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. In addition, our method removes the exponential dependence on the eigengap that appears in the sample-size lower bound required by prior utility guarantees, and improves the dependence on spectral parameters to match 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 7 and 8).

Preprint.

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, because their deflation-based algorithm estimates the eigenvectors one at a time, with parameters reselected at each step, it is 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 7) that achieves the optimal dependence on k in sample complexity. Unlike deflation-based methods such as k-DP-PCA, which reduce k-PCA to k private one-dimensional problems, our algorithm estimates the target k-dimensional subspace jointly. Moreover, its adaptive privacy mechanism uses a single block-level scale estimate and clipping center, rather than k separate range estimators, before adding calibrated symmetry-preserving Gaussian noise. 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 6. 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 $\lambda_k - \lambda_{k+1} > 0$.

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

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

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

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

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. 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 $\mathcal{E}$,

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

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

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

A.4 $\max_{\|u\|_2=1, \|v\|_2=1} \mathbb{E}\left[\exp\left(\left(\frac{\left|u^\top (A_i - \Sigma) v\right|^2}{K^2 \lambda_1^2 \|H_u\|_2}\right)^{1/2a}\right)\right] \leq 2$*, 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 9 and Algorithm 7. 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. 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 $A_1, \dots, A_n$ satisfy Assumption A. Set*

$$R = C\lambda_1\sqrt{dk}\big(K\gamma\log^a(ndk/\zeta) + 1\big), \qquad B = n/\log n, \qquad T = \lfloor n/B \rfloor.$$

Using appropriate learning rates, if

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

Algorithm 1 Private Oja's

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

Set $T \leftarrow \lfloor n/B \rfloor$, $\sigma \leftarrow 2R\sqrt{2\log(1.25/\delta)}/(B\varepsilon)$.
Choose $Q_0' \in \mathbb{R}^{d\times k}$ uniformly at random and set $Q_0 \leftarrow \mathrm{QR}(Q_0')$.
for $t = 1, \ldots, T$ **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.

then with probability at least $1-\zeta$*, Algorithm 1 outputs* Q_T *satisfying*

$$\left\|Q_TQ_T^\top - V_kV_k^\top\right\|_F \le \tilde{O}\left(\frac{\lambda_1\sqrt{kV}}{(\lambda_k-\lambda_{k+1})\sqrt{n}} + \frac{\lambda_1(K\gamma+1)kd}{\varepsilon(\lambda_k-\lambda_{k+1})n}\right).$$

The first term is the stochastic error of Oja's algorithm, and the second term 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* $M = \tilde{O}(\lambda_1 d)$*,* $V = O(d)$*,* $K = O(1)$*,* $a = 1$*, and* $\gamma = O(1)$*. Thus, applying Theorem 1 on this high-probability event gives*

$$\left\|Q_TQ_T^\top - V_kV_k^\top\right\|_F \le \tilde{O}\left(\frac{\lambda_1}{(\lambda_k-\lambda_{k+1})}\left(\frac{\sqrt{kd}}{\sqrt{n}} + \frac{dk}{\varepsilon n}\right)\right).$$

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* $s_i \sim \mathrm{Unif}\{v, -v\}$*,* $n_i \sim \mathcal{N}(0, \sigma^2 I_d)$*, and* $v \in \mathbb{R}^d$ *is unit norm. Let* $A_i = x_ix_i^\top$*, so* $\Sigma = \mathbb{E}[A_i] = vv^\top + \sigma^2 I_d$*. 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

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

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 size B, privacy parameters (ε, δ), learning rates $\{\eta_t\}_{t=1}^{\lfloor n/B \rfloor}$, and failure probability ζ

Choose $Q_0' \in \mathbb{R}^{d\times k}$ uniformly at random, $Q_0 \leftarrow \mathrm{QR}[Q_0']$
for $t = 1, 2, \dots, T = \lfloor n/B \rfloor$ **do**
  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 $\overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma_t^2)$, $\quad N_{\mathrm{sym},t} \sim GOE_k(\sigma_t^2)$
  $Q_t \leftarrow \mathrm{QR}\left[Q_{t-1} + \eta_t\left(\mathsf{P}_{\mathcal{S}_{Q_{t-1}}}(\bar{G}_t) + Q_{t-1}N_{\mathrm{sym},t} + (I - Q_{t-1}Q_{t-1}^\top)Z_t\right)\right]$
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 and setting the batch size to*

$$B = n/\log(n), \qquad T = \lfloor n/B \rfloor,$$

then for

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

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

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

Corollary 2 (Rank-k spiked covariance)**.** *Let $V_k \in \mathbb{R}^{d\times k}$ have orthonormal columns, let $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_k)$ with $\lambda_1 \ge \cdots \ge \lambda_k > 0$, and let $X_i = V_k\Lambda^{1/2} + \sigma Z_i$, $(Z_i)_{ab} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Set $A_i = X_i X_i^\top$, then $\Sigma = \mathbb{E}[A_i] = V_k\Lambda V_k^\top + k\sigma^2 I_d$. Running Algorithm 2 with input $A_1, \dots, A_n$ outputs Q_T with high probability satisfying*

$$\|Q_T Q_T^\top - V_k V_k^\top\|_F \le \tilde{O}\left(\frac{\sigma\sqrt{(\lambda_1 + k\sigma^2)kd}}{\lambda_k\sqrt{n}} + \frac{Kdk\,\sigma\sqrt{\lambda_1 + \sigma^2 k}}{\varepsilon\lambda_k n}\right)$$

Remark. The privacy term in Theorem 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 6). 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.

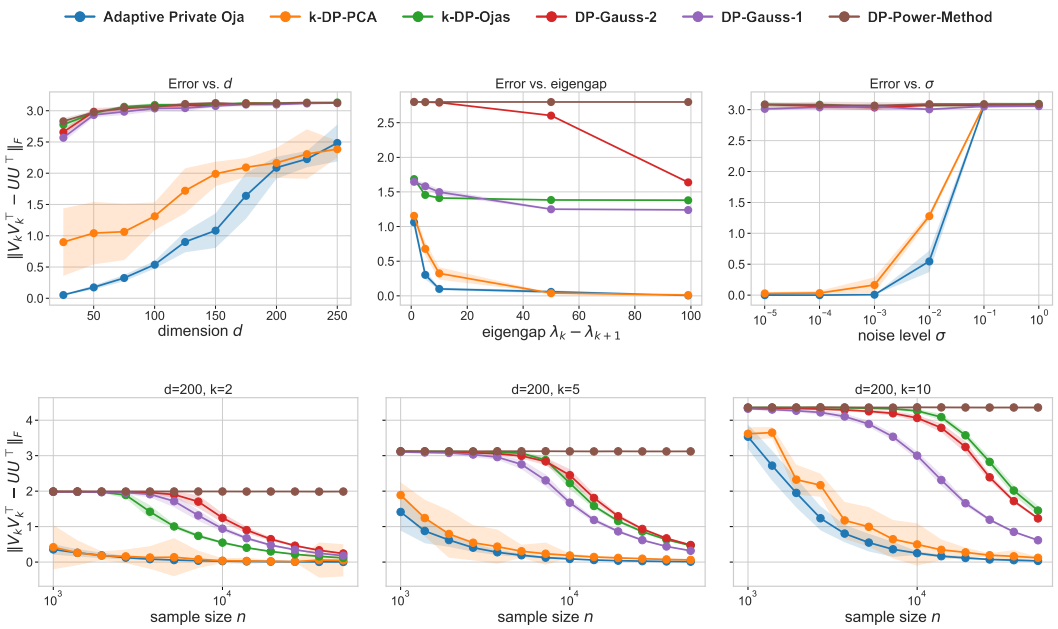

Figure 1: Comparison of ADADPO with baselines on the spiked covariance model. We plot the mean over 10 trials, with shaded regions representing 95% confidence intervals. We set $\varepsilon = 1$, $\delta = 0.01$. For the top row figures we set $k = 5$.

[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 $\sqrt{\sum_i \|C_{i,:}\|_1^2} \le 1$. 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 $\|x_i\|_2 \le R$ 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 $\|a\|_2 \le R$ (same R as for `DP-Gauss`) and $\|a\|_1 \le R'$, 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 $\|x_i\|_1 \le R'$ 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

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 $\{\eta_t\}_{t=1}^{\lfloor n/B \rfloor}$, 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 $\sigma_t \leftarrow \frac{8r_t\sqrt{dk}}{B\varepsilon}\sqrt{2\log(2.5/\delta)}$
  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

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 F.

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 projected update. In the Gaussian spiked covariance model of Cai et al. [2024], however, the data have additional structure. Namely, $A_i = x_ix_i^\top$ with $x_i \sim \mathcal{N}(0, \Sigma)$, and $\Sigma = U\Theta U^\top + \sigma^2 I_d$, 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 flat-tail structure permits a sharper dependence on the spectrum of Σ. The minimax lower bound (Theorem 8) scales as $\frac{\sqrt{\lambda_1\lambda_{k+1}}}{\lambda_k - \lambda_{k+1}}\left(\sqrt{\frac{dk}{n}} + \frac{dk}{\varepsilon n}\right)$, up to logarithmic factors. By contrast, applying Theorem 2 directly to $A_i = x_ix_i^\top$ yields

$$\left\|Q_TQ_T^\top - V_kV_k^\top\right\|_F \le \widetilde{O}\left(\frac{\lambda_1}{\lambda_k - \lambda_{k+1}}\sqrt{\frac{dk}{n}} + \frac{K\lambda_1}{\lambda_k - \lambda_{k+1}}\frac{dk}{\varepsilon n}\right).$$

Thus ADADPO has the correct dependence on the statistical and privacy parameters, but its spectral factor is $\lambda_1/(\lambda_k - \lambda_{k+1})$ rather than the sharper $\sqrt{\lambda_1\lambda_{k+1}}/(\lambda_k - \lambda_{k+1})$ available in the flat-tail Gaussian family.

The reason is geometric. Once an iterate Q is already close to the target subspace, not every component of the update A_iQ contributes equally to improving the subspace estimate. The component lying inside $\mathrm{span}(Q)$ changes the coordinates of the basis Q, but not the subspace it spans; the subspace is changed by the orthogonal, or tangent, component. Writing $\Pi_Q := I - QQ^\top$, this tangent component is Π_QA_iQ. Accordingly, our second algorithm uses ADADPO only as a warm-up phase. After obtaining a private estimate Q_0 within constant subspace error of the true top-k eigenspace, it runs a second-stage refinement on fresh samples using only the projected tangent updates

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

This removes the component of the update that merely rotates the current basis and focuses the private mean estimation on the residual correction that moves the subspace.

The benefit of the tangent update can be seen through the scale of the residual fluctuations. For $Q \in \mathbb{S}_{d,k}$, define

$$e(Q) = \left\|QQ^\top - V_kV_k^\top\right\|_F,$$

where $V_k V_k^\top$ is the projector onto the target eigenspace. The centered tangent fluctuation is

$$Z_Q(A) := \Pi_Q(A - \Sigma)Q.$$

We measure its variance by

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

This parameter controls concentration of the tangent minibatch mean. We also define the one-column tangent scale

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

where $q_1, \ldots, q_k$ are the columns of Q. This is the scale used for private range estimation and truncation. These quantities explain the improved spectral dependence. For Gaussian covariance data, the fourth-moment identity implies that, in a local basin around $V_k V_k^\top$,

$$v_Q \lesssim \sqrt{\lambda_1(\mathrm{tr}_{>k}(\Sigma) + k\lambda_{k+1})} + \lambda_1\sqrt{k}\, e(Q),$$

and

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

Thus, once Q is close to the target subspace, the tangent fluctuation scale is governed by the interaction between signal directions and noise directions. In the flat-tail spiked covariance model, $\mathrm{tr}_{>k}(\Sigma) = (d-k)\lambda_{k+1}$, so the leading scales become $v_Q \approx \sqrt{d\lambda_1\lambda_{k+1}}$, and $g_Q \approx \sqrt{\lambda_1\lambda_{k+1}}$. Substituting these scales into the tangent refinement analysis gives the spectral factor

$$\frac{\sqrt{\lambda_1\lambda_{k+1}}}{\lambda_k - \lambda_{k+1}},$$

rather than $\lambda_1/(\lambda_k - \lambda_{k+1})$. Intuitively, the full projected update $A_i Q$ still contains large fluctuations internal to the signal subspace, while the tangent update removes those directions and retains only the signal-noise interaction that actually moves the subspace.

We now state the local model used to analyze this refinement phase. Unlike the global model for ADADPO, this model only needs to hold after the warm-up phase has entered a constant-accuracy neighborhood of the target subspace.

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 $\left\|QQ^\top - V_k V_k^\top\right\|_2 \leq c_0$.*

B.1 $\|\Pi_Q(A_i - \Sigma)Q\|_2 \leq M_\perp$,

B.2 $v_Q \leq v_0 + v_1 e(Q)$,

B.3 $g_Q \leq g_0 + g_1 e(Q)$

B.4 *For every column q_r of Q, every unit vector $u \perp \mathrm{span}(Q)$, and every $\vartheta \in (0,1)$,*

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

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 $v_1 = 0$, but only 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.

Theorem 3 (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 $\left\|Q_0 Q_0^\top - V_k V_k^\top\right\|_F \leq c_0/2$ 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

$$n \geq \tilde{\Omega}\left(\max\left\{\frac{K^2 d + Kdk}{\varepsilon}, \frac{k(v_0^2 + v_1^2)}{(\lambda_k - \lambda_{k+1})^2}, \frac{\sqrt{k}M_\perp}{(\lambda_k - \lambda_{k+1})}, \frac{K(g_0+g_1)dk}{\varepsilon(\lambda_k - \lambda_{k+1})}\right\}\right),$$

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*

$$\left\|Q_T Q_T^\top - V_k V_k^\top\right\|_F \le \mathrm{Opt}_T + \tilde{O}\left(\frac{\sqrt{k}v_0}{(\lambda_k-\lambda_{k+1})\sqrt{n}} + \frac{\sqrt{k}M_\perp}{(\lambda_k-\lambda_{k+1})n} + \frac{Kg_0dk}{(\lambda_k-\lambda_{k+1})\varepsilon n}\right),$$

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

The leading statistical and privacy terms are governed by the tangent scales v_0 and g_0. Thus, when these are smaller than their global counterparts, the refinement phase improves the spectral dependence without changing the dependence on d, k, n, or ε.

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

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

where $\mathrm{Opt}_T := \exp\left(-c(\lambda_k-\lambda_{k+1})\sum_{t=1}^T \eta_t\right)\left\|Q_0Q_0^\top - V_kV_k^\top\right\|_F$*,* η_t *are learning rates. 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.*

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 guarantee and the full proofs for TAdaDPO, see Section C.5.

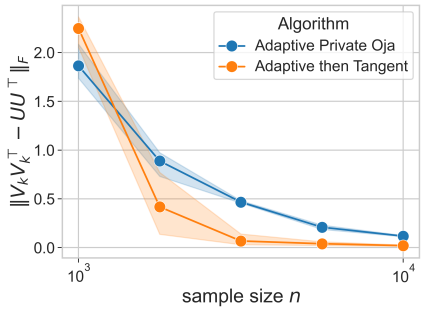

Figure 2: Comparison of AdaDPO with a two-stage variant that warm-starts with AdaDPO and then runs TAdaDPO.

