A. Proofs

A.1. Proof of Theorem 1

Although the proof structure generally mimics the proof of Theorem 1 in (Shamir, 2015) for the $k = 1$ special case, it is more intricate and requires several new technical tools. To streamline the presentation of the proof, we begin with proving a series of auxiliary lemmas in Subsection A.1.1, and then move to the main proof in Subsection A.1. The main proof itself is divided into several steps, each constituting one or more lemmas.

Throughout the proof, we use the well-known facts that for all matrices B, C, D of suitable dimensions, $\mathrm{Tr}(B + C) = \mathrm{Tr}(B) + \mathrm{Tr}(C)$, $\mathrm{Tr}(BC) = \mathrm{Tr}(CB)$, $\mathrm{Tr}(BCD) = \mathrm{Tr}(DBC)$, and $\mathrm{Tr}(B^\top B) = \|B\|_F^2$. Moreover, since Tr is a linear operation, $\mathbb{E}[\mathrm{Tr}(B)] = \mathbb{E}[\mathrm{Tr}(B)]$ for a random matrix B.

A.1.1. Auxiliary Lemmas

Lemma 2. *For any $B, C, D \succeq 0$, it holds that $\mathrm{Tr}(BC) \geq \mathrm{Tr}(B(C - D))$ and $\mathrm{Tr}(BC) \geq \mathrm{Tr}((B - D)C)$.*

Proof. It is enough to prove that for any positive semidefinite matrices E, G, it holds that $\mathrm{Tr}(EG) \geq 0$. The lemma follows by taking either $E = B, G = D$ (in which case, $\mathrm{Tr}(BC) = \mathrm{Tr}(B(C - D)) + \mathrm{Tr}(BD) \geq \mathrm{Tr}(B(C - D))$), or $E = D, G = C$ (in which case, $\mathrm{Tr}(BC) = \mathrm{Tr}((B - D)C) + \mathrm{Tr}(DC) \geq \mathrm{Tr}((B - D)C)$).

Any positive semidefinite matrix M can be written as the product $M^{1/2}M^{1/2}$ for some symmetric matrix $M^{1/2}$ (known as the matrix square root of M). Therefore,

$$\begin{aligned}\mathrm{Tr}(EG) &= \mathrm{Tr}(E^{1/2}E^{1/2}G^{1/2}G^{1/2}) = \mathrm{Tr}(G^{1/2}E^{1/2}E^{1/2}G^{1/2}) \\ &= \mathrm{Tr}((E^{1/2}G^{1/2})^\top (E^{1/2}G^{1/2})) = \|E^{1/2}G^{1/2}\|_F^2 \geq 0.\end{aligned}$$

□

Lemma 3. *If $B \succeq 0$ and $C \succ 0$, then*

$$\mathrm{Tr}(BC^{-1}) \geq \mathrm{Tr}(B(2I - C)),$$

where I is the identity matrix.

Proof. We begin by proving the one-dimensional case, where B, C are scalars $b \geq 0, c > 0$. The inequality then becomes $bc^{-1} \geq b(2 - c)$, which is equivalent to $1 \geq c(2 - c)$, or upon rearranging, $(c - 1)^2 \geq 0$, which trivially holds.

Turning to the general case, we note that by Lemma 2, it is enough to prove that $C^{-1} - (2I - C) \succeq 0$. To prove this, we make a couple of observations. The positive definite matrix C (like any positive definite matrix) has a singular value decomposition which can be written as $USU^\top$, where U is an orthogonal matrix, and S is a diagonal matrix with positive entries. Its inverse is $US^{-1}U^\top$, and $2I - C = 2I - USU^\top = U(2I - S)U^\top$. Therefore,

$$C^{-1} - (2I - C) = US^{-1}U^\top - U(2I - S)U^\top = U(S^{-1} - (2I - S))U^\top.$$

To show this matrix is positive semidefinite, it is enough to show that each diagonal entry of $S^{-1} - (2I - S)$ is non-negative. But this reduces to the one-dimensional result we already proved, when $b = 1$ and $c > 0$ is any diagonal entry in S. Therefore, $C^{-1} - (2I - C) \succeq 0$, from which the result follows. □

Lemma 4. *For any matrices B, C,*

$$\mathrm{Tr}(BC) \leq \|B\|_F \|C\|_F$$

and

$$\|BC\|_F \leq \|B\|_2 \|C\|_F.$$

Proof. The first inequality is immediate from Cauchy-Shwartz. As to the second inequality, letting $\mathbf{c}_i$ denote the i-th column of C, and $\|\cdot\|_2$ the Euclidean norm for vectors,

$$\|BC\|_F = \sqrt{\sum_i \|B\mathbf{c}_i\|_2^2} \leq \sqrt{\sum_i \left(\|B\|_2 \|\mathbf{c}_i\|_2\right)^2} = \|B\|_2 \sqrt{\sum_i \|\mathbf{c}_i\|_2^2} = \|B\|_2 \|C\|_F.$$

□

Lemma 5. *Let B_1, B_2, Z_1, Z_2 be $k \times k$ square matrices, where B_1, B_2 are fixed and Z_1, Z_2 are stochastic and zero-mean (i.e. their expectation is the all-zeros matrix). Furthermore, suppose that for some fixed $\alpha, \gamma, \delta > 0$, it holds with probability 1 that*

- *For all $\nu \in [0,1]$, $B_2 + \nu Z_2 \succeq \delta I$.*
- $\max\{\|Z_1\|_F, \|Z_2\|_F\} \leq \alpha$.
- $\|B_1 + \eta Z_1\|_2 \leq \gamma$.

Then

$$\mathbb{E}\left[\mathrm{Tr}\left((B_1+Z_1)(B_2+Z_2)^{-1}\right)\right] \geq \mathrm{Tr}(B_1 B_2^{-1}) - \frac{\alpha^2(1+\gamma/\delta)}{\delta^2}.$$

Proof. Define the function

$$f(\nu) = \mathrm{Tr}\left((B_1+\nu Z_1)(B_2+\nu Z_2)^{-1}\right) \ , \quad \nu \in [0,1].$$

Since $B_2 + \nu Z_2$ is positive definite, it is always invertible, hence $f(\nu)$ is indeed well-defined. Moreover, it can be differentiated with respect to ν, and we have

$$f'(\nu) = \mathrm{Tr}\left(Z_1(B_2+\nu Z_2)^{-1} - (B_1+\nu Z_1)(B_2+\nu Z_2)^{-1}Z_2(B_2+\nu Z_2)^{-1}\right).$$

Again differentiating with respect to ν, we have

$$\begin{aligned} f''(\nu) \ &= \ \mathrm{Tr}\Big(-2Z_1(B_2+\nu Z_2)^{-1}Z_2(B_2+\nu Z_2)^{-1} \\ &\qquad + 2(B_1+\nu Z_1)(B_2+\nu Z_2)^{-1}Z_2(B_2+\nu Z_2)^{-1}Z_2(B_2+\nu Z_2)^{-1}\Big) \\ &= 2\,\mathrm{Tr}\Big(\Big(-Z_1 + (B_1+\nu Z_1)(B_2+\nu Z_2)^{-1}Z_2\Big)(B_2+\nu Z_2)^{-1}Z_2(B_2+\nu Z_2)^{-1}\Big). \end{aligned}$$

Using Lemma 4 and the triangle inequality, this is at most

$$\begin{aligned} &2\| -Z_1 + (B_1+\nu Z_1)(B_2+\nu Z_2)^{-1}Z_2\|_F \|(B_2+\nu Z_2)^{-1}Z_2(B_2+\nu Z_2)^{-1}\|_F \\ &\leq 2\left(\|Z_1\|_F + \|(B_1+\nu Z_1)(B_2+\nu Z_2)^{-1}Z_2\|_F\right)\|(B_2+\nu Z_2)^{-1}\|_2^2\|Z_2\|_F \\ &\leq 2\left(\|Z_1\|_F + \|B_1+\nu Z_1\|_2\|\left(B_2+\nu Z_2\right)^{-1}\|_2\|Z_2\|_F\right)\|(B_2+\nu Z_2)^{-1}\|_2^2\|Z_2\|_F \\ &\leq 2\left(\alpha + \gamma\frac{1}{\delta}\alpha\right)\frac{1}{\delta^2}\alpha \ = \ \frac{2\alpha^2(1+\gamma/\delta)}{\delta^2}. \end{aligned}$$

Applying a Taylor expansion to $f(\cdot)$ around $\nu = 0$, with a Lagrangian remainder term, and substituting the values for $f'(\nu), f''(\nu)$, we can lower bound $f(1)$ as follows:

$$\begin{aligned} f(1) &\geq f(0) + f'(0)*(1-0) - \frac{1}{2}\max_\nu |f''(\nu)|*(1-0)^2 \\ &= \mathrm{Tr}\left(B_1B_2^{-1}\right) + \mathrm{Tr}\left(Z_1B_2^{-1} - B_1B_2^{-1}Z_2B_2^{-1}\right) - \frac{\alpha^2(1+\gamma/\delta)}{\delta^2}. \end{aligned}$$

Taking expectation over Z_1, Z_2, and recalling they are zero-mean, we get that

$$\mathbb{E}[f(1)] \geq \mathrm{Tr}\left(B_1B_2^{-1}\right) - \frac{\alpha^2(1+\gamma/\delta)}{\delta^2}.$$

Since $\mathbb{E}[f(1)] = \mathbb{E}\left[\mathrm{Tr}\left((B_1+Z_1)(B_2+Z_2)^{-1}\right)\right]$, the result in the lemma follows. □

Lemma 6. *Let $U_1, \ldots, U_k$ and R_1, R_2 be positive semidefinite matrices, such that $R_2 - R_1 \succeq 0$, and define the function*

$$f(x_1 \ldots x_k) = \mathrm{Tr}\left(\left(\sum_{i=1}^k x_i U_i + R_1\right)\left(\sum_{i=1}^k x_i U_i + R_2\right)^{-1}\right).$$

over all $(x_1 \ldots x_k) \in [\alpha, \beta]^d$ for some $\beta \geq \alpha \geq 0$. Then $\min_{(x_1 \ldots x_k) \in [\alpha,\beta]^d} f(\mathbf{x}) = f(\alpha, \ldots, \alpha)$.

