A. Proof of Lemma 1

Since we focus on a particular epoch \( s \), let us drop the subscript from \( \mathbf{w}_{s-1} \), and denote it simply at \( \mathbf{w} \). Rewriting the update equations from the algorithm, we have that

\[
\mathbf{w}_{t+1} = \frac{\mathbf{w}'_{t+1}}{\|\mathbf{w}'_{t+1}\|}, \quad \text{where} \quad \mathbf{w}'_{t+1} = (I + \eta A)\mathbf{w}_t + \eta (\mathbf{x}\mathbf{x}^\top - A)(\mathbf{w}_t - \mathbf{w}),
\]

where \( \mathbf{x} \) is the random instance chosen at iteration \( t \).

It is easy to verify that

\[
\langle \mathbf{w}'_{t+1}, \mathbf{v}_i \rangle = a_i + z_i, \quad (15)
\]

where

\[
a_i = (1 + \eta s_i)\langle \mathbf{w}_t, \mathbf{v}_i \rangle, \quad z_i = \eta \mathbf{v}_i^\top (\mathbf{x}\mathbf{x}^\top - A)(\mathbf{w}_t - \mathbf{w}).
\]

Moreover, since \( \mathbf{v}_1, \ldots, \mathbf{v}_d \) form an orthonormal basis in \( \mathbb{R}^d \), we have

\[
\|\mathbf{w}'_{t+1}\|^2 = \sum_{i=1}^d (\mathbf{v}_i, \mathbf{w}'_{t+1})^2 = \sum_{i=1}^d (a_i + z_i)^2. \quad (16)
\]

Let \( \mathbb{E} \) denote expectation with respect to \( \mathbf{x} \), conditioned on \( \mathbf{w}_t \). Combining Eq. (15) and Eq. (16), we have

\[
\mathbb{E} [\langle \mathbf{w}_{t+1}, \mathbf{v}_1 \rangle^2] = \mathbb{E} \left[ \frac{\mathbf{w}'_{t+1}}{\|\mathbf{w}'_{t+1}\|}, \mathbf{v}_1 \right]^2 = \mathbb{E} \left[ \frac{\mathbf{w}'_{t+1}}{\|\mathbf{w}'_{t+1}\|}, \mathbf{v}_1 \right]^2 = \mathbb{E} \left[ \sum_{i=1}^d (a_i + z_i)^2 \right]. \quad (17)
\]

Note that conditioned on \( \mathbf{w}_t \), the quantities \( a_1 \ldots a_d \) are fixed, whereas \( z_1 \ldots z_d \) are random variables (depending on the random choice of \( \mathbf{x} \)) over which we take an expectation.

The first step of the proof is to simplify Eq. (17), by pushing the expectations inside the numerator and the denominator. Of course, this may change the value of the expression, so we need to account for this change with some care. To do so, define the auxiliary non-negative random variables \( x, y \) and a function \( f(x, y) \) as follows:

\[
x = (a_1 + z_1)^2, \quad y = \sum_{i=2}^d (a_i + z_i)^2, \quad f(x, y) = \frac{x}{x+y}.
\]

Then we can write Eq. (17) as \( \mathbb{E}_{x, y}[f(x, y)] \). We now use a second-order Taylor expansion to relate it to \( f(\mathbb{E}[x], \mathbb{E}[y]) = \frac{\mathbb{E}(a_1 + z_1)}{\sum_{i=1}^d (a_i + z_i)^2} \). Specifically, we have that \( \mathbb{E}_{x, y}[f(x, y)] \) can be lower bounded by

\[
\mathbb{E}_{x, y} \left[ f(\mathbb{E}[x], \mathbb{E}[y]) + \nabla f(\mathbb{E}[x], \mathbb{E}[y])^\top \left( \begin{array}{c} x \\ y \end{array} \right) - \mathbb{E}_x \mathcal{E}_y \left( \begin{array}{c} \mathbb{E}[x] \\ \mathbb{E}[y] \end{array} \right) \right] \]

\[
\quad \geq f(\mathbb{E}[x], \mathbb{E}[y]) - \mathbb{E}_x \mathcal{E}_y \left\| \nabla^2 f(x, y) \right\| \max_{x, y} \left\| \left( \begin{array}{c} x - \mathbb{E}[x] \\ y - \mathbb{E}[y] \end{array} \right) \right\|^2,
\]

where \( \nabla^2 f(x, y) \) is the Hessian of \( f \) at \( (x, y) \).