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.

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. *arXiv preprint arXiv:2508.10879*, 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.

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.

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 $t \in [T]$,

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

simultaneously for all $t \in [T]$, up to changing the universal constant C.

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

First, by Assumption A.2,

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

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

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

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

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

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

with almost-sure bound M and variance proxy

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

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

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

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,

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

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

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

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

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

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

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

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

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

We next lower bound the same quantity in terms of the distance between projectors. Let $v_1, \ldots, v_d$ 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, $0 \leq a_i \leq 1$ and

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

Therefore,

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

Since

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

we get

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

Moreover,

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

so

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

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

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

Hence

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

Combining this with the upper bound from the assumption gives

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

Therefore,

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

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

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

□

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

$$\Pr\left(\sup_{P \in \mathbb{S}_{d,k}} \|P^* G\|_2 \;\leq\; \sqrt{k}\left(2\sigma\sqrt{d} + t\right)\right) \geq 1 - 2\exp\left(-\frac{t^2}{4\sigma^2}\right).$$

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

$$\Pr\left(\sup_{P \in \mathbb{S}_{d,k}} \|P^* G\|_2 \;\leq\; 2\sigma\sqrt{kd} + 2\sigma\sqrt{k\log\left(\frac{2}{\delta}\right)}\right) \geq 1 - \delta.$$

Since

$$\sup_{P \in \mathbb{S}_{d,k}} \|P^* G\|_2 = \left(\sum_{i=1}^{k} \sigma_i(G)^2\right)^{1/2},$$

the same bound may be written as

$$\Pr\left(\left(\sum_{i=1}^{k} \sigma_i(G)^2\right)^{1/2} \;\leq\; 2\sigma\sqrt{kd} + 2\sigma\sqrt{k\log\left(\frac{2}{\delta}\right)}\right) \geq 1 - \delta.$$

Proof. By the variational characterization,

$$\sup_{P \in \mathbb{S}_{d,k}} \|P^* G\|_2 = \left(\sum_{i=1}^{k} \sigma_i(G)^2\right)^{1/2}.$$

Hence

$$\sup_{P \in \mathbb{S}_{d,k}} \|P^* G\|_2 \leq \sqrt{k}\,\sigma_1(G) = \sqrt{k}\,\|G\|_{\mathrm{op}}.$$

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

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

Therefore,

$$\Pr\left(\sup_{P \in \mathbb{S}_{d,k}} \|P^* G\|_2 \geq \sqrt{k}\left(2\sigma\sqrt{d} + t\right)\right) \leq 2\exp\left(-\frac{t^2}{4\sigma^2}\right),$$

which proves the first claim. Setting

$$t = 2\sigma\sqrt{\log\left(\frac{2}{\delta}\right)}$$

yields the second claim. □

Lemma 7 (Lemma H.1. in Huang et al. [2021])**.** *For any deterministic matrices A, B and any standard Gaussian matrix Z of suitable sizes, it holds that*

$$\mathbb{P}\left\{\|AZB\|_2 \geq \|A\|_2\|B\|_2(1 + t)\right\} \leq e^{-t^2/2}.$$

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 in Section D 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 4 (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 9, 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 the symmetry-preserving Gaussian mechanism)**.** *Fix $R > 0$ and $Q \in \mathbb{R}^{d \times k}$ with $Q^\top Q = I_k$, and define*

$$f_Q(A) := \mathrm{clip}_R(AQ).$$

Consider the mechanism

$$M(A, Q) := f_Q(A) + W_Q,$$

where

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

with $N \sim \mathrm{GOE}_k(\sigma^2/2)$ 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

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

Proof. We proceed in three steps.

First, note that for every A,

$$f_Q(A) = \mathrm{clip}_R(AQ) \in \mathcal{S}_Q.$$

Indeed, $AQ \in \mathcal{S}_Q$ because

$$Q^\top(AQ) = Q^\top AQ$$

is symmetric when $A = A^\top$, and clipping only rescales the matrix by a nonnegative scalar, so it remains in the same linear subspace.

Second, we verify that W_Q is an isotropic Gaussian on $\mathcal{S}_Q$. Write any element of $\mathcal{S}_Q$ uniquely as

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

The noise has exactly this form:

$$W_Q = QN_{\mathrm{sym}} + N_\perp, \qquad N_{\mathrm{sym}} = N_{\mathrm{sym}}^\top, \quad Q^\top N_\perp = 0.$$

By construction, N_{sym} 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 in Frobenius inner product, W_Q is a centered Gaussian with covariance $\sigma^2 I_{\mathcal{S}_Q}$ on the subspace $\mathcal{S}_Q$.

Third, the mechanism is exactly the Gaussian mechanism applied to the query f_Q, but inside the Hilbert space $(\mathcal{S}_Q, \langle\cdot,\cdot\rangle_F)$. For neighboring $A \sim A'$, the sensitivity is

$$\Delta_Q = \sup_{A\sim A'} \|f_Q(A) - f_Q(A')\|_F.$$

Therefore the standard Gaussian mechanism theorem implies that $M(\cdot, Q)$ is (ε, δ)-DP whenever

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

Finally, because clipping enforces

$$\|f_Q(A)\|_F \leq R \quad \text{for all } A,$$

we have for neighbors $A \sim A'$,

$$\|f_Q(A) - f_Q(A')\|_F \leq \|f_Q(A)\|_F + \|f_Q(A')\|_F \leq 2R.$$

Hence $\Delta_Q \leq 2R$, and the sufficient condition

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

follows. □

This immediately implies the privacy of the update step in Algorithm 1, as we simply add this amount of noise plus more independent noise on top

Lemma 9 (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 $\mathrm{clip}_R(A_iQ)$ lies in

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

Indeed, $A_iQ \in \mathcal{S}_Q$ 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

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

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. Set*

$$R = C\lambda_1 \sqrt{dk}\big(K\gamma \log^a(ndk/\zeta) + 1\big), \qquad B = n/\log n, \qquad T = \lfloor n/B \rfloor.$$

Using appropriate learning rates, if

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

then with probability at least $1 - \zeta$, Algorithm 1 outputs Q_T satisfying

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

Proof. Let $T = \lfloor n/B \rfloor$. Lemma 10 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 C_t satisfies the necessary assumptions of Theorem 9. We see straight away that

$$\mathbb{E}[C_t] = \mathbb{E}\left[\frac{1}{B} \sum_{i=B\cdot(t-1)+1}^{tB} A_i \right] = \Sigma$$

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 $t \in [T]$,

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

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

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

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

$$\operatorname{dist}(Q_T, V) \le C' \frac{M'}{\Delta_k} \sqrt{\frac{\log(M'k/\Delta_k \delta)}{T - n_0}}$$

with probability at least $1 - \delta$.

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

$$\operatorname{dist}(Q_T, V) \le \tilde{O}\left(\frac{M'}{\Delta_k \sqrt{T}}\right).$$

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

$$\operatorname{dist}(Q_T, V) \le \tilde{O}\left(\frac{\mathsf{E}_B}{\Delta_k \sqrt{T}} + \frac{\mathsf{N}_B}{\Delta_k \sqrt{T}}\right).$$

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

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

while, using $B = n/\log n$,

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

Therefore,

$$\operatorname{dist}(Q_T, V) \le \tilde{O}\left(\frac{\lambda_1 \sqrt{kV}}{\Delta_k \sqrt{n}} + \frac{\sqrt{k} M}{\Delta_k n} + \frac{\lambda_1 (K\gamma + 1) kd}{\varepsilon \Delta_k n}\right).$$

Since $\Delta_k = \lambda_k - \lambda_{k+1}$, this is

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

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

Lemma 10 (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,

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

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 5, 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 $A_i Q$ 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 9 adding the noise matrix

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

calibrated to two 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.

Finally, the PRIVRANGE and TRUNCATEDMEAN act on a disjoint set of inputs, so by parallel composition the whole update is (ε, δ)-DP. 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 and setting the batch size to*

$$B = n/\log(n), \qquad T = \lfloor n/B \rfloor,$$

then for

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

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

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

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

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

Lemma 11 together with Theorem 5 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 C_t satisfies the necessary assumptions of Theorem 15 conditoned on a high probability event. 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 $t \in [T]$,

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

Moreover, by Theorem 5,

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

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 $(\mathcal{F}_{t-1}, \mathcal{R}_t)$, while the mean-half samples and the GOE base noise at time t are fresh. Hence the sequence satisfies the hypotheses of Theorem 15 in the usual conditional form: each update matrix is symmetric, has conditional expectation Σ, and obeys the uniform bound M' on $\mathcal{E}$. Applying Theorem 15 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

$$\operatorname{dist}(Q_T, V) \le C' \frac{M'}{\Delta_k} \sqrt{\frac{\log(M'k/\Delta_k\delta)}{T - n_0}}$$

with probability at least $1-\delta$.

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

$$\operatorname{dist}(Q_T, V) \le \tilde{O}\left(\frac{M'}{\Delta_k \sqrt{T}} \right).$$

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

$$\operatorname{dist}(Q_T, V) \le \tilde{O}\left(\frac{\mathsf{E}_B}{\Delta_k \sqrt{T}} + \frac{\mathsf{N}_B}{\Delta_k \sqrt{T}} \right).$$

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

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

while, using $B = n/\log n$,

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

Therefore,

$$\mathrm{dist}(Q_T, V) \le \tilde{O}\left(\frac{\lambda_1\sqrt{kV}}{\Delta_k\sqrt{n}} + \frac{\sqrt{k}M}{\Delta_k n} + \frac{\lambda_1 K\gamma\, kd}{\varepsilon\Delta_k n}\right).$$

Since $\Delta_k = \lambda_k - \lambda_{k+1}$, this is

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

If $M \gtrsim \lambda_1\sqrt{kV}$, then the first two statistical terms are bounded by

$$\tilde{O}\left(\frac{M}{\Delta_k\sqrt{n}}\right),$$

which gives the simplified display. □

Lemma 11 (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*

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

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*

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

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 $1 - \zeta/4$,

$$|(g_i^{(r)})_j - \mu_j^{(r)}| \le K\sqrt{\widehat{\Lambda}}\log^a(mdk/\zeta) \qquad \forall\, i \in [m],\ j \in [d],\ r \in [k]. \tag{3}$$

Next fix (j, r). 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 $\sqrt{\widehat{\Lambda}}$. This coordinate histogram is run with privacy budget $(\varepsilon_C/(dk), \delta_C/(dk))$. Let I_κ denote the bin containing $\mu_j^{(r)}$. Because the interval

$$\Big[\mu_j^{(r)} - \tau,\ \mu_j^{(r)} + \tau\Big], \qquad \tau = 2^{1/4}K\sqrt{\widehat{\Lambda}}\log^a(25),$$

is contained in the union of at most three adjacent bins, the population mass of the three bins $I_{\kappa-1} \cup I_\kappa \cup I_{\kappa+1}$ is at least 0.96. Hence at least one of these three bins has population mass at least 0.32, while any bin outside this set has population mass at most 0.04.

By uniform concentration of empirical bin frequencies and the accuracy guarantee of the private histogram learner, for the stated lower bound on m, with probability at least $1 - \zeta/4$ the released histogram differs from the population histogram by at most 0.01 uniformly over all coordinates (j, r) and all bins. On this event, a good bin among $I_{\kappa-1}, I_\kappa, I_{\kappa+1}$ still has released mass at least 0.31, whereas every bad bin has released mass at most 0.05. Therefore the privately selected maximizing bin must belong to $\{I_{\kappa-1}, I_\kappa, I_{\kappa+1}\}$.

Since $\bar{g}_{jr}$ is the left endpoint of the selected bin, it follows that

$$|\bar{g}_{jr} - \mu_j^{(r)}| \le 2K\sqrt{\widehat{\Lambda}}\log^a(mdk/\zeta) \tag{4}$$

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 (3) and (4), 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 $\widetilde{G}_i = G_i = A_iQ$ for all i. Since each A_i is symmetric, $Q^\top A_i Q$ is symmetric, and hence $A_iQ \in \mathcal{S}_Q$. Therefore the raw mean also belongs to $\mathcal{S}_Q$, so

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

□

Theorem 5 (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 $b \ge \tilde{O}\big(K^2 d\log(km/\zeta)\big)$, equivalently, $B \ge \tilde{O}\Big(\frac{K^2 d\log(k/(\delta\zeta))\log(1/(\delta\zeta))}{\varepsilon}\Big)$, 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 12). 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,

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

Since the m blocks are independent, a Chernoff bound implies that with probability at least $1 - \zeta/2$, at least a $3/4$-fraction of the values $\{m_j\}_{j=1}^m$ lie in the interval

$$\left[\frac{1}{\sqrt{2}}\,\Lambda(Q),\ \sqrt{2}\,\Lambda(Q)\right].$$

By construction of the geometric partition Ω, this interval intersects at most two consecutive bins, and therefore one of these bins contains a constant fraction of the points m_j. Lemma 12 then implies that, with probability at least $1 - \zeta/2$, the DP histogram learner returns a non-empty bin $[l, r]$ containing this heavy cluster.