Proof. Taking a partial derivative of f with respect to some x_j, we have

$$\begin{aligned}
&\frac{\partial}{\partial x_j} f(\mathbf{x}) \\
&= \mathrm{Tr}\left(U_j \left(\sum_{i=1}^{k} x_i U_i + R_2 \right)^{-1} - \left(\sum_{i=1}^{k} x_i U_i + R_1 \right) \left(\sum_{i=1}^{k} x_i U_i + R_2 \right)^{-1} U_j \left(\sum_{i=1}^{k} x_i U_i + R_2 \right)^{-1} \right) \\
&= \mathrm{Tr}\left(\left(I - \left(\sum_{i=1}^{k} U_i + R_1 \right) \left(\sum_{i=1}^{k} x_i U_i + R_2 \right)^{-1} \right) U_j \left(\sum_{i=1}^{k} x_i U_i + R_2 \right)^{-1} \right) \\
&= \mathrm{Tr}\left(\left(\left(\sum_{i=1}^{k} x_i U_i + R_2 \right) - \left(\sum_{i=1}^{k} x_i U_i + R_1 \right) \right) \left(\sum_{i=1}^{k} x_i U_i + R_2 \right)^{-1} U_j \left(\sum_{i=1}^{k} x_i U_i + R_2 \right)^{-1} \right) \\
&= \mathrm{Tr}\left((R_2 - R_1) \left(\sum_{i=1}^{k} x_i U_i + R_2 \right)^{-1} U_j \left(\sum_{i=1}^{k} x_i U_i + R_2 \right)^{-1} \right).
\end{aligned}$$

By the lemma's assumptions, each matrix in the product above is positive semidefinite, hence the product is positive semidefinite, and the trace is non-negative. Therefore, $\frac{\partial}{\partial x_j} f(\mathbf{x}) \geq 0$, which implies that the function is minimized when each x_j takes its smallest possible value, i.e. α. □

Lemma 7. *Let B be a $k \times k$ matrix with minimal singular value δ. Then*

$$1 - \frac{\|B^\top B\|_F^2}{\|B\|_F^2} \;\geq\; \max\left\{ 1 - \|B\|_F^2 \,,\; \frac{\delta^2}{k}\left(k - \|B\|_F^2\right) \right\}.$$

Proof. We have

$$1 - \frac{\|B^\top B\|_F^2}{\|B\|_F^2} \geq 1 - \frac{\|B\|_F^2 \|B\|_F^2}{\|B\|_F^2} = 1 - \|B\|_F^2,$$

so it remains to prove $1 - \frac{\|B^\top B\|_F^2}{\|B\|_F^2} \geq \frac{\delta^2}{k}\left(k - \|B\|_F^2\right)$. Let $\sigma_1, \ldots, \sigma_k$ denote the vector of singular values of B. The singular values of $B^\top B$ are $\sigma_1^2, \ldots, \sigma_k^2$, and the Frobenius norm of a matrix equals the Euclidean norm of its vector of singular values. Therefore, the lemma is equivalent to requiring

$$1 - \frac{\sum_{i=1}^k \sigma_i^4}{\sum_{i=1}^k \sigma_i^2} \;\geq\; \frac{\delta^2}{k}\left(k - \sum_{i=1}^k \sigma_i^2 \right),$$

assuming $\sigma_i \in [\delta, 1]$ for all i. This holds since

$$1 - \frac{\sum_i \sigma_i^4}{\sum_i \sigma_i^2} \;=\; \frac{\sum_i \sigma_i^2 - \sum_i \sigma_i^4}{\sum_i \sigma_i^2} \;=\; \frac{\sum_i \sigma_i^2 \left(1 - \sigma_i^2\right)}{\sum_i \sigma_i^2} \;\geq\; \frac{\delta^2 \sum_i \left(1 - \sigma_i^2\right)}{k} \;=\; \frac{\delta^2}{k}\left(k - \sum_i \sigma_i^2 \right).$$

□

Lemma 8. *For any $d \times k$ matrices C, D with orthonormal columns, let*

$$D_C = \arg \min_{DB \,:\, (DB)^\top (DB) = I} \|C - DB\|_F^2$$

be the nearest orthonormal-columns matrix to C in the column space of D (where B is a $k \times k$ matrix). Then the matrix B minimizing the above equals $B = VU^\top$, where $C^\top D = USV^\top$ is the SVD decomposition of $C^\top D$, and it holds that

$$\|C - D_C\|_F^2 \leq 2(k - \|C^\top D\|_F^2).$$

Proof. Since D has orthonormal columns, we have $D^\top D = I$, so the definition of B is equivalent to

$$B = \arg \min_{B \,:\, B^\top B = I} \|C - DB\|_F^2.$$

This is the orthogonal Procrustes problem (see e.g. (Golub & Van Loan, 2012)), and the solution is easily shown to be $B = VU^\top$ where $USV^\top$ is the SVD decomposition of $C^\top D$. In this case, and using the fact that $\|C\|_F^2 = \|D\|_F^2 = k$ (as C, D have orthonormal columns), we have that $\|C - D_C\|_F^2$ equals

$$\|C - DB\|_F^2 = \|C\|_F^2 + \|D\|_F^2 - 2\,\mathrm{Tr}(C^\top DB) = 2\left(k - \mathrm{Tr}(USV^\top(VU^\top))\right) = 2\left(k - \mathrm{Tr}(USU^\top)\right).$$

Since the trace function is similarity-invariant, this equals $2k - \mathrm{Tr}(S)$. Let $s_1 \ldots, s_k$ be the diagonal elements of S, and note that they can be at most 1 (since they are the singular values of $C^\top D$, and both C and D have orthonormal columns). Recalling that the Frobenius norm equals the Euclidean norm of the singular values, we can therefore upper bound the above as follows:

$$2\left(k - \mathrm{Tr}(USU^\top)\right) = 2\left(k - \mathrm{Tr}(S)\right) = 2\left(k - \sum_{i=1}^k s_i\right) \leq 2\left(k - \sum_{i=1}^k s_i^2\right) = 2\left(k - \|C^\top D\|_F^2\right).$$

□

Lemma 9. *Let W_t, W_t' be as defined in Algorithm 2, where we assume $\eta < \frac{1}{3}$. Then for any $d \times k$ matrix V_k with orthonormal columns, it holds that*

$$\left|\|V_k^\top W_t\|_F^2 - \|V_k^\top W_{t-1}\|_F^2\right| \leq \frac{12k\eta}{1-3\eta}.$$

Proof. Letting $\mathbf{s}_t, \mathbf{s}_{t-1}$ denote the vectors of singular values of $V_k^\top W_t$ and $V_k^\top W_{t-1}$, and noting that they are both in $[0,1]^k$ (as V_k, W_{t-1}, W_t all have orthonormal columns), the left hand side of the inequality in the lemma statement equals

$$|\|\mathbf{s}_t\|^2 - \|\mathbf{s}_{t-1}\|^2| = (\|\mathbf{s}_t\|_2 + \|\mathbf{s}_{t-1}\|_2)\,|\,\|\mathbf{s}_t\|_2 - \|\mathbf{s}_{t-1}\|_2\,| \leq 2\sqrt{k}\|\mathbf{s}_t - \mathbf{s}_{t-1}\|_2 \leq 2k\|\mathbf{s}_t - \mathbf{s}_{t-1}\|_\infty,$$

where $\|\cdot\|_\infty$ is the infinity norm. By Weyl's matrix perturbation theorem[4] (Horn & Johnson, 2012), this is upper bounded by

$$2k\|V_k^\top W_t - V_k^\top W_{t-1}\|_2 \leq 2k\|V_k\|_2\|W_t - W_{t-1}\|_2 \leq 2k\|W_t - W_{t-1}\|_2. \tag{8}$$

Recalling the relationship between W_t and W_{t-1} from Algorithm 2, we have that

$$W_t' = W_{t-1} + \eta N,$$

where

$$\|N\|_2 \leq \|\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top W_{t-1}\|_2 + \|\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top \tilde{W}_{s-1}B_{t-1}\|_2 + \|\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top \tilde{W}_{s-1}B_{t-1}\|_2 \leq 3,$$

as $W_{t-1}, \tilde{W}_{s-1}, B_{t-1}$ all have orthonormal columns, and $\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top$ and $\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$ have spectral norm at most 1. Therefore, W_t' equals W_{t-1}, up to a matrix perturbation of spectral norm at most 3η. Again by Weyl's theorem, this implies that the k non-zero singular values of the $d \times k$ matrix W_t' are different from those of W_{t-1} (which has orthonormal columns) by at most 3η, and hence all lie in $[1-3\eta, 1+3\eta]$. As a result, the singular values of $\left(W_t'^\top W_t'\right)^{-1/2}$ all lie in $\left[\frac{1}{1+3\eta}, \frac{1}{1-3\eta}\right]$. Collecting these observations, we have

$$\begin{aligned}
\|W_t - W_{t-1}\|_2 &= \|(W_{t-1} + \eta N)\left(W_{t-1}'^\top W_{t-1}'\right)^{-1/2} - W_{t-1}\|_2 \\
&\leq \|W_{t-1}\left(\left(W_{t-1}'^\top W_{t-1}'\right)^{-1/2} - I\right) + \eta N\left(W_{t-1}'^\top W_{t-1}'\right)^{-1/2}\|_2 \\
&\leq \|\left(W_{t-1}'^\top W_{t-1}'\right)^{-1/2} - I\|_2 + \eta\|N\|_2\|\left(W_{t-1}'^\top W_{t-1}'\right)^{-1/2}\|_2 \\
&\leq \frac{3\eta}{1-3\eta} + \frac{3\eta}{1-3\eta} \;=\; \frac{6\eta}{1-3\eta}.
\end{aligned}$$

Plugging back to Eq. (8), the result follows. □

[4] Using its version for singular values, which implies that the singular values of matrices B and $B+E$ are different by at most $\|E\|_2$.

A.1.2. MAIN PROOF

To simplify the technical derivations, note that the algorithm remains the same if we divide each $\mathbf{x}_i$ by $\sqrt{r}$, and multiply η by r. Since $\max_i \|\mathbf{x}_i\|^2 \leq r$, this corresponds to running the algorithm with step-size ηr rather than η, on a re-scaled dataset of points with squared norm at most 1, and with an eigengap of λ/r instead of λ. Therefore, we can simply analyze the algorithm assuming that $\max_i \|\mathbf{x}_i\|^2 \leq 1$, and in the end plug in λ/r instead of λ, and ηr instead of η, to get a result which holds for data with squared norm at most r.

PART I: ESTABLISHING A STOCHASTIC RECURRENCE RELATION

We begin by focusing on a single iteration t of the algorithm, and analyze how $\|V_k^\top W_t\|_F^2$ (which measures the similarity between the column spaces of V_k and W_t) evolves during that iteration. The key result we need is Lemma 10 below, which is specialized for our algorithm in Lemma 11.

Lemma 10. *Let A be a $d \times d$ symmetric matrix with all eigenvalues $s_1 \geq s_2 \geq \ldots \geq s_d$ in $[0,1]$, and suppose that $s_k - s_{k+1} \geq \lambda$ for some $\lambda > 0$.*

Let N be a $d \times k$ zero-mean random matrix such that $\|N\|_F \leq \sigma_N^F$ and $\|N\|_2 \leq \sigma_N^{sp}$ with probability 1, and define

$$r_N = 46\,(\sigma_N^F)^2\left(1 + \frac{8}{3}\left(\frac{1}{4}\sigma_N^{sp} + 2\right)^2\right)$$

Let W be a $d \times k$ matrix with orthonormal columns, and define

$$W' = (I + \eta A)W + \eta N \;\;,\;\; W'' = W'(W'^\top W')^{-1/2},$$

for some $\eta \in \left[0, \frac{1}{4\max\{1,\sigma_N^F\}}\right]$.

