pi\Sigma\Pi\|_2 \le \lambda_{k+1} + \lambda_1 e_2^2$, and $\|\Pi\Sigma P\|_2 \le 2\lambda_1 e_2$, we obtain

$$\left\|\mathbb{E}\left[YY^\top\right]\right\|_2 \le Ck\lambda_1(\lambda_{k+1} + \lambda_1 e_2^2).$$

Similarly,

$$Y^\top Y = Q^\top(A-\Sigma)\Pi(A-\Sigma)Q.$$

Using (8) with $B = \Pi$,

$$\mathbb{E}\left[Y^\top Y\right] = Q^\top\{\mathrm{tr}(\Sigma\Pi)\Sigma + \Sigma\Pi\Sigma\}Q.$$

Now

$$\mathrm{tr}(\Sigma\Pi) \le \mathrm{tr}_{>k}(\Sigma) + k\lambda_1 e_2^2,$$

and $\|\Pi\Sigma Q\|_2^2 \le 4\lambda_1^2 e_2^2$. Thus

$$\left\|\mathbb{E}\left[Y^\top Y\right]\right\|_2 \le C\left(\lambda_1\,\mathrm{tr}_{>k}(\Sigma) + k\lambda_1^2 e_2^2\right).$$

Combining the two variance bounds yields

$$v_Q^2 \le C\left(\lambda_1[\mathrm{tr}_{>k}(\Sigma) + k\lambda_{k+1}] + k\lambda_1^2 e_2^2\right),$$

so

$$v_Q \le C\left(\sqrt{\lambda_1[\mathrm{tr}_{>k}(\Sigma) + k\lambda_{k+1}]} + \lambda_1\sqrt{k}\,e(Q)\right).$$

This gives the stated v_0, v_1. □

Proof of Corollary 3. For flat-tail covariance, $\mathrm{tr}_{>k}(\Sigma) = (d-k)\lambda_{k+1}$. Hence Lemma 18 gives

$$v_0 \asymp \sqrt{d\lambda_1\lambda_{k+1}}, \qquad g_0 \asymp \sqrt{\lambda_1\lambda_{k+1}}.$$

Substituting these values into Theorem 5 yields

$$\left\|Q_T Q_T^\top - P_\star\right\|_F \le \mathrm{Opt}_T + \widetilde{O}\left(\frac{\sqrt{\lambda_1\lambda_{k+1}}}{\Delta}\left[\sqrt{\frac{dk}{n}} + \frac{Kdk}{\varepsilon n}\right] + \frac{\sqrt{k}M_\perp}{\Delta n}\right).$$

Since $\Delta = \lambda_k - \lambda_{k+1}$, this gives the inequality. □

D Results from Prior work

D.1 Existing Lower Bounds

Theorem 6 (Lower bound, Gaussian distribution, Theorem 5.3 in Liu et al. [2022])**.** *Let $\mathcal{M}_\varepsilon$ be a class of $(\varepsilon, 0)$-DP estimators that map n i.i.d. samples to an estimate $\hat{v} \in \mathbb{R}^d$. A set of Gaussian distributions with (λ_1, λ_2) as the first and second eigenvalues of the covariance matrix is denoted by $\mathcal{P}_{(\lambda_1,\lambda_2)}$. There exists a universal constant $C > 0$ such that*

$$\inf_{\hat{v}\in\mathcal{M}_\varepsilon} \sup_{P\in\mathcal{P}_{(\lambda_1,\lambda_2)}} \mathbb{E}_{S\sim P^n}[\sin(\hat{v}(S), v_1)] \ge C \min\left(\kappa\left(\sqrt{\frac{d}{n}} + \frac{d}{\varepsilon n}\right)\sqrt{\frac{\lambda_2}{\lambda_1}}, 1\right)$$

where v_1 is the true top eigenvector of the expectation of the i.i.d. samples and $\kappa = \frac{\lambda_1}{\lambda_1-\lambda_2}$.

Theorem 7 (Theorem 4.2 in Cai et al. [2024])**.** *Let the $d\times n$ data matrix X have i.i.d. columns sampled from a distribution $P = \mathcal{N}(0, U\Lambda U^\top + \sigma^2\mathbf{I}_d) \in \mathcal{P}(\lambda, \sigma^2)$, where $U \in \mathbb{R}^{d\times k}$ has orthonormal columns. Suppose $\delta \le c_0' \exp\{2\varepsilon - c_0(\varepsilon\sqrt{ndk} + dk)\}$ for some small constants $c_0, c_0' > 0$. Then, there exists an absolute constant $c_1 > 0$ such that*

$$\inf_{\tilde{U}\in\mathcal{U}_{\varepsilon,\delta}} \sup_{P\in\mathcal{P}(\lambda,\sigma^2)} \frac{\mathbb{E}\|\tilde{U}\tilde{U}^\top - UU^\top\|_F}{\sqrt{k}} \ge c_1\left(\left(\frac{\sigma\sqrt{\lambda+\sigma^2}}{\lambda}\right)\left(\sqrt{\frac{d}{n}} + \frac{d\sqrt{k}}{n\varepsilon}\right)\bigwedge 1\right)$$

where the infimum is taken over all the possible (ε, δ)-DP algorithms, denoted by $\mathcal{U}_{\varepsilon,\delta}$ and the expectation is taken with respect to both $\tilde{U}$ and P and

$$\mathcal{P}(\lambda, \sigma^2) := \{\mathcal{N}(0, \Sigma) : \Sigma = U\Lambda U^\top + \sigma^2\mathbf{I}_d, U \in \mathbb{O}_{d,k}, \Lambda = \mathit{diag}(\lambda_1, \ldots, \lambda_k), c_0\lambda \le \lambda_k \le \lambda_1 \le C_0\lambda\}$$

Corollary 5 (Corollary 3 in Düngler and Sanyal [2025])**.** *Let the $d \times n$ data matrix X have i.i.d. columns sampled from a distribution $P = \mathcal{N}(0, U\Lambda U^\top + \sigma^2\mathbf{I}_d) \in \mathcal{P}(\lambda, \sigma^2)$ where $\mathcal{P}(\lambda, \sigma^2) = \{\mathcal{N}(0, \Sigma), \Sigma = U\Lambda U^\top + \sigma^2\mathbf{I}_d, c\lambda \le \lambda_k \le \cdots \le \lambda_1 \le C\lambda\}$. Suppose $\delta \le c_0' \exp\{2\varepsilon - c_0(\varepsilon\sqrt{ndk} + dk)\}$ for some small constants $c_0, c_0' > 0$. Then, there exists an absolute constant $c_1 > 0$ such that*

$$\inf_{\tilde{U}\in\mathcal{U}_{\varepsilon,\delta}} \sup_{P\in\mathcal{P}(\lambda,\sigma^2)} \mathbb{E}[\zeta] \ge c_1\left(\left(\frac{\sigma\sqrt{\lambda_1+\sigma^2}}{\sum_{i=1}^k(\lambda_i+\sigma^2)}\right)\left(\sqrt{\frac{dk}{n}} + \frac{dk}{n\varepsilon}\right)\bigwedge 1\right).$$

D.2 Existing Algorithms

Algorithm 6 Oja's Algorithm, Huang et al. [2021]

Input: symmetric matrices $S = \{A_1, \ldots, A_n\}$ in $\mathbb{R}^{d\times d}$, $k \in [d]$, learning rates $\{\eta_t\}_{t=1}^n$

1: Choose $Q_0' \in \mathbb{R}^{d\times k}$ uniformly at random, $Q_0 \leftarrow \mathrm{QR}[Q_0']$
2: **for** $t = 1, 2, \ldots, T = n$ **do**
3: $\quad Q_t' \leftarrow Q_{t-1} + \eta_t A_t Q_{t-1}$
4: $\quad Q_t \leftarrow \mathrm{QR}[Q_t']$
5: **end for**
6: **return** Q_T

Theorem 8 (Main Theorem, Huang et al. [2021]). *Given matrices $A_1, \dots, A_n \in \mathbb{R}^{d\times d}$ that are i.i.d., symmetric random matrices, satisfying $\mathbb{E}[A_i] = \Sigma$ for all i with $\rho_k = \lambda_k - \lambda_{k+1} > 0$ the kth eigengap of Σ and the uniform bound*

$$\sup_i \sup_{P\in\mathcal{S}_{d,k}} \|P^\top (A_i - \Sigma)\|_F \le M \text{ a.s.}$$

For every $\delta \in (0,1)$, define learning rates

$$n_0 = \tilde{\Theta}\left(\frac{kM^2}{\delta^2\rho_k^2}\right), \qquad \beta = \tilde{\Theta}\left(\frac{M^2}{\rho_k^2}\right),$$

and

$$\eta_t = \begin{cases} \tilde{\Theta}\left(\dfrac{1}{\rho_k n_0}\right), & t \le n_0, \\ \Theta\left(\dfrac{1}{\rho_k(\beta + t - n_0)}\right), & t > n_0. \end{cases}$$

Let $V_k \in \mathbb{R}^{d\times k}$ be the orthogonal matrix whose columns are the k leading eigenvectors of Σ. Then for any $n > n_0$, the output Q_n of Oja's algorithm satisfies

$$\left\|Q_n Q_n^\top - V_k V_k^\top\right\|_2 \;\le\; C' \,\frac{M}{\rho_k}\sqrt{\frac{\log(Mk/(\rho_k\delta))}{n - n_0}}$$

with probability at least $1-\delta$, where C' is a universal positive constant.

Algorithm 7 Private Top Eigenvalue Estimation, Algorithm 4 in [Liu et al., 2022]

Input: $S = \{g_i\}_{i=1}^B$, (ε,δ)-DP, failure probability ζ

1: Let $\tilde{g}_i \leftarrow g_{2i} - g_{2i-1}$ for $i \in 1, 2, \dots, \lfloor B/2 \rfloor$. Let $\tilde{S} = \{\tilde{g}_i\}_{i=1}^{\lfloor B/2 \rfloor}$
2: Partition $\tilde{S}$ into $k = C_1 \log(1/(\delta\zeta))/\varepsilon$ subsets and denote each dataset as $G_j \in \mathbb{R}^{d\times b}$, where each dataset is of size $b = \lfloor B/2k \rfloor$
3: Let $\lambda_1^{(j)}$ be the top eigenvalue of $(1/b)G_j G_j^\top$ for $\forall j \in [k]$
4: Partition $[0,\infty)$ into $\Omega \leftarrow \{\dots, [2^{-2/4}, 2^{-1/4}), [2^{-1/4}, 1), [1, 2^{1/4}), [2^{1/4}, 2^{2/4}), \dots\} \cup \{[0,0]\}$
5: Run (ε,δ)-DP histogram learner of Lemma 11 on $\{\lambda_1^{(j)}\}_{j=1}^k$ over Ω
6: **if** all the bins are empty **then**
7: **return** $\perp$
8: **end if**
9: Let $[l, r]$ be a non-empty bin that contains the maximum number of points in the DP histogram
10: **return** $\hat{\Lambda} = l$

E A Frobenius-norm analysis of Oja's algorithm

Theorem 8 is stated in the operator norm, while Theorems 1, 2 and 5 and the lower bound of Theorem 7 are stated in the Frobenius norm. Using Theorem 8 as a black box would cost a conversion factor: by Lemma 19,

$$\|Q_n Q_n^\top - V_k V_k^\top\|_F = \sqrt{2}\,\|\sin\Theta\|_F \le \sqrt{2k}\,\|Q_n Q_n^\top - V_k V_k^\top\|_2,$$

with equality when the k principal angles coincide. The Frobenius bound obtained this way is loose by a factor $\sqrt{k}$ compared to Theorem 7. This appendix shows that the factor is an artifact of the conversion: under the same assumptions as Huang et al. [2021], carrying both phases of their analysis out directly in the Frobenius norm gives the bound without it. The conclusion is Theorem 13.

The obstruction is the norm in which the recursion runs. Huang et al. [2021] work in $\|X\|_{p,p} = (\mathbb{E}\|X\|_{S_p}^p)^{1/p}$, where the single index p is at once the Schatten exponent inside the expectation and the moment exponent outside it. Exponential tails force $p \asymp \log(1/\delta)$, and at that exponent $\|W\|_F \le k^{1/2-1/p}\|W\|_{S_p}$ loses $\sqrt{k}$; the inequality is an identity at $p = 2$, but p cannot be small

and large at once. Definition 5 pins the inner norm at Frobenius and lets the moment index run independently, which is what makes a Frobenius conclusion with exponential tails available, and since $\|\cdot\|_F$ is a Hilbert norm the martingale step costs only $\sqrt{p-1}$, with no dimensional factor. The analysis below follows Section 4 and Appendices A-D of Huang et al. [2021] closely.

Phase II requires nothing beyond Assumption C. Phase I requires in addition that the updates be i.i.d., which the algorithms arrange by running their first T_0 steps non adaptively. Section F lifts the finite support results below to that setting.

E.1 Setting and notation

Assumption C ((Σ, M)-predictable Oja model)**.** *Let $Z_0 \in \mathbb{R}^{d\times k}$ be the initialisation matrix and let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration with*

$$\sigma(Z_0) \subseteq \mathcal{F}_0.$$

The matrices $C_1, \dots, C_T \in \mathbb{R}^{d\times d}$ are symmetric and satisfy:

- *(C.1) C_t is $\mathcal{F}_t$-measurable and $\mathbb{E}[C_t \mid \mathcal{F}_{t-1}] = \Sigma$ for every $t \leq T$. The matrix $\Sigma \succeq 0$ has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ with corresponding eigenvectors $v_1, \dots, v_d$, and eigengap $\rho_k = \lambda_k - \lambda_{k+1} > 0$.*
- *(C.2) The centred updates are uniformly bounded in the Ky Fan 2-k norm: $\|C_t - \Sigma\|_{(2,k)} \leq M$ almost surely, for every $t \leq T$.*

The requirement $\sigma(Z_0) \subseteq \mathcal{F}_0$ is not an independence hypothesis and must not be replaced by one: for the adaptive algorithms the updates depend on Z_0 from the first step onwards, and (C.1) demands the conditional mean be Σ *given* the initialisation. (C.2) bounds the total fluctuation energy seen through any k-dimensional subspace.

Throughout, $C_1, \dots, C_T$ satisfy Assumption C, and we write $V := [v_1, \dots, v_k] \in \mathbb{R}^{d\times k}$ and $U := [v_{k+1}, \dots, v_d] \in \mathbb{R}^{d\times(d-k)}$ for the leading and trailing eigenvector blocks of Σ, so that $[\,V\ U\,]$ is orthogonal and $V^\top U = 0$. Given the initialisation Z_0 of Assumption C, put $Q_0 := \mathrm{QR}(Z_0)$, and let $\eta_t > 0$ be step sizes. The Oja iterates are

$$Q_t := \mathrm{QR}\big(Q_{t-1} + \eta_t C_t Q_{t-1}\big), \qquad Z_t := (I + \eta_t C_t)\cdots(I + \eta_1 C_1)\, Q_0, \qquad t \geq 1,$$

so that Q_t is an orthonormal basis for the column span of Z_t, and whenever $V^\top Z_t$ is invertible we set

$$W_t := U^\top Z_t (V^\top Z_t)^{-1} = U^\top Q_t (V^\top Q_t)^{-1}.$$

The analysis is carried out in W_t rather than in the subspace error itself, because the two are equivalent up to an absolute constant.

Lemma 19 (Frobenius $\sin\Theta$ bound)**.** *Let $Q \in \mathbb{R}^{d\times k}$ have orthonormal columns and let $\theta_1, \dots, \theta_k \in [0, \pi/2]$ be the principal angles between $\mathrm{range}(Q)$ and $\mathrm{range}(V)$, so that $\cos\theta_1, \dots, \cos\theta_k$ are the singular values of $V^\top Q$, and write $\Theta := \mathrm{diag}(\theta_1, \dots, \theta_k)$. Then*

$$\big\|QQ^\top - VV^\top\big\|_F^2 = 2\|\sin\Theta\|_F^2, \qquad \big\|QQ^\top - VV^\top\big\|_2 = \|\sin\Theta\|_2.$$

If moreover $V^\top Q$ is invertible and $W := U^\top Q (V^\top Q)^{-1}$, then the singular values of W are $\tan\theta_1, \dots, \tan\theta_k$, and consequently

$$\big\|QQ^\top - VV^\top\big\|_F \leq \sqrt{2}\,\|W\|_F.$$

Proof. Both $QQ^\top$ and $VV^\top$ are rank-k orthogonal projectors, so $(QQ^\top)^2 = QQ^\top$, $(VV^\top)^2 = VV^\top$ and $\mathrm{tr}(QQ^\top) = \mathrm{tr}(VV^\top) = k$. Expanding,

$$\big\|QQ^\top - VV^\top\big\|_F^2 = \mathrm{tr}\,\big((QQ^\top - VV^\top)^2\big) = 2k - 2\,\mathrm{tr}(QQ^\top VV^\top) = 2k - 2\|V^\top Q\|_F^2,$$

since $\mathrm{tr}(QQ^\top VV^\top) = \mathrm{tr}(Q^\top VV^\top Q) = \|V^\top Q\|_F^2$. As $\|V^\top Q\|_F^2 = \sum_i \cos^2\theta_i$, the right-hand side is $2\sum_i(1 - \cos^2\theta_i) = 2\|\sin\Theta\|_F^2$. For the second identity, the nonzero eigenvalues of $QQ^\top - VV^\top$ are $\pm\sin\theta_1, \dots, \pm\sin\theta_k$, each appearing once, so the operator norm is $\max_i \sin\theta_i = \|\sin\Theta\|_2$.