Consequently,

$$l \leq \sqrt{2}\,\Lambda(Q) \qquad \text{and} \qquad l \geq \frac{1}{2}\Lambda(Q),$$

so the output

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

satisfies

$$\widehat{\Lambda} \in [\Lambda(Q),\ C_\Lambda \Lambda(Q)].$$

Combining the above events proves the theorem. □

Lemma 12 (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

$$\mathbb{P}\big(|\tilde{p}_k - \hat{p}_k| \leq \beta\big) \geq 1 - \alpha.$$

Corollary 2 (Rank-k spiked covariance)**.** *Let $V_k \in \mathbb{R}^{d\times k}$ have orthonormal columns, let $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ with $\lambda_1 \geq \cdots \geq \lambda_k > 0$, and let $X_i = V_k \Lambda^{1/2} + \sigma Z_i$, $(Z_i)_{ab} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Set $A_i = X_i X_i^\top$, then $\Sigma = \mathbb{E}[A_i] = V_k \Lambda V_k^\top + k\sigma^2 I_d$. Running Algorithm 2 with input $A_1, \ldots, A_n$ outputs Q_T with high probability satisfying*

$$\|Q_T Q_T^\top - V_k V_k^\top\|_F \leq \tilde{O}\left(\frac{\sigma\sqrt{(\lambda_1 + k\sigma^2)kd}}{\lambda_k \sqrt{n}} + \frac{Kdk\,\sigma\sqrt{\lambda_1 + \sigma^2 k}}{\varepsilon \lambda_k n} \right)$$

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

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

Define

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

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,

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

Therefore

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

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

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

Hence

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

Combining the previous displays,

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

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

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

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

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

We now apply Theorem 2. Its first term is

$$\frac{\mu_1\sqrt{kV_{\mathrm{thm}}}}{(\mu_k - \mu_{k+1})\sqrt{n}}.$$

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

$$\frac{\mu_1 K dk\gamma}{\varepsilon(\mu_k - \mu_{k+1})n}.$$

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

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

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 8 Algorithm 2 gives*

$$\|Q_T Q_T^\top - V_k V_k^\top\|_F \le \widetilde{O}\left(\frac{\lambda_1\sqrt{kd}}{\Delta_k\sqrt{n}} + \frac{Kdk\lambda_1}{\varepsilon\Delta_k n}\right).$$

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

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

We now apply Theorem 2. Its first term becomes

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

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

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

Combining these three bounds proves the result. □

C.5 Proofs for the tangent refinement theorem

Theorem 6 (Restated Theorem 3)**.** *Suppose $A_1, \dots, A_n$ satisfy Assumption A. Suppose further that, together with the starting point Q_0, they satisfy Assumption B, and that*

$$\left\|Q_0Q_0^\top - V_kV_k^\top\right\|_F \le c_0/2$$

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

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

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

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

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 13 (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 14 (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 $S = (A_1, \dots, A_m)$, let $\widetilde{Y}_i(C)$ be ΠA_iQ 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

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

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 13. 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 14, 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 6. Let $e_t := \left\|Q_tQ_t^\top - P_\star\right\|_F$. We first collect the high-probability events. By Lemma 15 and a union bound over $t \in [T]$,

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

holds for all t with probability at least $1 - \zeta/8$. On this event, Lemma 16 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,

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

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

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

This is exactly the form needed for the local tangent contraction lemma. By Lemma 17, 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 Assumptions B.2 and B.3, this gives

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

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 18 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 logarithmic T, implies $D \le c\Delta$ for a small enough universal constant. 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 T is logarithmic and $m \asymp n/T$,

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

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 $\left\|Q_0 Q_0^\top - P_\star\right\|_F \le c_0/2$, induction gives $\left\|Q_t Q_t^\top - P_\star\right\|_F \le c_0$ for every $t \le T$. Combining the initializer event with the range, center, and perturbation events gives probability at least $1 - \zeta$. □

Lemma 15 (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.4 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 13; 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 $q_1, \ldots, q_k$. 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,

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

Thus the empirical matrix $\widehat{C}_{j,r} := \frac{1}{2b}\sum_{s\in\mathcal{G}_j} \widetilde{Y}_s^{(r)}\big(\widetilde{Y}_s^{(r)}\big)^\top$ 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

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

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

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

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,

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

Therefore

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

For every $r \in [k]$,

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

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 12, 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,

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

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

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

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

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

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

□

Lemma 16 (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 16. 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.4, together with $\widehat{\Lambda} \ge g_Q^2$, implies that for

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

we have

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

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

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

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

$$C_{\mathrm{bin}} \le C K \log^a\left(\frac{mdkT}{\alpha}\right)$$

consecutive bins. By the stability-based histogram guarantee, equivalently by Lemma 12, 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

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

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},

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

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

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

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 17 (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 $m = |\mathcal{B}_t^{\mathrm{mean}}|$, and define*

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

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

$$\|Y_i\|_2 \le M_\perp \qquad \text{a.s.}$$

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

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

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 $\operatorname{range}(\Pi)$, 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,

$$r_t \le CKg_Q \operatorname{polylog}(BdkT/\zeta),$$

completes the proof. □

Lemma 18 (Local contraction of the noisy tangent step)**.** *There exist universal constants $c_0, c, c', C > 0$ such that the following holds. Let $P_\star = V_kV_k^\top$, 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 $\eta \le c/\lambda_1$ 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

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

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.

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

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,

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

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

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

Therefore, if $\eta \leq c/\lambda_1$,

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

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

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

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

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

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

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

where

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

The triangle inequality gives

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

This proves the lemma. □

Lemma 19 (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 6,

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

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

$$\mathbb{E}\left[(A-\Sigma)B(A-\Sigma)\right] = \mathrm{tr}(\Sigma B)\Sigma + \Sigma B\Sigma. \tag{5}$$

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, (5) 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 (5) 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 (5) 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 19 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 6 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 $\Delta = \lambda_k - \lambda_{k+1}$, taking n large enough to absorb the last term and the Opt_T, gives the inequality. □

D Existing Lower Bounds

Theorem 7 (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*

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

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 8 (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 $\delta \leq c_0' \exp\{2\varepsilon - c_0(\varepsilon\sqrt{ndk} + dk)\}$ for some small constants $c_0, c_0' > 0$. Then, there exists an absolute constant $c_1 > 0$ such that*

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

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

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

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 $\mathcal{P}(\lambda, \sigma^2) = \{\mathcal{N}(0, \Sigma), \Sigma = U\Lambda U^\top + \sigma^2\mathbf{I}_d, c\lambda \leq \lambda_k \leq \cdots \leq \lambda_1 \leq C\lambda\}$. Suppose $\delta \leq c_0' \exp\{2\varepsilon - c_0(\varepsilon\sqrt{ndk} + dk)\}$ for some small constants $c_0, c_0' > 0$. Then, there exists an absolute constant $c_1 > 0$ such that*

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

E Further Algorithms

Theorem 9 (Main Theorem, Huang et al. [2021]). *Given matrices $A_1, \ldots, A_n \in \mathbb{R}^{d\times d}$ that are i.i.d., symmetric random matrices, satisfying $\mathbb{E}[A_i] = \Sigma$ for all i and the uniform bound*

$$\sup_i \sup_{P\in\mathcal{S}_{d,k}} \|P^\top(A_i - \Sigma)\|_2 \leq M \text{ a.s.}$$

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

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

Let $V \in \mathbb{R}^{d\times k}$ 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

$$\|Q_nQ_n^\top - VV^\top\|_F \leq C'\frac{M}{\rho_k}\sqrt{\frac{\log(Mk/(\rho_k\delta))}{n - n_0}}$$

with probability at least $1 - \delta$, where C' is a universal positive constant.

Note that Huang et al. [2021] requires the matrices to be i.i.d. However, their main recursion Theorem 3.1 does not require identically distributed nor independent. These requirements arise from their Phase I proof, that first proofs the result for finite support and then lifts the result to general distributions using i.i.d. We show that this result can be extended to our use case (of non identically distributed matrices) in Theorem 15.

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 $\lambda_{j,r} \leftarrow \lambda_1\Big(\frac{1}{2b} G_j^{(r)} \big(G_j^{(r)}\big)^\top\Big)$
  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 12 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 $(\varepsilon_C, \delta_C)$, failure probability ζ

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 $\widetilde{\mu} \leftarrow \frac{1}{B}\sum_{\ell=1}^{B} \widetilde{G}_\ell$

F 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

Algorithm 6 Private Top Eigenvalue Estimation, Algorithm 4 in [Liu et al., 2022]

Input: $S = \{g_i\}_{i=1}^{B}$, (ε, δ)-DP, failure probability ζ

Let $\tilde{g}_i \leftarrow g_{2i} - g_{2i-1}$ for $i \in 1, 2, \ldots, \lfloor B/2 \rfloor$. Let $\tilde{S} = \{\tilde{g}_i\}_{i=1}^{\lfloor B/2 \rfloor}$
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 $b = \lfloor B/2k \rfloor$
Let $\lambda_1^{(j)}$ be the top eigenvalue of $(1/b)G_j G_j^\top$ for $\forall j \in [k]$
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 12 on $\{\lambda_1^{(j)}\}_{j=1}^{k}$ 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 $\hat{\Lambda} = l$

Algorithm 7 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 $\{\eta_t\}_{t=1}^{n}$

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

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

$$g_k = \lambda_k - \lambda_{k+1} + z, \qquad z \sim \operatorname{Lap}(2/\varepsilon).$$

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

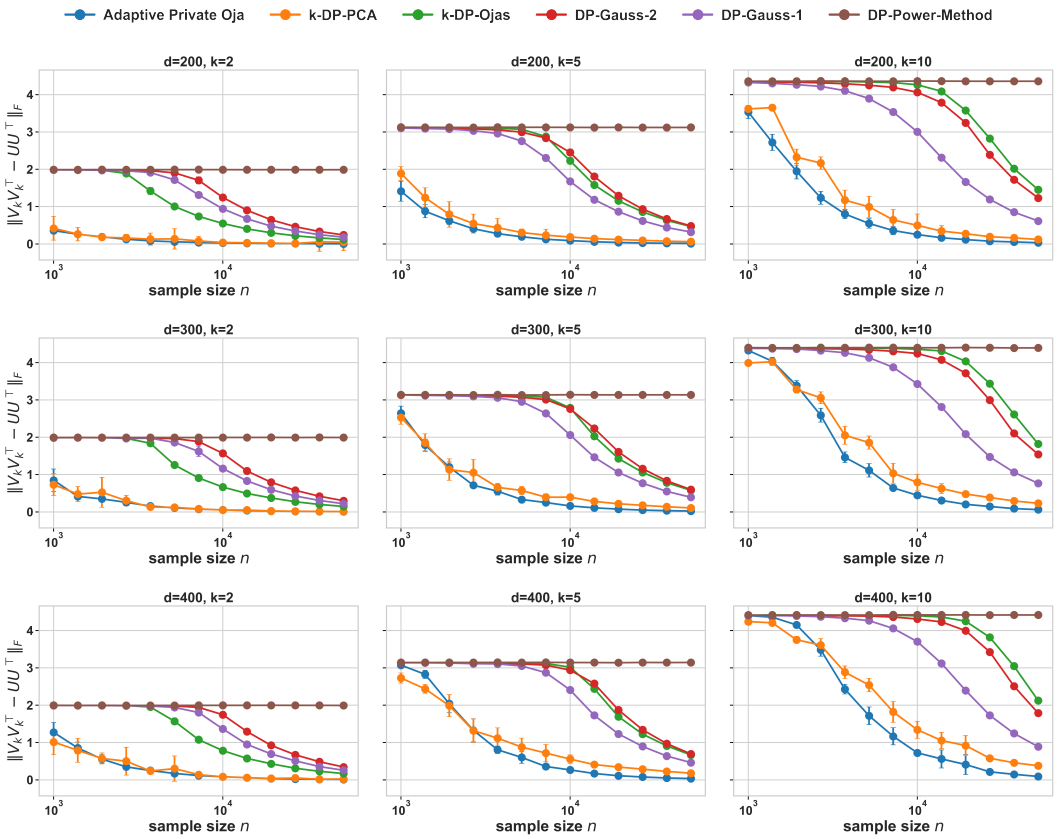

Figure 3: Comparison of ADADPO with baselines for varying k and d on the spiked covariance model. We plot the mean over 10 trials, with the bars representing the standard deviation.

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.

F.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

$$s_i \sim \text{Unif}\{\lambda_1 v, -\lambda_1 v\},$$

$v \in \mathbb{R}^d$ 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, including our algorithms MODIFIEDDP-PCA and k-DP-Ojas, as well as the two Gaussian baselines. 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.

F.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 $j = 1, \dots, k$. 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

F.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.

G Finite support Phase I result for non identical distributions

Let $\Sigma \succeq 0$ have eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$$

with corresponding orthonormal eigenvectors $v_1, \dots, v_d$. Fix $k \in [d-1]$ and write

$$\rho_k := \lambda_k - \lambda_{k+1} > 0.$$

Let

$$V := [v_1 \quad \cdots \quad v_k], \qquad U := [v_{k+1} \quad \cdots \quad v_d].$$

For a matrix $X \in \mathbb{R}^{d\times d}$, define

$$\|X\|_{(2,k)} := \sup_{P \in \mathbb{S}_{d,k}} \left\|P^\top X\right\|_F.$$

Assumption C ($(\Sigma, \{\lambda_i\}_{i=1}^d, M)$-predictable Oja model)**.** *Let $Z_0 \in \mathbb{R}^{d\times k}$ be the initialization matrix and let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration such that*

$$Z_0 \in \mathcal{F}_0.$$

The matrices $C_1, \ldots, C_{T_0} \in \mathbb{R}^{d\times d}$ are symmetric and satisfy the following conditions.

C.1 *C_t is $\mathcal{F}_t$-measurable and $\mathbb{E}[C_t \mid \mathcal{F}_{t-1}] = \Sigma$ for every $t \leq T_0$. The matrix $\Sigma \succeq 0$ has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$, corresponding eigenvectors $v_1, \ldots, v_d$, and eigengap $\rho_k = \lambda_k - \lambda_{k+1} > 0$.*

C.2 *The centered updates are uniformly bounded in the $(2,k)$-norm: $\|C_t - \Sigma\|_{(2,k)} = \sup_{P\in\mathbb{S}_{d,k}} \left\|P^\top (C_t - \Sigma)\right\|_F \leq M$ a.s. for every $t \leq T_0$.*

Definition 5 (Oja iterates)**.** Given an initial matrix $Q_0 \in \mathbb{R}^{d\times k}$ with orthonormal columns, define

$$Q_t := \mathrm{QR}\left(Q_{t-1} + \eta_t C_t Q_{t-1}\right), \qquad t = 1, \ldots, T_0.$$

Equivalently, if

$$Z_t := \prod_{s=1}^{t} (I + \eta_s C_s) Q_0,$$

then Q_t is an orthonormal basis for the column span of Z_t. Whenever $V^\top Z_t$ is invertible, define

$$W_t := U^\top Z_t (V^\top Z_t)^{-1}.$$

Remark. For stopped estimates below, we use the following convention. On the event where $V^\top Z_t$ is not invertible, W_t may be defined arbitrarily. All quantities involving W_t in the analysis are multiplied by $\mathbf{1}_t$ or $\mathbf{1}_{t-1}$, and the small-step/good-event conditions below ensure that $V^\top Z_t$ is invertible on the relevant stopped events.

For the Phase I analysis below, we take a constant step size

$$\eta_t \equiv \eta \qquad \text{for } t = 1, \ldots, T_0.$$

Let $\gamma > 0$ be the good-event threshold. For a random matrix X, define

$$\|X\|_{p,p} := \left(\mathbb{E}\,\mathrm{tr}\big((X^\top X)^{p/2}\big)\right)^{1/p}.$$

Assume that there is a deterministic finite set $\mathcal{A}_\star \subseteq \mathbb{R}^{d\times d}$ such that

$$C_t \in \mathcal{A}_\star \qquad \text{a.s. for every } t \leq T_0.$$

Define

$$\mathcal{E} := \left\{M^{-1}(A - \Sigma)UU^\top : A \in \mathcal{A}_\star\right\}.$$

Here M is the scalar boundedness constant from Assumption Assumption C.2.

For $r, \ell \geq 1$, define

$$\mathcal{E}_{r,\ell} := \left\{V^\top F_1 \cdots F_r U : F_i \in \mathcal{E} \text{ for at most } \ell \text{ distinct indices } i \in [r], \text{ and } F_i = (1+\eta\lambda_{k+1})^{-1}(I + \eta\Sigma)UU^\top \text{ otherwise}\right\}.$$

Define the good events G_t recursively by

$$G_t := G_{t-1} \cap \left\{\max_{E\in\mathcal{E}} \left\|V^\top E U W_t\right\|_2 \leq \gamma\right\}, \qquad t \geq 1,$$

with G_0 the corresponding initialization good event, and write

$$\mathbf{1}_t := \mathbf{1}_{G_t}.$$

For the Phase I analysis, fix a good-event threshold $\gamma \geq 2$ and a constant step size $\eta > 0$.