If $V_k = [\mathbf{v}_1, \mathbf{v}_2 \ldots, \mathbf{v}_k]$ is the $d \times k$ matrix of A's first k eigenvectors, then the following holds:

- $\mathbb{E}\left[1 - \|V_k^\top W''\|_F^2\right] \leq \left(1 - \frac{4}{5}\eta\lambda\|V_k^\top W\|_F^2\right)\left(1 - \|V_k^\top W\|_F^2\right) + \eta^2 r_N$
- *If $\|V_k^\top W\|_F^2 \geq k - \frac{1}{2}$, then*

$$\mathbb{E}_N\left[k - \|V_k^\top W''\|_F^2\right] \leq \left(k - \|V_k^\top W\|_F^2\right)\left(1 - \frac{1}{10}\eta\lambda\right) + \eta^2 r_N.$$

Proof. Using the fact that $\mathrm{Tr}(BCD) = \mathrm{Tr}(CDB)$ for any matrices B, C, D, we have

$$\begin{aligned}
\mathbb{E}\left[\|V_k^\top W''\|_F^2\right] &= \mathbb{E}\left[\mathrm{Tr}\left(W''^\top V_k V_k^\top W''\right)\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left(\left(W'^\top W'\right)^{-1/2} W'^\top V_k V_k^\top W' \left(W'^\top W'\right)^{-1/2}\right)\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left(\left(W'^\top V_k V_k^\top W'\right)\left(W'^\top W'\right)^{-1}\right)\right]. \qquad (9)
\end{aligned}$$

By definition of W', we have

$$\begin{aligned}
W'^\top V_k V_k^\top W' &= \left((I + \eta A)W + \eta N\right)^\top V_k V_k^\top \left((I + \eta A)W + \eta N\right) \\
&= B_1 + Z_1,
\end{aligned}$$

where we define

$$\begin{aligned}
B_1 &= W^\top (I + \eta A) V_k V_k^\top (I + \eta A) W + \eta^2 N^\top V_k V_k^\top N \\
Z_1 &= \eta N^\top V_k V_k^\top (I + \eta A) W + \eta W^\top (I + \eta A) V_k V_k^\top N.
\end{aligned}$$

Also, we have

$$\begin{aligned} W'^{\top}W' &= \left((I+\eta A)W+\eta N\right)^{\top}\left((I+\eta A)W+\eta N\right) \\ &= B_2 + Z_2, \end{aligned}$$

where

$$\begin{aligned} B_2 &= W^{\top}(I+\eta A)(I+\eta A)W + \eta^2 N^{\top}N \\ Z_2 &= \eta N^{\top}(I+\eta A)W + \eta W^{\top}(I+\eta A)N. \end{aligned}$$

With these definitions, we can rewrite Eq. (9) as $\mathbb{E}\left[\mathrm{Tr}((B_1+Z_1)(B_2+Z_2)^{-1})\right]$. We now wish to remove Z_1, Z_2, by applying Lemma 5. To do so, we check the lemma's conditions:

- Z_1, Z_2 *are zero mean*: This holds since they are linear in N, and N is assumed to be zero-mean.
- $B_2 + \nu Z_2 \succeq \frac{3}{8}I$ *for all* $\nu \in [0,1]$: Recalling the definition of B_2, Z_2, and the facts that $A \succeq 0$, $N^{\top}N \succeq 0$ (by construction), and $W^{\top}W = I$, we have that $B_2 \succeq I$. Moreover, the spectral norm of Z_2 is at most
$$2\eta\|N^{\top}(I+\eta A)W\|_2 \leq 2\eta\|N\|_2\|I+\eta A\|_2\|W\|_2 \leq 2\eta\sigma_N^{sp}(1+\eta) \leq 2\eta\sigma_N^{F}(1+\eta),$$
which by the assumption on η is at most $2\frac{1}{4}\left(1+\frac{1}{4}\right) = \frac{5}{8}$. This implies that the smallest singular value of $B_2 + \nu Z_2$ is at least $1 - \nu(5/8) \geq 3/8$.
- $\max\{\|Z_1\|_F, \|Z_2\|_F\} \leq \frac{5}{2}\eta\sigma_N^F$: By definition of Z_1, Z_2, and using Lemma 4, the Frobenius norm of these two matrices is at most
$$2\eta\|N\|_F\|(I+\eta A)\|_2\|W\|_2 \leq 2\eta\sigma_N^F(1+\eta),$$
which by the assumption on η is at most $2\eta\sigma_N^F\left(1+\frac{1}{4}\right) = \frac{5}{2}\eta\sigma_N^F$.
- $\|B_1 + \eta Z_1\|_2 \leq \left(\frac{1}{4}\sigma_N^{sp} + 2\right)^2$: Using the definition of B_1, Z_1 and the assumption $\eta \leq \frac{1}{4}$,
$$\begin{aligned} \|B_1+\eta Z_1\|_2 &\leq \|B_1\|_2 + \eta\|Z_1\|_2 \\ &\leq (1+\eta)^2 + \eta^2(\sigma_N^{sp})^2 + 2\eta\sigma_N^{sp}(1+\eta) \\ &\leq \left(\frac{5}{4}\right)^2 + \frac{1}{16}(\sigma_N^{sp})^2 + \frac{5}{8}\sigma_N^{sp} \\ &< \left(\frac{1}{4}\sigma_N^{sp} + 2\right)^2. \end{aligned}$$

Applying Lemma 5 and plugging back to Eq. (9), we get

$$\begin{aligned} \mathbb{E}\left[\|V_k^{\top}W''\|_F^2\right] &\geq \mathbb{E}\left[\mathrm{Tr}((B_1+Z_1)(B_2+Z_2)^{-1})\right] \\ &\geq \mathrm{Tr}\left(B_1B_2^{-1}\right) - \frac{400}{9}(\eta\sigma_N^F)^2\left(1+\frac{8}{3}\left(\frac{1}{4}\sigma_N^{sp}+2\right)^2\right). \end{aligned} \tag{10}$$

We now turn to lower bound $\mathrm{Tr}\left(B_1B_2^{-1}\right)$, by first re-writing B_1, B_2 in a different form. For $i = 1, \ldots, d$, let

$$U_i = W^{\top}\mathbf{v}_i\mathbf{v}_i^{\top}W,$$

where $\mathbf{v}_i$ is the eigenvector of A corresponding to the eigenvalue s_i. Note that each U_i is positive semidefinite, and $\sum_{i=1}^d U_i = W^{\top}W = I$. We have

$$\begin{aligned} B_1 &= W^{\top}(I+\eta A)V_kV_k^{\top}(I+\eta A)W + \eta^2 N^{\top}V_kV_k^{\top}N \\ &= W^{\top}\left((I+\eta A)V_k\right)\left((I+\eta A)V_k\right)^{\top}W + \eta^2 N^{\top}V_kV_k^{\top}N \\ &= \sum_{i=1}^k (1+\eta s_i)^2 W^{\top}\mathbf{v}_i\mathbf{v}_i^{\top}W + \eta^2 N^{\top}V_kV_k^{\top}N \\ &= \sum_{i=1}^k (1+\eta s_i)^2 U_i + \eta^2 N^{\top}V_kV_k^{\top}N. \end{aligned} \tag{11}$$

Similarly,

$$\begin{aligned} B_2 &= W^\top (I+\eta A)(I+\eta A)W + \eta^2 N^\top N \\ &= \sum_{i=1}^{d} (1+\eta s_i)^2 W^\top \mathbf{v}_i \mathbf{v}_i^\top W + \eta^2 N^\top N \\ &= \sum_{i=1}^{d} (1+\eta s_i)^2 U_i + \eta^2 N^\top N. \end{aligned} \tag{12}$$

Plugging Eq. (11) and Eq. (12) back into Eq. (10), we get

$$\begin{aligned} \mathbb{E}\left[\|V_k^\top W''\|_F^2\right] \geq \operatorname{Tr}\left(\left(\sum_{i=1}^{k}(1+\eta s_1)^2 U_i + \eta^2 N^\top V_k V_k^\top N\right)\left(\sum_{i=1}^{d}(1+\eta s_i)^2 U_i + \eta^2 N^\top N\right)^{-1}\right) \\ - \frac{400}{9}(\eta\sigma_N^F)^2\left(1+\frac{8}{3}\left(\frac{1}{4}\sigma_N^{sp}+2\right)^2\right). \end{aligned} \tag{13}$$

Recalling that $s_1 \geq s_2 \geq \ldots \geq s_k$ and letting $\alpha = (1+\eta s_k)^2, \beta = (1+\eta s_1)^2$, the trace term can be lower bounded by

$$\min_{x_1,\ldots,x_k \in [\alpha,\beta]} \operatorname{Tr}\left(\left(\sum_{i=1}^{k} x_i U_i + \eta^2 N^\top V_k V_k^\top N\right)\left(\sum_{i=1}^{k} x_i U_i + \sum_{i=k+1}^{d}(1+\eta s_i)^2 U_i + \eta^2 N^\top N\right)^{-1}\right).$$

Applying Lemma 6 (noting that as required by the lemma, $\sum_{i=k+1}^{d}(1+\eta s_i)^2 U_i + \eta^2 N^\top N - \eta^2 N^\top V_k V_k^\top N = \sum_{i=k+1}^{d}(1+\eta s_i)^2 U_i + \eta^2 N^\top \left(I - V_k V_k^\top\right) N \succeq 0$), we can lower bound the above by

$$\operatorname{Tr}\left(\left((1+\eta s_k)^2 \sum_{i=1}^{k} U_i + \eta^2 N^\top V_k V_k^\top N\right)\left((1+\eta s_k)^2 \sum_{i=1}^{k} U_i + \sum_{i=k+1}^{d}(1+\eta s_i)^2 U_i + \eta^2 N^\top N\right)^{-1}\right).$$

Using Lemma 2, this can be lower bounded by

$$\begin{aligned} &\operatorname{Tr}\left(\left((1+\eta s_k)^2 \sum_{i=1}^{k} U_i\right)\left((1+\eta s_k)^2 \sum_{i=1}^{k} U_i + \sum_{i=k+1}^{d}(1+\eta s_i)^2 U_i + \eta^2 N^\top N\right)^{-1}\right) \\ &= \operatorname{Tr}\left(\left(\sum_{i=1}^{k} U_i\right)\left(\sum_{i=1}^{k} U_i + \sum_{i=k+1}^{d}\left(\frac{1+\eta s_i}{1+\eta s_k}\right)^2 U_i + \left(\frac{\eta}{1+\eta s_k}\right)^2 N^\top N\right)^{-1}\right) \end{aligned}$$

Applying Lemma 3, this is at least

$$\operatorname{Tr}\left(\left(\sum_{i=1}^{k} U_i\right)\left(2I - \sum_{i=1}^{k} U_i - \sum_{i=k+1}^{d}\left(\frac{1+\eta s_i}{1+\eta s_k}\right)^2 U_i - \left(\frac{\eta}{1+\eta s_k}\right)^2 N^\top N\right)\right).$$

Recalling that $I = \sum_{i=1}^{d} U_i = \sum_{i=1}^{k} U_i + \sum_{i=k+1}^{d} U_i$, this can be simplified to

$$\operatorname{Tr}\left(\left(\sum_{i=1}^{k} U_i\right)\left(\sum_{i=1}^{k} U_i + \sum_{i=k+1}^{d}\left(2 - \left(\frac{1+\eta s_i}{1+\eta s_k}\right)^2\right) U_i - \left(\frac{\eta}{1+\eta s_k}\right)^2 N^\top N\right)\right). \tag{14}$$

Since $U_i \succeq 0$, then using Lemma 3, we can lower bound the expression above by shrinking each of the $\left(2 - \left(\frac{1+\eta s_i}{1+\eta s_k}\right)^2\right)$ terms. In particular, since $s_i \leq s_k - \lambda$ for each $i \geq k+1$,

$$2 - \left(\frac{1+\eta s_i}{1+\eta s_k}\right)^2 \;\geq\; 2 - \frac{1+\eta s_i}{1+\eta s_k} \;\geq\; 2 - \frac{1+\eta(s_k-\lambda)}{1+\eta s_k} \;=\; 1 + \frac{\eta\lambda}{1+\eta s_k},$$

which by the assumption that $\eta \leq 1/4$ and $s_k \leq s_1 \leq 1$, is at least $1 + \frac{4}{5}\eta\lambda$. Plugging this back into Eq. (14), and recalling that $\sum_{i=1}^{d} U_i = I$, we get the lower bound

$$\begin{aligned}
&\mathrm{Tr}\left(\left(\sum_{i=1}^{k} U_i\right)\left(\sum_{i=1}^{k} U_i + \sum_{i=k+1}^{d}\left(1+\frac{4}{5}\eta\lambda\right)U_i - \left(\frac{\eta}{1+\eta s_k}\right)^2 N^\top N\right)\right) \\
&= \mathrm{Tr}\left(\left(\sum_{i=1}^{k} U_i\right)\left(I + \frac{4}{5}\eta\lambda\left(I - \sum_{i=1}^{k} U_i\right) - \left(\frac{\eta}{1+\eta s_k}\right)^2 N^\top N\right)\right).
\end{aligned}$$