We now upper bound the two max-terms in the expression above. First, it is easily verified that

\[
\nabla^2 f(x, y) = \frac{1}{(x+y)^3} \left( \begin{array}{cc} -2y & x-y \\ x-y & 2x \end{array} \right).
\]

Since the spectral norm is upper bounded by the Frobenius norm, which for \( 2 \times 2 \) matrices is upper bounded by \( 2 \) times the magnitude of the largest entry in the matrix (which in our case is at most \( 2(x+y)/(x+y)^3 = 2/(x+y)^2 \leq 2/x^2 \)), we have

\[
\max_{x, y} \left\| \nabla^2 f(x, y) \right\| \leq \max_x \frac{4}{x^2} = \max_{z_1} \frac{4}{(a_1 + z_1)^2}.
\]
Now, recall that \( a_1 \geq \frac{1}{2} \) by the Lemma’s assumptions, and in contrast \( |z_1| \leq \eta \|v_1^T(xx^T - A)(w_t - \tilde{w})\| \leq \eta\|v_1\|\|xx^T - A\|\|w_t - \tilde{w}\| \leq c\eta \), so for \( \eta \) sufficiently small, \( |z_1| \leq \frac{1}{2}|a_1| \), and we can upper bound \( \frac{4}{(a_1 + z_1)^2} \) (and hence \( \max_{x,y}\|\nabla^2 f(x,y)\| \)) by some numerical constant \( c \).

Turning to the \( \max_{x,y}[(x - E[x])^2 + (y - E[y])^2] \) term in Eq. (18), and recalling that \( x = (a_1 + z_1)^2 \), \( y = \sum_{i=2}^{d}(a_i + z_i)^2 \), and the \( z_i \)'s are zero-mean, we have

\[
\max_{x,y} ((x - E[x])^2 + (y - E[y])^2) = \max_{z_1,...,z_d} 4 \left( (a_1 z_1)^2 + \left( \sum_{i=2}^{d} a_i z_i \right)^2 \right)
\]

By definition of \( a_i, z_i \), and recalling that \( \|w_t\|, \|v_1\|, \eta s_i \) and \( \|xx^T - A\| \) are all bounded by constants, this expression equals

\[
4\eta^2 \left( (1 + \eta s_1)\langle w_t, v_1 \rangle v_1^T(xx^T - A)(w_t - \tilde{w})^2 + \left( \sum_{i=2}^{d} (1 + \eta s_i)\langle v_i, w_t \rangle v_i^T(xx^T - A)(w_t - \tilde{w}) \right)^2 \right)
\]

\[
\leq c\eta^2 \left( \|w_t - \tilde{w}\|^2 + \left( \left( \sum_{i=2}^{d} (1 + \eta s_i)\langle v_i, w_t \rangle \right) \|w_t - \tilde{w}\| \right)^2 \right)
\]

\[
= c\eta^2 \|w_t - \tilde{w}\|^2 \left( 1 + \left( \sum_{i=2}^{d} (1 + \eta s_i)\|v_i v_i^T w_t\| \right)^2 \right)
\]

\[
\leq c\eta^2 \|w_t - \tilde{w}\|^2 \left( 1 + \left( \sum_{i=2}^{d} (1 + \eta s_i)\|v_i v_i^T\| \right)^2 \right)
\]

where in the last inequality we used the fact that \( v_2 \ldots v_d \) are orthonormal vectors, and \( (1 + \eta s_i) \) is bounded by a constant.