This is the analogue of the construction in [Huang et al., 2021, Section 5]. In that paper the population matrix is denoted by $\mathbf{M} = \mathbb{E}[A_t]$, while the scalar M denotes the boundedness constant. In the present notation the population matrix is Σ, and M remains the scalar boundedness constant. Thus the normalized fluctuation $M^{-1}(A_t - \mathbf{M})UU^\top$ in Huang et al. [2021] is replaced by $M^{-1}(C_t - \Sigma)UU^\top$.

Lemma 20 (Stopped well-definedness of W_t)**.** *Assume the Phase I small-step conditions*

$$\varepsilon := 2\eta M(1+\gamma) \le \frac{1}{2}, \qquad \eta\|\Sigma\|_2 \le \frac{1}{2}.$$

Then, for each $1 \le t \le T_0$*,* $W_t\mathbf{1}_{t-1}$ *is well-defined. Equivalently,* $V^\top Z_t$ *is invertible on* G_{t-1}*.*

Proof. The claim holds at $t = 0$ because $V^\top Z_0$ is a square Gaussian matrix and is invertible almost surely. Suppose inductively that $V^\top Z_{t-1}$ is invertible on G_{t-1}. On G_{t-1}, define

$$\Delta_t := \eta V^\top (C_t - \Sigma) Z_{t-1} \big(V^\top (I + \eta\Sigma) Z_{t-1}\big)^{-1}.$$

As in the one-step analysis of Huang et al. [2021], the good-event condition and the deterministic $(2, k)$-norm bound imply

$$\|\Delta_t \mathbf{1}_{t-1}\| \le \varepsilon \le \frac{1}{2}.$$

Moreover,

$$V^\top (I + \eta\Sigma) Z_{t-1} = (I + \eta\Lambda_V) V^\top Z_{t-1},$$

where $\Lambda_V = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$. Since $\Sigma \succeq 0$, the matrix $I + \eta\Lambda_V$ is invertible. Hence $V^\top (I + \eta\Sigma) Z_{t-1}$ is invertible on G_{t-1}.

Finally,

$$V^\top Z_t = V^\top (I + \eta C_t) Z_{t-1} = (I + \Delta_t) V^\top (I + \eta\Sigma) Z_{t-1}.$$

Since $\|\Delta_t\| \le 1/2$ on G_{t-1}, $I + \Delta_t$ is invertible there. Thus $V^\top Z_t$ is invertible on G_{t-1}, completing the induction. □

Note the main recursive bound of [Huang et al., 2021] applies verbatim to the present predictable setting: its proof uses only the conditional centering of the first-order fluctuation term, the deterministic $(2, k)$-norm bound, and the good-event condition (2.4). In our notation, $\mathbb{E}[C_t - \Sigma \mid \mathcal{F}_{t-1}] = 0$, and the Phase I good events ensure

$$\|V^\top (C_t - \Sigma) U W_{t-1} \mathbf{1}_{t-1}\|_2 \le M\gamma \qquad \text{a.s.}$$

Thus, with constant step size, the hypotheses of Theorem 3.1 are satisfied under the stated small-step assumptions.

Theorem 10 (Predictable version of Huang et al. Theorem 3.1)**.** *Let* $t \ge 1$*. For* $1 \le i \le t$*, define*

$$\varepsilon_i := 2\eta_i M(1+\gamma), \qquad s_t := \sum_{i=1}^{t} \eta_i.$$

Let $G_0 \supseteq G_1 \supseteq \cdots \supseteq G_t$ *be good events with indicators*

$$\mathbf{1}_i := \mathbf{1}_{G_i}.$$

Assume that for every $1 \le i \le t$*,*

$$\varepsilon_i \le \frac{1}{2}, \qquad \eta_i\|\Sigma\|_2 \le \frac{1}{2}, \qquad e^{-\eta_i\rho_k/4} \le \frac{\varepsilon_i}{\varepsilon_{i-1}}, \tag{6}$$

where the last condition is vacuous when $i = 1$*.*

Assume also that, for every $1 \le i \le t$*,*

$$\|V^\top (C_i - \Sigma) U W_{i-1} \mathbf{1}_{i-1}\|_2 \le M\gamma \qquad \textit{a.s.}$$

Then, for every $p \ge 2$*,*

$$\begin{aligned} \|W_t\mathbf{1}_t\|_{p,p}^2 &\le \|W_t\mathbf{1}_{t-1}\|_{p,p}^2 \\ &\le e^{-s_t\rho_k}\|W_0\mathbf{1}_0\|_{p,p}^2 + C_1 p\varepsilon_t^2 \sum_{i=0}^{t-1} \|W_i\mathbf{1}_i\|_{p,p}^2 + C_2 p k^{2/p}\varepsilon_t^2 t, \end{aligned} \tag{7}$$

where $C_1 = 21$ *and* $C_2 = 5$*.*

Moreover, if in addition

$$p\varepsilon_i^2 \le \frac{\eta_i \rho_k}{50} \qquad \textit{for every } 1 \le i \le t, \tag{8}$$

then

$$\begin{aligned} \|W_t \mathbf{1}_t\|_{p,p}^2 &\le \|W_t \mathbf{1}_{t-1}\|_{p,p}^2 \\ &\le e^{-s_t \rho_k/2} \|W_0 \mathbf{1}_0\|_{p,p}^2 + C_2 p k^{2/p} \varepsilon_t^2 t. \end{aligned} \tag{9}$$

Lemma 21 (Gaussian initialization for predictable Phase I)**.** *Let $Z_0 \in \mathbb{R}^{d\times k}$ have i.i.d. $N(0,1)$ entries and set*

$$Q_0 := \mathrm{QR}(Z_0).$$

Let $\mathcal{S}$ be a sigma-field independent of Z_0. Suppose that, conditional on $\mathcal{S}$, the set $\mathcal{A}_\star \subseteq \mathbb{R}^{d\times d}$ is finite and satisfies

$$|\mathcal{A}_\star| \le m.$$

Let

$$W_0 := U^\top Q_0 (V^\top Q_0)^{-1}.$$

Equivalently, almost surely,

$$W_0 = U^\top Z_0 (V^\top Z_0)^{-1}.$$

Define

$$\gamma_{\mathrm{init}} := \frac{\gamma}{\sqrt{2}e},$$

and let

$$G_0 := \left\{ \max_{1\le r,\ell \le T_0+1} \max_{E\in\mathcal{E}_{r,\ell}} \|EW_0\|_F \le \sqrt{\ell}\,\gamma_{\mathrm{init}} \right\} \cap \left\{ \|W_0\|_F \le \sqrt{d}\gamma \right\}.$$

There is a universal constant $C_\gamma > 0$ such that, for every $\delta_{\mathrm{init}} \in (0,1)$, if

$$\gamma \ge C_\gamma \min\left\{ \frac{\sqrt{k \log\left(\frac{emT_0}{\delta_{\mathrm{init}}}\right)}}{\delta_{\mathrm{init}}}, \frac{d}{\delta_{\mathrm{init}}^2} \right\},$$

then

$$\mathbb{P}(G_0^c \mid \mathcal{S}) \le \delta_{\mathrm{init}} \qquad \textit{a.s.}$$

Consequently,

$$\mathbb{P}(G_0^c) \le \delta_{\mathrm{init}}.$$

Proof. Condition on $\mathcal{S}$. Then the set $\mathcal{A}_\star$, and hence all product classes $\mathcal{E}_{r,\ell}$, are deterministic. Since Z_0 has i.i.d. Gaussian entries and $[V\ U]$ is orthogonal,

$$X := V^\top Z_0 \in \mathbb{R}^{k\times k}, \qquad Y := U^\top Z_0 \in \mathbb{R}^{(d-k)\times k}$$

are independent standard Gaussian matrices. Moreover X is invertible almost surely and

$$W_0 = YX^{-1}.$$

For $E \in \mathcal{E}_{r,\ell}$, the Phase I construction gives $\|E\|_F \le 1$. Also, $|\mathcal{E}_{r,\ell}| \le \big((m+1)(T_0+1)\big)^\ell$. Conditioning further on X, the matrix X^{-1} is deterministic, and Y remains a standard Gaussian matrix independent of X. For each fixed $E \in \mathcal{E}_{r,\ell}$, we apply Lemma 7 with

$$A = E, \qquad Z = Y, \qquad B = X^{-1}.$$

Since the Phase I product-class construction gives $\|E\|_F \le 1$, conditionally on X,

$$\mathbb{P}\left(\|EYX^{-1}\|_F > \|X^{-1}\|_F (1+s) \,\middle|\, X \right) \le e^{-s^2/2}.$$

Choosing

$$s = \sqrt{8\ell \log\left(\frac{emT_0}{\delta_{\mathrm{init}}}\right)}$$

and using $1 + s \le 2s$, followed by a union bound over $E \in \mathcal{E}_{r,\ell}$, gives

$$\max_{E\in\mathcal{E}_{r,\ell}} \|EW_0\|_F = \max_{E\in\mathcal{E}_{r,\ell}} \|EYX^{-1}\|_F \le 2\|X^{-1}\|_F \sqrt{8\ell \log\left(\frac{emT_0}{\delta_{\text{init}}}\right)}.$$

Taking a union bound over $1 \le r, \ell \le T_0 + 1$, this bound holds simultaneously for all r, ℓ with probability at least $1 - \delta_{\text{init}}/3$.

By the standard smallest-singular-value bound for a $k \times k$ Gaussian matrix,

$$\|X^{-1}\|_F \le \frac{18\sqrt{k}}{\delta_{\text{init}}}$$

with probability at least $1 - \delta_{\text{init}}/3$. Therefore, with probability at least $1 - 2\delta_{\text{init}}/3$,

$$\max_{1\le r,\ell\le T_0+1} \max_{E\in\mathcal{E}_{r,\ell}} \|EW_0\|_F \le 36\sqrt{\frac{k\ell \log(emT_0/\delta_{\text{init}})}{\delta_{\text{init}}^2}}.$$

Thus the first defining inequality of G_0 holds whenever

$$\gamma \ge C_\gamma \frac{\sqrt{k \log(emT_0/\delta_{\text{init}})}}{\delta_{\text{init}}}.$$

The same high-probability bounds on Y and X^{-1} also give

$$\|W_0\|_F = \|YX^{-1}\|_F \le \|Y\|_{\text{op}} \|X^{-1}\|_F \le C\sqrt{d}\,\frac{\sqrt{k}}{\delta_{\text{init}}}.$$

Increasing C_γ if necessary, the condition

$$\gamma \ge C_\gamma \frac{\sqrt{k \log(emT_0/\delta_{\text{init}})}}{\delta_{\text{init}}}$$

therefore implies

$$\|W_0\|_F \le \sqrt{d}\,\gamma.$$

The preceding argument shows that G_0 holds with probability at least $1 - \delta_{\text{init}}$ if

$$\gamma \ge C_\gamma \frac{\sqrt{k \log(emT_0/\delta_{\text{init}})}}{\delta_{\text{init}}}.$$

On the other hand, the crude bound

$$\|W_0\|_F \le Cd/\delta_{\text{init}}^2$$

holds with probability at least $1 - \delta_{\text{init}}$. Increasing C_γ by an absolute factor if necessary, the condition

$$\gamma \ge C_\gamma \frac{d}{\delta_{\text{init}}^2}$$

implies

$$Cd/\delta_{\text{init}}^2 \le \gamma_{\text{init}} = \frac{\gamma}{\sqrt{2}e}.$$

Hence, on this event,

$$\|W_0\|_F \le \gamma_{\text{init}}.$$

Since $\|E\|_{\text{op}} \le \|E\|_F \le 1$ for every $E \in \mathcal{E}_{r,\ell}$, it follows that, for every $1 \le r, \ell \le T_0 + 1$,

$$\|EW_0\|_F \le \|W_0\|_F \le \gamma_{\text{init}} \le \sqrt{\ell}\,\gamma_{\text{init}}.$$

Moreover,

$$\|W_0\|_F \le \gamma_{\text{init}} \le \sqrt{d}\,\gamma.$$

Thus G_0 also holds with probability at least $1 - \delta_{\text{init}}$ if

$$\gamma \ge C_\gamma \frac{d}{\delta_{\text{init}}^2}.$$

Thus G_0 holds whenever

$$\gamma \ge C_\gamma \min\left\{\frac{\sqrt{k \log(emT_0/\delta_{\text{init}})}}{\delta_{\text{init}}}, \frac{d}{\delta_{\text{init}}^2}\right\}.$$

□

Proposition 1 (Predictable one-step product estimate, D.1 analogue of Huang et al. [2021])**.** *Let $t \geq 1$. Throughout this proposition, $(\mathcal{F}_t)_{t\geq 0}$ is the filtration from Assumption C. In particular, $Z_0 \in \mathcal{F}_0$, $C_t \in \mathcal{F}_t$, and*

$$\mathbb{E}[C_t - \Sigma \mid \mathcal{F}_{t-1}] = 0.$$

Assume that $G_{t-1} \in \mathcal{F}_{t-1}$, and write $\mathbf{1}_{t-1} := \mathbf{1}_{G_{t-1}}$ and that the good-event condition

$$\|V^\top (C_t - \Sigma) U W_{t-1} \mathbf{1}_{t-1}\| \leq M\gamma \qquad \text{a.s.}$$

holds. Let $p \geq 2$, $r, \ell \geq 1$, and fix $E \in \mathcal{E}_{r,\ell}$. Since $\mathcal{A}_\star$ is deterministic, the product class $\mathcal{E}_{r,\ell}$ is deterministic. Thus E is deterministic, and in particular E is $\mathcal{F}_{t-1}$-measurable.

Define

$$H_t := U^\top (I + \eta\Sigma) Z_{t-1} \big(V^\top (I + \eta\Sigma) Z_{t-1}\big)^{-1},$$
$$\Delta_t := \eta V^\top (C_t - \Sigma) Z_{t-1} \big(V^\top (I + \eta\Sigma) Z_{t-1}\big)^{-1},$$

and

$$\widehat{\Delta}_t := \eta U^\top (C_t - \Sigma) Z_{t-1} \big(V^\top (I + \eta\Sigma) Z_{t-1}\big)^{-1}.$$

Let $\varepsilon := 2\eta M(1+\gamma)$ and $\overline{E}_t := 1 + 2\max_{E'' \in \mathcal{E}_{r+1,\ell+1}} \|E'' W_{t-1} \mathbf{1}_{t-1}\|_{p,p}$. Then

$$\|\Delta_t \mathbf{1}_{t-1}\| \leq \varepsilon \qquad \text{a.s.},$$

and

$$EW_t(I - \Delta_t^2) = EH_t + EJ_{t,1} + EJ_{t,2},$$

where

$$J_{t,1} := \widehat{\Delta}_t - H_t\Delta_t, \qquad J_{t,2} := -\widehat{\Delta}_t \Delta_t.$$

Moreover,

$$\|EJ_{t,1}\mathbf{1}_{t-1}\|_{p,p} \leq \overline{E}_t \varepsilon,$$
$$\|EJ_{t,2}\mathbf{1}_{t-1}\|_{p,p} \leq \overline{E}_t \varepsilon^2,$$

and

$$\mathbb{E}[EJ_{t,1}\mathbf{1}_{t-1} \mid \mathcal{F}_{t-1}] = 0.$$

Proof. The algebraic identity is the same as Lemma 2.7 of Huang et al. [2021], with $A_t - M$ replaced by $C_t - \Sigma$. Indeed,

$$V^\top Z_t = V^\top (I + \eta C_t) Z_{t-1} = (I + \Delta_t) V^\top (I + \eta\Sigma) Z_{t-1},$$

and expanding $U^\top Z_t (V^\top Z_t)^{-1}(I - \Delta_t^2)$ gives

$$W_t(I - \Delta_t^2) = H_t + \widehat{\Delta}_t - H_t\Delta_t - \widehat{\Delta}_t\Delta_t.$$

Multiplying by E gives the displayed decomposition.