For the third claim put $Y_1 := V^\top Q$, $Y_2 := U^\top Q$. Since $[\,V\ U\,]$ is orthogonal, $Y_1^\top Y_1 + Y_2^\top Y_2 = Q^\top Q = I_k$. Let $Y_1 = ACB^\top$ be a singular value decomposition, so $C = \cos\Theta$ after ordering; then $Y_2^\top Y_2 = I_k - BC^2B^\top = B\sin^2\Theta\, B^\top$. As $V^\top Q$ is invertible, every $\cos\theta_i > 0$, so C is invertible, $W = Y_2 Y_1^{-1} = Y_2 BC^{-1}A^\top$ and

$$W^\top W = AC^{-1}B^\top Y_2^\top Y_2 BC^{-1}A^\top = AC^{-1}\sin^2\Theta\, C^{-1}A^\top = A\tan^2\Theta\, A^\top.$$

Hence the singular values of W are the $\tan\theta_i$, and since $\sin\theta \le \tan\theta$ on $[0,\pi/2)$, $\|W\|_F^2 = \sum_i \tan^2\theta_i \ge \sum_i \sin^2\theta_i = \frac{1}{2}\|QQ^\top - VV^\top\|_F^2$. □

Applied at $Q = Q_t$, Lemma 19 gives $\|Q_tQ_t^\top - VV^\top\|_F \le \sqrt{2}\,\|W_t\|_F$ for every t and for our analysis is therefore suffices to bound $\|W\|_F$. We write $\Lambda_1 := \operatorname{diag}(\lambda_1,\dots,\lambda_k)$ and $\Lambda_2 := \operatorname{diag}(\lambda_{k+1},\dots,\lambda_d)$, so $\Sigma = V\Lambda_1V^\top + U\Lambda_2U^\top$, and, for a threshold $\gamma \ge 1$ to be fixed in each phase,

$$\varepsilon_t := 2\eta_t M(1+\gamma), \qquad \chi_t := \frac{1+\eta_t\lambda_{k+1}}{1+\eta_t\lambda_k} \in (0,1].$$

When the step size is constant we drop the subscripts and write ε, χ, η. Each phase conditions on a sequence of decreasing *good events* $G_0 \supseteq G_1 \supseteq \cdots$, adapted to the filtration $(\mathcal{F}_t)$ of Assumption C, with indicators $\mathbf{1}_t := \mathbf{1}_{G_t}$; the two phases use different events, specified in Sections E.5 and E.6. We abbreviate the small-step conditions

$$\varepsilon_i \le \tfrac{1}{2}, \tag{S1}$$

$$\eta_i\|\Sigma\|_2 \le \tfrac{1}{2}, \tag{S2}$$

$$e^{-\eta_i\rho_k/4} \le \varepsilon_i/\varepsilon_{i-1} \quad \text{(vacuous at } i=1 \text{ and for constant step size)}, \tag{S3}$$

$$p\varepsilon_i^2 \le \eta_i\rho_k/400. \tag{S4}$$

$C, C', C_0, C_1, \dots$ denote absolute positive constants. On the event where $V^\top Z_t$ fails to be invertible, W_t may be defined arbitrarily, as every quantity involving W_t in the analysis is multiplied by $\mathbf{1}_t$ or $\mathbf{1}_{t-1}$, and by Lemma 25 $V^\top Z_t$ is invertible on theses events.

E.2 Preliminaries

Definition 5 (Mixed norm)**.** For a random matrix X and $p \ge 2$ set $\|X\|_{F,p} := \big(\mathbb{E}\|X\|_F^p\big)^{1/p}$.

Both the Frobenius norm and the p-th power sit inside the expectation; $\|X\|_{F,p}$ is the $L^p(\Omega)$ norm of the random variable $\|X\|_F$. In terms of singular values we have,

$$\|X\|_{F,p}^p = \mathbb{E}\Big(\sum_i \sigma_i(X)^2\Big)^{p/2},$$

as opposed to

$$\|X\|_{p,p}^p = \mathbb{E}\sum_i \sigma_i(X)^p,$$

which is the norm used by Huang et al. [2021]. In $\|\cdot\|_{p,p}$ the index p serves both as the Schatten exponent inside and as the moment exponent outside, while in $\|\cdot\|_{F,p}$ the inner norm is pinned at Frobenius and only the outer moment index varies. The two agree at $p = 2$. Decoupling the two indices is what allows a Frobenius-norm conclusion (inner norm fixed) together with exponential tails (outer index large).

Lemma 20 (Properties of $\|\cdot\|_{F,p}$)**.** *Let $p \ge 2$ and let X, Y, Z be random matrices of the same size.*

(i) $\|\cdot\|_{F,p}$ *is a norm.*

(ii) If A, B are random matrices with $\|A\|_2 \le a$ and $\|B\|_2 \le b$ almost surely, then $\|AXB\|_{F,p} \le ab\,\|X\|_{F,p}$.

(iii) $\|X\mathbf{1}_{G'}\|_{F,p} \le \|X\mathbf{1}_G\|_{F,p} \le \|X\|_{F,p}$ *whenever* $G' \subseteq G$.

(iv) $\|X\|_{p,p} \le \|X\|_{F,p}$.

(v) $\mathbb{P}(\|X\|_F > s) \le s^{-p}\|X\|_{F,p}^p$ *for every* $s > 0$.

(vi) (One-step smoothness.) *Let* $\mathcal{F}$ *be a* σ*-algebra, let* X *be* $\mathcal{F}$*-measurable and let* $\mathbb{E}[Y \mid \mathcal{F}] = 0$. *Then* $\|X+Y\|_{F,p}^2 \le \|X\|_{F,p}^2 + 2\cdot(p-1)\|Y\|_{F,p}^2$.

(vii) Under the hypotheses of (vi), for every Z *and every* $\lambda > 0$*:*

$$\|X+Y+Z\|_{F,p}^2 \le (1+\lambda)(\|X\|_{F,p}^2 + 2(p-1)\|Y\|_{F,p}^2 + \lambda^{-1}\|Z\|_{F,p}^2)$$

(viii) (Random class index.) *Let* $\mathcal{D}$ *be a finite set of deterministic matrices with* $|\mathcal{D}| = m$*, let* D *be a random matrix taking values in* $\mathcal{D}$*, and let* Y *be any random matrix of conforming size. Then* $\|DY\|_{F,p} \le m^{1/p}\max_{D'\in\mathcal{D}}\|D'Y\|_{F,p}$*; in particular* $\|DY\|_{F,p} \le e\max_{D'\in\mathcal{D}}\|D'Y\|_{F,p}$ *whenever* $p \ge \log m$.

Proof. (i) Note $\|X+Y\|_F \le \|X\|_F + \|Y\|_F$, so by monotonicity of $t \mapsto t^p$ and of the expectation, by Minkowski's inequality in $L^p(\Omega)$,

$$\|X+Y\|_{F,p} = \big(\mathbb{E}\|X+Y\|_F^p\big)^{1/p} \le \big(\mathbb{E}(\|X\|_F + \|Y\|_F)^p\big)^{1/p} \le \big(\mathbb{E}\|X\|_F^p\big)^{1/p} + \big(\mathbb{E}\|Y\|_F^p\big)^{1/p}.$$

Homogeneity and definiteness are immediate.

(ii) holds point-wise from $\|AXB\|_F \le \|A\|_2\|X\|_F\|B\|_2$.
(iii) from $\|X\mathbf{1}_G\|_F = \|X\|_F\mathbf{1}_G$ and since $\mathbf{1}_G^p = \mathbf{1}_G$,

$$\|X\mathbf{1}_G\|_{F,p}^p = \mathbb{E}\big[\|X\|_F^p\,\mathbf{1}_G\big].$$

This is monotone in the event: $G' \subseteq G$ implies $\mathbf{1}_{G'} \le \mathbf{1}_G$ pointwise, hence

$$\mathbb{E}\big[\|X\|_F^p\,\mathbf{1}_{G'}\big] \le \mathbb{E}\big[\|X\|_F^p\,\mathbf{1}_G\big] \le \mathbb{E}\big[\|X\|_F^p\big].$$

(iv) is $\|X\|_{S_p} \le \|X\|_{S_2} = \|X\|_F$ for $p \ge 2$.
(v) is Markov's inequality applied to $\|X\|_F^p$.

(vi) follows from the 2-smoothness of $L^p(\Omega; H)$ for a Hilbert space H; here $H = \mathbb{R}^{(d-k)\times k}$ with the Frobenius inner product. Recall that a Banach space $(\mathcal{X}, \|\cdot\|)$ is $(2, D)$-*smooth* if

$$\|u+v\|^2 + \|u-v\|^2 \le 2\|u\|^2 + 2D^2\|v\|^2 \qquad \text{for all } u, v \in \mathcal{X}.$$

A Hilbert space is $(2,1)$-smooth, with equality: this is the parallelogram law. By Pinelis [1994, Proposition 2.1], L^p is $(2, \sqrt{p-1})$-smooth for $p \ge 2$. That proposition is stated for scalar-valued $L^p(T, \mathcal{A}, \nu)$, but its proof verifies the sufficient condition $(\|x\|^2)''(v,v) \le 2D^2\|v\|^2$ by differentiating $\int_T |x|^p \mathrm{d}\nu$ twice and applying Hölder's inequality, and for H-valued integrands the same computation applies, Cauchy-Schwarz giving $\frac{\mathrm{d}^2}{\mathrm{d}s^2}\|x+sv\|_H^p|_{s=0} \le p(p-1)\|x\|_H^{p-2}\|v\|_H^2$ where the scalar case has equality. Hence $L^p(\Omega; H)$ is $(2, \sqrt{p-1})$-smooth, and taking $u = X$ and $v = Y$,

$$\|X+Y\|_{F,p}^2 \le 2\|X\|_{F,p}^2 + 2(p-1)\|Y\|_{F,p}^2 - \|X-Y\|_{F,p}^2.$$

Since $\mathbb{E}[Y \mid \mathcal{F}] = 0$ and X is $\mathcal{F}$-measurable, $-Y$ is again a martingale increment and $\mathbb{E}[X-Y \mid \mathcal{F}] = X$, so conditional Jensen applied to the convex map $u \mapsto \|u\|_F^p$ gives

$$\|X-Y\|_{F,p}^p = \mathbb{E}\|X-Y\|_F^p \ge \mathbb{E}\big\|\mathbb{E}[X-Y \mid \mathcal{F}]\big\|_F^p = \mathbb{E}\|X\|_F^p = \|X\|_{F,p}^p,$$

whence $\|X-Y\|_{F,p} \ge \|X\|_{F,p}$. Substituting gives (vi). At $p = 2$ the mechanism is visible directly:

$$\mathbb{E}\|X+Y\|_F^2 = \mathbb{E}\|X\|_F^2 + 2\,\mathbb{E}\langle X, Y\rangle_F + \mathbb{E}\|Y\|_F^2, \qquad \mathbb{E}\langle X, Y\rangle_F = \mathbb{E}\big\langle X, \mathbb{E}[Y \mid \mathcal{F}]\big\rangle_F = 0,$$

the last step because X is $\mathcal{F}$-measurable.

(vii): Using

$$(a+b)^2 \le (1+\lambda)\left(a^2 + \lambda^{-1}b^2\right),$$

we obtain

$$\begin{aligned}
\|X+Y+Z\|_{F,p}^2 &\le \Big(\|X+Y\|_{F,p} + \|Z\|_{F,p}\Big)^2 \\
&\le (1+\lambda)\left(\|X+Y\|_{F,p}^2 + \lambda^{-1}\|Z\|_{F,p}^2\right) \\
&\le (1+\lambda)\left(\|X\|_{F,p}^2 + 2(p-1)\|Y\|_{F,p}^2 + \lambda^{-1}\|Z\|_{F,p}^2\right),
\end{aligned}$$

where the last inequality follows from (vi).

(viii): pointwise $\|DY\|_F^p \le \sum_{D' \in \mathcal{D}} \|D'Y\|_F^p$, since the left-hand side is one of the summands and all are non-negative; taking expectations and the p-th root gives $\|DY\|_{F,p} \le m^{1/p} \max_{D'} \|D'Y\|_{F,p}$, and $m^{1/p} = e^{(\log m)/p} \le e$ when $p \ge \log m$. No measurability relation between D and Y is required. The factor $m^{1/p}$ cannot be removed: if Y takes two values y_1, y_2 with probability $\frac{1}{2}$ each and $\mathcal{D} = \{D_1, D_2\}$ with $\|D_1 y_1\|_F = \|D_2 y_2\|_F = 1$ and $\|D_1 y_2\|_F = \|D_2 y_1\|_F = 0$, then choosing $D = D_i$ on $\{Y = y_i\}$ gives $\|DY\|_{F,p} = 1$ while $\max_i \|D_i Y\|_{F,p} = 2^{-1/p}$. □

E.3 The one-step decomposition

Lemma 21 (Frobenius bound on the U-block)**.** *Under Assumption C.2, almost surely* $\|U^\top (C_t - \Sigma) V\|_F \le M$ *and* $\|C_t - \Sigma\|_2 \le M$.

Proof. Since U has orthonormal columns, $\|U^\top (C_t - \Sigma) V\|_F \le \|(C_t - \Sigma) V\|_F = \|V^\top (C_t - \Sigma)\|_F$. As $V \in S_{d,k}$, the right-hand side is one of the quantities over which the supremum defining $\|\cdot\|_{(2,k)}$ is taken, so it is at most $\|C_t - \Sigma\|_{(2,k)} \le M$. The second claim is $\|X\|_2 \le \|X\|_{(2,k)}$. □

Lemma 22 (One-step decomposition, in factored form)**.** *Fix $t \ge 1$ and suppose $V^\top Z_{t-1}$ and $I - \Delta_t^2$ are invertible, where*

$$H_t := U^\top (I + \eta_t \Sigma) Z_{t-1} \big(V^\top (I + \eta_t \Sigma) Z_{t-1}\big)^{-1}, \quad \begin{aligned} \Delta_t &:= \eta_t V^\top (C_t - \Sigma) Z_{t-1} \big(V^\top (I + \eta_t \Sigma) Z_{t-1}\big)^{-1}, \\ \widehat{\Delta}_t &:= \eta_t U^\top (C_t - \Sigma) Z_{t-1} \big(V^\top (I + \eta_t \Sigma) Z_{t-1}\big)^{-1}. \end{aligned}$$

Then $V^\top Z_t = (I + \Delta_t) V^\top (I + \eta_t \Sigma) Z_{t-1}$,

$$W_t (I - \Delta_t^2) = H_t + J_{t,1} + J_{t,2}, \qquad J_{t,1} := \widehat{\Delta}_t - H_t \Delta_t, \quad J_{t,2} := -\widehat{\Delta}_t \Delta_t,$$

and, writing $Z_{t-1}(V^\top Z_{t-1})^{-1} = V + U W_{t-1}$,

$$H_t = (I + \eta_t \Lambda_2) W_{t-1} (I + \eta_t \Lambda_1)^{-1}, \quad \begin{aligned} \Delta_t &= \eta_t V^\top (C_t - \Sigma)(V + U W_{t-1})(I + \eta_t \Lambda_1)^{-1}, \\ \widehat{\Delta}_t &= \eta_t U^\top (C_t - \Sigma)(V + U W_{t-1})(I + \eta_t \Lambda_1)^{-1}. \end{aligned}$$

Proof. The identity $W_t(I - \Delta_t^2) = H_t + J_{t,1} + J_{t,2}$ is Lemma 2.7 of Huang et al. [2021], proved in their Appendix G.1, where $J_{t,1} = \widehat{\Delta}_t - H_t \Delta_t$ and $J_{t,2} = -\widehat{\Delta}_t \Delta_t$ are defined as above; their statement records only that these two matrices have norms $O(\eta_t)$ and $O(\eta_t^2)$. The argument is short, so we give it in full. The closed forms for H_t, Δ_t and $\widehat{\Delta}_t$ are not stated there.

$V^\top Z_t = V^\top (I + \eta_t C_t) Z_{t-1} = V^\top (I + \eta_t \Sigma) Z_{t-1} + \eta_t V^\top (C_t - \Sigma) Z_{t-1} = (I + \Delta_t) V^\top (I + \eta_t \Sigma) Z_{t-1}$, by the definition of Δ_t. Hence $(V^\top Z_t)^{-1} = (V^\top (I + \eta_t \Sigma) Z_{t-1})^{-1} (I + \Delta_t)^{-1}$, and since $U^\top Z_t = U^\top (I + \eta_t \Sigma) Z_{t-1} + \eta_t U^\top (C_t - \Sigma) Z_{t-1}$,

$$W_t (I + \Delta_t) = U^\top Z_t \big(V^\top (I + \eta_t \Sigma) Z_{t-1}\big)^{-1} = H_t + \widehat{\Delta}_t.$$

Multiplying on the right by $I - \Delta_t$ gives the decomposition.

For the closed forms, $U^\top (I + \eta_t \Sigma) = (I + \eta_t \Lambda_2) U^\top$ and $V^\top (I + \eta_t \Sigma) = (I + \eta_t \Lambda_1) V^\top$, so $V^\top (I + \eta_t \Sigma) Z_{t-1} = (I + \eta_t \Lambda_1) V^\top Z_{t-1}$; multiplying each of $U^\top (I + \eta_t \Sigma) Z_{t-1}$, $\eta_t V^\top (C_t - \Sigma) Z_{t-1}$ and $\eta_t U^\top (C_t - \Sigma) Z_{t-1}$ on the right by $(V^\top Z_{t-1})^{-1} (I + \eta_t \Lambda_1)^{-1}$ and using $Z_{t-1}(V^\top Z_{t-1})^{-1} = V + U W_{t-1}$ gives the three expressions. □

Lemma 23 (Contraction)**.** *Whenever $V^\top Z_{t-1}$ is invertible, $\|H_t \mathbf{1}_{t-1}\|_F \le \chi_t \|W_{t-1} \mathbf{1}_{t-1}\|_F$ pointwise, and hence $\|H_t \mathbf{1}_{t-1}\|_{F,p} \le \chi_t \|W_{t-1} \mathbf{1}_{t-1}\|_{F,p}$.*

Proof. By Lemma 22, $H_t = (I + \eta_t \Lambda_2) W_{t-1} (I + \eta_t \Lambda_1)^{-1}$, so $\|AXB\|_F \le \|A\|_2 \|X\|_F \|B\|_2$ gives, pointwise,

$$\|H_t \mathbf{1}_{t-1}\|_F \le \|I + \eta_t \Lambda_2\|_2 \, \|(I + \eta_t \Lambda_1)^{-1}\|_2 \, \|W_{t-1} \mathbf{1}_{t-1}\|_F.$$

Both Λ_1 and Λ_2 are diagonal with non-negative entries, since $\Sigma \succeq 0$ by Assumption C.1, so $\|I + \eta_t \Lambda_2\|_2 = 1 + \eta_t \lambda_{k+1}$ and $\|(I + \eta_t \Lambda_1)^{-1}\|_2 = (1 + \eta_t \lambda_k)^{-1}$, the largest entry of Λ_2 being λ_{k+1} and the smallest of Λ_1 being λ_k. Their product is χ_t, and taking L^p norms gives the second claim. □

Lemma 24 (Step bounds)**.** *Under the hypotheses of Lemma 25, put* $\mathcal{E}_t := 1 + 2\|W_{t-1}\mathbf{1}_{t-1}\|_{F,p}$. *Then*

$$\|\widehat{\Delta}_t \mathbf{1}_{t-1}\|_{F,p} \le \eta_t M\big(1 + \|W_{t-1}\mathbf{1}_{t-1}\|_{F,p}\big), \tag{9}$$

$$\|J_{t,1}\mathbf{1}_{t-1}\|_{F,p} \le \eta_t M + \varepsilon_t \|W_{t-1}\mathbf{1}_{t-1}\|_{F,p} \;\le\; \varepsilon_t \mathcal{E}_t, \tag{10}$$

$$\|J_{t,2}\mathbf{1}_{t-1}\|_{F,p} \le \varepsilon_t^2 \mathcal{E}_t. \tag{11}$$

Proof. All three bounds rest on one substitution. By Lemma 25, $V^\top Z_{t-1}$ is invertible on G_{t-1}; since $UU^\top + VV^\top = I$ and $U^\top Z_{t-1}(V^\top Z_{t-1})^{-1} = W_{t-1}$,

$$Z_{t-1} = \big(VV^\top + UU^\top\big)Z_{t-1} = (V + UW_{t-1})\,V^\top Z_{t-1},$$

while $V^\top(I + \eta_t\Sigma)U = 0$ and $V^\top(I + \eta_t\Sigma)V = I + \eta_t\Lambda_1$ give $V^\top(I + \eta_t\Sigma)Z_{t-1} = (I + \eta_t\Lambda_1)V^\top Z_{t-1}$. The right factor common to Δ_t and $\widehat{\Delta}_t$ therefore collapses:

$$Z_{t-1}\big(V^\top(I+\eta_t\Sigma)Z_{t-1}\big)^{-1} = (V+UW_{t-1})\,V^\top Z_{t-1}\,(V^\top Z_{t-1})^{-1}(I+\eta_t\Lambda_1)^{-1} = (V+UW_{t-1})(I+\eta_t\Lambda_1)^{-1},$$

which is the closed form of Lemma 22. Since $\Lambda_1 \succeq 0$ by Assumption C.1, $\|(I + \eta_t\Lambda_1)^{-1}\|_2 \le 1$.