Again using Lemma 2, this is at least

$$\begin{aligned}
&\mathrm{Tr}\left(\left(\sum_{i=1}^{k} U_i\right)\left(I + \frac{4}{5}\eta\lambda\left(I - \sum_{i=1}^{k} U_i\right)\right)\right) - \left(\frac{\eta}{1+\eta s_k}\right)^2 \mathrm{Tr}\left(\left(\sum_{i=1}^{k} U_i\right) N^\top N\right) \\
&\geq \mathrm{Tr}\left(\left(\sum_{i=1}^{k} U_i\right)\left(I + \frac{4}{5}\eta\lambda\left(I - \sum_{i=1}^{k} U_i\right)\right)\right) - \left(\frac{\eta}{1+\eta s_k}\right)^2 \mathrm{Tr}\left(N^\top N\right) \\
&\geq \mathrm{Tr}\left(\left(\sum_{i=1}^{k} U_i\right)\left(I + \frac{4}{5}\eta\lambda\left(I - \sum_{i=1}^{k} U_i\right)\right)\right) - \eta^2\left(\sigma_N^F\right)^2.
\end{aligned}$$

Recall that this is a lower bound on the trace term in Eq. (13). Plugging it back and slightly simplifying, we get

$$\mathbb{E}\left[\|V_k^\top W''\|_F^2\right] \geq \mathrm{Tr}\left(\left(\sum_{i=1}^{k} U_i\right)\left(I + \frac{4}{5}\eta\lambda\left(I - \sum_{i=1}^{k} U_i\right)\right)\right) - \eta^2 r_N,$$

where

$$r_N = 46\,(\sigma_N^F)^2\left(1 + \frac{8}{3}\left(\frac{1}{4}\sigma_N^{sp} + 2\right)^2\right).$$

The trace term above can be re-written (using the definition of U_i and the fact that $\mathrm{Tr}(B^\top B) = \|B\|_F^2$) as

$$\begin{aligned}
&\mathrm{Tr}\left(\left(W^\top \sum_{i=1}^{k} \mathbf{v}_i\mathbf{v}_i^\top W\right)\left(I + \frac{4}{5}\eta\lambda\left(I - W^\top \sum_{i=1}^{k} \mathbf{v}_i\mathbf{v}_i^\top W\right)\right)\right) \\
&= \left(1+\frac{4}{5}\eta\lambda\right)\mathrm{Tr}\left(W^\top V_k V_k^\top W\right) - \frac{4}{5}\eta\lambda\,\mathrm{Tr}\left(\left(W^\top V_k V_k^\top W\right)\left(W^\top V_k V_k^\top W\right)\right) \\
&= \left(1+\frac{4}{5}\eta\lambda\right)\|V_k^\top W\|_F^2 - \frac{4}{5}\eta\lambda\|W^\top V_k V_k^\top W\|_F^2 \\
&= \|V_k^\top W\|_F^2\left(1 + \frac{4}{5}\eta\lambda\left(1 - \frac{\|W^\top V_k V_k^\top W\|_F^2}{\|V_k^\top W\|_F^2}\right)\right).
\end{aligned}$$

Applying Lemma 7, and letting δ denote the minimal singular value of $V_k^\top W$, this is lower bounded by

$$\|V_k^\top W\|_F^2\left(1 + \frac{4}{5}\eta\lambda\max\left\{1 - \|V_k^\top W\|_F^2\,,\ \frac{\delta^2}{k}\left(k - \|V_k^\top W\|_F^2\right)\right\}\right).$$

Overall, we get that

$$\mathbb{E}\left[\|V_k^\top W''\|_F^2\right] \geq \|V_k^\top W\|_F^2\left(1 + \frac{4}{5}\eta\lambda\max\left\{1 - \|V_k^\top W\|_F^2\,,\ \frac{\delta^2}{k}\left(k - \|V_k^\top W\|_F^2\right)\right\}\right) - \eta^2 r_N. \tag{15}$$

We now consider two options:

- Taking the first argument of the max term in Eq. (15), we get

$$\mathbb{E}\left[\|V_k^\top W''\|_F^2\right] \geq \|V_k^\top W\|_F^2 \left(1 + \frac{4}{5}\eta\lambda\left(1 - \|V_k^\top W\|_F^2\right)\right) - \eta^2 r_N.$$

  Subtracting 1 from both sides and simplifying, we get

$$\mathbb{E}\left[1 - \|V_k^\top W''\|_F^2\right] \leq \left(1 - \frac{4}{5}\eta\lambda\|V_k^\top W\|_F^2\right)\left(1 - \|V_k^\top W\|_F^2\right) + \eta^2 r_N.$$

- Suppose that $\|V_k^\top W\|_F^2 \geq k - \frac{1}{2}$. Taking the second argument of the max term in Eq. (15), we get

$$\mathbb{E}\left[\|V_k^\top W''\|_F^2\right] \geq \|V_k^\top W\|_F^2 \left(1 + \frac{4\eta\lambda\delta^2}{5k}\left(k - \|V_k^\top W\|_F^2\right)\right) - \eta^2 r_N.$$

  Subtracting both sides from k, , we get

$$\begin{aligned}
\mathbb{E}\left[k - \|V_k^\top W''\|_F^2\right] &\leq \left(k - \|V_k^\top W\|_F^2\right) - \frac{4\eta\lambda\delta^2}{5k}\|V_k^\top W\|_F^2\left(k - \|V_k^\top W\|_F^2\right) + \eta^2 r_N \\
&= \left(k - \|V_k^\top W\|_F^2\right)\left(1 - \frac{4\eta\lambda\delta^2}{5k}\|V_k^\top W\|_F^2\right) + \eta^2 r_N \\
&\leq \left(k - \|V_k^\top W\|_F^2\right)\left(1 - \frac{4\eta\lambda\delta^2}{5k}\left(k - \frac{1}{2}\right)\right) + \eta^2 r_N
\end{aligned}$$

  Since $k \geq 1$, we can lower bound the $\left(k - \frac{1}{2}\right)$ term by $\frac{k}{2}$. Moreover, the condition $k - \|V_k^\top W\|_F^2 \leq \frac{1}{2}$ implies that the singular values $\sigma_1, \ldots, \sigma_k$ of $V_k^\top W$ satisfy $k - \sum_{i=1}^k \sigma_i^2 \leq \frac{1}{2}$. But each σ_i is in $[0, 1]$ (as V_k, W have orthonormal columns), so no σ_i can be less than $\frac{1}{2}$. This implies that $\delta \geq \frac{1}{2}$. Plugging the lower bounds $k - \frac{1}{2} \geq \frac{k}{2}$ and $\delta \geq \frac{1}{2}$ into the above, we get

$$\mathbb{E}\left[k - \|V_k^\top W''\|_F^2\right] \leq \left(k - \|V_k^\top W\|_F^2\right)\left(1 - \frac{1}{10}\eta\lambda\right) + \eta^2 r_N.$$

□

Lemma 11. *Let A, W_t be as defined in Algorithm 2, and suppose that $\eta \in \left[0, \frac{1}{23\sqrt{k}}\right]$. Then the following holds for some positive numerical constants c_1, c_2, c_3:*

- $\mathbb{E}\left[1 - \|V_k^\top W''\|_F^2\right] \leq \left(1 - c_1\eta\lambda\|V_k^\top W\|_F^2\right)\left(1 - \|V_k^\top W\|_F^2\right) + c_2 k\eta^2$
- *If $\|V_k^\top W_t\|_F^2 \geq k - \frac{1}{2}$, then*

$$\mathbb{E}\left[k - \|V_k^\top W_{t+1}\|_F^2\right] \;\leq\; \left(k - \|V_k^\top W_t\|_F^2\right)\left(1 - c_1\eta\left(\lambda - c_2\eta\right)\right) \;+\; c_3\eta^2(k - \|V_k^\top \tilde{W}_{s-1}\|_F^2).$$

In the above, the expectation is over the random draw of the index i_t, conditioned on W_t and $\tilde{W}_{s-1}$.

Proof. To apply Lemma 10, we need to compute upper bounds σ_N^F and σ_N^{sp} on the Frobenius and spectral norms of N, which in our case equals $(\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top - A)(W_t - \tilde{W}_{s-1}B_t)$. Since $\|A\|_2, \|\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top\|_2 \leq 1$, and $W_t, \tilde{W}_{s-1}, B_t$ have orthonormal columns, the spectral norm of N is at most

$$\|(\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top - A)(W_t - \tilde{W}_{s-1}B_t)\|_2 \leq \left(\|\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top\|_2 + \|A\|_2\right)\left(\|W_t\|_2 + \|\tilde{W}_{s-1}\|_2\|B_t\|_2\right) \leq 4,$$

so we may take $\sigma_N^{sp} = 4$. As to the Frobenius norm, using Lemma 4 and a similar calculation, we have

$$\|N\|_F^2 \leq 4\|W_t - \tilde{W}_{s-1}B_t\|_F^2.$$

To upper bound this, define

$$V_{W_t} = \arg \min_{V_k B:(V_k B)^\top (V_k B)=I} \|W_t - V_k B\|_F^2$$

to be the nearest orthonormal-columns matrix to W_t in the column space of V_k, and

$$\tilde{W}_V = \arg \min_{\tilde{W}_{s-1} B:(\tilde{W}_{s-1} B)^\top (\tilde{W}_{s-1} B)=I} \|V_{W_t} - \tilde{W}_{s-1} B\|_F^2$$

to be the nearest orthonormal-columns matrix to V_{W_t} in the column space of $\tilde{W}_{s-1}$. Also, recall that by definition,

$$\tilde{W}_{s-1} B_t = \arg \min_{\tilde{W}_{s-1} B:(\tilde{W}_{s-1} B)^\top (\tilde{W}_{s-1} B)=I} \|W_t - \tilde{W}_{s-1} B\|_F^2$$

is the nearest orthonormal-columns matrix to W_t in the column space of $\tilde{W}_{s-1}$. Therefore, we must have $\|W_t - \tilde{W}_{s-1} B_t\|_F^2 \leq \|W_t - \tilde{W}_V\|_F^2$. Using this and Lemma 8, we have

$$\begin{aligned}
\|W_t - \tilde{W}_{s-1} B_t\|_F^2 &\leq \|W_t - \tilde{W}_V\|_F^2 \\
&= \|(W_t - V_{W_t}) - (\tilde{W}_V - V_{W_t})\|_F^2 \\
&\leq 2\|W_t - V_{W_t}\|_F^2 + 2\|\tilde{W}_V - V_{W_t}\|_F^2 \\
&= 4\left(k - \|V_k^\top W_t\|_F^2\right) + 4\left(k - \|V_{W_t}^\top \tilde{W}_{s-1}\|_F^2\right).
\end{aligned}$$

By definition of V_{W_t}, we have $V_{W_t} = V_k B$ where $B^\top B = B^\top V_k^\top V_k B = (V_k B)^\top (V_k B) = I$. Therefore B is an orthogonal $k \times k$ matrix, and $\|V_{W_t}^\top \tilde{W}_{s-1}\|_F^2 = \|B^\top V_k^\top \tilde{W}_{s-1}\|_F^2 = \|V_k^\top \tilde{W}_{s-1}\|_F^2$, so the above equals $4(k - \|V_k^\top W_t\|_F^2) + 4(k - \|V_k^\top \tilde{W}_{s-1}\|_F^2)$. Overall, we get that the squared Frobenius norm of N can be upper bounded by

$$(\sigma_N^F)^2 = 16\left((k - \|V_k^\top W_t\|_F^2) + (k - \|V_k^\top \tilde{W}_{s-1}\|_F^2)\right).$$