Plugging the bounds we have derived into Eq. (18), we get a lower bound of

\[
f(E[x], E[y]) - c\eta^2 \|w_t - \tilde{w}\|^2 = \frac{a_1^2 + z_1^2}{\sum_{i=1}^{d} (a_i^2 + z_i^2)} - c\eta^2 \|w_t - \tilde{w}\|^2
\]  

(19)

By definition of \( z_i \) and the fact that \( v_1, \ldots, v_d \) are orthonormal (hence \( \sum_i v_i v_i^T \) is the identity matrix), we have

\[
\sum_{i=1}^{d} z_i^2 = \eta^2 (w_t - \tilde{w})^T(xx^T - A) \left( \sum_{i=1}^{d} v_i v_i^T \right)(xx^T - A)(w_t - \tilde{w})
\]

\[
= \eta^2 (w_t - \tilde{w})^T(xx^T - A)(xx^T - A)(w_t - \tilde{w})
\]

\[
= \eta^2 \|xx^T - A\| \|w_t - \tilde{w}\|^2 \leq c\eta^2 \|w_t - \tilde{w}\|^2,
\]

so we can lower bound Eq. (19) by

\[
\frac{a_1^2}{\sum_{i=1}^{d} a_i^2 + c\eta^2 \|w_t - \tilde{w}\|^2} - c\eta^2 \|w_t - \tilde{w}\|^2.
\]  

(20)
Focusing on the first term in Eq. (20) for the moment, and substituting in the definition of $a_s$, we can write it as

$$
\frac{(1 + \eta s_1)^2 \langle w_t, v_1 \rangle^2}{(1 + \eta s_2)^2 \langle w_t, v_1 \rangle^2 + \sum_{i=2}^{d} (1 + \eta s_i)^2 \langle v_i, w_t \rangle^2 + c \eta^2 \| w_t - \tilde{w} \|^2} \geq \frac{\langle w_t, v_1 \rangle^2}{\langle w_t, v_1 \rangle^2 + (1 + \frac{\eta s_2}{1 + \eta s_1})^2 (1 - \langle w_t, v_1 \rangle^2) + c \eta^2 \| w_t - \tilde{w} \|^2} = 1 - \left(1 - \left(1 + \frac{\eta s_2}{1 + \eta s_1}\right)^2 \right) (1 - \langle w_t, v_1 \rangle^2) + c \eta^2 \| w_t - \tilde{w} \|^2
$$

where in the last step we used the elementary inequality $\frac{1}{1+x} \geq 1 + x$ for all $x \leq 1$ (and this is indeed justified since $\langle w_t, v_1 \rangle \leq 1$ and $\frac{\eta s_2}{1 + \eta s_1} \leq 1$). This can be further lower bounded by

$$
\langle w_t, v_1 \rangle^2 \left(1 + \left(1 - \left(1 + \frac{\eta s_2}{1 + \eta s_1}\right)^2 \right) (1 - \langle w_t, v_1 \rangle^2) - c \eta^2 \| w_t - \tilde{w} \|^2\right) \geq \langle w_t, v_1 \rangle^2 \left(1 + \eta \frac{s_1 - s_2}{1 + \eta s_1} (1 - \langle w_t, v_1 \rangle^2) - c \eta^2 \| w_t - \tilde{w} \|^2\right)
$$

where in the last inequality we used the fact that $s_1 - s_2 = \lambda$ and that $\eta s_1 \leq \eta$ which is at most 1 (again using the assumption that $\eta$ is sufficiently small).