(9): splitting $V + UW_{t-1}$ and discarding the right factor gives, pointwise,

$$\begin{aligned}\|\widehat{\Delta}_t\mathbf{1}_{t-1}\|_F &\le \eta_t\big\|U^\top(C_t - \Sigma)V\big\|_F + \eta_t\big\|U^\top(C_t - \Sigma)U\big\|_2\,\big\|W_{t-1}\mathbf{1}_{t-1}\big\|_F \\ &\le \eta_t M\big(1 + \|W_{t-1}\mathbf{1}_{t-1}\|_F\big),\end{aligned}$$

both terms by Lemma 21. Taking L^p norms gives (9); in particular $\|U^\top(C_t - \Sigma)V\|_{F,p} \le M$, since the bound holds almost surely. *This is the only step at which the factor $k^{1/p}$ of Huang et al. arises*: their Proposition A.1 bounds the first term by $k^{1/p}\|C_t - \Sigma\|_2$ rather than by Lemma 21. That is a sharpening of Assumption C.2 and not a consequence of the choice of norm; no product class stands in front of the block here, which is why the point arises at all.

(10): since $J_{t,1} = \widehat{\Delta}_t - H_t\Delta_t$, pointwise

$$\begin{aligned}\|J_{t,1}\mathbf{1}_{t-1}\|_F &\le \|\widehat{\Delta}_t\mathbf{1}_{t-1}\|_F + \|H_t\mathbf{1}_{t-1}\|_F\,\|\Delta_t\mathbf{1}_{t-1}\|_2 \\ &\le \eta_t M\big(1 + \|W_{t-1}\mathbf{1}_{t-1}\|_F\big) + \tfrac{1}{2}\varepsilon_t\|W_{t-1}\mathbf{1}_{t-1}\|_F,\end{aligned}$$

by Lemma 23 with $\chi_t \le 1$ and (12). Taking L^p norms and using $\eta_t M \le \frac{1}{2}\varepsilon_t$,

$$\|J_{t,1}\mathbf{1}_{t-1}\|_{F,p} \le \eta_t M + \varepsilon_t\|W_{t-1}\mathbf{1}_{t-1}\|_{F,p} \le \varepsilon_t\mathcal{E}_t.$$

(11): since $J_{t,2} = -\widehat{\Delta}_t\Delta_t$, pointwise

$$\|J_{t,2}\mathbf{1}_{t-1}\|_F \le \|\widehat{\Delta}_t\mathbf{1}_{t-1}\|_F\,\|\Delta_t\mathbf{1}_{t-1}\|_2 \le \tfrac{1}{2}\varepsilon_t\big(1 + \|W_{t-1}\mathbf{1}_{t-1}\|_F\big)\cdot\tfrac{1}{2}\varepsilon_t,$$

again by (12), and taking L^p norms gives $\|J_{t,2}\mathbf{1}_{t-1}\|_{F,p} \le \frac{1}{4}\varepsilon_t^2\mathcal{E}_t \le \varepsilon_t^2\mathcal{E}_t$.

No upper bound on ε_t is used in (9)–(11): they hold at every step size. The condition (S1) enters only through Lemma 25, which supplies the invertibility and (12). □

Lemma 25 (Well-definedness and the bound on Δ_t)**.** *Assume* (S1) *and* $\|V^\top(C_t - \Sigma)UW_{t-1}\mathbf{1}_{t-1}\|_2 \le M\gamma$ *almost surely, for every* $t \le T$. *Then for every* $t \le T$, *on* G_{t-1} *the matrix* $V^\top Z_{t-1}$ *is invertible, so that* H_t, Δ_t *and* $\widehat{\Delta}_t$ *are defined and have the closed forms of Lemma 22;*

$$\|\Delta_t\mathbf{1}_{t-1}\|_2 \le \tfrac{1}{2}\varepsilon_t \le \varepsilon_t \le \tfrac{1}{2} \qquad \textit{almost surely}; \tag{12}$$

and $I - \Delta_t^2$ *and* $V^\top Z_t$ *are invertible on* G_{t-1}, *with* $\|(I - \Delta_t^2)^{-1}\mathbf{1}_{t-1}\|_2 \le (1 - \varepsilon_t^2)^{-1}$.

Proof. By induction on t. At $t = 0$ the matrix $V^\top Z_0 = V^\top Q_0$ is invertible almost surely.

Suppose $V^\top Z_{t-1}$ is invertible on G_{t-1}. Since $\Lambda_1 \succeq 0$ by Assumption C.1, $V^\top(I + \eta_t\Sigma)Z_{t-1} = (I + \eta_t\Lambda_1)V^\top Z_{t-1}$ is invertible there as well, so H_t, Δ_t and $\widehat{\Delta}_t$ are defined on G_{t-1} and Lemma 22

gives $\Delta_t = \eta_t V^\top (C_t - \Sigma)(V + UW_{t-1})(I + \eta_t \Lambda_1)^{-1}$. Discarding the right factor, whose operator norm is at most 1, and splitting $V + UW_{t-1}$,

$$\|\Delta_t \mathbf{1}_{t-1}\|_2 \le \eta_t \big\|V^\top (C_t - \Sigma) V\big\|_2 + \eta_t \big\|V^\top (C_t - \Sigma) U W_{t-1} \mathbf{1}_{t-1}\big\|_2 \le \eta_t M + \eta_t M \gamma = \tfrac{1}{2}\varepsilon_t,$$

the first term by Lemma 21 and the second by hypothesis; with (S1) this is (12). In particular $\|\Delta_t \mathbf{1}_{t-1}\|_2 \le \frac{1}{2} < 1$, so Lemma 7 with $A = \Delta_t$, $G = G_{t-1}$ and $\varepsilon = \varepsilon_t$ shows that $I + \Delta_t$ and $I - \Delta_t^2$ are invertible on G_{t-1} and, by (3), that $\|(I - \Delta_t^2)^{-1} \mathbf{1}_{t-1}\|_2 \le (1 - \varepsilon_t^2)^{-1}$. Finally $V^\top Z_t = (I + \Delta_t) V^\top (I + \eta_t \Sigma) Z_{t-1}$ by Lemma 22 is a product of three matrices each invertible on G_{t-1}, and $G_t \subseteq G_{t-1}$, so the induction continues. □

Remark. The base case is supplied by the initialisation: $Q_0 = \mathrm{QR}(Q_0')$ with Q_0' having i.i.d. $\mathcal{N}(0,1)$ entries, so $V^\top Q_0 = V^\top Q_0' R^{-1}$ with $V^\top Q_0'$ a $k \times k$ standard Gaussian matrix, which is invertible almost surely.

E.4 The main recursive bound

The two results below bound $\|W_t \mathbf{1}_t\|_{F,p}$ itself and they are stated for an arbitrary decreasing sequence of good events (G_t) and an arbitrary threshold γ; each phase supplies its own events and its own γ, and with them the hypothesis $\|V^\top (C_t - \Sigma) U W_{t-1} \mathbf{1}_{t-1}\|_2 \le M\gamma$. Phase I applies Theorem 9 at a constant step size in Proposition 4, Phase II at a decaying one in Proposition 5.

Theorem 9 (Unrolled recursion)**.** *Let $t \ge 1$, set $s_t := \sum_{i \le t} \eta_i$, and assume* (S1)–(S3) *and $\|V^\top (C_i - \Sigma) U W_{i-1} \mathbf{1}_{i-1}\|_2 \le M\gamma$ almost surely, for every $i \le t$. Then for every $p \ge 2$,*

$$\|W_t \mathbf{1}_t\|_{F,p}^2 \le \|W_t \mathbf{1}_{t-1}\|_{F,p}^2 \le e^{-s_t \rho_k} \|W_0 \mathbf{1}_0\|_{F,p}^2 + C_1 p \varepsilon_t^2 \sum_{i=0}^{t-1} \|W_i \mathbf{1}_i\|_{F,p}^2 + C_2 p \varepsilon_t^2 t, \tag{13}$$

with $C_1 = 39$ and $C_2 = 10$; and if (S4) *also holds for every $i \le t$, then*

$$\|W_t \mathbf{1}_t\|_{F,p}^2 \le \|W_t \mathbf{1}_{t-1}\|_{F,p}^2 \le e^{-s_t \rho_k / 2} \|W_0 \mathbf{1}_0\|_{F,p}^2 + C_2 p \varepsilon_t^2 t. \tag{14}$$

These are the two conclusions of Theorem 3.1 of Huang et al. with $k^{2/p}$ replaced by 1.

Proof. Bounding $K_{1,i}$. Expanding Proposition 1,

$$K_{1,i} = \chi_i^2 + 16 p \varepsilon_i^2 + 5 \varepsilon_i^2 \chi_i^2 + 80 p \varepsilon_i^4,$$

and $\chi_i \le 1$ together with $\varepsilon_i^2 \le \frac{1}{4}$ from (S1) give $5\varepsilon_i^2 \chi_i^2 \le 5\varepsilon_i^2$ and $80 p \varepsilon_i^4 \le 20 p \varepsilon_i^2$, so that

$$K_{1,i} \le \chi_i^2 + (5 + 36p) \varepsilon_i^2.$$

For χ_i, write $\lambda_{k+1} = \lambda_k - \rho_k$ and use $1 + \eta_i \lambda_k \le 2$ from (S2):

$$\chi_i = \frac{1 + \eta_i \lambda_{k+1}}{1 + \eta_i \lambda_k} = \frac{1 + \eta_i \lambda_k - \eta_i \rho_k}{1 + \eta_i \lambda_k} = 1 - \frac{\eta_i \rho_k}{1 + \eta_i \lambda_k} \le 1 - \tfrac{1}{2} \eta_i \rho_k,$$

whence $\chi_i^2 \le (1 - \frac{1}{2}\eta_i \rho_k)^2 \le e^{-\eta_i \rho_k}$. Finally $5 + 36p \le 39p$ for $p \ge 2$, so $K_{1,i} \le e^{-\eta_i \rho_k} + C_1 p \varepsilon_i^2$. Condition (S2) in fact gives $\eta_i \lambda_k \le \frac{1}{2}$, hence $1 + \eta_i \lambda_k \le \frac{3}{2}$ and the sharper $\chi_i \le 1 - \frac{2}{3}\eta_i \rho_k$, which is what the proof of Theorem 10 uses; the weaker form is kept here because it delivers the conclusion in exactly the form of Theorem 3.1 of Huang et al.

Proof of (13)*, by induction on t.* At $t = 1$, Proposition 1 gives $\|W_1 \mathbf{1}_1\|_{F,p}^2 \le K_{1,1} \|W_0 \mathbf{1}_0\|_{F,p}^2 + K_{2,1} \le e^{-\eta_1 \rho_k} \|W_0 \mathbf{1}_0\|_{F,p}^2 + C_1 p \varepsilon_1^2 \|W_0 \mathbf{1}_0\|_{F,p}^2 + C_2 p \varepsilon_1^2$. Assume (13) at $t - 1$. Then

$$\begin{aligned}
\|W_t \mathbf{1}_t\|_{F,p}^2 &\le e^{-\eta_t \rho_k} \|W_{t-1} \mathbf{1}_{t-1}\|_{F,p}^2 + C_1 p \varepsilon_t^2 \|W_{t-1} \mathbf{1}_{t-1}\|_{F,p}^2 + C_2 p \varepsilon_t^2 \\
&\le e^{-\eta_t \rho_k} \Big(e^{-s_{t-1} \rho_k} \|W_0 \mathbf{1}_0\|_{F,p}^2 + C_1 p \varepsilon_{t-1}^2 \sum_{i=0}^{t-2} \|W_i \mathbf{1}_i\|_{F,p}^2 + C_2 p \varepsilon_{t-1}^2 (t-1) \Big) \\
&\qquad + C_1 p \varepsilon_t^2 \|W_{t-1} \mathbf{1}_{t-1}\|_{F,p}^2 + C_2 p \varepsilon_t^2 \\
&\le e^{-s_t \rho_k} \|W_0 \mathbf{1}_0\|_{F,p}^2 + C_1 p \varepsilon_t^2 \sum_{i=0}^{t-1} \|W_i \mathbf{1}_i\|_{F,p}^2 + C_2 p \varepsilon_t^2 t,
\end{aligned}$$

the last step using $e^{-\eta_t\rho_k}\varepsilon_{t-1}^2 \le \varepsilon_t^2$, which is (S3) squared together with $e^{-\eta_t\rho_k} \le e^{-\eta_t\rho_k/2}$.

Proof of (14). Here $\eta_i\rho_k \le \eta_i\lambda_1 \le \frac{1}{2} \le 1$ by (S2), so $\chi_i^2 \le 1 - \eta_i\rho_k + \frac{1}{4}(\eta_i\rho_k)^2 \le 1 - \frac{3}{4}\eta_i\rho_k$, and using $1 + x \le e^x$ twice,

$$K_{1,i} \le (1 + 5\varepsilon_i^2)\big(1 - \tfrac{3}{4}\eta_i\rho_k + 16p\varepsilon_i^2\big) \le \exp\big(-\tfrac{3}{4}\eta_i\rho_k + (5 + 16p)\varepsilon_i^2\big) \le e^{-\eta_i\rho_k/2},$$

since $5 + 16p \le \frac{37}{2}p$ for $p \ge 2$ and (S4) give $(5 + 16p)\varepsilon_i^2 \le \frac{1}{4}\eta_i\rho_k$. The induction is then immediate, again by $e^{-\eta_t\rho_k/2}\varepsilon_{t-1}^2 \le \varepsilon_t^2$. □

Proposition 1 (One-step recursion without product classes)**.** *Assume* (S1)*,* (S2) *and* $\|V^\top(C_t - \Sigma)UW_{t-1}\mathbf{1}_{t-1}\|_2 \le M\gamma$ *almost surely. Then for* $p \ge 2$*,*

$$\|W_t\mathbf{1}_t\|_{F,p}^2 \le \|W_t\mathbf{1}_{t-1}\|_{F,p}^2 \le K_{1,t}\|W_{t-1}\mathbf{1}_{t-1}\|_{F,p}^2 + K_{2,t},$$
$$K_{1,t} = (1 + 5\varepsilon_t^2)\big(\chi_t^2 + 16p\varepsilon_t^2\big), \qquad K_{2,t} = 10p\varepsilon_t^2.$$

This is Proposition A.2 of Huang et al. with $K_{2,t} = 5pk^{2/p}\varepsilon_t^2$ *replaced by* $10p\varepsilon_t^2$*.*

Proof. The first inequality is Lemma 20(iii), since $G_t \subseteq G_{t-1}$.

By Lemma 22,

$$W_t\mathbf{1}_{t-1}(I - \Delta_t^2) = H_t\mathbf{1}_{t-1} + J_{t,1}\mathbf{1}_{t-1} + J_{t,2}\mathbf{1}_{t-1}, \qquad J_{t,1} = \widehat{\Delta}_t - H_t\Delta_t.$$

Assumption C.1 gives $\mathbb{E}[C_t - \Sigma \mid \mathcal{F}_{t-1}] = 0$, and both Δ_t and $\widehat{\Delta}_t$ are linear in $C_t - \Sigma$ with $\mathcal{F}_{t-1}$-measurable coefficients, so

$$\mathbb{E}\big[\widehat{\Delta}_t \mid \mathcal{F}_{t-1}\big] = \eta_t U^\top \mathbb{E}\big[C_t - \Sigma \mid \mathcal{F}_{t-1}\big]\, Z_{t-1}\big(V^\top(I + \eta_t\Sigma)Z_{t-1}\big)^{-1} = 0,$$

and likewise $\mathbb{E}[\Delta_t \mid \mathcal{F}_{t-1}] = 0$. The matrix H_t is $\mathcal{F}_{t-1}$-measurable, being a function of Z_{t-1}, as is $\mathbf{1}_{t-1}$; hence

$$\mathbb{E}\big[J_{t,1}\mathbf{1}_{t-1} \mid \mathcal{F}_{t-1}\big] = \Big(\mathbb{E}\big[\widehat{\Delta}_t \mid \mathcal{F}_{t-1}\big] - H_t\,\mathbb{E}\big[\Delta_t \mid \mathcal{F}_{t-1}\big]\Big)\mathbf{1}_{t-1} = 0,$$

which is the hypothesis of Lemma 20(vii) with $\mathcal{F} = \mathcal{F}_{t-1}$ and $X = H_t\mathbf{1}_{t-1}$.