Plugging σ_N^{sp} and $(\sigma_N^F)^2$ into the r_N as defined in Lemma 10, and picking any $\eta \in [0, \frac{1}{23\sqrt{k}}]$ (which satisfies the condition in Lemma 10 that $\eta \in \left[0, \frac{1}{4\max\{1, \sigma_N^F\}}\right]$, since $4\max\{1, \sigma_n^F\} \leq 4\max\{1, \sqrt{16 * 2k}\} < 23\sqrt{k}$), we get

$$\begin{aligned}
r_N &= 736\left((k - \|V_k^\top W_t\|_F^2) + (k - \|V_k^\top \tilde{W}_{s-1}\|_F^2)\right)\left(1 + \frac{8}{3}\left(\frac{1}{4}4 + 2\right)^2\right) \\
&\leq 18400\ \left((k - \|V_k^\top W_t\|_F^2) + (k - \|V_k^\top \tilde{W}_{s-1}\|_F^2)\right).
\end{aligned}$$

This implies that $r_N \leq 36800k$ always, which by application of Lemma 10, gives the first part of our lemma. As to the second part, assuming $\|V_k^\top W_t\|_F^2 \geq k - \frac{1}{2}$ and applying Lemma 10, we get that

$$\begin{aligned}
\mathbb{E}\left[k - \|V_k^\top W_{t+1}\|_F^2\right] &\leq \left(k - \|V_k^\top W_t\|_F^2\right)\left(1 - \frac{1}{10}\eta\lambda\right) \\
&\quad + 18400\,\eta^2\left((k - \|V_k^\top W_t\|_F^2) + (k - \|V_k^\top \tilde{W}_{s-1}\|_F^2)\right) \\
&= \left(k - \|V_k^\top W_t\|_F^2\right)\left(1 - \eta\left(\frac{1}{10}\lambda - 18400\eta\right)\right) \\
&\quad + 18400\,\eta^2(k - \|V_k^\top \tilde{W}_{s-1}\|_F^2).
\end{aligned}$$

This corresponds to the lemma statement. □

PART II: SOLVING THE RECURRENCE RELATION FOR A SINGLE EPOCH

Since we focus on a single epoch, we drop the subscript from $\tilde{W}_{s-1}$ and denote it simply as $\tilde{W}$.

Suppose that $\eta = \alpha\lambda$, where α is a sufficiently small constant to be chosen later. Also, let

$$b_t = k - \|V_k^\top W_t\|_F^2 \quad \text{and} \quad \tilde{b} = k - \|V_k^\top \tilde{W}\|_F^2.$$

Then Lemma 11 tells us that if α is a sufficiently small constant, $b_t \leq \frac{1}{2}$, then

$$\mathbb{E}\left[b_{t+1}|W_t\right] \;\leq\; \left(1 - c\alpha\lambda^2\right) b_t + c'\alpha^2\lambda^2\tilde{b} \tag{16}$$

for some numerical constants c, c'.

Lemma 12. *Let B be the event that $b_t \leq \frac{1}{2}$ for all $t = 0, 1, 2, \ldots, m$. Then for certain positive numerical constants c_1, c_2, c_3, if $\alpha \leq c_1$, then*

$$\mathbb{E}[b_m|B] \leq \left(\left(1 - c_2\alpha\lambda^2\right)^m + c_3\alpha\right)\tilde{b},$$

where the expectation is over the randomness in the current epoch.

Proof. Recall that b_t is a deterministic function of the random variable W_t, which depends in turn on W_{t-1} and the random instance chosen at round t. We assume that W_0 (and hence $\tilde{b}$) are fixed, and consider how b_t evolves as a function of t. Using Eq. (16), we have

$$\mathbb{E}[b_{t+1}|W_t, B] = \mathbb{E}\left[b_{t+1}|W_t, b_{t+1} \leq \frac{1}{2}\right] \;\leq\; \mathbb{E}[b_{t+1}|W_t] \;\leq\; \left(1 - c\alpha\lambda^2\right) b_t \;+\; c'\alpha^2\lambda^2\tilde{b}.$$

Note that the first equality holds, since conditioned on W_t, b_{t+1} is independent of $b_1, \ldots, b_t$, so the event B is equivalent to just requiring $b_{t+1} \leq 1/2$.

Taking expectation over W_t (conditioned on B), we get that

$$\begin{aligned}\mathbb{E}[b_{t+1}|B] \;&\leq\; \mathbb{E}\left[\left(1 - c\alpha\lambda^2\right) b_t + c'\alpha^2\lambda^2\tilde{b}\middle| B\right] \\ &= \left(1 - c\alpha\lambda^2\right)\mathbb{E}\left[b_t|B\right] + c'\alpha^2\lambda^2\tilde{b}.\end{aligned}$$

Unwinding the recursion, and using that $b_0 = \tilde{b}$, we therefore get that

$$\begin{aligned}\mathbb{E}[b_m|B] \;&\leq\; \left(1 - c\alpha\lambda^2\right)^m\tilde{b} + c'\alpha^2\lambda^2\tilde{b}\sum_{i=0}^{m-1}\left(1 - c\alpha\lambda^2\right)^i \\ &\leq\; \left(1 - c\alpha\lambda^2\right)^m\tilde{b} + c'\alpha^2\lambda^2\tilde{b}\sum_{i=0}^{\infty}\left(1 - c\alpha\lambda^2\right)^i \\ &=\; \left(1 - c\alpha\lambda^2\right)^m\tilde{b} + c'\alpha^2\lambda^2\tilde{b}\frac{1}{c\alpha\lambda^2} \\ &=\; \left(\left(1 - c\alpha\lambda^2\right)^m + \frac{c'}{c}\alpha\right)\tilde{b}.\end{aligned}$$

as required. □

We now turn to prove that the event B assumed in Lemma 12 indeed holds with high probability:

Lemma 13. *The following holds for certain positive numerical constants c_1, c_2, c_3: If $\alpha \leq c_1$, then for any $\beta \in (0, 1)$ and m, if*

$$\tilde{b} + c_2 km\alpha^2\lambda^2 + c_3 k\sqrt{m\alpha^2\lambda^2\log(1/\beta)} \leq \frac{1}{2}, \tag{17}$$

then it holds with probability at least $1 - \beta$ that

$$b_t \;\leq\; \tilde{b} + c_2 km\alpha^2\lambda^2 + c_3 k\sqrt{m\alpha^2\lambda^2\log(1/\beta)} \;\leq\; \frac{1}{2}$$

for all $t = 0, 1, 2, \ldots, m$.

Proof. To prove the lemma, we analyze the stochastic process $b_0(= \tilde{b}), b_1, b_2, \ldots, b_m$, and use a concentration of measure argument. First, we collect the following facts:

- $\tilde{b} = b_0 \leq \frac{1}{2}$: This directly follows from the assumption stated in the lemma.
- *As long as* $b_t \leq \frac{1}{2}$, $\mathbb{E}\left[b_{t+1}|W_t\right] \leq b_t + c_2\alpha^2\lambda^2\tilde{b}$ *for some constant* c_2: Supposing α is sufficiently small, then by Eq. (16),

$$\mathbb{E}\left[b_{t+1}|W_t\right] \;\leq\; \left(1 - c\alpha\lambda^2\right) b_t + c'\alpha^2\lambda^2\tilde{b} \;\leq\; b_t + c'\alpha^2\lambda^2\tilde{b}.$$

- $|b_{t+1} - b_t|$ *is bounded by* $c_3'k\alpha\lambda$ *for some constant* c_3': Applying Lemma 9, and assuming that α is at most some sufficiently small constant c_1 (e.g. $\alpha \leq \frac{1}{12}$, so $\eta = \alpha\lambda \leq \frac{1}{12}$),

$$|b_{t+1} - b_t| \;=\; \left|\|V_k^\top W_{t+1}\|_F^2 - \|V_k^\top W_t\|_F^2\right| \;\leq\; \frac{12k\eta}{1-3\eta} \leq \frac{12k\alpha\lambda}{3/4} \;=\; 16k\alpha\lambda.$$

Armed with these facts, and using the maximal version of the Hoeffding-Azuma inequality (Hoeffding, 1963), it follows that with probability at least $1-\beta$, it holds simultaneously for all $t = 1, \ldots, m$ (and for $t = 0$ by assumption) that

$$b_t \leq \tilde{b} + c_2 m\alpha^2\lambda^2\tilde{b} + c_3 k\sqrt{m\alpha^2\lambda^2\log(1/\beta)}$$

for some constants c_2, c_3, as long as the expression above is less than $\frac{1}{2}$. If the expression is indeed less than $\frac{1}{2}$, then we get that $b_t \leq \frac{1}{2}$ for all t. Upper bounding $\tilde{b}$ by k and slightly simplifying, we get the statement in the lemma. □

Combining Lemma 12 and Lemma 13, and using Markov's inequality, we get the following corollary:

Lemma 14. *Let confidence parameters* $\beta, \gamma \in (0,1)$ *be fixed. Suppose that* m, α *are chosen such that* $\alpha \leq c_1$ *and*

$$\tilde{b} + c_2 km\alpha^2\lambda^2 + c_3 k\sqrt{m\alpha^2\lambda^2\log(1/\beta)} \leq \frac{1}{2},$$

where c_1, c_2, c_3 *are certain positive numerical constants. Then with probability at least* $1-(\beta+\gamma)$*, it holds that*

$$b_m \leq \frac{1}{\gamma}\left(\left(1 - c\alpha\lambda^2\right)^m + c'\alpha\right)\tilde{b}.$$

for some positive numerical constants c, c'.

PART III: ANALYZING THE ENTIRE ALGORITHM'S RUN

Given the analysis in Lemma 14 for a single epoch, we are now ready to prove our theorem. Let

$$\tilde{b}_s = k - \|V_k^\top \tilde{W}_s\|_F^2.$$

By assumption, at the beginning of the first epoch, we have $\tilde{b}_0 = k - \|V_k^\top \tilde{W}_0\|_F^2 \leq \frac{1}{2}$. Therefore, by Lemma 14, for any $\beta, \gamma \in \left(0, \frac{1}{2}\right)$, if we pick any

$$\alpha \leq \min\left\{c_1, \frac{1}{2c'}\gamma^2\right\} \quad \text{and} \quad m \geq \frac{3\log(1/\gamma)}{c\alpha\lambda^2} \quad \text{such that} \quad \frac{1}{2} + c_2 km\alpha^2\lambda^2 + c_3 k\sqrt{m\alpha^2\lambda^2\log(1/\beta)} \leq \frac{1}{2}, \tag{18}$$

then we get with probability at least $1-(\beta+\gamma)$ that

$$b_m \;\leq\; \frac{1}{\gamma}\left(\left(1 - c\alpha\lambda^2\right)^{\frac{3\log(1/\gamma)}{c\alpha\lambda^2}} + \frac{1}{2}\gamma^2\right)\tilde{b}_0$$

Using the inequality $(1-(1/x))^{ax} \leq \exp(-a)$, which holds for any $x > 1$ and any a, and taking $x = 1/(c\alpha\lambda^2)$ and $a = 3\log(1/\gamma)$, we can upper bound the above by

$$\begin{aligned}&\frac{1}{\gamma}\left(\exp\left(-3\log\left(\frac{1}{\gamma}\right)\right) + \frac{1}{2}\gamma^2\right)\tilde{b}_0 \\ &= \frac{1}{\gamma}\left(\gamma^3 + \frac{1}{2}\gamma^2\right)\tilde{b}_0 \;\leq\; \gamma\tilde{b}_0.\end{aligned}$$

Since b_m equals the starting point $\tilde{b}_1$ for the next epoch, we get that $\tilde{b}_1 \leq \gamma \tilde{b}_0 \leq \gamma \frac{1}{2}$. Again applying Lemma 14, and performing the same calculation we have that with probability at least $1-(\beta+\gamma)$ over the next epoch, $\tilde{b}_2 \leq \gamma \tilde{b}_1 \leq \gamma^2 \tilde{b}_0$. Repeatedly applying Lemma 14 and using a union bound, we get that after T epochs, with probability at least $1-T(\beta+\gamma)$,

$$k - \|V_k^\top \tilde{W}_T\|_F^2 \;=\; \tilde{b}_T \;\leq\; \gamma^T \tilde{b}_0 \;<\; \gamma^T.$$

Therefore, for any desired accuracy parameter ϵ, we simply need to use $T = \left\lceil \frac{\log(1/\epsilon)}{\log(1/\gamma)} \right\rceil$ epochs, and get $k - \|V_k^\top \tilde{W}_s\|_F^2 \leq \epsilon$ with probability at least $1 - T(\beta+\gamma) = 1 - \left\lceil \frac{\log(1/\epsilon)}{\log(1/\gamma)} \right\rceil (\beta+\gamma)$.