The bound on $\Delta_t \mathbf{1}_{t-1}$ follows as in Huang et al.: using the small-step assumption for $I + \eta\Sigma$,

$$\begin{aligned}\|\Delta_t \mathbf{1}_{t-1}\| &\leq 2\eta \left\|V^\top (C_t - \Sigma)(UU^\top + VV^\top) Z_{t-1} (V^\top Z_{t-1})^{-1} \mathbf{1}_{t-1}\right\| \\ &\leq 2\eta \|V^\top (C_t - \Sigma) U W_{t-1} \mathbf{1}_{t-1}\| + 2\eta \|V^\top (C_t - \Sigma) V\| \\ &\leq 2\eta M(\gamma + 1) = \varepsilon.\end{aligned}$$

Here the final term uses the deterministic $(2, k)$-norm bound.

The bounds on $EJ_{t,1}\mathbf{1}_{t-1}$ and $EJ_{t,2}\mathbf{1}_{t-1}$ follow from the same product-class bookkeeping as in Huang et al. [2021]: the matrices generated by

$$EU^\top (I + \eta\Sigma) U \quad \text{and} \quad EU^\top (C_t - \Sigma) UU^\top$$

belong respectively to the enlarged classes $\mathcal{E}_{r+1,\ell}$ and $\mathcal{E}_{r+1,\ell+1}$, after the normalizations appearing in the definition of the product class. Therefore the corresponding terms are controlled by $\overline{E}_t\varepsilon$ and $\overline{E}_t\varepsilon^2$.

We first record the relevant measurability. Since $Z_0 \in \mathcal{F}_0$ and $C_s \in \mathcal{F}_s$ for $s \leq t-1$, the iterate Z_{t-1} is $\mathcal{F}_{t-1}$-measurable. Consequently, under the stopped convention for inverses, W_{t-1} and H_t

are $\mathcal{F}_{t-1}$-measurable. Moreover, the good events are adapted: $G_j \in \mathcal{F}_j$ for each j, by induction from their recursive definition. Hence $\mathbf{1}_{t-1}$ is $\mathcal{F}_{t-1}$-measurable. It remains to check centering. The matrix E is deterministic, hence $\mathcal{F}_{t-1}$-measurable. Moreover, Δ_t and $\widehat{\Delta}_t$ are linear in $C_t - \Sigma$, with all other factors $\mathcal{F}_{t-1}$-measurable. Hence

$$\mathbb{E}[\widehat{\Delta}_t \mathbf{1}_{t-1} \mid \mathcal{F}_{t-1}] = 0, \qquad \mathbb{E}[\Delta_t \mathbf{1}_{t-1} \mid \mathcal{F}_{t-1}] = 0.$$

Since E and H_t are $\mathcal{F}_{t-1}$-measurable,

$$\begin{aligned}\mathbb{E}[EJ_{t,1}\mathbf{1}_{t-1} \mid \mathcal{F}_{t-1}] &= E\,\mathbb{E}[\widehat{\Delta}_t \mathbf{1}_{t-1} \mid \mathcal{F}_{t-1}] - EH_t\,\mathbb{E}[\Delta_t \mathbf{1}_{t-1} \mid \mathcal{F}_{t-1}] \\ &= 0.\end{aligned}$$

□

Proposition 2 (Predictable one-step Phase I recursion; adapted from Proposition D.2 of Huang et al. [2021])**.** *Let $C_1, \dots, C_{T_0}$ satisfy Assumption C. Assume that there is a deterministic finite set $\mathcal{A}_\star \subseteq \mathbb{R}^{d\times d}$ such that*

$$C_t \in \mathcal{A}_\star \qquad \text{a.s. for every } t \le T_0.$$

Let $p \ge 2$, let $r, \ell \ge 1$, and define

$$\varepsilon := 2\eta M(1+\gamma).$$

If

$$\varepsilon \le \frac{1}{2},$$

then, for every $t \le T_0$,

$$\begin{aligned}\max_{E\in\mathcal{E}_{r,\ell}} \|EW_t\mathbf{1}_{t-1}\|_{p,p}^2 \le \overline{K}_1 \max_{E'\in\mathcal{E}_{r+1,\ell}} \|E'W_{t-1}\mathbf{1}_{t-1}\|_{p,p}^2 \\ + \overline{K}_2 \max_{E''\in\mathcal{E}_{r+1,\ell+1}} \|E''W_{t-1}\mathbf{1}_{t-1}\|_{p,p}^2 + \overline{K}_2,\end{aligned}$$

where

$$\overline{K}_1 := (1+5\varepsilon^2)\left(\frac{1+\eta\lambda_{k+1}}{1+\eta\lambda_k}\right)^2$$

and

$$\overline{K}_2 := (1+5\varepsilon^2)8p\varepsilon^2.$$

Proof of Proposition 2. Apply Proposition 1. Since $\|\Delta_t\mathbf{1}_{t-1}\| \le \varepsilon \le 1/2$, the matrix $I - \Delta_t^2$ is invertible on G_{t-1}, and

$$\|(I-\Delta_t^2)^{-1}\mathbf{1}_{t-1}\| \le \frac{1}{1-\varepsilon^2}.$$

Therefore

$$\|(I-\Delta_t^2)^{-1}\mathbf{1}_{t-1}\|^2 \le \frac{1}{(1-\varepsilon^2)^2} \le 1+5\varepsilon^2.$$

Thus the passage from $EW_t(I-\Delta_t^2)$ to EW_t contributes the common factor $1+5\varepsilon^2$ in the squared norm recursion, which is absorbed into $\overline{K}_1$ and $\overline{K}_2$. For each fixed $E \in \mathcal{E}_{r,\ell}$, use the matrix smoothness inequality with

$$X = EH_t\mathbf{1}_{t-1}, \qquad Y = EJ_{t,1}\mathbf{1}_{t-1}, \qquad Z = EJ_{t,2}\mathbf{1}_{t-1}.$$

The preceding proposition gives

$$\mathbb{E}[Y \mid \mathcal{F}_{t-1}] = 0,$$

and the required L_p-bounds on Y and Z. The deterministic term satisfies

$$\|EH_t\mathbf{1}_{t-1}\|_{p,p}^2 \le \left(\frac{1+\eta\lambda_{k+1}}{1+\eta\lambda_k}\right)^2 \max_{E'\in\mathcal{E}_{r+1,\ell}} \|E'W_{t-1}\mathbf{1}_{t-1}\|_{p,p}^2.$$

Taking the maximum over the finite class $\mathcal{E}_{r,\ell}$ gives the claimed recursion. □

Remark. The matrices E, E', E'' are deterministic matrices ranging over the finite product classes $\mathcal{E}_{r,\ell}$, $\mathcal{E}_{r+1,\ell}$, and $\mathcal{E}_{r+1,\ell+1}$, respectively. They are dummy variables inside the displayed maxima and are not random variables.

Theorem 11 (Iteration of the predictable Phase I recursion, analogue of Thoerem D.3 in [Huang et al., 2021])**.** *Assume the hypotheses of Proposition 2. Let* $\varepsilon := 2\eta M(1+\gamma)$*, and suppose that, for some* $p \geq 2$*,*

$$\varepsilon \leq \frac{1}{2}, \qquad \eta \left\|\Sigma\right\|_2 \leq \frac{1}{2}, \qquad p\varepsilon^2 \leq \frac{\eta\rho_k}{50}, \qquad \gamma \geq 2. \tag{10}$$

Set

$$\gamma_{\text{init}} := \frac{\gamma}{\sqrt{2}\, e}.$$

Assume that the initial good event G_0 *is chosen so that, on* G_0*,*

$$\max_{1\leq r,\ell\leq T_0+1} \max_{E\in\mathcal{E}_{r,\ell}} \left\|EW_0\right\|_F \leq \sqrt{\ell}\,\gamma_{\text{init}}. \tag{11}$$

Then, for every $1 \leq t \leq T_0$ *and every* $1 \leq r, \ell \leq T_0 - t + 1$*,*

$$\max_{E\in\mathcal{E}_{r,\ell}} \left\|EW_t\mathbf{1}_t\right\|_{p,p}^2 \leq \max_{E\in\mathcal{E}_{r,\ell}} \left\|EW_t\mathbf{1}_{t-1}\right\|_{p,p}^2 \leq \ell\gamma_{\text{init}}^2 e^{-t\eta\rho_k/2} + 6p\gamma^2\varepsilon^2 t.$$

Proof. Since $G_t \subseteq G_{t-1}$, we have $\mathbf{1}_t \leq \mathbf{1}_{t-1}$. Therefore

$$\max_{E\in\mathcal{E}_{r,\ell}} \left\|EW_t\mathbf{1}_t\right\|_{p,p}^2 \leq \max_{E\in\mathcal{E}_{r,\ell}} \left\|EW_t\mathbf{1}_{t-1}\right\|_{p,p}^2.$$

It remains to prove the second inequality.

Let

$$A_{t,r,\ell} := \max_{E\in\mathcal{E}_{r,\ell}} \left\|EW_t\mathbf{1}_t\right\|_{p,p}^2$$

and

$$B_{t,r,\ell} := \max_{E\in\mathcal{E}_{r,\ell}} \left\|EW_t\mathbf{1}_{t-1}\right\|_{p,p}^2.$$

By Proposition 2,

$$B_{t,r,\ell} \leq \overline{K}_1 A_{t-1,r+1,\ell} + \overline{K}_2 A_{t-1,r+1,\ell+1} + \overline{K}_2, \tag{12}$$

where

$$\overline{K}_1 = (1+5\varepsilon^2)\left(\frac{1+\eta\lambda_{k+1}}{1+\eta\lambda_k}\right)^2$$

and

$$\overline{K}_2 = (1+5\varepsilon^2)8p\varepsilon^2.$$

We first record two elementary consequences of (10). Since $\rho_k = \lambda_k - \lambda_{k+1}$ and $\eta\left\|\Sigma\right\|_2 \leq 1/2$,

$$\frac{1+\eta\lambda_{k+1}}{1+\eta\lambda_k} = 1 - \frac{\eta\rho_k}{1+\eta\lambda_k} \leq 1 - \frac{2}{3}\eta\rho_k.$$

Writing $x := \eta\rho_k$, we have $0 \leq x \leq 1/2$. Also $p\varepsilon^2 \leq x/50$, and since $p \geq 2$, $\varepsilon^2 \leq x/100$. Hence

$$\overline{K}_1 + \overline{K}_2 = (1+5\varepsilon^2)\left[\left(\frac{1+\eta\lambda_{k+1}}{1+\eta\lambda_k}\right)^2 + 8p\varepsilon^2\right] \leq e^{-x/2} = e^{-\eta\rho_k/2}.$$

Indeed, the preceding display follows from the estimates

$$\left(1-\frac{2x}{3}\right)^2 \leq e^{-4x/3}, \qquad 8p\varepsilon^2 \leq \frac{4x}{25}, \qquad 1+5\varepsilon^2 \leq e^{x/20},$$

together with the elementary inequality

$$e^{x/20}\left(e^{-4x/3} + \frac{4x}{25}\right) \leq e^{-x/2}, \qquad 0 \leq x \leq \frac{1}{2}.$$

Moreover, since $\varepsilon \leq 1/2$,

$$\overline{K}_2 = (1+5\varepsilon^2)8p\varepsilon^2 \leq 18p\varepsilon^2. \tag{13}$$

We now prove by induction on t that, for all $1 \le r, \ell \le T_0 - t + 1$,

$$A_{t,r,\ell} \le B_{t,r,\ell} \le \ell \gamma_{\text{init}}^2 e^{-t\eta\rho_k/2} + \frac{\gamma^2}{3}\,\overline{K}_2\, t. \tag{14}$$

The first inequality in (14) follows from $\mathbf{1}_t \le \mathbf{1}_{t-1}$, so it suffices to prove the bound for $B_{t,r,\ell}$.

Consider first $t = 1$. By (12) and the initial condition (11),

$$\begin{aligned} B_{1,r,\ell} &\le \overline{K}_1 \max_{E' \in \mathcal{E}_{r+1,\ell}} \|E' W_0 \mathbf{1}_0\|_{p,p}^2 + \overline{K}_2 \max_{E'' \in \mathcal{E}_{r+1,\ell+1}} \|E'' W_0 \mathbf{1}_0\|_{p,p}^2 + \overline{K}_2 \\ &\le \overline{K}_1\, \ell \gamma_{\text{init}}^2 + \overline{K}_2\, (\ell+1)\gamma_{\text{init}}^2 + \overline{K}_2 \\ &\le \ell \gamma_{\text{init}}^2 (\overline{K}_1 + \overline{K}_2) + (1 + \gamma_{\text{init}}^2)\overline{K}_2. \end{aligned}$$

Since

$$\overline{K}_1 + \overline{K}_2 \le e^{-\eta\rho_k/2}$$

and $\gamma \ge 2$ implies

$$1 + \gamma_{\text{init}}^2 = 1 + \frac{\gamma^2}{2e^2} \le \frac{\gamma^2}{3},$$

we obtain

$$B_{1,r,\ell} \le \ell \gamma_{\text{init}}^2 e^{-\eta\rho_k/2} + \frac{\gamma^2}{3}\overline{K}_2.$$

Thus (14) holds for $t = 1$.

Now assume (14) holds at time $t-1$. Let $1 \le r, \ell \le T_0 - t + 1$. Then $1 \le r+1, \ell+1 \le T_0 - (t-1) + 1$, so the induction hypothesis applies to the two terms on the right-hand side of (12). Hence

$$\begin{aligned} B_{t,r,\ell} &\le \overline{K}_1 \left[\ell \gamma_{\text{init}}^2 e^{-(t-1)\eta\rho_k/2} + \frac{\gamma^2}{3}\overline{K}_2 (t-1) \right] \\ &\quad + \overline{K}_2 \left[(\ell+1)\gamma_{\text{init}}^2 e^{-(t-1)\eta\rho_k/2} + \frac{\gamma^2}{3}\overline{K}_2 (t-1) \right] + \overline{K}_2 \\ &\le \ell \gamma_{\text{init}}^2 (\overline{K}_1 + \overline{K}_2) e^{-(t-1)\eta\rho_k/2} + \gamma_{\text{init}}^2 \overline{K}_2 e^{-(t-1)\eta\rho_k/2} \\ &\quad + \frac{\gamma^2}{3}\overline{K}_2 (t-1)(\overline{K}_1 + \overline{K}_2) + \overline{K}_2. \end{aligned}$$

Using

$$\overline{K}_1 + \overline{K}_2 \le e^{-\eta\rho_k/2} \le 1,$$

we get

$$\begin{aligned} B_{t,r,\ell} &\le \ell \gamma_{\text{init}}^2 e^{-t\eta\rho_k/2} + \gamma_{\text{init}}^2 \overline{K}_2 + \frac{\gamma^2}{3}\overline{K}_2 (t-1) + \overline{K}_2 \\ &\le \ell \gamma_{\text{init}}^2 e^{-t\eta\rho_k/2} + \frac{\gamma^2}{3}\overline{K}_2 t. \end{aligned}$$

The last inequality again uses $1 + \gamma_{\text{init}}^2 \le \gamma^2/3$. This proves (14) for time t.