Lemma 20(vii) with $X = H_t\mathbf{1}_{t-1}$, $Y = J_{t,1}\mathbf{1}_{t-1}$ and $Z = J_{t,2}\mathbf{1}_{t-1}$ gives, for every $\lambda > 0$,

$$\|W_t\mathbf{1}_{t-1}(I - \Delta_t^2)\|_{F,p}^2 \le (1 + \lambda)\Big(\|H_t\mathbf{1}_{t-1}\|_{F,p}^2 + 2(p-1)\|J_{t,1}\mathbf{1}_{t-1}\|_{F,p}^2 + \lambda^{-1}\|J_{t,2}\mathbf{1}_{t-1}\|_{F,p}^2\Big),$$

and inserting $\|J_{t,1}\mathbf{1}_{t-1}\|_{F,p} \le \varepsilon_t\mathcal{E}_t$ and $\|J_{t,2}\mathbf{1}_{t-1}\|_{F,p} \le \varepsilon_t^2\mathcal{E}_t$ from Lemma 24,

$$\|W_t\mathbf{1}_{t-1}(I - \Delta_t^2)\|_{F,p}^2 \le (1 + \lambda)\Big(\|H_t\mathbf{1}_{t-1}\|_{F,p}^2 + 2(p-1)\varepsilon_t^2\mathcal{E}_t^2 + \lambda^{-1}\varepsilon_t^4\mathcal{E}_t^2\Big).$$

Choosing $\lambda = \varepsilon_t^2$ collapses the last two terms, since $2(p-1) + 1 = 2p - 1 \le 2p$:

$$\|W_t\mathbf{1}_{t-1}(I - \Delta_t^2)\|_{F,p}^2 \le (1 + \varepsilon_t^2)\Big(\|H_t\mathbf{1}_{t-1}\|_{F,p}^2 + 2p\,\mathcal{E}_t^2\varepsilon_t^2\Big).$$

By (12) and (S1), $\|\Delta_t\mathbf{1}_{t-1}\|_2 \le \varepsilon_t \le \frac{1}{2}$, and $W_t\mathbf{1}_{t-1} = (W_t\mathbf{1}_{t-1})\mathbf{1}_{t-1}$, so (5) of Lemma 7 with $A = \Delta_t$ and $G = G_{t-1}$ gives, pointwise and hence in $\|\cdot\|_{F,p}$,

$$\|W_t\mathbf{1}_{t-1}\|_{F,p} \le (1 - \varepsilon_t^2)^{-1}\|W_t\mathbf{1}_{t-1}(I - \Delta_t^2)\|_{F,p}.$$

Squaring and combining,

$$\|W_t\mathbf{1}_{t-1}\|_{F,p}^2 \le \frac{1 + \varepsilon_t^2}{(1 - \varepsilon_t^2)^2}\Big(\|H_t\mathbf{1}_{t-1}\|_{F,p}^2 + 2p\,\mathcal{E}_t^2\varepsilon_t^2\Big) \le (1 + 5\varepsilon_t^2)\Big(\|H_t\mathbf{1}_{t-1}\|_{F,p}^2 + 2p\,\mathcal{E}_t^2\varepsilon_t^2\Big),$$

the last step because $(1 + \varepsilon^2)(1 - \varepsilon^2)^{-2} \le 1 + 5\varepsilon^2$ for $\varepsilon \le \frac{1}{2}$.

Lemma 23 gives $\|H_t\mathbf{1}_{t-1}\|_{F,p}^2 \le \chi_t^2\|W_{t-1}\mathbf{1}_{t-1}\|_{F,p}^2$, and $(a + b)^2 \le 2a^2 + 2b^2$ applied to $\mathcal{E}_t = 1 + 2\|W_{t-1}\mathbf{1}_{t-1}\|_{F,p}$ gives $\mathcal{E}_t^2 \le 2 + 8\|W_{t-1}\mathbf{1}_{t-1}\|_{F,p}^2$. Hence

$$\|W_t\mathbf{1}_{t-1}\|_{F,p}^2 \le (1 + 5\varepsilon_t^2)\big(\chi_t^2 + 16p\varepsilon_t^2\big)\|W_{t-1}\mathbf{1}_{t-1}\|_{F,p}^2 + 4(1 + 5\varepsilon_t^2)p\varepsilon_t^2,$$

and $4(1 + 5\varepsilon_t^2) \le 10$ for $\varepsilon_t \le \frac{1}{2}$, which is $K_{2,t}$. □

E.5 Phase I

Phase I uses a constant step size η, a threshold $\gamma \geq 2$, and the product-class construction of Section 5 of Huang et al. [2021], which we recall. Suppose there is a deterministic finite set $\mathcal{A}_\star \subseteq \mathbb{R}^{d\times d}$ with $C_t \in \mathcal{A}_\star$ almost surely for every $t \leq T_0$, and put $m := |\mathcal{A}_\star|$. Set

$$\mathcal{E} := \{M^{-1}(A-\Sigma)UU^\top : A \in \mathcal{A}_\star\},$$

$$\mathcal{E}_{r,\ell} := \big\{V^\top F_1 \cdots F_r U : \ F_i \in \mathcal{E} \text{ for at most } \ell \text{ distinct } i \in [r], \text{ and } F_i = (1+\eta\lambda_{k+1})^{-1}(I+\eta\Sigma)UU^\top \text{ otherwise}\big\},$$

Lemma 26 (Properties of the product classes)**.** *Let $r, \ell \geq 1$ and let $E = V^\top F_1 \cdots F_r U \in \mathcal{E}_{r,\ell}$. Then*

$$\|E\|_F \leq 1,$$

$$EU^\top = V^\top F_1 \cdots F_r UU^\top = V^\top F_1 \cdots F_r, \tag{15}$$

and $\mathcal{E}_{r,\ell} \subseteq \mathcal{E}_{r,\ell+1}$.

Proof. Every factor is a contraction in operator norm. For a noise factor, $\|M^{-1}(A-\Sigma)UU^\top\|_2 \leq M^{-1}\|A-\Sigma\|_2 \leq 1$ by Assumption C.2; for the drift factor, $(1+\eta\lambda_{k+1})^{-1}\|(I+\eta\Sigma)UU^\top\|_2 = 1$, since $\Sigma \succeq 0$ and its largest eigenvalue on $\mathrm{range}(U)$ is λ_{k+1}.

The leading factor is a contraction in Frobenius norm. If F_1 is a noise factor then

$$\|V^\top F_1\|_F = M^{-1}\|V^\top(A-\Sigma)UU^\top\|_F \leq M^{-1}\|V^\top(A-\Sigma)\|_F \leq 1$$

by Assumption C.2; and if F_1 is the drift factor then $V^\top F_1 = 0$, since $V^\top\Sigma U = \Lambda_1 V^\top U = 0$ and $V^\top U = 0$.

The bound. Splitting $E = \big(V^\top F_1\big)\big(F_2 \cdots F_r U\big)$ and applying Lemma 20(ii) to the second group,

$$\|E\|_F \leq \|V^\top F_1\|_F \prod_{i\geq 2} \|F_i\|_2 \, \|U\|_2 \leq 1.$$

One factor has to be bounded in Frobenius norm: operator-norm bounds alone would give only $\|E\|_F \leq \sqrt{k}\,\|E\|_2$.

Every factor ends in $UU^\top$, and $U^\top U = I_{d-k}$ gives $F_r UU^\top = F_r$, which is (15). And finally an element with at most ℓ noise factors has at most $\ell+1$. □

Define the Phase I good events

$$G_t := G_{t-1} \cap \Big\{ \max_{E\in\mathcal{E}} \|V^\top EUW_t\|_2 \leq \gamma \Big\}, \qquad t \geq 1,$$

with G_0 the initialisation event

$$G_0 := \bigcap_{r,\ell=1}^{T_0+1} \Big\{ \max_{E\in\mathcal{E}_{r,\ell}} \|EW_0\|_F \leq \sqrt{\ell}\,\gamma_{\mathrm{init}} \Big\} \cap \big\{\|W_0\|_F \leq \sqrt{d}\,\gamma\big\}, \qquad \gamma_{\mathrm{init}} := \frac{\gamma}{e\sqrt{2}}.$$

These events satisfy $\|V^\top(C_t-\Sigma)UW_{t-1}\mathbf{1}_{t-1}\|_2 \leq M\gamma$ almost surely, which is needed for the main recursive bound (Theorem 9): for $F_t := M^{-1}(C_t-\Sigma)UU^\top \in \mathcal{E}$ one has $\|V^\top F_t UW_{t-1}\|_2 \leq \gamma$ on G_{t-1}.

Proposition 2 (One-step product estimate)**.** *Assume* (S1)*,* (S2)*, let $E \in \mathcal{E}_{r,\ell}$ and let $p \geq \max\{2, \log m\}$. Put*

$$\overline{E}_t := 1 + 2e \max_{E''\in\mathcal{E}_{r+1,\ell+1}} \|E''W_{t-1}\mathbf{1}_{t-1}\|_{F,p}.$$

Then $EW_t(I-\Delta_t^2) = EH_t + EJ_{t,1} + EJ_{t,2}$, and

$$\|EJ_{t,1}\mathbf{1}_{t-1}\|_{F,p} \leq \overline{E}_t\varepsilon, \tag{16}$$

$$\|EJ_{t,2}\mathbf{1}_{t-1}\|_{F,p} \leq \overline{E}_t\varepsilon^2, \tag{17}$$

$$\|EH_t\mathbf{1}_{t-1}\|_{F,p} \leq \chi\,\|E'W_{t-1}\mathbf{1}_{t-1}\|_{F,p} \quad \text{for } E' := (1+\eta\lambda_{k+1})^{-1}EU^\top(I+\eta\Sigma)U \in \mathcal{E}_{r+1,\ell}, \tag{18}$$

$$\mathbb{E}\big[EJ_{t,1}\mathbf{1}_{t-1} \mid \mathcal{F}_{t-1}\big] = 0. \tag{19}$$

Proof. The decomposition is the identity of Lemma 22 multiplied on the left by E.

(18): by Lemma 22, $H_t = (I+\eta\Lambda_2)W_{t-1}(I+\eta\Lambda_1)^{-1}$. The columns of U are the eigenvectors of Σ with eigenvalues $\lambda_{k+1},\dots,\lambda_d$, so $U^\top\Sigma U = \Lambda_2$ and $U^\top U = I_{d-k}$, giving $I+\eta\Lambda_2 = U^\top(I+\eta\Sigma)U$ and

$$EH_t = \underbrace{EU^\top(I+\eta\Sigma)U}_{=(1+\eta\lambda_{k+1})E'}\, W_{t-1}(I+\eta\Lambda_1)^{-1}.$$

Since $\Lambda_1 \succeq 0$ with smallest entry λ_k, $\|(I+\eta\Lambda_1)^{-1}\|_2 = (1+\eta\lambda_k)^{-1}$, so Lemma 20(ii) yields

$$\|EH_t\mathbf{1}_{t-1}\|_{F,p} \le \frac{1+\eta\lambda_{k+1}}{1+\eta\lambda_k}\|E'W_{t-1}\mathbf{1}_{t-1}\|_{F,p} = \chi\,\|E'W_{t-1}\mathbf{1}_{t-1}\|_{F,p}.$$

That E' lies in $\mathcal{E}_{r+1,\ell}$ follows from (15): with $F_{r+1} := (1+\eta\lambda_{k+1})^{-1}(I+\eta\Sigma)UU^\top$, a drift factor of the class, $V^\top F_1\cdots F_rF_{r+1}U = (1+\eta\lambda_{k+1})^{-1}EU^\top(I+\eta\Sigma)U = E'$, a product of $r+1$ factors whose new factor is not a noise factor, so the count is still at most ℓ.

(16): by Lemma 22, $J_{t,1} = \widehat{\Delta}_t - H_t\Delta_t$, so

$$\|EJ_{t,1}\mathbf{1}_{t-1}\|_{F,p} \le \underbrace{\|E\widehat{\Delta}_t\mathbf{1}_{t-1}\|_{F,p}}_{(*)} + \underbrace{\|EH_t\mathbf{1}_{t-1}\|_{F,p}}_{(**)}\,\|\Delta_t\mathbf{1}_{t-1}\|_2.$$

For $(*)$, the closed form $\widehat{\Delta}_t = \eta U^\top(C_t-\Sigma)(V+UW_{t-1})(I+\eta\Lambda_1)^{-1}$ of Lemma 22 gives

$$(*) \le \eta\Big(\|EU^\top(C_t-\Sigma)V\|_{F,p} + \|EU^\top(C_t-\Sigma)UW_{t-1}\mathbf{1}_{t-1}\|_{F,p}\Big),$$

the right factor having operator norm at most 1 and being discarded by Lemma 20(ii). The first summand is at most M, since $\|E\|_2 \le \|E\|_F \le 1$ by Lemma 26 and $\|U^\top(C_t-\Sigma)V\|_F \le M$ almost surely by Lemma 21. For the second, put $F_{r+1} := M^{-1}(C_t-\Sigma)UU^\top$, which lies in $\mathcal{E}$ because $C_t \in \mathcal{A}_\star$ almost surely, and $E'' := V^\top F_1\cdots F_rF_{r+1}U$; then (15) and $U^\top U = I_{d-k}$ give

$$E'' = M^{-1}V^\top F_1\cdots F_r(C_t-\Sigma)UU^\top U = M^{-1}EU^\top(C_t-\Sigma)U \;\in\; \mathcal{E}_{r+1,\ell+1},$$

the new factor being a noise factor, so the count rises from ℓ to $\ell+1$, and hence $EU^\top(C_t-\Sigma)UW_{t-1} = M\,E''W_{t-1}$. Here E'' is *random*: as C_t ranges over $\mathcal{A}_\star$ it takes m values, all of them in the deterministic class $\mathcal{E}_{r+1,\ell+1}$. Lemma 20(viii) with $\mathcal{D} = \{M^{-1}EU^\top(A-\Sigma)U : A\in\mathcal{A}_\star\}$ and $Y = W_{t-1}\mathbf{1}_{t-1}$ therefore gives

$$\|E''W_{t-1}\mathbf{1}_{t-1}\|_{F,p} \le m^{1/p}\max_{E'''\in\mathcal{E}_{r+1,\ell+1}}\|E'''W_{t-1}\mathbf{1}_{t-1}\|_{F,p} \le e\max_{E'''\in\mathcal{E}_{r+1,\ell+1}}\|E'''W_{t-1}\mathbf{1}_{t-1}\|_{F,p}, \tag{20}$$

using $p \ge \log m$; the maximum is over a finite set of *deterministic* matrices, so $\overline{E}_t$ is deterministic. Combining, and using $\eta M \le \frac{1}{2}\varepsilon$ from $\varepsilon = 2\eta M(1+\gamma)$ with $\gamma \ge 1$,

$$(*) \le \eta M\Big(1 + e\max_{E'''}\|E'''W_{t-1}\mathbf{1}_{t-1}\|_{F,p}\Big) \le \tfrac{1}{2}\varepsilon\,\overline{E}_t.$$

For $(**)$, since $\chi \le 1$ by Lemma 23 and $\mathcal{E}_{r+1,\ell} \subseteq \mathcal{E}_{r+1,\ell+1}$ by Lemma 26,

$$(**) \le \chi\max_{E'''\in\mathcal{E}_{r+1,\ell+1}}\|E'''W_{t-1}\mathbf{1}_{t-1}\|_{F,p} \le e\max_{E'''\in\mathcal{E}_{r+1,\ell+1}}\|E'''W_{t-1}\mathbf{1}_{t-1}\|_{F,p} = \frac{\overline{E}_t-1}{2} \le \tfrac{1}{2}\overline{E}_t.$$

With $\|\Delta_t\mathbf{1}_{t-1}\|_2 \le \varepsilon$ from (12) and Lemma 20(ii),

$$\|EJ_{t,1}\mathbf{1}_{t-1}\|_{F,p} \le \tfrac{1}{2}\varepsilon\overline{E}_t + \tfrac{1}{2}\overline{E}_t\varepsilon = \overline{E}_t\varepsilon.$$

(17): by Lemma 22, $J_{t,2} = -\widehat{\Delta}_t\Delta_t$, so

$$\|EJ_{t,2}\mathbf{1}_{t-1}\|_{F,p} \le \underbrace{\|E\widehat{\Delta}_t\mathbf{1}_{t-1}\|_{F,p}}_{\le\frac{1}{2}\varepsilon\overline{E}_t}\underbrace{\|\Delta_t\mathbf{1}_{t-1}\|_2}_{\le\varepsilon} \le \overline{E}_t\varepsilon^2.$$

(19): E is deterministic, and H_t, W_{t-1} and $\mathbf{1}_{t-1}$ are $\mathcal{F}_{t-1}$-measurable, so

$$\mathbb{E}\big[EJ_{t,1}\mathbf{1}_{t-1} \mid \mathcal{F}_{t-1}\big] = E\Big(\mathbb{E}\big[\widehat{\Delta}_t \mid \mathcal{F}_{t-1}\big] - H_t\,\mathbb{E}\big[\Delta_t \mid \mathcal{F}_{t-1}\big]\Big)\mathbf{1}_{t-1}.$$

Both $\widehat{\Delta}_t$ and Δ_t are linear in $C_t - \Sigma$ with $\mathcal{F}_{t-1}$-measurable coefficients, so by Assumption C.1,

$$\mathbb{E}\big[\widehat{\Delta}_t \mid \mathcal{F}_{t-1}\big] = \eta\, U^\top \underbrace{\mathbb{E}\big[C_t - \Sigma \mid \mathcal{F}_{t-1}\big]}_{=0} Z_{t-1}\big(V^\top (I+\eta\Sigma) Z_{t-1}\big)^{-1} = 0,$$

and likewise $\mathbb{E}[\Delta_t \mid \mathcal{F}_{t-1}] = 0$. □

Proposition 3 (One-step recursion with product classes)**.** *Under the hypotheses of Proposition 2, for every $t \le T_0$,*

$$\max_{E\in\mathcal{E}_{r,\ell}} \|EW_t\mathbf{1}_{t-1}\|_{F,p}^2 \le \overline{K}_1 \max_{E'\in\mathcal{E}_{r+1,\ell}} \|E'W_{t-1}\mathbf{1}_{t-1}\|_{F,p}^2 + \overline{K}_2 \max_{E''\in\mathcal{E}_{r+1,\ell+1}} \|E''W_{t-1}\mathbf{1}_{t-1}\|_{F,p}^2 + \overline{K}_2,$$

with $\overline{K}_1 = (1+5\varepsilon^2)\chi^2$ and $\overline{K}_2 = (1+5\varepsilon^2)16e^2p\varepsilon^2$. These are the constants of Huang et al. with $\overline{K}_2$ multiplied by two independent factors: e^2, supplied by (20)*, and* 2*, from the proof of Lemma 20(vi).*

Proof. The argument is that of Proposition 1, with Proposition 2 in place of Lemma 24. Fix $E \in \mathcal{E}_{r,\ell}$. By Proposition 2,

$$\|EW_t\mathbf{1}_{t-1}(I-\Delta_t^2)\|_{F,p}^2 = \|EH_t\mathbf{1}_{t-1} + EJ_{t,1}\mathbf{1}_{t-1} + EJ_{t,2}\mathbf{1}_{t-1}\|_{F,p}^2,$$

where $EH_t\mathbf{1}_{t-1}$ is $\mathcal{F}_{t-1}$-measurable and $\mathbb{E}[EJ_{t,1}\mathbf{1}_{t-1} \mid \mathcal{F}_{t-1}] = 0$ by (19), so Lemma 20(vii) applies with $\lambda = \varepsilon^2$, and (16) and (17) bound the two fluctuation terms:

$$\begin{aligned}\|EW_t\mathbf{1}_{t-1}(I-\Delta_t^2)\|_{F,p}^2 &\le (1+\varepsilon^2)\Big(\|EH_t\mathbf{1}_{t-1}\|_{F,p}^2 + 2(p-1)\underbrace{\|EJ_{t,1}\mathbf{1}_{t-1}\|_{F,p}^2}_{\le \overline{E}_t^{\,2}\varepsilon^2} + \frac{1}{\varepsilon^2}\underbrace{\|EJ_{t,2}\mathbf{1}_{t-1}\|_{F,p}^2}_{\le \overline{E}_t^{\,2}\varepsilon^4}\Big)\\ &\le (1+\varepsilon^2)\Big(\|EH_t\mathbf{1}_{t-1}\|_{F,p}^2 + 2p\,\overline{E}_t^{\,2}\varepsilon^2\Big),\end{aligned}$$

since $2(p-1)+1 = 2p-1 \le 2p$.