Using a confidence parameter δ, we pick $\beta = \gamma = \frac{\delta}{2}$, which ensures that the accuracy bound above holds with probability at least

$$1 - \left\lceil \frac{\log(1/\epsilon)}{\log(2/\delta)} \right\rceil \delta \;\geq\; 1 - \left\lceil \frac{\log(1/\epsilon)}{\log(2)} \right\rceil \delta \;=\; 1 - \left\lceil \log_2\left(\frac{1}{\epsilon}\right) \right\rceil \delta.$$

Substituting this choice of β, γ into Eq. (18), and recalling that the step size η equals $\alpha\lambda$, we get that $k - \|V_k^\top \tilde{W}_T\|_F^2 \leq \epsilon$ with probability at least $1 - \lceil \log_2(1/\epsilon) \rceil \delta$, provided that

$$\eta \leq c\delta^2\lambda \quad , \quad m \geq \frac{c' \log(2/\delta)}{\eta\lambda} \quad , \quad km\eta^2 + k\sqrt{m\eta^2 \log(2/\delta)} \leq c''$$

for suitable positive constants c, c', c''.

To get the theorem statement, recall that the analysis we performed pertains to data whose squared norm is bounded by 1. By the reduction discussed at the beginning of the proof, we can apply it to data with squared norm at most r, by replacing λ with λ/r, and η with ηr, leading to the condition

$$\eta \leq \frac{c\delta^2}{r^2}\lambda \quad , \quad m \geq \frac{c' \log(2/\delta)}{\eta\lambda} \quad , \quad km\eta^2 r^2 + rk\sqrt{m\eta^2 \log(2/\delta)} \leq c''$$

and establishing the theorem.

A.2. Proof of Theorem 2

The proof relies mainly on the techniques and lemmas of Section A.1, used to prove Theorem 1. As done in Section A.1, we will assume without loss of generality that $r = \max_i \|\mathbf{x}_i\|^2$ is at most 1, and then transform the bound to a bound for general r (see the discussion at the beginning of Subsection A.1.2)

First, we extract the following result, which is essentially the first part of Lemma 11 (for $k = 1$):

Lemma 15. *Let $A, \mathbf{w}_t$ be as defined in Algorithm 1, and suppose that $\eta \in \left[0, \frac{1}{23}\right]$. Then*

$$\mathbb{E}_{i_t}\left[1 - \langle \mathbf{v}_1, \mathbf{w}_{t+1}\rangle^2 \middle| \mathbf{w}_t, \tilde{\mathbf{w}}_{s-1}\right] \leq \left(1 - c\eta\lambda\langle \mathbf{v}_1, \mathbf{w}_t\rangle^2\right)\left(1 - \langle \mathbf{v}_1, \mathbf{w}_t\rangle^2\right) + c'\eta^2,$$

for some positive numerical constants c, c'.

Note that this bound holds regardless of what is $\tilde{\mathbf{w}}_{s-1}$, and in particular holds across different epochs of Algorithm 1. Therefore, it is enough to show that starting from some initial point $\mathbf{w}_0$, after sufficiently many stochastic updates as specified in line 6-10 of the algorithm (or in terms of the analysis, sufficiently many applications of Lemma 15), we end up with a point $\mathbf{w}_T$ for which $1 - \langle \mathbf{v}_1, \mathbf{w}_T\rangle \leq \frac{1}{2}$, as required. Note that to simplify the notation, we will use here a single running index $\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_T$ (whereas in the algorithm we restarted the indexing after every epoch).

The proof is based on martingale arguments, quite similar to the ones in Subsection A.1.2 but with slight changes. First, we let

$$b_t = 1 - \langle \mathbf{v}_1, \mathbf{w}_t\rangle^2$$

to simplify notation. We note that $b_0 = 1 - \langle \mathbf{v}_1, \mathbf{w}_0\rangle^2$ is assumed fixed, whereas $b_1, b_2, \ldots$ are random variables based on the sampling process. Lemma 11 tells us that if η is sufficiently small, and $b_t \leq 1 - \xi$ for some $\xi \in (0,1)$, then

$$\mathbb{E}\left[b_{t+1} | b_t\right] \;\leq\; (1 - c\eta\lambda\xi)\, b_t + c'\eta^2. \tag{19}$$

for some numerical constants c, c'.

Lemma 16. *Let B be the event that $b_t \le 1-\xi$ for all $t = 0, 1, \ldots, T$. Then for certain positive numerical constants c_1, c_2, c_3, if $\eta \le c_1\lambda$, then*

$$\mathbb{E}[b_T|B] \le \left((1-c_2\eta\lambda\xi)^T + c_3\frac{\eta}{\lambda\xi}\right).$$

Proof. Using Eq. (19), we have for any b_t satisfying event B that

$$\mathbb{E}[b_{t+1}|b_t, B] = \mathbb{E}\left[b_{t+1}|b_t, b_{t+1} \le 1-\xi\right] \;\le\; \mathbb{E}[b_{t+1}|b_t] \;\le\; (1-c\eta\lambda\xi)\, b_t \;+\; c'\eta^2.$$

Taking expectation over b_t (conditioned on B), we get that

$$\begin{aligned}\mathbb{E}[b_{t+1}|B] \;&\le\; \mathbb{E}\left[(1-c\eta\lambda\xi)\, b_t + c'\eta^2 \big| B\right] \\ &= (1-c\eta\lambda\xi)\,\mathbb{E}\left[b_t|B\right] + c'\eta^2.\end{aligned}$$

Unwinding the recursion, we get

$$\begin{aligned}\mathbb{E}[b_T|B] \;&\le\; (1-c\eta\lambda\xi)^T\, b_0 + c'\eta^2 \sum_{i=0}^{T-1} (1-c\eta\lambda\xi)^i \\ &\le\; (1-c\eta\lambda\xi)^T + c'\eta^2 \sum_{i=0}^{\infty} (1-c\eta\lambda\xi)^i \\ &=\; (1-c\eta\lambda\xi)^T + c'\eta^2 \frac{1}{c\eta\lambda\xi} \;\le\; (1-c\eta\lambda\xi)^T + \frac{c'}{c}\frac{\eta}{\lambda\xi}.\end{aligned}$$

□

We now turn to prove that the event B assumed in Lemma 12 indeed holds with high probability:

Lemma 17. *The following holds for certain positive numerical constants c_1, c_2, c_3: If $\eta \le c_1\lambda$, then for any $\beta \in (0,1)$, if*

$$b_0 + c_2 T\eta^2 + c_3\sqrt{T\eta^2\log(1/\beta)} \le 1-\xi, \tag{20}$$

then it holds with probability at least $1-\beta$ that

$$b_t \;\le\; b_0 + c_2 T\eta^2 + c_3\sqrt{T\eta^2\log(1/\beta)} \;\le\; 1-\xi$$

for all $t = 0, 1, \ldots, T$.

Proof. To prove the lemma, we analyze the stochastic process $b_1, b_2, \ldots, b_T$, and use a concentration of measure argument. First, we collect the following facts:

- $b_0 \le 1-\xi$: This directly follows from the assumption stated in the lemma.
- $\mathbb{E}\left[b_{t+1}|b_t\right] \le b_t + c'\eta^2$ *for some constant c'*: By Eq. (19),

$$\mathbb{E}\left[b_{t+1}|W_t\right] \;\le\; (1-c\eta\lambda\xi)\, b_t + c'\eta^2 \;\le\; b_t + c'\eta^2.$$

- $|b_{t+1} - b_t|$ *is bounded by $c\eta$ for some constant c*: Applying Lemma 9 for the case $k=1$, and assuming $\eta \le 1/12$,

$$|b_{t+1} - b_t| \;=\; \left|\langle \mathbf{v}_1, \mathbf{w}_{t+1}\rangle^2 - \langle \mathbf{v}, \mathbf{w}_t\rangle^2\right| \;\le\; \frac{12\eta}{1-3\eta} \le \frac{12\eta}{3/4} \;=\; 16\eta.$$

Armed with these facts, and using the maximal version of the Hoeffding-Azuma inequality (Hoeffding, 1963), it follows that with probability at least $1-\beta$, it holds simultaneously for all $t = 0, 1, \ldots, T$ that

$$b_t \leq b_0 + c_2 T\eta^2 + c_3\sqrt{T\eta^2 \log(1/\beta)}$$

for some constants c_2, c_3. If the expression is indeed less than $1-\xi$, then we get that $b_t \leq 1-\xi$ for all t, from which the lemma follows. □

Combining Lemma 16 and Lemma 17, and using Markov's inequality, we get the following corollary:

Lemma 18. *Let confidence parameters $\beta, \gamma \in (0,1)$ be fixed. Then for some positive numerical constants c_1, c_2, c_3, c, c', if $\eta \leq c_1\lambda$ and*

$$b_0 + c_2 T\eta^2 + c_3\sqrt{T\eta^2\log(1/\beta)} \leq 1-\xi,$$

then with probability at least $1-(\beta+\gamma)$, it holds that

$$b_T \leq \frac{1}{\gamma}\left((1-c\eta\lambda\xi)^T + c'\frac{\eta}{\lambda\xi}\right).$$

We are now ready to prove our theorem. By Lemma 18, for any $\beta, \gamma \in \left(0, \frac{1}{2}\right)$ and any

$$\begin{aligned}\eta \leq \min\left\{c_1, \frac{1}{2c'}\gamma^2\right\}\lambda\xi \quad \text{and} \quad T \geq \frac{3\log(1/\gamma)}{c\eta\lambda\xi} \\ \text{such that} \quad b_0 + c_2 T\eta^2 + c_3\sqrt{T\eta^2\log(1/\beta)} \leq 1-\xi,\end{aligned} \tag{21}$$

we get with probability at least $1-(\beta+\gamma)$ that

$$b_T \;\leq\; \frac{1}{\gamma}\left((1-c\eta\lambda\xi)^{\frac{3\log(1/\gamma)}{c\eta\lambda\xi}} + \frac{1}{2}\gamma^2\right).$$

Using the inequality $(1-(1/x))^{ax} \leq \exp(-a)$, which holds for any $x > 1$ and any a, and taking $x = 1/(c\eta\lambda\xi)$ and $a = 3\log(1/\gamma)$, we can upper bound the above by

$$\frac{1}{\gamma}\left(\exp\left(-3\log\left(\frac{1}{\gamma}\right)\right) + \frac{1}{2}\gamma^2\right) \;=\; \frac{1}{\gamma}\left(\gamma^3 + \frac{1}{2}\gamma^2\right),$$

and since we assume $\gamma < \frac{1}{2}$, this is at most $\frac{1}{2}$. Overall, we got that with probability at least $1-\beta-\gamma$, $b_T \leq \frac{1}{2}$, and therefore $1-\langle \mathbf{v}_1, \mathbf{w}_T\rangle^2 \leq \frac{1}{2}$ as required.