Finally, applying (13) gives

$$\frac{\gamma^2}{3}\overline{K}_2 t \le 6p\gamma^2\varepsilon^2 t.$$

Therefore

$$B_{t,r,\ell} \le \ell \gamma_{\text{init}}^2 e^{-t\eta\rho_k/2} + 6p\gamma^2\varepsilon^2 t,$$

and the theorem follows. □

Proposition 3 (Predictable Phase I moment estimates, analgoue of D.4 in Huang et al. [2021])**.** *Assume that $C_1, \dots, C_{T_0}$ satisfy Assumption C. Assume further that there exists a deterministic finite set $\mathcal{A}_\star \subseteq \mathbb{R}^{d\times d}$ such that*

$$C_t \in \mathcal{A}_\star \qquad \text{a.s. for every } t \le T_0,$$

and set $m_\star := |\mathcal{A}_\star|$. Define

$$\mathcal{E} := \left\{ M^{-1}(A - \Sigma)UU^\top : A \in \mathcal{A}_\star \right\}.$$

Let

$$\varepsilon := 2\eta M(1+\gamma).$$

For $\delta \in (0,1)$, define

$$p_0 := \left\lceil \log \frac{6k}{\delta} \right\rceil, \qquad p_1 := \left\lceil \log \frac{12T_0(m_\star+1)}{\delta} \right\rceil,$$

and $p_\star := \max\{p_0, p_1\}$. Assume that the parameters satisfy

$$\varepsilon \le \frac{1}{2}, \qquad \eta \left\|\Sigma\right\|_2 \le \frac{1}{2}, \qquad p_\star \varepsilon^2 \le \frac{\eta\rho_k}{50}, \qquad \gamma \ge 2.$$

and

$$6p_\star \varepsilon^2 T_0 \le \frac{1}{2e^2}.$$

Let $Z_0 \in \mathbb{R}^{d\times k}$ have i.i.d. $N(0,1)$ entries, (independent of $C_1, \dots, C_{T_0}$) and set $Q_0 := \mathrm{QR}(Z_0)$. Define G_0 as in Lemma 21 with $m = m_\star$, $\delta_{\mathrm{init}} = \frac{\delta}{12}$. Assume that

$$\gamma \ge C_\gamma \min\left\{ \frac{\sqrt{k\log\left(\frac{12em_\star T_0}{\delta}\right)}}{\delta/12}, \frac{d}{(\delta/12)^2} \right\}.$$

Then

$$\mathbb{P}(G_0^c) \le \frac{\delta}{12}.$$

Assume also that

$$d\gamma^2 e^{-T_0\eta\rho_k/2} \le \frac{1}{2e^2} k^{2/p_0}.$$

Then

$$\left\|W_{T_0}\mathbf{1}_{T_0}\right\|_{p_0,p_0} \le e^{-1}k^{1/p_0},$$

and

$$\max_{1\le j\le T_0} \max_{F\in\mathcal{E}} \left\|V^\top F U W_j \mathbf{1}_{j-1}\right\|_{p_1,p_1} \le \frac{\gamma}{e}.$$

Proof. We first prove the terminal bound for W_{T_0}. The Phase I good events ensure the good-event condition required by Theorem 10; indeed, for each $1 \le i \le T_0$, if

$$F_i := M^{-1}(C_i - \Sigma)UU^\top,$$

then $F_i \in \mathcal{E}$ almost surely, and on G_{i-1},

$$\|V^\top F_i U W_{i-1}\|_2 \le \gamma.$$

Therefore

$$\|V^\top (C_i - \Sigma)U W_{i-1}\mathbf{1}_{i-1}\|_2 \le M\gamma \qquad \text{a.s.}$$

Since the step size is constant, $\varepsilon_i \equiv \varepsilon$ and $s_{T_0} = T_0\eta$. The hypotheses

$$\varepsilon \le \frac{1}{2}, \qquad \eta\|\Sigma\|_2 \le \frac{1}{2}, \qquad p_\star\varepsilon^2 \le \frac{\eta\rho_k}{50}$$

therefore imply the hypotheses of Theorem 10 with $p = p_0$. Hence

$$\|W_{T_0}\mathbf{1}_{T_0}\|_{p_0,p_0}^2 \le \|W_{T_0}\mathbf{1}_{T_0-1}\|_{p_0,p_0}^2 \le e^{-T_0\eta\rho_k/2}\|W_0\mathbf{1}_0\|_{p_0,p_0}^2 + 5p_0k^{2/p_0}\varepsilon^2 T_0.$$

By the definition of G_0,

$$\|W_0\mathbf{1}_0\|_F^2 \le d\gamma^2.$$

Since the Schatten p_0-norm is bounded by the Frobenius norm for $p_0 \ge 2$,

$$\|W_0\mathbf{1}_0\|_{p_0,p_0}^2 \le d\gamma^2.$$

Therefore, by the assumed Phase I length condition,

$$d\gamma^2 e^{-T_0\eta\rho_k/2} \le \frac{1}{2e^2}k^{2/p_0},$$

we have

$$e^{-T_0\eta\rho_k/2}\|W_0\mathbf{1}_0\|_{p_0,p_0}^2 \le \frac{1}{2e^2}k^{2/p_0}.$$

Moreover, since $p_0 \le p_\star$, the parameter condition

$$6p_\star\varepsilon^2 T_0 \le \frac{1}{2e^2}$$

implies

$$5p_0k^{2/p_0}\varepsilon^2T_0 \le 6p_\star k^{2/p_0}\varepsilon^2T_0 \le \frac{1}{2e^2}k^{2/p_0}.$$

Combining the two preceding displays gives

$$\|W_{T_0}\mathbf{1}_{T_0}\|_{p_0,p_0}^2 \le e^{-2}k^{2/p_0}.$$

Equivalently,

$$\|W_{T_0}\mathbf{1}_{T_0}\|_{p_0,p_0} \le e^{-1}k^{1/p_0}.$$

We now prove the good-event moment bound. Fix $1 \le j \le T_0$ and $F \in \mathcal{E}$. By definition of the product class,

$$V^\top FU \in \mathcal{E}_{1,1}.$$

By the definition of G_0, on G_0, for every $E \in \mathcal{E}_{1,1}$,

$$\|EW_0\|_F \le \gamma_{\text{init}} = \frac{\gamma}{\sqrt{2}e}.$$

Thus, since the Schatten p_1-norm is bounded by the Frobenius norm,

$$\max_{E\in\mathcal{E}_{1,1}}\|EW_0\mathbf{1}_0\|_{p_1,p_1}^2 \le \frac{\gamma^2}{2e^2}.$$

Therefore the initialization condition required by Theorem 11 is satisfied for $r = 1$, $\ell = 1$, and $p = p_1$. Applying Theorem 11 with these choices gives

$$\|V^\top FUW_j\mathbf{1}_{j-1}\|_{p_1,p_1}^2 \le \frac{\gamma^2}{2e^2}e^{-j\eta\rho_k/2} + 6p_1\gamma^2\varepsilon^2 j.$$

Since $e^{-j\eta\rho_k/2} \le 1$, $p_1 \le p_\star$, and $j \le T_0$,

$$\|V^\top FUW_j\mathbf{1}_{j-1}\|_{p_1,p_1}^2 \le \frac{\gamma^2}{2e^2} + 6p_\star\gamma^2\varepsilon^2T_0.$$

Using the parameter condition

$$6p_\star\varepsilon^2T_0 \le \frac{1}{2e^2},$$

we obtain

$$6p_\star\gamma^2\varepsilon^2T_0 \le \frac{\gamma^2}{2e^2}.$$

Hence

$$\|V^\top FUW_j\mathbf{1}_{j-1}\|_{p_1,p_1}^2 \le \frac{\gamma^2}{e^2}.$$

Taking square roots and then maximizing over $1 \le j \le T_0$ and $F \in \mathcal{E}$ proves

$$\max_{1\le j\le T_0}\max_{F\in\mathcal{E}}\|V^\top FUW_j\mathbf{1}_{j-1}\|_{p_1,p_1} \le \frac{\gamma}{e}.$$

□

Theorem 12 (Finite-support predictable Phase I theorem, D.5 analogue)**.** *Assume that $C_1, \dots, C_{T_0}$ satisfy Assumption C. Assume further that there exists a deterministic finite set $\mathcal{A}_\star \subseteq \mathbb{R}^{d\times d}$ such that*

$$C_t \in \mathcal{A}_\star \qquad \text{a.s. for every } t \le T_0,$$

and set $m_\star := |\mathcal{A}_\star|$. Assume that $m_\star \le T_0^3$. Let $Z_0 \in \mathbb{R}^{d\times k}$ have i.i.d. $N(0,1)$ entries, independent of $C_1, \dots, C_{T_0}$, and set $Q_0 := \mathrm{QR}(Z_0)$. Let Q_t be the Oja iterates from Definition 5 with constant Phase I step size $\eta_t \equiv \eta$. Let

$$p_0 := \left\lceil \log \frac{6k}{\delta} \right\rceil, \qquad p_1 := \left\lceil \log \frac{12T_0(m_\star+1)}{\delta} \right\rceil, \qquad p_\star := \max\{p_0, p_1\}.$$

Choose

$$\gamma := \max\left\{e,\ \frac{144C_\gamma d}{\delta^2}\right\}.$$

Let

$$L_\gamma := \max\left\{1, \log\left(\frac{2e^2 d\gamma^2}{k^{2/p_0}}\right)\right\}.$$

Fix a universal constant $C_\eta \ge 2$, and set

$$\eta := \frac{C_\eta L_\gamma}{\rho_k T_0}.$$

Then for

$$T_0 \ge \max\left\{\frac{4C_\eta L_\gamma M(1+\gamma)}{\rho_k}, \frac{2C_\eta L_\gamma \|\Sigma\|_2}{\rho_k}, \frac{200C_\eta p_\star L_\gamma M^2(1+\gamma)^2}{\rho_k^2}, \frac{48e^2 C_\eta^2 p_\star \gamma^2 M^2(1+\gamma)^2 L_\gamma^2}{\rho_k^2 k^{2/p_0}}\right\}.$$

we have

$$\mathbb{P}\left(\left\|U^\top Q_{T_0}(V^\top Q_{T_0})^{-1}\right\| > 1\right) \le \frac{\delta}{3}.$$

Proof. First,

$$\varepsilon = 2\eta M(1+\gamma) = \frac{2C_\eta L_\gamma M(1+\gamma)}{\rho_k T_0} \le \frac{1}{2}$$

by the first lower bound on T_0. Similarly,

$$\eta\|\Sigma\|_2 = \frac{C_\eta L_\gamma \|\Sigma\|_2}{\rho_k T_0} \le \frac{1}{2}$$

by the second lower bound on T_0.

Next,

$$\begin{aligned} p_\star \varepsilon^2 &= 4p_\star \eta^2 M^2 (1+\gamma)^2 \\ &= \frac{4p_\star C_\eta^2 L_\gamma^2 M^2 (1+\gamma)^2}{\rho_k^2 T_0^2}. \end{aligned}$$

The desired inequality

$$p_\star \varepsilon^2 \le \frac{\eta\rho_k}{50}$$

is equivalent to

$$T_0 \ge \frac{200C_\eta p_\star L_\gamma M^2(1+\gamma)^2}{\rho_k^2},$$

which is exactly the third lower bound on T_0.

For the accumulated-error condition, we compute

$$\begin{aligned} 6p_\star \gamma^2 \varepsilon^2 T_0 &= 24 p_\star \gamma^2 \eta^2 M^2 (1+\gamma)^2 T_0 \\ &= \frac{24 p_\star \gamma^2 C_\eta^2 L_\gamma^2 M^2 (1+\gamma)^2}{\rho_k^2 T_0}. \end{aligned}$$

Thus

$$6p_\star \gamma^2 \varepsilon^2 T_0 \le \frac{1}{2e^2} k^{2/p_0}$$

follows from the fourth lower bound on T_0.

Moreover, since

$$p_0 = \left\lceil \log \frac{6k}{\delta} \right\rceil \geq \log k,$$

we have

$$k^{2/p_0} \leq e^2.$$

Because $\gamma \geq e$, this implies

$$k^{2/p_0} \leq \gamma^2.$$

Therefore the preceding bound also yields

$$6p_\star \varepsilon^2 T_0 \leq \frac{1}{2e^2}.$$

It remains to check the Gaussian initialization requirements. By the assumed choice of γ, with

$$\delta_{\text{init}} = \frac{\delta}{12}, \qquad m = m_\star,$$

Lemma 21 gives

$$\mathbb{P}(G_0^c) \leq \frac{\delta}{12}.$$

Finally, the contraction condition for the random initialization follows from the definition of L_γ. Indeed,

$$T_0 \eta \rho_k = C_\eta L_\gamma,$$

and since $C_\eta \geq 2$,

$$e^{-T_0 \eta \rho_k / 2} = e^{-C_\eta L_\gamma / 2} \leq e^{-L_\gamma}.$$

Since

$$L_\gamma \geq \log\left(\frac{2e^2 d \gamma^2}{k^{2/p_0}} \right),$$

we get

$$d\gamma^2 e^{-T_0 \eta \rho_k / 2} \leq d\gamma^2 e^{-L_\gamma} \leq \frac{1}{2e^2} k^{2/p_0}.$$

Thus all hypotheses of $Proposition$ 3 hold.