Note $\|\Delta_t\mathbf{1}_{t-1}\|_2 \le \varepsilon \le \frac{1}{2}$ by (12) and (S1), and $EW_t\mathbf{1}_{t-1} = (EW_t\mathbf{1}_{t-1})\mathbf{1}_{t-1}$, so (5) of Lemma 7 with $A = \Delta_t$ and $G = G_{t-1}$ gives $\|EW_t\mathbf{1}_{t-1}\|_{F,p} \le (1-\varepsilon^2)^{-1}\|EW_t\mathbf{1}_{t-1}(I-\Delta_t^2)\|_{F,p}$, whence

$$\|EW_t\mathbf{1}_{t-1}\|_{F,p}^2 \le \underbrace{\frac{1+\varepsilon^2}{(1-\varepsilon^2)^2}}_{\le 1+5\varepsilon^2}\Big(\|EH_t\mathbf{1}_{t-1}\|_{F,p}^2 + 2p\,\overline{E}_t^{\,2}\varepsilon^2\Big).$$

Now $\overline{E}_t = 1 + 2e\max_{E''\in\mathcal{E}_{r+1,\ell+1}}\|E''W_{t-1}\mathbf{1}_{t-1}\|_{F,p}$, so $(a+b)^2 \le 2a^2+2b^2$ gives

$$\overline{E}_t^{\,2} \le 2 + 8e^2 \max_{E''\in\mathcal{E}_{r+1,\ell+1}} \|E''W_{t-1}\mathbf{1}_{t-1}\|_{F,p}^2,$$

and therefore

$$(1+5\varepsilon^2)2p\,\overline{E}_t^{\,2}\varepsilon^2 \le (1+5\varepsilon^2)16e^2p\varepsilon^2 \max_{E''\in\mathcal{E}_{r+1,\ell+1}} \|E''W_{t-1}\mathbf{1}_{t-1}\|_{F,p}^2 + 4(1+5\varepsilon^2)p\varepsilon^2 \le \overline{K}_2 \max_{E''\in\mathcal{E}_{r+1,\ell+1}} \|E''W_{t-1}\mathbf{1}_{t-1}\|_{F,p}^2 + \overline{K}_2$$

the last step because $4 \le 16e^2$.

Finally $\|EH_t\mathbf{1}_{t-1}\|_{F,p}^2 \le \chi^2 \max_{E'\in\mathcal{E}_{r+1,\ell}}\|E'W_{t-1}\mathbf{1}_{t-1}\|_{F,p}^2$ by (18). The right-hand side does not depend on $E \in \mathcal{E}_{r,\ell}$, so the maximum over E may be taken on the left. □

Theorem 10 (Iteration of the product-class recursion)**.** *Assume* (S1)*,* (S2)*,* (S4) *with constant step size, $p \ge \max\{2, \log m\}$, $\gamma \ge 2$, and that on G_0 one has $\max_{1\le r,\ell\le T_0+1}\max_{E\in\mathcal{E}_{r,\ell}}\|EW_0\|_F \le \sqrt{\ell}\,\gamma_{\mathrm{init}}$. Then for every $1 \le t \le T_0$ and every $1 \le r, \ell \le T_0 - t + 1$,*

$$\max_{E\in\mathcal{E}_{r,\ell}} \|EW_t\mathbf{1}_t\|_{F,p}^2 \le \max_{E\in\mathcal{E}_{r,\ell}} \|EW_t\mathbf{1}_{t-1}\|_{F,p}^2 \le \ell\gamma_{\mathrm{init}}^2 e^{-t\eta\rho_k/2} + 12e^2p\gamma^2\varepsilon^2 t.$$

Proof. The first inequality is Lemma 20(iii). Writing $B_{t,r,\ell} := \max_{E\in\mathcal{E}_{r,\ell}}\|EW_t\mathbf{1}_{t-1}\|_{F,p}^2$ and $A_{t,r,\ell} := \max_{E\in\mathcal{E}_{r,\ell}}\|EW_t\mathbf{1}_t\|_{F,p}^2 \le B_{t,r,\ell}$, Proposition 3 reads

$$B_{t,r,\ell} \le \overline{K}_1 A_{t-1,r+1,\ell} + \overline{K}_2 A_{t-1,r+1,\ell+1} + \overline{K}_2, \qquad \overline{K}_1 = (1+5\varepsilon^2)\chi^2, \quad \overline{K}_2 = (1+5\varepsilon^2)16e^2p\varepsilon^2.$$

Put $x := \eta\rho_k$, so $x \le \frac{1}{2}$ by (S2). As in the proof of Theorem 9, $\chi = 1 - \frac{\eta\rho_k}{1+\eta\lambda_k}$, and (S2) in the form $\eta\lambda_k \le \frac{1}{2}$ gives $1 + \eta\lambda_k \le \frac{3}{2}$, so $\chi \le 1 - \frac{2}{3}x$. Further, (S4) with $p \ge 2$ gives $\varepsilon^2 \le x/800$, hence

$$16e^2 p\varepsilon^2 \le \frac{16e^2}{400}\, x \le \tfrac{3}{10}x, \qquad 1 + 5\varepsilon^2 \le e^{5\varepsilon^2} \le e^{x/160},$$

and for $x \le \frac{1}{2}$, $(1 - \frac{2}{3}x)^2 \le e^{-4x/3} \le 1 - \frac{4}{3}x + \frac{8}{9}x^2 \le 1 - \frac{8}{9}x$. Hence

$$\overline{K}_1 + \overline{K}_2 = (1+5\varepsilon^2)\big(\chi^2 + 16e^2 p\varepsilon^2\big) \le e^{x/160}\big(1 - \tfrac{8}{9}x + \tfrac{3}{10}x\big) \le e^{x/160} e^{-(8/9 - 3/10)x} \le e^{-x/2} = e^{-\eta\rho_k/2} \le 1,$$

using $1 - u \le e^{-u}$ and $\frac{8}{9} - \frac{3}{10} - \frac{1}{160} \ge \frac{1}{2}$; and $\overline{K}_2 \le \frac{9}{4} \cdot 16e^2 p\varepsilon^2 = 36e^2 p\varepsilon^2$ from $\varepsilon \le \frac{1}{2}$.

We show by induction that $B_{t,r,\ell} \le \ell\gamma_{\mathrm{init}}^2 e^{-t\eta\rho_k/2} + \frac{\gamma^2}{3}\overline{K}_2 t$ for all $1 \le r, \ell \le T_0 - t + 1$.

At $t = 1$, the hypothesis on G_0 gives

$$B_{1,r,\ell} \le \overline{K}_1 \underbrace{\max_{E' \in \mathcal{E}_{r+1,\ell}} \|E' W_0 \mathbf{1}_0\|_{F,p}^2}_{\le \ell\gamma_{\mathrm{init}}^2} + \overline{K}_2 \underbrace{\max_{E'' \in \mathcal{E}_{r+1,\ell+1}} \|E'' W_0 \mathbf{1}_0\|_{F,p}^2}_{\le (\ell+1)\gamma_{\mathrm{init}}^2} + \overline{K}_2 \le \ell\gamma_{\mathrm{init}}^2\big(\overline{K}_1 + \overline{K}_2\big) + \big(1 + \gamma_{\mathrm{init}}^2\big)\overline{K}_2,$$

and $\gamma \ge 2$ gives $1 + \gamma_{\mathrm{init}}^2 = 1 + \gamma^2/2e^2 \le \gamma^2/3$, which is the claim.

For $t > 1$, note $r + 1, \ell + 1 \le T_0 - (t-1) + 1$, so the induction hypothesis applies to both terms on the right, and

$$\begin{aligned} B_{t,r,\ell} &\le \overline{K}_1\Big(\ell\gamma_{\mathrm{init}}^2 e^{-(t-1)\eta\rho_k/2} + \tfrac{\gamma^2}{3}\overline{K}_2(t-1)\Big) + \overline{K}_2\Big((\ell+1)\gamma_{\mathrm{init}}^2 e^{-(t-1)\eta\rho_k/2} + \tfrac{\gamma^2}{3}\overline{K}_2(t-1)\Big) + \overline{K}_2 \\ &\le \ell\gamma_{\mathrm{init}}^2\big(\overline{K}_1 + \overline{K}_2\big)e^{-(t-1)\eta\rho_k/2} + \big(1 + \gamma_{\mathrm{init}}^2\big)\overline{K}_2 + \tfrac{\gamma^2}{3}\overline{K}_2(t-1)\big(\overline{K}_1 + \overline{K}_2\big) \\ &\le \ell\gamma_{\mathrm{init}}^2 e^{-t\eta\rho_k/2} + \tfrac{\gamma^2}{3}\overline{K}_2 t, \end{aligned}$$

using $\overline{K}_1 + \overline{K}_2 \le e^{-\eta\rho_k/2} \le 1$ and $1 + \gamma_{\mathrm{init}}^2 \le \gamma^2/3$ again. Finally $\frac{\gamma^2}{3}\overline{K}_2 t \le 12e^2 p\gamma^2\varepsilon^2 t$. □

Proposition 4 (Phase I moment estimates)**.** *Let the updates $C_1, \dots, C_{T_0}$ satisfy Assumption C and lie almost surely in a deterministic finite set $\mathcal{A}_\star$ with $m := |\mathcal{A}_\star|$, let the step size be constant equal to η, let $\gamma \ge 2$, let $\mathcal{E}$ and $\mathcal{E}_{r,\ell}$ be the classes built from $\mathcal{A}_\star$ at the start of this section, and let $(G_t)_{t \ge 0}$ be the good events defined there with $\gamma_{\mathrm{init}} = \gamma/(e\sqrt{2})$. Let $\delta \in (0,1)$, put*

$$p_0 := \max\Big\{2,\ \log\frac{6}{\delta}\Big\}, \qquad p_1 := \max\Big\{2,\ \log\frac{12T_0(m+1)}{\delta}\Big\}, \qquad p_\star := \max\{p_0, p_1\},$$

and assume (S1)*,* (S2)*,* (S4) *with $p = p_\star$, together with*

$$12e^2 p_\star \varepsilon^2 T_0 \le \frac{1}{2e^2}, \qquad d\gamma^2 e^{-T_0\eta\rho_k/2} \le \frac{1}{2e^2}.$$

Then

$$\|W_{T_0}\mathbf{1}_{T_0}\|_{F,p_0} \le e^{-1} \qquad \textit{and} \qquad \max_{1 \le j \le T_0} \max_{F \in \mathcal{E}} \|V^\top F U W_j \mathbf{1}_{j-1}\|_{F,p_1} \le \frac{\gamma}{e}.$$

Proof. Admissible exponents. Definition 5 defines $\|\cdot\|_{F,p}$ only for $p \ge 2$, and Lemma 20 and Theorems 9–10 are stated under that restriction, so both exponents are truncated at 2; this matches the convention $p_\delta = \max\{2, \log(2/\delta)\}$ used in Phase II. The truncation is active only for p_0, and only when $\delta > 6e^{-2}$: since $12T_0(m+1) \ge 24$ we have $p_1 = \log\frac{12T_0(m+1)}{\delta} > \log 24 > 2$ always, whence also $p_\star = p_1$. The truncation costs nothing downstream, because the only property of p_0 used outside this proof is $e^{-p_0} \le \delta/6$ (Theorem 11), which holds with equality when $\log\frac{6}{\delta} \ge 2$ and, when $\log\frac{6}{\delta} < 2$, holds because $\delta > 6e^{-2}$ gives $e^{-p_0} = e^{-2} < \delta/6$.

Terminal bound. The Phase I good events supply the hypothesis $\|V^\top(C_i - \Sigma)UW_{i-1}\mathbf{1}_{i-1}\|_2 \le M\gamma$ of Theorem 9, and (S3) is vacuous for a constant step size, so (14) at $p = p_0$ gives

$$\|W_{T_0}\mathbf{1}_{T_0}\|_{F,p_0}^2 \le e^{-T_0\eta\rho_k/2}\|W_0\mathbf{1}_0\|_{F,p_0}^2 + 10p_0\varepsilon^2 T_0.$$

By the definition of G_0, $\|W_0\mathbf{1}_0\|_F^2 \le d\gamma^2$ pointwise, hence $\|W_0\mathbf{1}_0\|_{F,p_0}^2 \le d\gamma^2$. The second displayed condition bounds the first term by $\frac{1}{2e^2}$, and $p_0 \le p_\star$ with the first condition bounds the second by $10p_0\varepsilon^2 T_0 \le 12e^2 p_\star \varepsilon^2 T_0 \le \frac{1}{2e^2}$. Hence $\|W_{T_0}\mathbf{1}_{T_0}\|_{F,p_0}^2 \le e^{-2}$.

Good-event moment bound. For $F \in \mathcal{E}$ one has $V^\top F U \in \mathcal{E}_{1,1}$, and the definition of G_0 gives $\max_{E\in\mathcal{E}_{1,1}} \|EW_0\mathbf{1}_0\|_{F,p_1}^2 \le \gamma_{\text{init}}^2 = \gamma^2/2e^2$. Theorem 10 with $r = \ell = 1$ and $p = p_1$ therefore gives $\|V^\top F U W_j \mathbf{1}_{j-1}\|_{F,p_1}^2 \le \frac{\gamma^2}{2e^2} e^{-j\eta\rho_k/2} + 12e^2 p_1 \gamma^2 \varepsilon^2 j \le \frac{\gamma^2}{2e^2} + \frac{\gamma^2}{2e^2}$, using $e^{-j\eta\rho_k/2} \le 1$, $p_1 \le p_\star$, $j \le T_0$ and the first condition. Theorem 10 is applicable at $p = p_1$ because $p_1 \ge \log \frac{12T_0(m+1)}{\delta} \ge \log m$, which is the hypothesis $p \ge \log m$ of Propositions 2–3; this is the only exponent at which the product-class recursion is used. □

Theorem 11 (Finite-support Phase I)**.** *Let the updates satisfy the hypotheses of Proposition 4, with deterministic finite support of cardinality m and constant step size η, and let $Q_0 = \mathrm{QR}(Z_0)$ for Z_0 with i.i.d. $\mathcal{N}(0,1)$ entries and $\sigma(Z_0) \subseteq \mathcal{F}_0$. Let $\delta \in (0,1)$, C_γ be a sufficiently large absolute constant, and set*

$$\gamma := \max\Big\{e,\, C_\gamma \min\Big\{\frac{\sqrt{k\log(emT_0/\delta)}}{\delta},\, \frac{d}{\delta^2}\Big\}\Big\}, \quad L_\gamma := \max\{1, \log(2e^2 d\gamma^2)\}, \quad \eta := \frac{C_\eta L_\gamma}{\rho_k T_0},$$

with $C_\eta \ge 2$. If

$$T_0 \ge C \max\Big\{\frac{L_\gamma M(1+\gamma)}{\rho_k},\, \frac{L_\gamma \|\Sigma\|_2}{\rho_k},\, \frac{p_\star L_\gamma M^2(1+\gamma)^2}{\rho_k^2},\, \frac{p_\star L_\gamma^2 M^2(1+\gamma)^2}{\rho_k^2}\Big\},$$

then $\mathbb{P}(\|W_{T_0}\|_F > 1) \le \delta/3$.

Proof. The first three lower bounds on T_0 give (S1), (S2) and (S4) with $p = p_\star$, and the fourth gives $12e^2 p_\star \varepsilon^2 T_0 \le \frac{1}{2e^2}$. The computation of Theorem D.5 of Huang et al. gives the same bound without the factor $2e^2$; the two factors are independent, e^2 coming from the random class index of (20) and 2 from the constant in Lemma 20(vi), and both are absorbed by the free constant in the lower bound on T_0. For the contraction condition, $T_0\eta\rho_k = C_\eta L_\gamma$ and $C_\eta \ge 2$ give $e^{-T_0\eta\rho_k/2} \le e^{-L_\gamma}$, and the definition of L_γ then gives $d\gamma^2 e^{-T_0\eta\rho_k/2} \le \frac{1}{2e^2}$, the second condition of Proposition 4. The corresponding definition in Huang et al. is $L_\gamma = \max\{1, \log(2e^2 d\gamma^2/k^{2/p_0})\}$, and the fourth lower bound on T_0 carries a denominator k^{2/p_0}; since $k^{2/p_0} \in [1, e^2]$ whenever $p_0 \ge \log k$, the two differ by at most 2 additively and by a factor at most e^2 respectively, so T_0 is unchanged in order.