It remains to show that the parameter choices in Eq. (21) can indeed be satisfied. First, we fix $\xi = \frac{1}{2}\zeta$ (where we recall that $0 < \zeta \leq \langle \mathbf{v}_1, \mathbf{w}_0\rangle^2$), which trivially ensures that $b_0 = 1-\langle \mathbf{v}_1, \mathbf{w}_0\rangle^2$ is at most $1-2\xi$. Moreover, suppose we pick $\beta = \gamma$ in $(0, \exp(-1))$, and η, T so that

$$\eta \leq \frac{c_*\gamma^2\lambda\xi^3}{\log^2(1/\gamma)} \;,\; T = \left\lfloor \frac{3\log(1/\gamma)}{c'_*\eta\lambda\xi}\right\rfloor, \tag{22}$$

where c_*, c'_* are sufficiently small constants so that the bounds on η, T in Eq. (21) are satisfied. This implies that the third bound in Eq. (21) is also satisfied, since by plugging in the values / bounds of T and η, and using the assumptions $\gamma = \beta \leq \exp(-1)$ and $\xi \leq 1$, we have

$$\begin{aligned}&b_0 + c_2 T\eta^2 + c_3\sqrt{T\eta^2\log(1/\gamma)} \\ &\leq 1-2\xi + c_2\frac{3\log(1/\gamma)}{c'_*\lambda\xi}\eta + c_3\sqrt{\frac{3\log(1/\gamma)}{c'_*\lambda\xi}\eta\log(1/\gamma)} \\ &\leq 1-2\xi + c_2\frac{3c_*\gamma^2\xi^2}{c'_*\log(1/\gamma)} + c_3\sqrt{\frac{3c_*\gamma^2\xi^2}{c'_*}} \\ &\leq 1-2\xi + \left(\frac{3c_2c_*}{c'_*} + c_3\sqrt{\frac{3c_*}{c'_*}}\right)\xi,\end{aligned}$$

which is less than $1-\xi$ if we pick c_* sufficiently small compared to c'_*.

To summarize, we get that for any $\gamma \in (0, \exp(-1))$, by picking η as in Eq. (22), we have that after T iterations (where T is specified in Eq. (22)), with probability at least $1-2\gamma$, we get $\mathbf{w}_T$ such that $1 - \langle \mathbf{v}_1, \mathbf{w}_T \rangle \leq \frac{1}{2}$. Substituting $\delta = 2\gamma$ and $\zeta = 2\xi$, we get that if

$$\langle \mathbf{v}_1, \tilde{\mathbf{w}}_0 \rangle^2 \geq \zeta > 0,$$

and η satisfies

$$\eta \leq \frac{c_1 \delta^2 \lambda \zeta^3}{\log^2(2/\delta)}$$

(for some universal constant c_1), then with probability at least $1-\delta$, after

$$T \;=\; \left\lfloor \frac{c_2 \log(2/\delta)}{\eta \lambda \zeta} \right\rfloor.$$

stochastic iterations, we get a satisfactory point $\mathbf{w}_T$.

As discussed at the beginning of the proof, this analysis is valid assuming $r = \max_i \|\mathbf{x}_i\|^2 \leq 1$. By the reduction discussed at the beginning of Subsection A.1.2, we can get an analysis for any r by substituting $\lambda \to \lambda/r$ and $\eta \to \eta r$. This means that we should pick η satisfying

$$\eta r \leq \frac{c_1 \delta^2 (\lambda/r) \zeta^3}{\log^2(2/\delta)} \quad \Rightarrow \quad \eta \leq \frac{c_1 \delta^2 \lambda \zeta^3}{r^2 \log^2(2/\delta)},$$

and getting the required point after

$$T \;=\; \left\lfloor \frac{c_2 \log(2/\delta)}{(\eta r)(\lambda/r)\zeta} \right\rfloor \;=\; \left\lfloor \frac{c_2 \log(2/\delta)}{\eta \lambda \zeta} \right\rfloor$$

iterations.

A.3. Proof of Theorem 4

For simplicity of notation, we drop the $_A$ subscript from F_A, and refer simply to F.

We first prove the following two auxiliary lemmas:

Lemma 19. *If A is a symmetric matrix, then the gradient of the function $F(\mathbf{w}) = -\frac{\mathbf{w}^\top A \mathbf{w}}{\|\mathbf{w}\|^2}$ at some $\mathbf{w}$ equals*

$$-\frac{2}{\|\mathbf{w}\|^2} \left(F(\mathbf{w}) I + A\right) \mathbf{w},$$

and its Hessian equals

$$-\frac{1}{\|\mathbf{w}\|^2} \left(\left(I - \frac{4}{\|\mathbf{w}\|^2} \mathbf{w}\mathbf{w}^\top \right) \left(F(\mathbf{w}) I + A \right) \right)^\perp,$$

where $B^\perp = B + B^\top$ (i.e., a matrix B plus its transpose).

Proof. By the product and chain rules (using the fact that $\frac{1}{\|\mathbf{w}\|^2}$ is a composition of $\mathbf{w} \mapsto \|\mathbf{w}\|^2$ and $z \mapsto \frac{1}{z}$), the gradient of $F(\mathbf{w}) = -\frac{1}{\|\mathbf{w}\|^2} \left(\mathbf{w}^\top A \mathbf{w}\right)$ equals

$$\mathbf{w} \frac{2}{\|\mathbf{w}\|^4} \left(\mathbf{w}^\top A \mathbf{w}\right) - \left(A\mathbf{w}\right) \frac{2}{\|\mathbf{w}\|^2}, \tag{23}$$

giving the gradient bound in the lemma statement after a few simplifications.

Differentiating the vector-valued Eq. (23) with respect to $\mathbf{w}$ (using the product and chain rules, and the fact that $\frac{1}{\|\mathbf{w}\|^4}$ is a

composition of $\mathbf{w} \mapsto \|\mathbf{w}\|^2$, $z \mapsto z^2$, and $z \mapsto \frac{1}{z}$), we get that the Hessian of F equals

$$\begin{aligned} & I\frac{2}{\|\mathbf{w}\|^4}(\mathbf{w}^\top A\mathbf{w}) \;+\; \mathbf{w}\left(-\frac{2}{\|\mathbf{w}\|^8} * 2\|\mathbf{w}\|^2 * 2\mathbf{w}\right)^\top \left(\mathbf{w}^\top A\mathbf{w}\right) \;+\; \mathbf{w}\frac{2}{\|\mathbf{w}^4\|}\left(2A\mathbf{w}\right)^\top \\ & \quad - A\frac{2}{\|\mathbf{w}\|^2} - (A\mathbf{w})\left(-\frac{2}{\|\mathbf{w}\|^4} * 2\mathbf{w}\right)^\top \\ & = \; -\frac{2F(\mathbf{w})}{\|\mathbf{w}\|^2}I \;+\; \frac{8F(\mathbf{w})}{\|\mathbf{w}\|^4}\mathbf{w}\mathbf{w}^\top \;+\; \frac{4}{\|\mathbf{w}\|^4}\mathbf{w}\mathbf{w}^\top A \;-\; \frac{2}{\|\mathbf{w}\|^2}A \;+\; \frac{4}{\|\mathbf{w}\|^4}A\mathbf{w}\mathbf{w}^\top \\ & = -\frac{1}{\|\mathbf{w}\|^2}\left(2F(\mathbf{w})I - \frac{8F(\mathbf{w})}{\|\mathbf{w}\|^2}\mathbf{w}\mathbf{w}^\top - \frac{4}{\|\mathbf{w}\|^2}\mathbf{w}\mathbf{w}^\top A + 2A - \frac{4}{\|\mathbf{w}\|^2}A\mathbf{w}\mathbf{w}^\top\right), \end{aligned}$$

which can be verified to equal the expression in the lemma statement (using the fact that $A, \mathbf{w}\mathbf{w}^\top$ and I are all symmetric matrices, hence equal their transpose). □

Lemma 20. *Let $\mathbf{w}_0, \mathbf{v}_1$ be two unit vectors such that $\|\mathbf{w}_0 - \mathbf{v}_1\| \le \epsilon < \frac{1}{2}$ (which implies $\langle \mathbf{w}_0, \mathbf{v}_1\rangle > 0$). Let $\mathbf{v}_1'$ be the intersection of the ray $\{a\mathbf{v}_1 : a \ge 0\}$ with the hyperplane $H_{\mathbf{w}_0} = \{\mathbf{w} : \langle \mathbf{w}, \mathbf{w}_0\rangle = 1\}$. Then $\|\mathbf{v}_1' - \mathbf{w}_0\| \le \frac{5}{4}\epsilon$.*

Proof. See Figure 2 in the main text for a graphical illustration.

Letting $\mathbf{v}_1' = a\mathbf{v}$, a must satisfy $\langle a\mathbf{v}_1, \mathbf{w}_0\rangle = 1$. Since $\mathbf{v}_1, \mathbf{w}_0$ are unit vectors, this implies

$$a \;=\; \frac{1}{\langle \mathbf{v}_1, \mathbf{w}_0\rangle} \;=\; \frac{2}{2 - \|\mathbf{v}_1 - \mathbf{w}_0\|^2},$$

and since $\|\mathbf{v}_1 - \mathbf{w}_0\| \le \epsilon$, this means that

$$a \in \left[1, \frac{2}{2-\epsilon^2}\right].$$

Therefore,

$$\|\mathbf{v}_1' - \mathbf{w}_0\| \;\le\; \|\mathbf{v}_1 - \mathbf{w}_0\| + \|\mathbf{v}_1' - \mathbf{v}_1\| \;\le\; \epsilon + \|a\mathbf{v}_1 - \mathbf{v}_1\| \;\le\; \epsilon + |a-1| \;\le\; \epsilon + \frac{2}{2-\epsilon^2} - 1 \;=\; \epsilon + \frac{\epsilon^2}{2+\epsilon^2},$$

and since $\epsilon < \frac{1}{2}$, this is at most $\frac{5}{4}\epsilon$. □

We now turn to prove the theorem. Let $\nabla^2(\mathbf{w})$ denote the Hessian at some point $\mathbf{w}$. To show smoothness and strong convexity as stated in the theorem, it is enough to fix some unit $\mathbf{w}_0$ which is ϵ-close to the leading eigenvector $\mathbf{v}_1$ (where ϵ is assumed to be sufficiently small), and show that for any point $\mathbf{w}$ on $H_{\mathbf{w}_0}$ which is $\mathcal{O}(\epsilon)$ close to $\mathbf{w}_0$, and any direction $\mathbf{g}$ along $H_{\mathbf{w}_0}$ (i.e. any unit $\mathbf{g}$ such that $\langle \mathbf{g}, \mathbf{w}_0\rangle = 0$), it holds that $\mathbf{g}^\top\nabla^2(\mathbf{w})\mathbf{g} \in [\lambda, 20]$. This implies that the second derivative in an $\mathcal{O}(\epsilon)$ neighborhood of $\mathbf{w}_0$ on $H_{\mathbf{w}_0}$ is always in $[\lambda, 20]$, hence the function is both λ-strongly convex in that neighborhood.

More formally, letting $\epsilon \in (0,1)$ be a small parameter to be chosen later, consider any $\mathbf{w}_0$ such that

$$\|\mathbf{w}_0\| = 1 \;\;,\;\; \|\mathbf{w}_0 - \mathbf{v}_1\| \le \epsilon,$$

any $\mathbf{w}$ such that

$$\langle \mathbf{w} - \mathbf{w}_0, \mathbf{w}_0\rangle = 0 \;\;,\;\; \|\mathbf{w} - \mathbf{w}_0\| \le 2\epsilon,$$

and any $\mathbf{g}$ such that

$$\|\mathbf{g}\| = 1 \;\;,\;\; \langle \mathbf{g}, \mathbf{w}_0\rangle = 0.$$

Our goal is to show that for an appropriate ϵ, we have $\mathbf{g}^\top\nabla^2(\mathbf{w})\mathbf{g} \in [\lambda, 20]$. Moreover, by Lemma 20, the neighborhood set $H_{\mathbf{w}_0} \cap B_{\mathbf{w}_0}(2\epsilon)$ would also contain a point $a\mathbf{v}_1$ for some a, which is a global optimum of F due to its scale-invariance. This would establish the theorem.