By Definition 5,

$$Z_{T_0} = \prod_{s=1}^{T_0} (I + \eta C_s) Q_0$$

and Q_{T_0} is an orthonormal basis for the column span of Z_{T_0}. Hence there exists an invertible $k \times k$ matrix R_{T_0} such that

$$Z_{T_0} = Q_{T_0} R_{T_0}.$$

On G_{T_0}, Lemma 20 implies that $V^\top Z_{T_0}$ is invertible. Therefore, on G_{T_0},

$$\begin{aligned} W_{T_0} &= U^\top Z_{T_0} (V^\top Z_{T_0})^{-1} \\ &= U^\top Q_{T_0} R_{T_0} \left(V^\top Q_{T_0} R_{T_0} \right)^{-1} \\ &= U^\top Q_{T_0} (V^\top Q_{T_0})^{-1}. \end{aligned}$$

Consequently,

$$\mathbb{P}\left(\left\| U^\top Q_{T_0} (V^\top Q_{T_0})^{-1} \right\| > 1 \right) \leq \mathbb{P}\left(\| W_{T_0} \mathbf{1}_{T_0} \| > 1 \right) + \mathbb{P}(G_{T_0}^c).$$

Let

$$p_0 := \left\lceil \log \frac{6k}{\delta} \right\rceil.$$

By Markov's inequality and the definition of $\|\cdot\|_{p,p}$,

$$\mathbb{P}\left(\| W_{T_0} \mathbf{1}_{T_0} \| > 1 \right) \leq \| W_{T_0} \mathbf{1}_{T_0} \|_{p_0, p_0}^{p_0}.$$

By Proposition 3,

$$\|W_{T_0}\mathbf{1}_{T_0}\|_{p_0,p_0} \le e^{-1}k^{1/p_0}.$$

Therefore

$$\mathbb{P}\left(\|W_{T_0}\mathbf{1}_{T_0}\| > 1\right) \le e^{-p_0}k \le \frac{\delta}{6}.$$

It remains to control $\mathbb{P}(G_{T_0}^c)$. Since the good events are decreasing,

$$G_{T_0}^c \subseteq G_0^c \cup \bigcup_{j=1}^{T_0} \left(G_j^c \cap G_{j-1}\right).$$

Hence

$$\mathbb{P}(G_{T_0}^c) \le \mathbb{P}(G_0^c) + \sum_{j=1}^{T_0} \mathbb{P}(G_j^c \cap G_{j-1}).$$

By Proposition 3,

$$\mathbb{P}(G_0^c) \le \frac{\delta}{12}.$$

Now fix $j \in \{1, \dots, T_0\}$. On G_{j-1}, the event G_j^c means that there exists $F \in \mathcal{E}$ such that

$$\left\|V^\top F U W_j\right\| > \gamma.$$

Equivalently,

$$\left\|V^\top F U W_j \mathbf{1}_{j-1}\right\| > \gamma.$$

Therefore, by a union bound over $F \in \mathcal{E}$,

$$\mathbb{P}(G_j^c \cap G_{j-1}) \le \sum_{F\in\mathcal{E}} \mathbb{P}\left(\left\|V^\top F U W_j \mathbf{1}_{j-1}\right\| > \gamma\right).$$

Let

$$p_1 := \left\lceil \log \frac{12T_0(m_\star + 1)}{\delta} \right\rceil.$$

For each fixed $F \in \mathcal{E}$, Markov's inequality gives

$$\mathbb{P}\left(\left\|V^\top F U W_j \mathbf{1}_{j-1}\right\| > \gamma\right) \le \gamma^{-p_1} \left\|V^\top F U W_j \mathbf{1}_{j-1}\right\|_{p_1,p_1}^{p_1}.$$

By Proposition 3,

$$\left\|V^\top F U W_j \mathbf{1}_{j-1}\right\|_{p_1,p_1} \le \frac{\gamma}{e}.$$

Hence

$$\mathbb{P}\left(\left\|V^\top F U W_j \mathbf{1}_{j-1}\right\| > \gamma\right) \le e^{-p_1}.$$

Therefore

$$\mathbb{P}(G_j^c \cap G_{j-1}) \le |\mathcal{E}| e^{-p_1}.$$

Since

$$|\mathcal{E}| \le |\mathcal{A}_\star| = m_\star,$$

we get

$$\sum_{j=1}^{T_0} \mathbb{P}(G_j^c \cap G_{j-1}) \le T_0 m_\star e^{-p_1}.$$

By the definition of p_1,

$$e^{-p_1} \le \frac{\delta}{12T_0(m_\star + 1)}.$$

Thus

$$T_0 m_\star e^{-p_1} \le \frac{\delta}{12}.$$

Consequently,

$$\mathbb{P}(G_{T_0}^c) \le \frac{\delta}{12} + \frac{\delta}{12} = \frac{\delta}{6}.$$

Combining the two estimates,

$$\mathbb{P}\left(\|W_{T_0}\| > 1\right) \le \frac{\delta}{6} + \frac{\delta}{6} = \frac{\delta}{3}.$$

Since

$$W_{T_0} = U^\top Q_{T_0} (V^\top Q_{T_0})^{-1},$$

the claim follows. □

H Lifting from finite support to bounded predictable distributions

We now remove the finite-support restriction. Throughout this section, $C_1, \dots, C_{T_0}$ satisfy Assumption C, but are not assumed to have finite support. For $t \le T_0$, let

$$K_t(\omega, \cdot) := \mathcal{L}(C_t \mid \mathcal{F}_{t-1})(\omega)$$

be a regular conditional law of C_t given the past. Thus, almost surely,

$$\int C \, K_t(\omega, dC) = \Sigma$$

and

$$K_t\left(\left\{C : \|C - \Sigma\|_{(2,k)} \le M\right\}\right) = 1.$$

The argument below reduces the general bounded predictable case to the finite-support case. The only finite-support input used is Theorem 12, applied uniformly over deterministic finite supports of size at most $T_0(L+1)$.

Lemma 22 (Matrix Bernstein input)**.** *Let $X_1, \dots, X_L \in \mathbb{R}^{d\times d}$ be independent, mean-zero, self-adjoint matrices. Suppose*

$$\|X_\ell\|_2 \le R \qquad \text{a.s.}$$

and

$$\left\|\sum_{\ell=1}^{L} \mathbb{E}[X_\ell^2]\right\|_2 \le \sigma^2.$$

Then, for every $s > 0$,

$$\mathbb{P}\left(\left\|\sum_{\ell=1}^{L} X_\ell\right\|_2 \ge s\right) \le 2d \exp\left(-\frac{s^2/2}{\sigma^2 + Rs/3}\right).$$

Lemma 23 (Conditional empirical-mean bound)**.** *Let K be a probability measure on symmetric $d \times d$ matrices such that*

$$\int C \, K(dC) = \Sigma$$

and

$$\|C - \Sigma\|_{(2,k)} \le M \qquad K\text{-a.s.}$$

Let

$$\widehat{C}_1, \dots, \widehat{C}_L \overset{\text{i.i.d.}}{\sim} K, \qquad \widehat{\Sigma}_L := \frac{1}{L}\sum_{\ell=1}^{L} \widehat{C}_\ell.$$

Then, for every $u > 0$,

$$\mathbb{P}\left(\left\|\widehat{\Sigma}_L - \Sigma\right\|_2 \ge u\right) \le 2d \exp\left(-\frac{Lu^2/2}{M^2 + Mu/3}\right).$$

Consequently, if $0 < \alpha \le 1/2$, $0 < \beta < 1$, and

$$L \ge C_0 \frac{k}{\alpha^2} \log \frac{2d}{\beta},$$

for a sufficiently large universal constant C_0, then

$$\mathbb{P}\left(\left\|\widehat{\Sigma}_L - \Sigma\right\|_2 > \frac{\alpha M}{(1-\alpha)\sqrt{k}}\right) \le \beta.$$

Proof. Set

$$X_\ell := \widehat{C}_\ell - \Sigma.$$

Then $X_1, \dots, X_L$ are independent, mean-zero, self-adjoint matrices. Moreover,

$$\|X_\ell\|_2 \le \|X_\ell\|_{(2,k)} \le M.$$

Hence

$$X_\ell^2 \preceq M^2 I_d \qquad \text{and} \qquad \left\| \sum_{\ell=1}^{L} \mathbb{E}[X_\ell^2] \right\|_2 \le LM^2.$$

Applying Lemma 22 to $\sum_{\ell=1}^{L} X_\ell$ gives

$$\mathbb{P}\left(\left\| \frac{1}{L} \sum_{\ell=1}^{L} X_\ell \right\|_2 \ge u \right) \le 2d \exp\left(-\frac{Lu^2/2}{M^2 + Mu/3} \right).$$

This is the first claim. Substituting

$$u = \frac{\alpha M}{(1-\alpha)\sqrt{k}}$$

and using $\alpha \le 1/2$, the second claim follows after increasing C_0. □

Lemma 24 (Mean-preserving finite-support replacement)**.** *Let*

$$\widehat{C}_1, \dots, \widehat{C}_L$$

be symmetric matrices satisfying

$$\left\| \widehat{C}_\ell - \Sigma \right\|_{(2,k)} \le M \qquad \textit{for every } \ell \le L.$$

Let

$$\widehat{\Sigma}_L := \frac{1}{L} \sum_{\ell=1}^{L} \widehat{C}_\ell.$$

Fix $0 < \alpha \le 1/2$*, and assume*

$$\left\| \widehat{\Sigma}_L - \Sigma \right\|_2 \le \frac{\alpha M}{(1-\alpha)\sqrt{k}}.$$

Define

$$\widehat{C}_0 := \Sigma - \frac{1-\alpha}{\alpha} \left(\widehat{\Sigma}_L - \Sigma \right)$$

and

$$\widehat{K}^\sharp := \alpha \delta_{\widehat{C}_0} + (1-\alpha) \frac{1}{L} \sum_{\ell=1}^{L} \delta_{\widehat{C}_\ell}.$$

Then

$$\int C \, \widehat{K}^\sharp(dC) = \Sigma,$$

and

$$\|C - \Sigma\|_{(2,k)} \le M \qquad \widehat{K}^\sharp\textit{-a.s.}$$

Moreover, $\widehat{K}^\sharp$ *has support of cardinality at most* $L + 1$*.*

Proof. The mean identity is immediate:

$$\begin{aligned} \int C \, \widehat{K}^\sharp(dC) &= \alpha \widehat{C}_0 + (1-\alpha) \widehat{\Sigma}_L \\ &= \alpha \Sigma - (1-\alpha)(\widehat{\Sigma}_L - \Sigma) + (1-\alpha) \widehat{\Sigma}_L \\ &= \Sigma. \end{aligned}$$

The original atoms satisfy the required $(2,k)$-bound by assumption. For the correction atom,

$$\begin{aligned}\left\|\widehat{C}_0 - \Sigma\right\|_{(2,k)} &= \frac{1-\alpha}{\alpha}\left\|\widehat{\Sigma}_L - \Sigma\right\|_{(2,k)} \\ &\le \frac{1-\alpha}{\alpha}\sqrt{k}\left\|\widehat{\Sigma}_L - \Sigma\right\|_2 \\ &\le M.\end{aligned}$$

Thus every atom of $\widehat{K}^\sharp$ is M-bounded in the $(2,k)$-norm. □

Lemma 25 (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 can be coupled so that the two next updates disagree with probability at most α. Then*

$$d_{\mathrm{TV}}(\mathsf{P}, \mathsf{P}^\sharp) \le T_0\alpha.$$

Consequently, for every measurable event $\mathcal{A}$ depending on the update path,

$$\mathsf{P}(\mathcal{A}) \le \mathsf{P}^\sharp(\mathcal{A}) + T_0\alpha.$$

Proof. Construct the two processes recursively. As long as the two histories agree, use the assumed one-step coupling. Once a disagreement occurs, continue the two processes arbitrarily. By a union bound, the probability that the two paths ever disagree is at most $T_0\alpha$. Total variation distance is bounded by the probability of disagreement under any coupling, which proves the claim. □

Proposition 4 (Finite-support reduction for predictable kernels)**.** *Let $C_1, \dots, C_{T_0}$ satisfy Assumption C, without assuming finite support. Let Q_t be the Oja iterates from Definition 5 with constant Phase I step size $\eta_t \equiv \eta$. Let $\delta, \xi \in (0,1)$. Set*

$$\alpha := \frac{\xi}{4T_0}, \qquad \beta := \frac{\xi}{4T_0}.$$

Let L be an integer satisfying

$$L \ge C_0 \frac{kT_0^2}{\xi^2}\log\frac{8dT_0}{\xi}, \qquad L+1 \le T_0^2.$$

Assume that the hypotheses of Theorem 12 hold uniformly for every deterministic finite support $\mathcal{A}_\star \subseteq \mathbb{R}^{d\times d}$ satisfying

$$|\mathcal{A}_\star| \le T_0(L+1)$$

with the same constants M, η, γ, δ. Then

$$\mathbb{P}\left(\left\|U^\top Q_{T_0}(V^\top Q_{T_0})^{-1}\right\| > 1\right) \le \frac{\delta}{3} + \xi.$$

Proof. For each $t \le T_0$, write

$$K_t(\omega, \cdot) = \mathcal{L}(C_t \mid \mathcal{F}_{t-1})(\omega).$$

Conditionally on the past, draw auxiliary atoms

$$\widehat{C}_{t,1}, \dots, \widehat{C}_{t,L} \overset{\text{i.i.d.}}{\sim} K_t(\omega, \cdot),$$

and define

$$\widehat{\Sigma}_t := \frac{1}{L}\sum_{\ell=1}^{L}\widehat{C}_{t,\ell}.$$

Let

$$\mathcal{G}_t := \left\{\left\|\widehat{\Sigma}_t - \Sigma\right\|_2 \le \frac{\alpha M}{(1-\alpha)\sqrt{k}}\right\}.$$

By Lemma 23 and the choice of L,

$$\mathbb{P}(\mathcal{G}_t^c \mid \mathcal{F}_{t-1}) \le \beta \qquad \text{a.s.}$$

Therefore, with

$$\mathcal{G} := \bigcap_{t=1}^{T_0} \mathcal{G}_t,$$

we have

$$\mathbb{P}(\mathcal{G}^c) \le T_0\beta = \frac{\xi}{4}.$$

On $\mathcal{G}_t$, define the correction atom

$$\widehat{C}_{t,0} := \Sigma - \frac{1-\alpha}{\alpha}\left(\widehat{\Sigma}_t - \Sigma\right),$$

and define

$$\widehat{K}_t^\sharp := \alpha\delta_{\widehat{C}_{t,0}} + (1-\alpha)\frac{1}{L}\sum_{\ell=1}^{L}\delta_{\widehat{C}_{t,\ell}}.$$

By Lemma 24,

$$\int C\,\widehat{K}_t^\sharp(dC) = \Sigma$$

and

$$\|C - \Sigma\|_{(2,k)} \le M \qquad \widehat{K}_t^\sharp\text{-a.s.}$$

Moreover, $\widehat{K}_t^\sharp$ has at most $L+1$ atoms.