By Lemma 25, on G_{T_0} the matrix $V^\top Z_{T_0}$ is invertible and $W_{T_0} = U^\top Q_{T_0}(V^\top Q_{T_0})^{-1}$, so

$$\mathbb{P}\big(\|W_{T_0}\|_F > 1\big) \le \mathbb{P}\big(\|W_{T_0}\mathbf{1}_{T_0}\|_F > 1\big) + \mathbb{P}(G_{T_0}^c),$$

and by Proposition 4 with Lemma 20(v) at $s = 1$,

$$\mathbb{P}\big(\|W_{T_0}\mathbf{1}_{T_0}\|_F > 1\big) \le \|W_{T_0}\mathbf{1}_{T_0}\|_{F,p_0}^{p_0} \le e^{-p_0} \le \frac{\delta}{6}.$$

The corresponding step of Huang et al. produces $k\,e^{-p_0}$ and so requires $p_0 = \log(6k/\delta)$; the factor k there is the k^{1/p_0} of their terminal bound, which Proposition 4 does not carry. For $\mathbb{P}(G_{T_0}^c)$, the event G_0 concerns W_0 alone, so a union bound over the finitely many elements of the classes $\mathcal{E}_{r,\ell}$, combined with Gaussian bounds on $\|U^\top Z_0\|_2$ and $\|(V^\top Z_0)^{-1}\|_F$, gives

$$\mathbb{P}(G_0^c) \le \frac{\delta}{12} \tag{21}$$

whenever $\gamma \ge C_\gamma \min\{\sqrt{k\log(emT_0/\delta)}/\delta,\ d/\delta^2\}$, which is the choice made in the statement; this is Lemma H.4 of Huang et al. [2021]. and for $j \ge 1$, Markov's inequality in $\|\cdot\|_2 \le \|\cdot\|_F$ together with the second bound of Proposition 4 gives $\mathbb{P}(G_j^c \cap G_{j-1}) \le \sum_{F\in\mathcal{E}} \gamma^{-p_1} \|V^\top F U W_j \mathbf{1}_{j-1}\|_{F,p_1}^{p_1} \le m\,e^{-p_1}$, so $\sum_{j\le T_0} \mathbb{P}(G_j^c \cap G_{j-1}) \le mT_0 e^{-p_1} \le \delta/12$ by the choice of p_1. Hence $\mathbb{P}(G_{T_0}^c) \le \delta/6$ and the two terms sum to $\delta/3$. □

E.6 Phase II

For Phase II the good events are taken in the Frobenius norm. Fix a starting time τ and relabel it as the index 0, so that $W_0 = W_\tau$; fix a threshold $\gamma \ge 1$ and set

$$G_0 := \{\|W_0\|_F \le 1\}, \qquad G_m := G_{m-1} \cap \{\|W_m\|_F \le \gamma\}, \qquad \mathbf{1}_m := \mathbf{1}_{G_m}.$$

These events are decreasing, and since $\|W_s \mathbf{1}_s\|_2 \le \|W_s \mathbf{1}_s\|_F \le \gamma$ they imply the hypothesis $\|V^\top (C_m - \Sigma) U W_{m-1} \mathbf{1}_{m-1}\|_2 \le M\gamma$ of Lemmas 24 and 25. No product class and no finite-support hypothesis is used in this subsection.

Phase II is Theorem 9 with a decaying step size, as in Section 4 of Huang et al. The one point to notice is that the hypotheses (S1)–(S3) of its first conclusion (13) do not involve p, while the conclusion holds for every $p \ge 2$. The good-event check at step j may therefore use its own exponent p_j. Choosing p_j to grow linearly in j, as large as the fluctuation term permits, makes the per-step failure probabilities decay geometrically, so the union bound over the run costs $O(1)$ rather than a factor $T - \tau$. This is the device of Theorem 4.2 of Huang et al.; in the Frobenius norm it is better still, because their Markov step pays $k\, e^{-p}$ where ours pays e^{-p}.

Proposition 5 (Phase II with decaying step size)**.** *Let $\|W_0\|_F \le 1$. Fix $c_\eta \ge 8$, put $\alpha := \frac{1}{2} c_\eta$ and $\gamma := 2e$, and for $m \ge 1$ set*

$$\eta_m = \frac{c_\eta}{\rho_k (\beta + m)}, \qquad \beta \ge \max\Big\{1,\ \frac{4 c_\eta \lambda_1}{\rho_k},\ \frac{27 c_\eta M}{\rho_k},\ \frac{C_3 c_\eta^2 M^2}{\rho_k^2} \Lambda_\delta \Big\},$$

$$\Lambda_\delta := \log\Big(e + \frac{c_\eta^2 M^2}{\rho_k^2 \delta}\Big) + \log \frac{2}{\delta}, \qquad p_\delta := \max\{2, \log(2/\delta)\}.$$

Then with probability at least $1 - \delta$,

$$\|W_T\|_F \le e \Big(\frac{\beta + 1}{\beta + T}\Big)^{\alpha} + C_4 \, \frac{c_\eta M}{\rho_k} \sqrt{\frac{p_\delta}{\beta + T}}.$$

Remark. Neither β nor the bound depends on T, d or k. The constant c_η is free subject to $c_\eta \ge 8$, and may be chosen as a function of the sample size, as Theorems 1 and 2 do; β then depends on it through the displayed branches, but still not on T. Every constant in the proof is independent of c_η: in particular $C_5 = 2(1 + \gamma)\sqrt{C_1 \gamma^2 + C_2}$ with $\gamma = 2e$, and $C_4 = e C_5$.

Proof. Hypotheses of Theorem 9. The good events are decreasing and supply $\|V^\top (C_m - \Sigma) U W_{m-1} \mathbf{1}_{m-1}\|_2 \le M\gamma$. Since η_m decreases it suffices to check (S1)–(S2) at $m = 1$. For (S2), $\eta_1 \lambda_1 = c_\eta \lambda_1 / (\rho_k(\beta + 1)) \le \frac{1}{4}$ by the second bound on β. For (S1), $\varepsilon_1 = 2 c_\eta M (1 + 2e)/(\rho_k(\beta + 1)) \le \frac{1}{2}$ by the third, since $4(1 + 2e) < 27$. For (S3), write $x := 1/(\beta + m) \le \frac{1}{2}$; then $\varepsilon_m / \varepsilon_{m-1} = \eta_m / \eta_{m-1} = 1 - x$, while $e^{-\eta_m \rho_k / 4} = e^{-c_\eta x / 4} \le e^{-2x} \le 1 - x$, the middle inequality because $c_\eta \ge 8$ and the last because $x \le \frac{1}{2}$. Condition (S4) is never used.

Moment bound. Fix $p \ge 2$ and write $S_m := \beta + m$. We have $\|W_0 \mathbf{1}_0\|_{F,p} \le 1$ and $\|W_i \mathbf{1}_i\|_{F,p} \le \gamma$ pointwise, so, with $\gamma \ge 1$, the sum in (13) is at most $m\gamma^2$ and

$$\|W_m \mathbf{1}_*\|_{F,p}^2 \le e^{-s_m \rho_k} + (C_1 \gamma^2 + C_2)\, p\, \varepsilon_m^2 m, \qquad \mathbf{1}_* \in \{\mathbf{1}_m, \mathbf{1}_{m-1}\}.$$

For the first term, $\rho_k s_m = c_\eta \sum_{i=1}^m (\beta + i)^{-1} \ge c_\eta \log \frac{\beta + m}{\beta + 1}$, so $e^{-s_m \rho_k} \le ((\beta + 1)/S_m)^{c_\eta}$. For the second, $\varepsilon_m^2 m = 4 c_\eta^2 M^2 (1 + \gamma)^2 m / (\rho_k^2 S_m^2) \le 4 c_\eta^2 M^2 (1 + \gamma)^2 / (\rho_k^2 S_m)$ because $m \le S_m$. Hence, by $\sqrt{u + v} \le \sqrt{u} + \sqrt{v}$,

$$\|W_m \mathbf{1}_*\|_{F,p} \le \Theta_m(p) := \Big(\frac{\beta + 1}{S_m}\Big)^{\alpha} + C_5 \, \frac{c_\eta M}{\rho_k} \sqrt{\frac{p}{S_m}}, \qquad C_5 := 2(1 + \gamma)\sqrt{C_1 \gamma^2 + C_2}, \tag{22}$$

with $\alpha = \frac{1}{2} c_\eta$ as stated.

Good event. Put $\nu := \rho_k^2 / (C_5^2 c_\eta^2 M^2)$ and $p_j := \nu S_j$. The fourth bound on β gives $\nu\beta \ge 2$, as $\Lambda_\delta \ge 1$ and $C_3 \ge 2 C_5^2$, so $p_j \ge 2$ and (22) applies at $p = p_j$. By construction the second term of $\Theta_j(p_j)$ equals 1, and the first is at most 1; hence $\Theta_j(p_j) \le 2 = \gamma / e$, and Markov's inequality gives

$$\mathbb{P}\big(\|W_j\|_F > \gamma,\ G_{j-1}\big) \le \gamma^{-p_j} \|W_j \mathbf{1}_{j-1}\|_{F,p_j}^{p_j} \le e^{-p_j} = e^{-\nu(\beta + j)}.$$

Summing and using $e^\nu - 1 \ge \nu$,

$$\mathbb{P}(G_T^c) \le \sum_{j \ge 1} e^{-\nu(\beta + j)} = \frac{e^{-\nu\beta}}{e^\nu - 1} \le \frac{e^{-\nu\beta}}{\nu} \le \frac{\delta}{2},$$

the last step because $\nu\beta \geq \log \frac{2}{\nu\delta}$, which is the fourth bound on β once $1/\nu = C_5^2 c_\eta^2 M^2/\rho_k^2$ is substituted and C_3 is taken large enough in terms of C_5. The number of steps never enters.

Conclusion. Apply (22) once more at $p = p_\delta$: Markov's inequality gives $\mathbb{P}\big(\|W_T\|_F > e\Theta_T(p_\delta),\, G_T\big) \leq e^{-p_\delta} \leq \delta/2$, and $e\Theta_T(p_\delta)$ is the asserted bound with $C_4 := eC_5$. □

No constant in this subsection depends on k, d or $k^{2/p}$.

E.7 From the error matrix to the subspace error

Theorem 12 (Phase II in Frobenius norm). *Let $T_0 < T$ and suppose Phase I delivers $\|W_{T_0}\|_F \leq 1$ with probability at least $1 - \delta_I$. Run Proposition 5 on $(T_0, T]$ with $\tau = T_0$, failure probability δ_{II} and the β defined there, and set $S := \beta + T - T_0$ and $\alpha := \frac{1}{2}c_\eta \geq 4$. Then for every $T > T_0$, with probability at least $1 - \delta_I - \delta_{II}$,*

$$\big\|Q_T Q_T^\top - VV^\top\big\|_F \;\leq\; \sqrt{2}\, e\Big(\frac{\beta+1}{S}\Big)^{\alpha} \;+\; C_8\, \frac{M}{\rho_k}\sqrt{\frac{\Lambda_{\delta_{II}}}{S}}.$$

No burn-in stage is required between the two phases, and no lower bound on $T - T_0$ is needed.

Proof. On the Phase I event $\|W_{T_0}\|_F \leq 1$, the entry hypothesis of Proposition 5 holds. By that proposition, with probability $1 - \delta_{II}$, $\|W_T\|_F \leq e(\frac{\beta+1}{S})^\alpha + C_4 \frac{c_\eta M}{\rho_k}\sqrt{p_{\delta_{II}}/S}$. For the second term, $\Lambda_\delta \geq 1$ and $\Lambda_\delta \geq \log\frac{2}{\delta}$ give $p_\delta = \max\{2, \log\frac{2}{\delta}\} \leq 2\Lambda_\delta$. Lemma 19 converts, multiplying both terms by $\sqrt{2}$, and the two failure probabilities add; take $C_8 := 2C_4 c_\eta$. □

E.8 The main theorem

Theorem 13 (Main theorem, Frobenius form; replaces Theorem 8). *Let $C_1, \dots, C_n$ satisfy Assumption C with boundedness constant M and $\rho_k > 0$, and let $C_1, \dots, C_{n_0}$ be in addition i.i.d. and independent of Z_0, as in Section F. Let Z_0 have i.i.d. $\mathcal{N}(0,1)$ entries with $\sigma(Z_0) \subseteq \mathcal{F}_0$, and let Z_0 be independent of the auxiliary array of Proposition 6; set $Q_0 := \mathrm{QR}(Z_0)$. Let V_k contain the k leading eigenvectors of Σ. Fix $\delta \in (0,1)$ and put $\delta_I := \delta/2$, $\xi := \delta/6$, $\delta_{II} := 2\delta/3$. Let n_0 be the Phase I length, let L and $m_\star = n_0(L+1)+1$ be as in Proposition 6 at parameters (δ_I, ξ), and let γ, L_γ, η and the lower bound on n_0 be as in Theorem 11 at failure probability δ_I and support size $m_\star$, so that $n_0 = \widetilde{\Theta}\big(kM^2/\delta^2\rho_k^2\big)$; and let $\beta, \Lambda := \Lambda_{\delta_{II}}$ and c_η be as in Proposition 5 at failure probability δ_{II}. Take*

$$\eta_t = \begin{cases} \dfrac{C_\eta L_\gamma}{\rho_k n_0}, & t \leq n_0, \\[2ex] \dfrac{c_\eta}{\rho_k(\beta + t - n_0)}, & t > n_0. \end{cases}$$

Then for every $n > n_0$, the output Q_n satisfies, with probability at least $1 - \delta$,

$$\boxed{\big\|Q_n Q_n^\top - V_k V_k^\top\big\|_F \;\leq\; \sqrt{2}\, e\Big(\frac{\beta+1}{\beta+n-n_0}\Big)^{\alpha} \;+\; C_8\, \frac{M}{\rho_k}\sqrt{\frac{\Lambda}{n-n_0}}} \tag{23}$$

where $\alpha = \frac{1}{2}c_\eta \geq 4$ and $\Lambda \asymp \log\frac{1}{\delta} + \log\big(e + \frac{M^2}{\rho_k^2\delta}\big)$.

Remark (on the two terms). The second term is the rate; the first is the cost of forgetting the Phase I output and decays like $(n-n_0)^{-\alpha}$ with $\alpha = c_\eta/2 \geq 4$, against $(n-n_0)^{-1/2}$ for the second. It falls below the second once

$$n - n_0 \;\gtrsim\; \beta\,\Gamma^{1/(2\alpha-1)}, \qquad \Gamma := \frac{\beta\rho_k^2}{C_3 c_\eta^2 M^2 \Lambda} \;\geq\; 1,$$

where Γ measures how far β exceeds its last branch; $\Gamma = O(1)$ exactly when that branch attains the maximum defining β, and $2\alpha - 1 \geq 7$, so the extra factor is at worst $\Gamma^{1/7}$. Both displays treat c_η as an absolute constant. Proposition 5 requires only $c_\eta \geq 8$, so c_η may instead be chosen as a function of the horizon, and this is what Theorems 1 and 2 do. The first term of (23) is at most $\varepsilon_\star$ as

soon as $\frac{\beta+n-n_0}{\beta+1} \geq (\sqrt{2}e/\varepsilon_\star)^{1/\alpha}$. Take $\varepsilon_\star = n^{-1/2}$ and $c_\eta \geq \max\{8, \log n\}$: the right-hand side is $(\sqrt{2}e)^{2/c_\eta} n^{1/c_\eta} \leq (\sqrt{2}e)^{1/4}\, e < 3.81$, since $n^{1/c_\eta} \leq n^{1/\log n} = e$, while $n - n_0 \geq 7\beta$ together with $\beta \geq 1$ gives $\frac{\beta+n-n_0}{\beta+1} \geq \frac{8\beta}{2\beta} = 4$. Hence

$$c_\eta \geq \max\{8, \log n\} \text{ and } n - n_0 \geq 7\beta \quad \Longrightarrow \quad \sqrt{2}\, e\Big(\frac{\beta+1}{\beta+n-n_0}\Big)^{\alpha} \;\leq\; \frac{1}{\sqrt{n}},$$

so the crossover is a fixed multiple of β and no Γ factor survives. This is what lets the batch count of Theorems 1 and 2 be $\Theta(\varkappa \log n)$ with $\varkappa := \lambda_1/\rho_k$, since the branch $4c_\eta\lambda_1/\rho_k$ of β then reads $4\varkappa\lceil \log n \rceil$. At constant c_η the first display would instead force $n - n_0 \gtrsim \beta\, n^{1/(2\alpha)}$, polynomial in n.

Remark (on the hypotheses and on $\widetilde{\Theta}$). Phase II uses nothing beyond Assumption C, and its only distributional input is the conditional mean of C.1, through Theorem 9; the almost-sure bound of C.2 enters only pointwise, through Lemma 21. Independence of the updates is required by Phase I alone, and only on the first n_0 updates, where it is what makes the atom array of Proposition 6 drawable in advance. It is the hypothesis of Huang et al., and on the batches where it is needed it holds because Algorithms 1 and 2 run their first n_0 batches with a noise multiplier fixed before the run.

The notation $\widetilde{\Theta}$ suppresses factors polylogarithmic in d, k, n_0, $1/\delta$ and $1/\xi$. The support size is $m_\star = n_0(L+1)+1$, and since $L = \widetilde{O}(kn_0^2/\xi^2)$ this gives $\log m_\star = O\big(\log \frac{kdn_0}{\xi}\big)$. That $m_\star$ depends on n_0 while the lower bound on n_0 depends on $m_\star$, through $p_\star$ and γ, is a fixed point rather than a circularity: the dependence is logarithmic, so the lower bound is of the form $n_0 \geq A\log^b(Cn_0)$ with A, b, C free of n_0, which is solved by any $n_0 \geq C'A\log^b(CA)$ and leaves the order $\widetilde{\Theta}\big(kM^2/(\delta^2\rho_k^2)\big)$ unchanged.

Proof. Proposition 6 gives $\|W_{n_0}\|_F \leq 1$ with probability at least $1 - \delta_I/3 - \xi = 1 - \delta/3$. On that event Theorem 12 applies with $T_0 = n_0$, Phase I failure probability $\delta/3$ and Phase II failure probability $\delta_{II} = 2\delta/3$, and its conclusion is (23), using $\beta + n - n_0 \geq n - n_0$ in the second term. The two failure probabilities sum to δ. □

Corollary 6 (Operator-norm form)**.** *Under the hypotheses of Theorem 13, with probability at least* $1 - \delta$*,*

$$\big\|Q_nQ_n^\top - V_kV_k^\top\big\|_2 \;\leq\; \sqrt{2}\, e\Big(\frac{\beta+1}{\beta+n-n_0}\Big)^{\alpha} \;+\; C_8\,\frac{M}{\rho_k}\sqrt{\frac{\Lambda}{n-n_0}}.$$

Proof. $\|X\|_2 \leq \|X\|_F$ for every matrix X. □

Remark. Corollary 6 shows that Theorem 13 implies Theorem 8 at no cost, since $\|\cdot\|_2 \leq \|\cdot\|_F$ pointwise, and with $\log(Mk/\rho_k\delta)$ improved to Λ. The converse implication is the expensive one, and is what this appendix exists to avoid.