The easier part is to show the upper bound on $\mathbf{g}^\top \nabla^2(\mathbf{w})\mathbf{g}$. Since $\mathbf{g}$ is a unit vector, it is enough to bound the spectral norm of $\nabla^2(\mathbf{w})$, which equals

$$\begin{aligned}
&\left\| \frac{1}{\|\mathbf{w}\|^2} \left(\left(I - \frac{4}{\|\mathbf{w}\|^2}\mathbf{w}\mathbf{w}^\top \right) \left(F(\mathbf{w})I + A \right) \right)^\perp \right\|_2 \\
&\leq \frac{2}{\|\mathbf{w}\|^2} \left\| \left(I - \frac{4}{\|\mathbf{w}\|^2}\mathbf{w}\mathbf{w}^\top \right) \left(F(\mathbf{w})I + A \right) \right\|_2 \\
&\leq \frac{2}{\|\mathbf{w}\|^2} \left\| I - \frac{4}{\|\mathbf{w}\|^2}\mathbf{w}\mathbf{w}^\top \right\|_2 \|F(\mathbf{w})I + A\|_2 \\
&\leq \frac{2}{\|\mathbf{w}\|^2} \left(\|I\|_2 + \left\| \frac{4}{\|\mathbf{w}\|^2}\mathbf{w}\mathbf{w}^\top \right\|_2 \right) \left(\|F(\mathbf{w})I\|_2 + \|A\|_2 \right).
\end{aligned}$$

Since the spectral norm of A is 1, and $\|\mathbf{w}\|^2 \geq 1$ (as $\mathbf{w}$ lies on a hyperplane $H_{\mathbf{w}_0}$ tangent to a unit vector $\mathbf{w}_0$), it is easy to verify that this is at most $2(1+4)(1+1) = 20$ as required.

We now turn to lower bound $\mathbf{g}^\top \nabla^2(\mathbf{w})\mathbf{g}$, which by Lemma 19 equals

$$-\frac{1}{\|\mathbf{w}\|^2}\mathbf{g}^\top \left(\left(I - \frac{4}{\|\mathbf{w}\|^2}\mathbf{w}\mathbf{w}^\top \right) \left(F(\mathbf{w})I + A \right) \right)^\perp \mathbf{g}.$$

Since $\mathbf{g}^\top B^\perp \mathbf{g} = \mathbf{g}^\top B\mathbf{g} + \mathbf{g}^\top B^\top \mathbf{g} = 2\mathbf{g}^\top B\mathbf{g}$, the above equals

$$-\frac{2}{\|\mathbf{w}\|^2}\mathbf{g}^\top \left(I - \frac{4}{\|\mathbf{w}\|^2}\mathbf{w}\mathbf{w}^\top \right) \left(F(\mathbf{w})I + A \right)\mathbf{g}. \tag{24}$$

Using the fact that $\mathbf{w} = \mathbf{w}_0 + (\mathbf{w} - \mathbf{w}_0)$, and $\langle \mathbf{g}, \mathbf{w}_0 \rangle = 0$, we get that $\langle \mathbf{g}, \mathbf{w} \rangle = \langle \mathbf{g}, \mathbf{w} - \mathbf{w}_0 \rangle$. Moreover, since A is positive semidefinite and has spectral norm of 1, $F(\mathbf{w}) = -\frac{\mathbf{w}^\top A \mathbf{w}}{\|\mathbf{w}\|^2} \in [-1, 0]$. Expanding Eq. (24) and plugging these in, we get

$$\begin{aligned}
&-\frac{2}{\|\mathbf{w}\|^2} \left(F(\mathbf{w})\mathbf{g}^\top \left(I - \frac{4}{\|\mathbf{w}\|^2}\mathbf{w}\mathbf{w}^\top \right) \mathbf{g} + \mathbf{g}^\top \left(I - \frac{4}{\|\mathbf{w}\|^2}\mathbf{w}\mathbf{w}^\top \right) A\mathbf{g} \right) \\
&= \frac{2}{\|\mathbf{w}\|^2} \left(-F(\mathbf{w})\|\mathbf{g}\|^2 + \frac{4F(\mathbf{w})}{\|\mathbf{w}\|^2}\langle \mathbf{g}, \mathbf{w} - \mathbf{w}_0 \rangle^2 \; - \; \mathbf{g}^\top A\mathbf{g} + \frac{4}{\|\mathbf{w}\|^2}\langle \mathbf{g}, \mathbf{w} - \mathbf{w}_0 \rangle \mathbf{w}^\top A\mathbf{g} \right) \\
&\geq \frac{2}{\|\mathbf{w}\|^2} \left(-F(\mathbf{w})\|\mathbf{g}\|^2 - \frac{4}{\|\mathbf{w}\|^2}\|\mathbf{g}\|^2\|\mathbf{w} - \mathbf{w}_0\|^2 \; - \; \mathbf{g}^\top A\mathbf{g} - \frac{4}{\|\mathbf{w}\|^2}\|\mathbf{g}\|\|\mathbf{w} - \mathbf{w}_0\|\|\mathbf{w}\|\|A\|_2\|\mathbf{g}\| \right).
\end{aligned}$$

Since $\|\mathbf{g}\| = 1$, $\|A\|_2 = 1$, $\|\mathbf{w} - \mathbf{w}_0\| \leq 2\epsilon$, and $\|\mathbf{w}\|^2 = \|\mathbf{w}_0\|^2 + \|\mathbf{w} - \mathbf{w}_0\|^2$ is between 1 and $1 + 4\epsilon^2$, this is at least

$$\frac{2}{\|\mathbf{w}\|^2} \left((-F(\mathbf{w})) - 16\epsilon^2 - \mathbf{g}^\top A\mathbf{g} - 8\epsilon\sqrt{1+4\epsilon^2} \right) \; = \; \frac{2}{\|\mathbf{w}\|^2} \left(-F(\mathbf{w}) - \mathbf{g}^\top A\mathbf{g} - 8\epsilon \left(2\epsilon + \sqrt{1+4\epsilon^2} \right) \right). \tag{25}$$

Let us now analyze $-F(\mathbf{w})$ and $\mathbf{g}^\top A\mathbf{g}$ more carefully. The idea will be to show that since we are close to the optimum, $-F(\mathbf{w})$ is very close to 1, and $\mathbf{g}$ (which is orthogonal to the near-optimal $\mathbf{w}_0$) is such that $\mathbf{g}^\top A\mathbf{g}$ is strictly smaller than 1. This would give us a positive lower bound on Eq. (25).

- By the triangle inequality and the assumptions $\|\mathbf{w}_0 - \mathbf{v}_1\| \leq \epsilon$, $\|\mathbf{w} - \mathbf{w}_0\| \leq 2\epsilon$, we have $\|\mathbf{w} - \mathbf{v}_1\| \leq 3\epsilon$. Also, we claim that $F(\cdot)$ is 4-Lipschitz outside the unit Euclidean ball (since the gradient of F at any point with norm ≥ 1, according to Lemma 19, has norm at most 4). Therefore, $|F(\mathbf{w}) + 1| = |F(\mathbf{w}) - F(\mathbf{v}_1)| \leq 4\|\mathbf{w} - \mathbf{v}_1\| \leq 12\epsilon$, so overall,
$$F(\mathbf{w}) \leq -1 + 12\epsilon. \tag{26}$$
- Since $\langle \mathbf{w}_0, \mathbf{g} \rangle = 0$, and $\|\mathbf{w}_0 - \mathbf{v}_1\| \leq \epsilon$, it follows that
$$|\langle \mathbf{v}_1, \mathbf{g} \rangle| \; \leq \; |\langle \mathbf{v}_1 - \mathbf{w}_0, \mathbf{g} \rangle| + |\langle \mathbf{w}_0, \mathbf{g} \rangle| \; \leq \; \|\mathbf{v}_1 - \mathbf{w}_0\|\|\mathbf{g}\| + 0 \; \leq \; \epsilon.$$

Letting $\mathbf{v}_1, \ldots, \mathbf{v}_d$ and $1 = s_1 > s_2 \geq .. \geq s_d \geq 0$ be the eigenvectors and eigenvalues of A in decreasing order (and recalling that $s_2 \leq s_1 - \lambda = 1 - \lambda$ for some eigengap $\lambda > 0$), we get

$$\begin{aligned}\mathbf{g}^\top A \mathbf{g} = \sum_{i=1}^{d} s_i \langle \mathbf{v}_i, \mathbf{g} \rangle^2 \;&\leq\; \langle \mathbf{v}_1, \mathbf{g} \rangle^2 + (1-\lambda) \sum_{i=1}^{d} \langle \mathbf{v}_i, \mathbf{g} \rangle^2 \\ &= \langle \mathbf{v}_1, \mathbf{g} \rangle^2 + (1-\lambda)(1 - \langle \mathbf{v}_1, \mathbf{g} \rangle^2) \;=\; \lambda \langle \mathbf{v}_1, \mathbf{g} \rangle^2 + (1-\lambda) \\ &\leq \lambda \epsilon^2 + (1-\lambda) \;=\; 1 - (1-\epsilon^2)\lambda. \end{aligned} \tag{27}$$

Plugging Eq. (26) and Eq. (27) back into Eq. (25), we get a lower bound of

$$\begin{aligned}&\frac{2}{\|\mathbf{w}\|^2} \left(1 - 12\epsilon - \left(1 - (1-\epsilon^2)\lambda\right) - 8\epsilon \left(2\epsilon + \sqrt{1+4\epsilon^2}\right)\right) \\ &= \frac{2}{\|\mathbf{w}\|^2} \left((1-\epsilon^2)\lambda - 8\epsilon \left(1.5 + 2\epsilon + \sqrt{1+4\epsilon^2}\right)\right) \\ &= \frac{2}{\|\mathbf{w}\|^2} \left(1 - \epsilon^2 - \frac{8\epsilon \left(1.5 + 2\epsilon + \sqrt{1+4\epsilon^2}\right)}{\lambda}\right) \lambda.\end{aligned}$$

Using the fact that $\sqrt{1+z^2} \leq 1 + z$, this can be loosely lower bounded by

$$\frac{2}{\|\mathbf{w}\|^2} \left(1 - \epsilon - \frac{8\epsilon\,(2.5 + 4\epsilon)}{\lambda}\right) \lambda.$$

Recalling that $\|\mathbf{w}\|^2 = \|\mathbf{w}_0\|^2 + \|\mathbf{w} - \mathbf{w}_0\|^2$ is at most $1 + 4\epsilon^2$, and picking ϵ sufficiently small compared to λ, (say $\epsilon = \lambda/44$), we get that the above is at least λ, which implies the required strong convexity condition.

To summarize, by picking $\epsilon = \lambda/44$, we have shown that the function $F(\mathbf{w})$ is λ-strongly convex and 20-smooth in a neighborhood of size $2\epsilon = \frac{\lambda}{22}$ around $\mathbf{w}_0$ on the hyperplane $H_{\mathbf{w}_0}$, provided that $\|\mathbf{w}_0 - \mathbf{v}_1\| \leq \epsilon = \frac{\lambda}{44}$. By Lemma 20, we are guaranteed that this neighborhood contains $\mathbf{v}_1$ up to some rescaling (which is immaterial for our scale-invariant function F), hence by optimizing F in that neighborhood, we will get a globally optimal solution.