Define the auxiliary finite-support process by drawing, conditionally on the past and on the auxiliary atoms,

$$C_t^\sharp \sim \widehat{K}_t^\sharp \qquad \text{on } \mathcal{G}_t.$$

On $\mathcal{G}_t^c$, draw instead from the empirical kernel

$$\widehat{K}_t := \frac{1}{L}\sum_{\ell=1}^{L}\delta_{\widehat{C}_{t,\ell}}.$$

Then set

$$Q_t^\sharp := \mathrm{QR}\left(Q_{t-1}^\sharp + \eta C_t^\sharp Q_{t-1}^\sharp\right).$$

Now condition on the auxiliary atoms and on $\mathcal{G}$. On this event, define the global auxiliary support

$$\mathcal{A}_\star^\sharp := \bigcup_{t=1}^{T_0}\mathrm{supp}(\widehat{K}_t^\sharp).$$

Then

$$|\mathcal{A}_\star^\sharp| \le T_0(L+1) \le T_0^3.$$

Furthermore,

$$C_t^\sharp \in \mathcal{A}_\star^\sharp \qquad \text{a.s. for every } t \le T_0,$$
$$\mathbb{E}[C_t^\sharp \mid \mathcal{F}_{t-1}^\sharp] = \Sigma,$$

and

$$\left\|C_t^\sharp - \Sigma\right\|_{(2,k)} \le M \qquad \text{a.s.}$$

Therefore Theorem 12 applies to the auxiliary process and gives

$$\mathbb{P}\left(\left\|U^\top Q_{T_0}^\sharp (V^\top Q_{T_0}^\sharp)^{-1}\right\| > 1 \,\middle|\, \text{auxiliary atoms}, \mathcal{G}\right) \le \frac{\delta}{3}.$$

Averaging over the auxiliary atoms and using $\mathbb{P}(\mathcal{G}^c) \le \xi/4$, we obtain

$$\mathbb{P}\left(\left\|U^\top Q_{T_0}^\sharp (V^\top Q_{T_0}^\sharp)^{-1}\right\| > 1\right) \le \frac{\delta}{3} + \frac{\xi}{4}.$$

It remains to compare the auxiliary process with the original process. At time t, couple one draw from K_t with one draw from $\widehat{K}_t^\sharp$ as follows. First draw the auxiliary atoms $\widehat{C}_{t,1}, \dots, \widehat{C}_{t,L}$. Then

draw J_t uniformly from $\{1, \dots, L\}$. After averaging over the atoms, $\widehat{C}_{t,J_t}$ has conditional law K_t. Use $\widehat{C}_{t,J_t}$ as the original update.

For the auxiliary update, use the same atom $\widehat{C}_{t,J_t}$ unless the α-mass correction atom is selected. Thus, whenever the two histories agree up to time $t-1$, the two updates disagree with conditional probability at most α. By Lemma 25,

$$d_{\mathrm{TV}}(\mathsf{P}, \mathsf{P}^\sharp) \le T_0\alpha = \frac{\xi}{4}.$$

Consequently,

$$\begin{aligned}\mathbb{P}\left(\left\|U^\top Q_{T_0}(V^\top Q_{T_0})^{-1}\right\| > 1\right) &\le \mathbb{P}\left(\left\|U^\top Q^\sharp_{T_0}(V^\top Q^\sharp_{T_0})^{-1}\right\| > 1\right) + \frac{\xi}{4} \\ &\le \frac{\delta}{3} + \frac{\xi}{2} \le \frac{\delta}{3} + \xi.\end{aligned}$$

This proves the proposition. □

Lemma 26 (Deterministic tangent-coordinate bound)**.** *Let $Q \in \mathbb{R}^{d\times k}$ have orthonormal columns. If $V^\top Q$ is invertible, then*

$$\operatorname{dist}(Q, V) \le \left\|U^\top Q(V^\top Q)^{-1}\right\|_2.$$

Proof. Let

$$W := U^\top Q(V^\top Q)^{-1}.$$

The column space of Q is the same as the column space of

$$Q(V^\top Q)^{-1} = V + UW.$$

Therefore the principal-angle tangent matrix is W. Since $\sin\theta \le \tan\theta$ for each principal angle,

$$\operatorname{dist}(Q, V) = \|\sin\Theta(Q, V)\|_2 \le \|\tan\Theta(Q, V)\|_2 = \|W\|_2.$$

□

Theorem 13 (Phase I for general bounded predictable Oja updates)**.** *Let $C_1, \dots, C_{T_0}$ satisfy Assumption C, without assuming finite support. Let Q_t be the Oja iterates from Definition 5 with constant Phase I step size $\eta_t \equiv \eta$. Let $\delta, \xi \in (0, 1)$. Suppose there exists an integer L satisfying*

$$L \ge C_0 \frac{kT_0^2}{\xi^2} \log \frac{8dT_0}{\xi}, \qquad L + 1 \le T_0^2.$$

Assume that the hypotheses of Theorem 12 hold uniformly for every deterministic finite support $\mathcal{A}_\star \subseteq \mathbb{R}^{d\times d}$ satisfying

$$|\mathcal{A}_\star| \le T_0(L + 1)$$

with the same constants M, η, γ, δ. Then

$$\mathbb{P}\left(\left\|U^\top Q_{T_0}(V^\top Q_{T_0})^{-1}\right\| > 1\right) \le \frac{\delta}{3} + \xi.$$

Consequently,

$$\operatorname{dist}(Q_{T_0}, V) \le 1$$

with probability at least $1 - \delta/3 - \xi$.

Proof. The probability bound is exactly Proposition 4. The distance statement follows from Lemma 26. □

Lemma 27 (Bounded-noise private updates satisfy the predictable model)**.** *Let $A_1, \dots, A_n \in \mathbb{R}^{d\times d}$ be independent, symmetric, and satisfy*

$$\mathbb{E}[A_i] = \Sigma, \qquad \|A_i - \Sigma\|_{(2,k)} \le M_A \qquad \text{a.s.}$$

Let $T = \lfloor n/B \rfloor$, let $m = B/2$, and define

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

Let

$$\mathcal{H}_t := \sigma(\mathcal{F}_{t-1}, \mathcal{R}_t).$$

Assume that, on the no-clipping event, the private update can be written as

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

where

$$C_t = \frac{2}{B}\sum_{i\in I_t} A_i + G_t, \qquad G_t \mid \mathcal{H}_t \sim \mathrm{GOE}_d(\sigma_t^2) \text{ conditioned on } \|G_t\|_{(2,k)} \le \mathsf{N}_B.$$

Assume σ_t is $\mathcal{H}_t$-measurable. Then, under the sequential bounded-noise law,

$$\mathbb{E}[C_t \mid \mathcal{H}_t] = \Sigma$$

and

$$\|C_t - \Sigma\|_{(2,k)} \le M_A + \mathsf{N}_B \qquad \text{a.s.}$$

Consequently, with

$$M_{\mathrm{priv}} := M_A + \mathsf{N}_B,$$

the updates $C_1, \dots, C_T$ satisfy Assumption C with $M = M_{\mathrm{priv}}$.

Proof. Since I_t is disjoint from the previous blocks and the matrices A_i are independent,

$$\begin{aligned}\mathbb{E}\left[\frac{2}{B}\sum_{i\in I_t} A_i \,\middle|\, \mathcal{H}_t\right] &= \frac{2}{B}\sum_{i\in I_t}\mathbb{E}[A_i] \\ &= \frac{2}{B}\cdot\frac{B}{2}\,\Sigma = \Sigma.\end{aligned}$$

For fixed $\mathcal{H}_t$, the conditional law of G_t is symmetric under $G \mapsto -G$, because both the GOE law and the event

$$\{\|G\|_{(2,k)} \le \mathsf{N}_B\}$$

are symmetric. Hence

$$\mathbb{E}[G_t \mid \mathcal{H}_t] = 0.$$

Therefore

$$\mathbb{E}[C_t \mid \mathcal{H}_t] = \Sigma.$$

Since $\mathcal{F}_{t-1} \subseteq \mathcal{H}_t$, the tower property gives

$$\mathbb{E}[C_t \mid \mathcal{F}_{t-1}] = \mathbb{E}[\mathbb{E}[C_t \mid \mathcal{H}_t] \mid \mathcal{F}_{t-1}] = \Sigma.$$

For the norm bound,

$$\begin{aligned}\|C_t - \Sigma\|_{(2,k)} &\le \left\|\frac{2}{B}\sum_{i\in I_t}(A_i - \Sigma)\right\|_{(2,k)} + \|G_t\|_{(2,k)} \\ &\le \frac{2}{B}\sum_{i\in I_t}\|A_i - \Sigma\|_{(2,k)} + \mathsf{N}_B \\ &\le \frac{2}{B}\cdot\frac{B}{2}M_A + \mathsf{N}_B \\ &= M_A + \mathsf{N}_B.\end{aligned}$$

This proves the claim. □

Theorem 14 (Phase I for bounded-noise private Oja updates)**.** *Work under the sequential bounded-noise law of Lemma 27, and set*

$$M_{\mathrm{priv}} := M_A + \mathsf{N}_B.$$

Let $T_0 = T$. Assume that the hypotheses of Theorem 13 hold with

$$M = M_{\mathrm{priv}}.$$

Equivalently, all Phase I small-step conditions are imposed with

$$\varepsilon = 2\eta M_{\mathrm{priv}}(1+\gamma).$$

Then

$$\mathbb{P}\left(\left\|U^\top Q_T (V^\top Q_T)^{-1}\right\| > 1\right) \le \frac{\delta}{3} + \xi.$$

Consequently,

$$\mathrm{dist}(Q_T, V) \le 1$$

with probability at least $1 - \delta/3 - \xi$.

Proof. By Lemma 27, the private updates satisfy Assumption C with $M = M_{\mathrm{priv}}$. Applying Theorem 13 with this value of M gives the claimed probability bound and the distance statement. □

Theorem 15 (Main theorem for bounded-noise predictable Oja updates)**.** *Let* $A_1, \dots, A_n \in \mathbb{R}^{d\times d}$ *be independent symmetric random matrices satisfying*

$$\mathbb{E}[A_i] = \Sigma, \qquad \|A_i - \Sigma\|_{(2,k)} \le M_A \qquad \textit{a.s.}$$

Let B *be even and define*

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

Let $V \in \mathbb{R}^{d\times k}$ *contain the leading* k *eigenvectors of* Σ*, and assume*

$$\rho_k := \lambda_k - \lambda_{k+1} > 0.$$

Work under the sequential bounded-noise law in which, conditionally on

$$\mathcal{H}_t := \sigma(\mathcal{F}_{t-1}, \mathcal{R}_t),$$

the update matrix is

$$C_t = \frac{2}{B}\sum_{i\in I_t} A_i + G_t, \qquad G_t \mid \mathcal{H}_t \sim \mathrm{GOE}_d(\sigma_t^2) \textit{ conditioned on } \|G_t\|_{(2,k)} \le \mathsf{N}_B,$$

where σ_t *is* $\mathcal{H}_t$*-measurable. Assume that, under this law, the Oja update is*

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

Define the effective boundedness parameter

$$M_{\mathrm{priv}} := M_A + \mathsf{N}_B.$$

Let $\delta \in (0,1)$*, and set*

$$\delta_{\mathrm{I}} := \frac{\delta}{2}, \qquad \delta_{\mathrm{II}} := \frac{2\delta}{3}, \qquad \xi := \frac{\delta}{6}.$$

Let $T_0 < T$ *be the Phase I length. Choose an integer* L *satisfying*

$$L \ge C_0 \frac{kT_0^2}{\xi^2} \log \frac{8dT_0}{\xi},$$

and define

$$m_{\mathrm{lift}} := T_0(L+1).$$

Set

$$p_0 := \left\lceil \log \frac{6k}{\delta_{\mathrm{I}}} \right\rceil, \qquad p_1 := \left\lceil \log \frac{12T_0(m_{\mathrm{lift}}+1)}{\delta_{\mathrm{I}}} \right\rceil, \qquad p_\star := \max\{p_0, p_1\}.$$

Let $C_\gamma > 0$ *be the universal constant from the Gaussian initialization lemma, and choose*

$$\gamma := \max\left\{e,\, C_\gamma \frac{d}{\delta_{\mathrm{I}}^2}\right\}.$$

Define

$$L_\gamma := \max\left\{1,\, \log\left(\frac{2e^2 d\gamma^2}{k^{2/p_0}}\right)\right\}.$$

Let $C_\eta \geq 2$ be a sufficiently large universal constant and set the Phase I step size

$$\eta_t \equiv \eta := \frac{C_\eta L_\gamma}{\rho_k T_0}, \qquad 1 \leq t \leq T_0.$$

Assume T_0 is large enough that

$$T_0 \geq \max\left\{ \frac{4C_\eta L_\gamma M_{\mathrm{priv}}(1+\gamma)}{\rho_k}, \frac{2C_\eta L_\gamma \|\Sigma\|_2}{\rho_k}, \right.$$
$$\left. \frac{200 C_\eta p_\star L_\gamma M_{\mathrm{priv}}^2 (1+\gamma)^2}{\rho_k^2}, \frac{48 e^2 C_\eta^2 p_\star L_\gamma^2 M_{\mathrm{priv}}^2 (1+\gamma)^2}{\rho_k^2} \right\}.$$

For $t > T_0$, use the Phase II step size

$$\eta_t = \Theta\left(\frac{1}{\rho_k(\beta + t - T_0)}\right), \qquad \beta = \widetilde{\Theta}\left(\frac{M_{\mathrm{priv}}^2}{\rho_k^2}\right),$$

with constants chosen as in the predictable Phase II theorem, with failure probability δ_{II}.

Then, for every $T > T_0$, the output Q_T satisfies

$$\left\|Q_T Q_T^\top - V V^\top\right\|_F \leq C' \frac{M_{\mathrm{priv}}}{\rho_k} \sqrt{\frac{\log\left(\frac{M_{\mathrm{priv}} k}{\rho_k \delta}\right)}{T - T_0}}$$

with probability at least $1 - \delta$, where $C' > 0$ is a universal constant.

Proof. By Lemma 27, under the sequential bounded-noise law,

$$\mathbb{E}[C_t \mid \mathcal{F}_{t-1}] = \Sigma$$

and

$$\|C_t - \Sigma\|_{(2,k)} \leq M_A + \mathsf{N}_B = M_{\mathrm{priv}} \qquad \text{a.s.}$$

Thus the updates $C_1, \ldots, C_T$ satisfy Assumption C with $M = M_{\mathrm{priv}}$.

The empirical lifting construction uses supports of size at most

$$m_{\mathrm{lift}} = T_0(L+1).$$

With the above definitions of $p_0, p_1, p_\star, \gamma, L_\gamma, \eta$, the displayed lower bound on T_0 implies the Phase I small-step and length conditions required by Theorem 13 with $M = M_{\mathrm{priv}}$, failure parameter δ_{I}, and lifting error ξ. Therefore

$$\mathbb{P}\left(\mathrm{dist}(Q_{T_0}, V) > 1\right) \leq \frac{\delta_{\mathrm{I}}}{3} + \xi = \frac{\delta}{6} + \frac{\delta}{6} = \frac{\delta}{3}.$$

On the event $\mathrm{dist}(Q_{T_0}, V) \leq 1$, the predictable Phase II theorem applies to the martingale-difference updates $C_t - \Sigma$, using

$$\mathbb{E}[C_t - \Sigma \mid \mathcal{F}_{t-1}] = 0, \qquad \|C_t - \Sigma\|_{(2,k)} \leq M_{\mathrm{priv}}.$$

With the stated decaying step sizes, it gives

$$\left\|Q_T Q_T^\top - V V^\top\right\|_F \leq C' \frac{M_{\mathrm{priv}}}{\rho_k} \sqrt{\frac{\log\left(\frac{M_{\mathrm{priv}} k}{\rho_k \delta}\right)}{T - T_0}}$$

with conditional failure probability at most

$$\delta_{\mathrm{II}} = \frac{2\delta}{3}.$$

Combining the Phase I and Phase II failure probabilities gives total failure probability at most

$$\frac{\delta}{3} + \frac{2\delta}{3} = \delta.$$

This proves the theorem. □