F Lifting from finite support to i.i.d. bounded updates

Phase I of Algorithm 1 and of the warm-up branch of Algorithm 2 is run at a fixed clipping radius and a fixed noise multiplier. On the no-clipping event the update is therefore

$$C_t = \frac{1}{B}\sum_{i=B(t-1)+1}^{tB} A_i + G_t, \qquad G_t \sim \mathrm{GOE}_d(\sigma^2),$$

which does not depend on Q_{t-1}: the Phase I updates $C_1, \ldots, C_{T_0}$ are i.i.d., independent of the initialisation Z_0, have mean Σ, and satisfy $\|C_t - \Sigma\|_{(2,k)} \leq M$ almost surely. Write K for their common law. They are not finitely supported, so Theorem 11 does not apply to them directly; this section supplies the replacement of K by a finitely supported law with the same mean and the same $(2,k)$-bound. Phase II needs nothing from this section, since it uses only Assumption C.

F.1 The adaptive updates and the pre-update filtration

The warm-up updates above are i.i.d., and nothing more is needed from them. The adaptive updates of Algorithms 2 and 3 are not, and they enter Phase II, which uses only Assumption C. The next lemma records that they satisfy it, with respect to a filtration that places the step-t multiplier randomness in the conditioning.

Lemma 27 (The private updates satisfy Assumption C)**.** *The blocks I_t are disjoint and deterministic with $|I_t| = B/2$, the matrices $A_1, \dots, A_n$ are independent and symmetric with $\mathbb{E}[A_i] = \Sigma$ and $\|A_i - \Sigma\|_{(2,k)} \le M_A$ almost surely,*

$$C_t = \frac{2}{B}\sum_{i\in I_t} A_i + G_t, \qquad G_t \mid \mathcal{H}_t \sim \mathrm{GOE}_d(\sigma_t^2) \text{ conditioned on } \|G_t\|_{(2,k)} \le \mathsf{N}_B,$$

with σ_t measurable with respect to $\mathcal{H}_t = \sigma(\mathcal{F}_{t-1}, \mathcal{R}_t)$ and the block I_t independent of $\mathcal{H}_t$, where $\mathcal{R}_t$ is the randomness the algorithm draws at step t to form σ_t and $\mathcal{F}_t := \sigma(\mathcal{F}_{t-1}, \mathcal{R}_t, C_t)$. Define the pre-update filtration

$$\widetilde{\mathcal{F}}_{t-1} := \mathcal{H}_t = \sigma(\mathcal{F}_{t-1}, \mathcal{R}_t), \qquad t \ge 1, \qquad \widetilde{\mathcal{F}}_t = \sigma(\mathcal{F}_t, \mathcal{R}_{t+1}). \tag{24}$$

Then Assumption C holds with respect to $(\widetilde{\mathcal{F}}_t)_{t\ge 0}$, with $M = M_A + \mathsf{N}_B$.

Proof. Write μ_t for the law of $\frac{2}{B}\sum_{i\in I_t} A_i$, P_σ for the law of $\mathrm{GOE}_d(\sigma^2)$ conditioned on $\{\|G\|_{(2,k)} \le \mathsf{N}_B\}$, and set $K_t^{(\sigma)} := \mu_t * P_\sigma$.

$(\widetilde{\mathcal{F}}_t)$ is a filtration, and the process is adapted to it. Since $\mathcal{F}_t = \sigma(\mathcal{F}_{t-1}, \mathcal{R}_t, C_t)$ we have $\mathcal{F}_{t-1} \subseteq \mathcal{F}_t$ and $\mathcal{R}_t \subseteq \mathcal{F}_t$, hence $\widetilde{\mathcal{F}}_{t-1} = \sigma(\mathcal{F}_{t-1}, \mathcal{R}_t) \subseteq \mathcal{F}_t \subseteq \sigma(\mathcal{F}_t, \mathcal{R}_{t+1}) = \widetilde{\mathcal{F}}_t$, so $(\widetilde{\mathcal{F}}_t)_{t\ge 0}$ is increasing. Moreover C_t is $\mathcal{F}_t$-measurable and hence $\widetilde{\mathcal{F}}_t$-measurable, and Q_t, being a function of Q_0 and $C_1, \dots, C_t$, is $\widetilde{\mathcal{F}}_t$-measurable as well. Finally $Z_0 \in \mathcal{F}_0 \subseteq \widetilde{\mathcal{F}}_0$, so the requirement $\sigma(Z_0) \subseteq \mathcal{F}_0$ of Assumption C is met for $(\widetilde{\mathcal{F}}_t)$.

The conditional law. The blocks are disjoint and the A_i independent, so $\frac{2}{B}\sum_{i\in I_t} A_i$ is independent of $\mathcal{H}_t = \widetilde{\mathcal{F}}_{t-1}$; and G_t is conditionally P_{σ_t} given $\widetilde{\mathcal{F}}_{t-1}$ with σ_t being $\widetilde{\mathcal{F}}_{t-1}$-measurable. Hence

$$\mathcal{L}\big(C_t \mid \widetilde{\mathcal{F}}_{t-1}\big) = \mu_t * P_{\sigma_t} = K_t^{(\sigma_t)}, \qquad \sigma_t \in \widetilde{\mathcal{F}}_{t-1},$$

so (σ_t) is predictable for $(\widetilde{\mathcal{F}}_t)$. The indexing matters: σ_t is computed from $\mathcal{R}_t$ and so is *not* $\mathcal{F}_{t-1}$-measurable, and conditioning on $\mathcal{F}_{t-1}$ alone would return the mixture $\mathbb{E}[K_t^{(\sigma_t)} \mid \mathcal{F}_{t-1}]$; shifting to the pre-update filtration (24) places the fresh randomness of step t in the conditioning and recovers the single kernel $K_t^{(\sigma_t)}$. The filtration $(\widetilde{\mathcal{F}}_t)$ is strictly larger than the one generated by $C_1, \dots, C_t$; Assumption C and the whole of Sections E.5–E.6 are to be read with respect to it. This is legitimate because the computation below gives $\mathbb{E}[C_t \mid \widetilde{\mathcal{F}}_{t-1}] = \Sigma$ directly, which is stronger than Assumption C.1 for the smaller filtration, while every quantity the recursions require to be predictable — W_{t-1}, $\mathbf{1}_{t-1}$ and the good events G_{t-1}, all functions of Z_0 and $C_1, \dots, C_{t-1}$ — is $\widetilde{\mathcal{F}}_{t-1}$-measurable.

Mean and bound. Since $|I_t| = B/2$ and $\mathbb{E}[A_i] = \Sigma$ for every i, $\int C\,\mu_t(dC) = \frac{2}{B}\cdot\frac{B}{2}\,\Sigma = \Sigma$, for every block. The truncated Gaussian is centred: $\mathrm{GOE}_d(\sigma^2)$ is symmetric under $G \mapsto -G$ and the truncation event is invariant under that map, so P_σ has mean 0. Hence $\int C\,K_t^{(\sigma)}(dC) = \Sigma$ for every t and every σ. For the bound, almost surely

$$\Big\|\frac{2}{B}\sum_{i\in I_t} A_i - \Sigma\Big\|_{(2,k)} = \Big\|\frac{2}{B}\sum_{i\in I_t}(A_i - \Sigma)\Big\|_{(2,k)} \le \frac{2}{B}\sum_{i\in I_t}\|A_i - \Sigma\|_{(2,k)} \le \frac{2}{B}\cdot\frac{B}{2}\,M_A = M_A,$$

and $\|G\|_{(2,k)} \le \mathsf{N}_B$ holds P_σ-almost surely by construction, so $\|C - \Sigma\|_{(2,k)} \le M_A + \mathsf{N}_B = M$ holds $K_t^{(\sigma)}$-almost surely, again for every t and every σ. Neither computation sees σ or t. □

F.2 The reduction

The reduction uses four standard ingredients, which we record first: a matrix Bernstein inequality, its consequence for the empirical mean of a $(2,k)$-bounded law, the mean-preserving replacement that turns such a law into a finitely supported one, and a sequential coupling bound.

Lemma 28 (Matrix Bernstein input)**.** *Let $X_1, \dots, X_L \in \mathbb{R}^{d\times d}$ be independent, mean-zero, self-adjoint matrices with $\|X_\ell\|_2 \le R$ almost surely and $\big\|\sum_{\ell=1}^{L} \mathbb{E}[X_\ell^2]\big\|_2 \le \sigma^2$. Then for every $s > 0$,*

$$\mathbb{P}\Big(\Big\|\sum_{\ell=1}^{L} X_\ell\Big\|_2 \ge s\Big) \le 2d \exp\Big(-\frac{s^2/2}{\sigma^2 + Rs/3}\Big).$$

Proof. Theorem 6.1.1 of Tropp [2015] gives $\mathbb{P}\big(\lambda_{\max}(\sum_\ell X_\ell) \ge s\big) \le d\exp\big(-\frac{s^2/2}{\sigma^2+Rs/3}\big)$; applying it to $\sum_\ell X_\ell$ and to $-\sum_\ell X_\ell$, whose summands satisfy the same hypotheses, and taking a union bound gives the factor $2d$. □

Lemma 29 (Conditional empirical-mean bound)**.** *Let K be a probability measure on symmetric $d\times d$ matrices with $\int C\,K(dC) = \Sigma$ and $\|C - \Sigma\|_{(2,k)} \le M$ for K-almost every C. Let $\widehat{C}_1, \dots, \widehat{C}_L$ be i.i.d. K and put $\widehat{\Sigma}_L := \frac{1}{L}\sum_{\ell\le L}\widehat{C}_\ell$. Then for every $u > 0$,*

$$\mathbb{P}\big(\|\widehat{\Sigma}_L - \Sigma\|_2 \ge u\big) \le 2d\exp\Big(-\frac{Lu^2/2}{M^2 + Mu/3}\Big).$$

Consequently, if $0 < \alpha \le \frac{1}{2}$, $0 < \beta < 1$ and

$$L \;\ge\; C_0\,\frac{k}{\alpha^2}\,\log\frac{2d}{\beta}$$

for a sufficiently large universal constant C_0, then

$$\mathbb{P}\Big(\|\widehat{\Sigma}_L - \Sigma\|_2 > \frac{\alpha M}{(1-\alpha)\sqrt{k}}\Big) \le \beta.$$

Proof. Set $X_\ell := \widehat{C}_\ell - \Sigma$; these are independent, mean-zero and self-adjoint, and $\|X_\ell\|_2 \le \|X_\ell\|_{(2,k)} \le M$. Hence $X_\ell^2 \preceq M^2 I_d$ and $\big\|\sum_{\ell\le L}\mathbb{E}[X_\ell^2]\big\|_2 \le LM^2$. Lemma 28 applied with $R = M$, $\sigma^2 = LM^2$ and $s = Lu$ gives the first claim. Substituting $u = \alpha M/((1-\alpha)\sqrt{k})$ and using $\alpha \le \frac{1}{2}$ gives the second, after increasing C_0. □

Lemma 30 (Mean-preserving finite-support replacement)**.** *Let $\widehat{C}_1, \dots, \widehat{C}_L$ be symmetric matrices with $\|\widehat{C}_\ell - \Sigma\|_{(2,k)} \le M$ for every $\ell \le L$, and put $\widehat{\Sigma}_L := \frac{1}{L}\sum_{\ell\le L}\widehat{C}_\ell$. Fix $0 < \alpha \le \frac{1}{2}$ and assume $\|\widehat{\Sigma}_L - \Sigma\|_2 \le \alpha M/((1-\alpha)\sqrt{k})$. Define*

$$\widehat{C}_0 := \Sigma - \frac{1-\alpha}{\alpha}\big(\widehat{\Sigma}_L - \Sigma\big), \qquad \widehat{K}^\sharp := \alpha\,\delta_{\widehat{C}_0} + (1-\alpha)\frac{1}{L}\sum_{\ell=1}^{L}\delta_{\widehat{C}_\ell}.$$

Then $\int C\,\widehat{K}^\sharp(dC) = \Sigma$, $\|C - \Sigma\|_{(2,k)} \le M$ for $\widehat{K}^\sharp$-almost every C, and $\widehat{K}^\sharp$ has support of cardinality at most $L + 1$.

Proof. The mean identity is immediate:

$$\int C\,\widehat{K}^\sharp(dC) = \alpha\widehat{C}_0 + (1-\alpha)\widehat{\Sigma}_L = \alpha\Sigma - (1-\alpha)\big(\widehat{\Sigma}_L - \Sigma\big) + (1-\alpha)\widehat{\Sigma}_L = \Sigma.$$

The original atoms obey the $(2,k)$-bound by hypothesis, and for the correction atom

$$\|\widehat{C}_0 - \Sigma\|_{(2,k)} = \frac{1-\alpha}{\alpha}\|\widehat{\Sigma}_L - \Sigma\|_{(2,k)} \le \frac{1-\alpha}{\alpha}\sqrt{k}\,\|\widehat{\Sigma}_L - \Sigma\|_2 \le M$$

by the assumed bound. The support bound is clear from the definition. □

Lemma 31 (Sequential coupling bound)**.** *Let P and $\mathsf{P}^\sharp$ be the path laws of two predictable processes on T_0 steps. Suppose that at every time t, whenever the two histories agree up to time $t-1$, the two one-step transition laws admit a coupling under which the next updates disagree with probability at most α. Then $d_{\mathrm{TV}}(\mathsf{P}, \mathsf{P}^\sharp) \le T_0\alpha$, and consequently $\mathsf{P}(A) \le \mathsf{P}^\sharp(A) + T_0\alpha$ for every measurable event A depending on the update path.*

Proof. Build the two processes recursively, using the assumed one-step coupling while the histories agree and continuing arbitrarily once they differ. By a union bound the probability that the paths ever disagree is at most $T_0\alpha$, and total variation distance is bounded by the disagreement probability under any coupling. □

Proposition 6 (Finite-support reduction for i.i.d. updates)**.** *Let $C_1, \dots, C_{T_0}$ be i.i.d. symmetric matrices with common law K, mean Σ and $\|C_t - \Sigma\|_{(2,k)} \le M$ almost surely, independent of Z_0, and let Q_t be the Oja iterates with constant Phase I step size $\eta_t \equiv \eta$. Let $\delta, \xi \in (0,1)$ and set*

$$\alpha := \frac{\xi}{4T_0}, \qquad \beta := \frac{\xi}{4T_0},$$

and let L be an integer with

$$L \;\ge\; C_0\,\frac{kT_0^2}{\xi^2}\,\log\frac{8dT_0}{\xi}.$$

Assume that the hypotheses of Theorem 11 hold uniformly for every deterministic finite support $\mathcal{A}_\star \subseteq \mathbb{R}^{d\times d}$ with

$$|\mathcal{A}_\star| \;\le\; m_\star := T_0(L+1)+1,$$

with the same constants M, η, γ, δ. Then

$$\mathbb{P}\big(\|W_{T_0}\|_F > 1\big) \le \frac{\delta}{3} + \xi.$$

Proof. The atom array, drawn in advance. On an auxiliary probability space, independent of $C_1, \dots, C_{T_0}$ and of Q_0, draw

$$\widehat{C}_{t,1}, \dots, \widehat{C}_{t,L} \overset{\text{i.i.d.}}{\sim} K, \qquad t \le T_0,$$

all independent, and write $\mathcal{A}$ for the array and $\{\widehat{C}_{t,\ell} : \ell \le L\}$ for the *time-t row*. The rows are independent of one another; in particular the time-t row is independent of all rows at times $s < t$.

Put $\widehat{\Sigma}_t := \frac{1}{L}\sum_{\ell\le L}\widehat{C}_{t,\ell}$ and

$$\mathcal{G}_t := \Big\{\big\|\widehat{\Sigma}_t - \Sigma\big\|_2 \le \frac{\alpha M}{(1-\alpha)\sqrt{k}}\Big\}.$$

Since K has mean Σ and is M-bounded in $\|\cdot\|_{(2,k)}$, Lemma 29 applies: the lower bound on L is $L \ge C_0\alpha^{-2}k\log(2d/\beta)$ after absorbing the factor 16 into C_0, and therefore

$$\mathbb{P}(\mathcal{G}_t^c) \le \beta \qquad \text{for every } t \le T_0. \tag{25}$$

On $\mathcal{G}_t$ define the correction atom and the replacement law of Lemma 30,

$$\widehat{C}_{t,0} := \Sigma - \frac{1-\alpha}{\alpha}\big(\widehat{\Sigma}_t - \Sigma\big), \qquad \widehat{K}_t^\sharp := \alpha\,\delta_{\widehat{C}_{t,0}} + (1-\alpha)\frac{1}{L}\sum_{\ell=1}^{L}\delta_{\widehat{C}_{t,\ell}},$$

and on $\mathcal{G}_t^c$ set $\widehat{K}_t^\sharp := \delta_\Sigma$. By Lemma 30 in the first case and trivially in the second, *in both cases*

$$\int C\,\widehat{K}_t^\sharp(dC) = \Sigma, \qquad \widehat{K}_t^\sharp\big(\{\|C-\Sigma\|_{(2,k)} \le M\}\big) = 1, \qquad \big|\mathrm{supp}(\widehat{K}_t^\sharp)\big| \le L+1. \tag{26}$$

The auxiliary process. Define $C_t^\sharp \sim \widehat{K}_t^\sharp$ independently across t, using randomisation independent of $\mathcal{A}$, and set $Q_t^\sharp := \mathrm{QR}(Q_{t-1}^\sharp + \eta C_t^\sharp Q_{t-1}^\sharp)$. Let $\mathcal{F}_t^\sharp := \sigma(\mathcal{A}, C_1^\sharp, \dots, C_t^\sharp)$. By (26), *with no conditioning on a good event*,

$$\mathbb{E}\big[C_t^\sharp \mid \mathcal{F}_{t-1}^\sharp\big] = \Sigma, \qquad \big\|C_t^\sharp - \Sigma\big\|_{(2,k)} \le M \ \text{ a.s.},$$

and $C_t^\sharp$ takes values in

$$\mathcal{A}_\star^\sharp := \{\Sigma\} \cup \bigcup_{t\le T_0}\mathrm{supp}\big(\widehat{K}_t^\sharp\big), \qquad |\mathcal{A}_\star^\sharp| \le T_0(L+1)+1 = m_\star,$$

which is $\mathcal{A}$-measurable.

Applying Theorem 11. Condition on $\mathcal{A}$. The auxiliary process is driven by randomisation independent of $\mathcal{A}$, so its conditional law given $\mathcal{A}$ is again that of a process with conditional mean Σ and $\|C_t^\sharp - \Sigma\|_{(2,k)} \le M$, now with the *deterministic* finite support $\mathcal{A}_\star^\sharp$ of cardinality at most $m_\star$. The initialisation Z_0 is independent of $\mathcal{A}$, so its law is unchanged by the conditioning. Theorem 11 therefore applies and gives

$$\mathbb{P}\Big(\big\|U^\top Q_{T_0}^\sharp (V^\top Q_{T_0}^\sharp)^{-1}\big\|_F > 1 \;\Big|\; \mathcal{A}\Big) \le \frac{\delta}{3},$$

and averaging over $\mathcal{A}$,

$$\mathbb{P}\Big(\big\|U^\top Q_{T_0}^\sharp (V^\top Q_{T_0}^\sharp)^{-1}\big\|_F > 1\Big) \le \frac{\delta}{3}. \tag{27}$$

Coupling. We realise the two processes on a common space and verify the hypothesis of Lemma 31 with per-step disagreement probability $\alpha + \beta = \xi/(2T_0)$. Draw J_t uniformly from $\{1, \dots, L\}$ and $B_t \sim \mathrm{Bernoulli}(\alpha)$, independently of $\mathcal{A}$ and of everything else, and realise *unconditionally*

$$C_t := \widehat{C}_{t,J_t}, \qquad C_t^\sharp := \begin{cases} \widehat{C}_{t,0}, & \text{on } \mathcal{G}_t \text{ with } B_t = 1, \\ \widehat{C}_{t,J_t}, & \text{on } \mathcal{G}_t \text{ with } B_t = 0, \\ \Sigma, & \text{on } \mathcal{G}_t^c. \end{cases}$$

That C_t is realised from the array on $\mathcal{G}_t^c$ as well as on $\mathcal{G}_t$ is what keeps its law exact: conditioning the realisation on $\mathcal{G}_t$, an event about the time-t row, would bias the row.

The two marginals. The common past $\mathcal{F}_{t-1}$ is a function of Q_0, of $(J_s, B_s)_{s<t}$ and of the rows at times $s < t$, so the time-t row is independent of $\mathcal{F}_{t-1}$. Two conditionings must be kept apart. Given $\mathcal{A}$ as well as the past, C_t is a uniform draw from the L realised atoms of row t, so its law is the *empirical* measure of that row,

$$\mathcal{L}\big(C_t \mid \mathcal{F}_{t-1}, \mathcal{A}\big) = \frac{1}{L} \sum_{\ell \le L} \delta_{\widehat{C}_{t,\ell}},$$

which is random and is not K. Averaging over the time-t row recovers the law: since J_t is uniform and independent and $\widehat{C}_{t,1}, \dots, \widehat{C}_{t,L}$ are i.i.d. K and independent of $\mathcal{F}_{t-1}$, for every bounded measurable f

$$\mathbb{E}\big[f(C_t) \mid \mathcal{F}_{t-1}\big] = \frac{1}{L} \sum_{\ell \le L} \mathbb{E}\big[f(\widehat{C}_{t,\ell})\big] = \int f \,\mathrm{d}K, \qquad \text{that is} \qquad \mathcal{L}\big(C_t \mid \mathcal{F}_{t-1}\big) = K,$$

which is the law of the original update; and $\mathcal{L}(C_t^\sharp \mid \mathcal{A}, \mathcal{F}_{t-1}) = \widehat{K}_t^\sharp$ by (26), in both of its cases. So the coupling has the two intended marginals, and in particular its $\sharp$-marginal is the process to which (27) applies.

Disagreement. The two updates differ only on $\{B_t = 1\} \cup \mathcal{G}_t^c$. The first has probability α by construction; the second has probability at most β by (25), and $\mathcal{G}_t$ is an event about the time-t row, hence independent of $\mathcal{F}_{t-1}$. Note that it is *not* true that $\mathcal{A}$ is independent of $\mathcal{F}_{t-1}$ — the past is built from the earlier rows — and only the independence of the time-t row from the past is used.

By Lemma 31, $d_{\mathrm{TV}}(\mathsf{P}, \mathsf{P}^\sharp) \le T_0 \cdot \frac{\xi}{2T_0} = \frac{\xi}{2}$, and with (27),

$$\mathbb{P}\big(\|U^\top Q_{T_0} (V^\top Q_{T_0})^{-1}\|_F > 1\big) \le \frac{\delta}{3} + \frac{\xi}{2} \le \frac{\delta}{3} + \xi. \qquad \square$$

Remark (what $m_\star$ costs, and the condition on L). Only $\log m_\star$ enters Theorem 11, through γ and through $p_1 = \log \frac{12T_0(m_\star+1)}{\delta}$. With $L = \widetilde{O}(kT_0^2/\xi^2)$,

$$\log m_\star = O\Big(\log \frac{kdT_0}{\xi}\Big),$$

so the order of T_0 is unchanged.

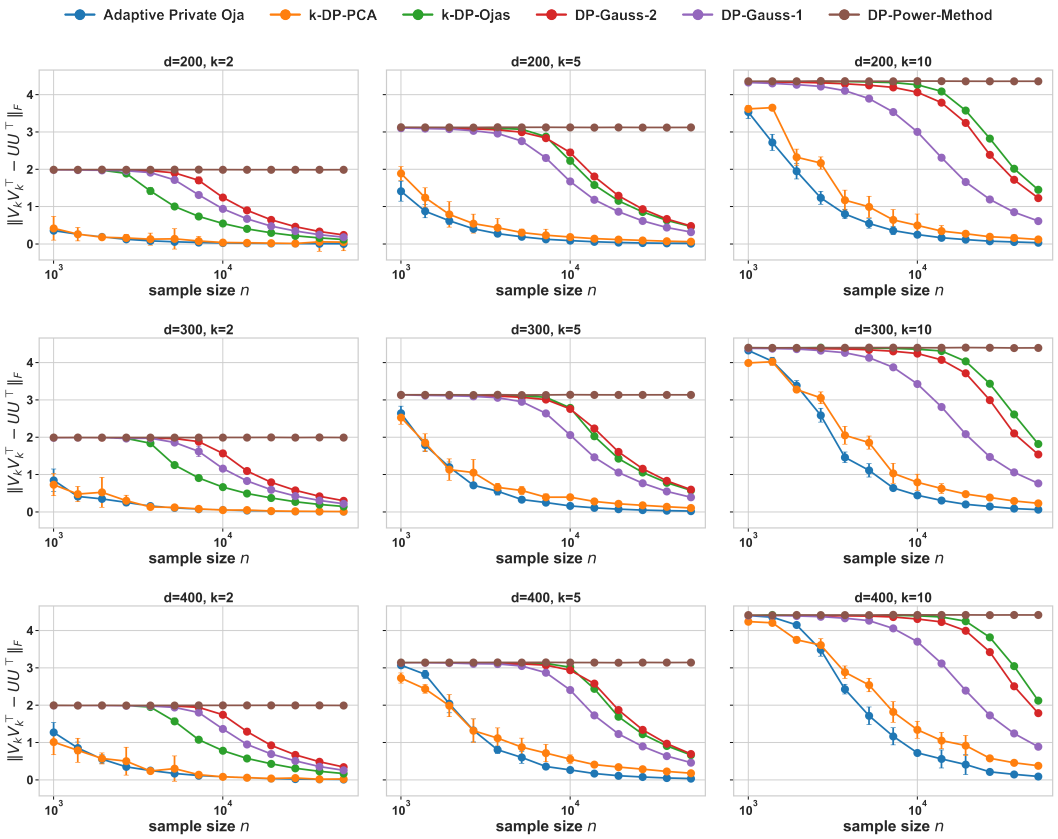


Figure 3: Comparison of AdaDPO with baselines for varying k and d on the spiked covariance model. We plot the mean over 10 trials, with the bars representing the standard deviation.

Remark. The array must be drawn in advance, and for an adaptive update it cannot instead be drawn along the path: the product class $\mathcal{E}_{r,\ell}$ of Theorem 11 must be deterministic, because the centring step of Proposition 2 needs E to be $\mathcal{F}_{t-1}$-measurable while the class elements are built from noise factors at times after t. Conditioning on a path-dependent array would make the class deterministic but would destroy the conditional mean, since the time-s atoms ($s > t$) are drawn from $K_s(\omega_{s-1}, \cdot)$ and ω_{s-1} contains C_t. Appendix E of Huang et al. draws its atoms from a single fixed law, so the obstruction does not arise there and their construction does not transfer verbatim.

G Experiments

This appendix documents the implementation choices and algorithmic details omitted in the main paper. All experiments were run under the replace-model (ε, δ)-DP setting with $\varepsilon = 1$ and $\delta = 0.01$, unless explicitly stated otherwise. In Figure 1a we give additional experiment results that had to be omitted from the main paper due to space constraints.

In Section 3.1 we compare the performance of AdaDPO to k-DP-PCA and k-DP-Ojas [Düngler and Sanyal, 2025], two adapted versions of the DP-Gauss algorithms of Dwork et al. [2014], we refer to as `DP-Gauss-1` and `DP-Gauss-2` respectively, and an adapted version of the noisy power method [Hardt and Price, 2014].

Given a stream of matrices $\{A_i\}$ and a clipping threshold β, chose n according to the input distribution, `DP-Gauss-1` first rescales each matrix so that its trace is at most β^2:

$$\tilde{A}_i = A_i \cdot \min\{1, \beta^2 / \operatorname{Tr}(A_i)\}.$$

It then forms the clipped sum $X = \sum_i \tilde{A}_i$ and applies the Gaussian mechanism,

$$X' = X + E,$$

where E is symmetric and its upper-triangular entries, including the diagonal, are sampled independently from $\mathcal{N}(0, \Delta_1^2 \mathbf{I}_d)$, with

$$\Delta_1 = \frac{\beta^2 \sqrt{2\log(1.25/\delta)}}{\varepsilon}.$$

The algorithm finally computes an eigendecomposition of X' and releases its top k eigenvectors. The second variant, `DP-Gauss-2`, uses the same clipping and summation steps to obtain X. It then computes the top k eigenvectors V_k of X, and privatizes the eigengap by setting

$$g_k = \lambda_k - \lambda_{k+1} + z, \qquad z \sim \mathrm{Lap}(2/\varepsilon).$$

Next, it applies the Gaussian mechanism to V_k:

$$W = V_k V_k^\top + E,$$

where E is symmetric with independent upper-triangular entries, including the diagonal, sampled from $\mathcal{N}(0, \Delta_2^2 \mathbf{I}_d)$, and

$$\Delta_2 = \frac{\beta^2 \left(1 + \sqrt{2\log(1/\delta)}/\varepsilon\right)}{|g_k - 2\left(1 + \log(1/\delta)/\varepsilon\right)|}.$$

Since the added noise may destroy orthogonality, we perform a final eigendecomposition of W and release the resulting top k eigenvectors. If $g_k \leq 0$, the procedure is not differentially private, even though it follows Algorithm 2 of Dwork et al. [2014]. A fully compliant version, also described in that work, would use the PTR mechanism, but this incurs additional privacy loss. For simplicity and to make the implementation more flexible, we instead resample the noise whenever $g_k \leq 0$.Finally, `DP-Power-Method` clips the matrices using both the square root of the trace and the trace of the square root of the diagonal. In the rank-one case $A = aa^\top$, with $a \in \mathbb{R}^d$, these correspond respectively to clipping $\|a\|_2 \leq \beta$ and $\|a\|_1 \leq \alpha$, where β is chosen as in the DP-Gauss baselines. The algorithm then sums the clipped matrices and runs the noisy power method, using the version described in Nicolas et al. [2024]. The Gaussian noise added at each power-iteration is scaled by an additional factor of $\beta\alpha$.

`k-DP-PCA` and `k-DP-Ojas` are developed for stochastic data, and therefore need no adaptation. However, they both require a learning rate schedule and k-DP-PCA requires a batch size. We discuss the hyperparameter tuning for this together with the hyperparameter tuning done for ADADPO.

G.1 Data Generation

We generate synthetic data from a spiked covariance model. Each matrix $A_i \in \mathbb{R}^{d\times d}$ contains a deterministic rank-k signal component together with a random noise component that makes the sample full rank. For $k = 1$, we draw

$$x_i = s_i + n_i,$$

where

$$s_i \sim \mathrm{Unif}\{\lambda_1 v, -\lambda_1 v\},$$

$v \in \mathbb{R}^d$ is a unit vector, $\lambda_1 \in \mathbb{R}$ is a signal-strength parameter, and

$$n_i \sim \mathcal{N}(0, \sigma^2 I_d).$$

We then set $A_i = x_i x_i^\top$. The quantities λ_1 and σ are supplied as inputs to the sampler, while v is obtained by sampling a standard Gaussian vector in $\mathbb{R}^d$ and normalizing it.

For $k > 1$, we use a different construction. We first sample a matrix $V \in \mathbb{R}^{d\times k}$ with i.i.d. standard Gaussian entries and orthonormalize its columns via Gram–Schmidt, obtaining $V_k \in \mathbb{R}^{d\times k}$. We then define

$$A_i = V_k \Lambda V_k^\top + z_i z_i^\top,$$

where

$$z_i \sim \mathcal{N}(0, \sigma^2 I_d),$$

and $\Lambda \in \mathbb{R}^{k\times k}$ is diagonal with user-specified eigenvalues. This $k > 1$ construction is not a direct generalization of the $k = 1$ sampling scheme. Indeed, independently drawing k vectors as in the rank-one case and summing their outer products would produce a mixture of Gaussians rather than a single spiked covariance model. We therefore fix the signal subspace and impose the rank-k structure deterministically through $V_k \Lambda V_k^\top$.

For `DP-Gauss-1` and `DP-Gauss-2`, we set

$$\beta = C\sqrt{\lambda_1} + \sigma\sqrt{d\log(n/\zeta)},$$

where n is the sample size and $1-\zeta$ is the target probability that no clipping occurs. Throughout all experiments, we use $\zeta = 0.01$ for every method. ADADPO, k-DP-PCA and k-DP-Ojas require the parameters K and a from Assumption A. For the synthetic model above, one has $a = 1$ and $K = O(1)$, so we take $a = 1$ and $K = 1$ in all experiments.

G.2 Hyperparameter Tuning

General protocol. Hyperparameters are tuned by running each candidate configuration across the full sweep grid, averaging the target subspace error over all settings and trials, and selecting the configuration with minimum mean error. Exact ties are broken uniformly at random.

Learning-rate schedules. Let T denote the number of update steps, let n denote the total sample size, let $\tilde{\lambda}_1$ denote the leading population eigenvalue used by the implementation, and let

$$\Delta_1 := \tilde{\lambda}_1 - \tilde{\lambda}_2$$

denote the leading eigengap. The learning-rate candidates used in the experiments are as follows.

1. For an offset $c > 0$, a cutoff fraction $\rho \in (0, 1]$, and an extra decay power $p \geq 0$, define
$$\tau := \lceil \rho T \rceil .$$
The schedule is
$$\eta_t = \frac{1}{t+c} \qquad \text{for } 1 \leq t \leq \tau,$$
and
$$\eta_t = \frac{1}{t+c}\left(\frac{\tau+c}{t+c}\right)^p \qquad \text{for } t > \tau.$$
The two step-decay candidates used in the experiments are ($\rho = \frac{1}{2}$, $p = \frac{1}{2}$) and ($\rho = \frac{1}{3}$, $p = 2$).
2. `lambda1_spiked`**.**
$$\eta_t = \frac{1}{20\,\sigma\,\tilde{\lambda}_1 + \Delta_1\, t/\log n}, \qquad t = 1, \ldots, T.$$
3. $$\eta_t = \frac{1}{20\,\sigma\,\tilde{\lambda}_1\, s_d + \Delta_1\, t/\log n}, \qquad t = 1, \ldots, T,$$
where s_d is a dimension factor. In the spiked-data implementation, $s_d = d$ is replaced by the ambient eigenvalue-array length; in the restrained-Gaussian implementation, $s_d = d$ exactly.

Component-wise k-DP-PCA schedules. For k-DP-PCA, the implementation maintains a separate schedule for each deflation step $j = 1, \ldots, k$. Let

$$\Delta_j := \tilde{\lambda}_j - \tilde{\lambda}_{j+1},$$

the component-wise learning rate is

$$\eta_t^{(j)} = \frac{1}{b_j + \Delta_j\, t/\log n},$$

where

$$b_j \in \{20\,\sigma\,\tilde{\lambda}_j, 20\,\sigma\,\tilde{\lambda}_j\, d\}.$$

Adaptive Private Oja tuning. ADADPO is tuned over:

- batch rules $B \in \{\sqrt{n}, n/\log n\}$ projected to an even batch size and then increased if necessary to satisfy the range-estimation feasibility condition;
- learning-rate offset 10.0;
- learning-rate families as described above

k-DP-PCA tuning. k-DP-PCA is tuned independently over:

- batch rules $B \in \{\sqrt{n}/k, n/(\log n \cdot k)\}$ together with an internal feasibility correction enforcing the truncation/range requirements of the implementation;
- the same learning-rate family set as Adaptive Private Oja.

The final sweep plots use the best Adaptive Private Oja configuration and best k-DP-PCA configuration selected independently in this way.

k-DP-Ojas tuning. For k-DP-Ojas, we found empirically that a simple decreasing schedule, independent of the eigenvalues, works well, and therefore set

$$\eta_j = \frac{1}{1+j}, \qquad j \in [n],$$

for all k iterations of k-DP-Ojas.

Finetuning in Tangent Space The hybrid method uses ADADPO for the warm start and then runs TADADPO on the remaining matrices. The two methods are both finetuned with the options described above, and allowed to have different batch sizes and learning-rates

G.3 Error Metric

All reported curves use the projection Frobenius error

$$\|UU^\top - V_k V_k^\top\|_F,$$

where U is the estimated k-dimensional subspace and V_k is the true population top-k eigenspace.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We introduce a non-deflation-based adaptive private Oja method with optimal dependence on d and k, removal of the prior exponential eigengap-dependent sample-size requirement, and a tangent-space refinement improving spectral dependence in the flat-tail spiked covariance setting.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the work and open questions in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper states the assumptions for each main result in the theorem hypotheses and provides complete proofs in the appendix (Section C), including the privacy and utility proofs for the fixed-radius, adaptive, and tangent-refinement algorithms.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe our algorithm in Section 3 and provide the experimental details in Section G.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will release the code publicly after we have cleaned it.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section G for full details and Section 3.1 for a general overview.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We ran a minimum of 10 trials for each experiment and included the variance of results in the plots.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments were run locally on a MacBook M3 Pro.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper is primarily theoretical. The experiments use synthetic data only, so they do not involve private or personal information. We therefore foresee no negative social impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not anticipate any significant risk that this work will be misused.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We have not released any new assets as part of this work.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not use crowdsourcing and does not involve human-subject research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve human subjects nor crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used only for routine assistance, such as writing, editing, and code refinement, and did not influence the scientific contributions or results.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.