Plugging this lower bound on the first term in Eq. (20), and recalling that $\langle w_t, v_1 \rangle^2$ is assumed to be at least $1/4$, we get the following lower bound on Eq. (20):

$$
\langle w_t, v_1 \rangle^2 \left(1 + \eta \frac{s_1 - s_2}{1 + \eta s_1} (1 - \langle w_t, v_1 \rangle^2) - c \eta^2 \| w_t - \tilde{w} \|^2\right) \geq \langle w_t, v_1 \rangle^2 \left(1 + \eta \frac{\lambda}{2} (1 - \langle w_t, v_1 \rangle^2) - c \eta^2 \| w_t - \tilde{w} \|^2\right).
$$

To summarize the derivation so far, starting from Eq. (17) and concatenating the successive lower bounds we have derived, we get that

$$
E[\langle w_{t+1}, v_1 \rangle^2] \geq \langle w_t, v_1 \rangle^2 \left(1 + \eta \frac{\lambda}{2} (1 - \langle w_t, v_1 \rangle^2) - c \eta^2 \| w_t - \tilde{w} \|^2\right). \tag{21}
$$

We now get rid of the $\| w_t - \tilde{w} \|^2$ term, by noting that since $(x + y)^2 \leq 2(x^2 + y^2)$ and $\| w_t \| = \| v_1 \| = 1$,

$$
\| w_t - \tilde{w} \|^2 \leq (\| w_t - v_1 \| + \| \tilde{w} - v_1 \|)^2 \leq 2 (\| w_t - v_1 \|^2 + \| \tilde{w} - v_1 \|^2) = 2 (2 - 2\langle w_t, v_1 \rangle + 2 - 2\langle \tilde{w}, v_1 \rangle).
$$

Since we assume that $\langle w_t, v_1 \rangle$, $\langle \tilde{w}, v_1 \rangle$ are both positive, and they are also at most 1, this is at most

$$
2 (2 - 2\langle w_t, v_1 \rangle^2 + 2 - 2\langle \tilde{w}, v_1 \rangle^2) = 4 (1 - \langle w_t, v_1 \rangle^2) + 4 (1 - \langle \tilde{w}, v_1 \rangle^2).
$$

Plugging this back into Eq. (21), we get that

$$
E[\langle w_{t+1}, v_1 \rangle^2] \geq \langle w_t, v_1 \rangle^2 \left(1 + \eta \frac{\lambda}{2} - c \eta^2 \right) (1 - \langle w_t, v_1 \rangle^2) - c \eta^2 (1 - \langle \tilde{w}, v_1 \rangle^2),
$$
and since we can assume \( \frac{n \lambda}{2} - c \eta^2 \geq \frac{n \lambda}{4} \) by picking \( \eta \) sufficiently smaller than \( \lambda \), this can be simplified to
\[
E[(w_{t+1}, v_1)^2] \geq (w_t, v_1)^2 \left( 1 + \frac{n \lambda}{4} (1 - \langle w_t, v_1 \rangle^2) - c \eta^2 (1 - \langle \tilde{w}, v_1 \rangle^2) \right).
\]

The final stage of the proof consists of converting the bound above to a bound on \( E[1 - \langle w_{t+1}, v_1 \rangle^2] \) in terms of \( 1 - \langle w_t, v_1 \rangle^2 \). To simplify the notation, let \( b = (1 - \langle w_t, v_1 \rangle^2) \) and \( \tilde{b} = c \eta^2 (1 - \langle \tilde{w}, v_1 \rangle^2) \), so the bound above implies
\[
E[1 - \langle w_{t+1}, v_1 \rangle^2] \leq 1 - (1 - b) \left( 1 + \frac{n \lambda}{4} b - \tilde{b} \right) = 1 - (1 - b) - \frac{n \lambda}{4} b (1 - b) + (1 - b) \tilde{b} = b - \frac{n \lambda}{4} b (1 - b) - \tilde{b} + \tilde{b} = b \left( 1 - \frac{n \lambda}{4} (1 - b) - \tilde{b} \right) + \tilde{b}.
\]

Plugging back the definitions of \( \tilde{b}, b \), we get that
\[
E[1 - \langle w_{t+1}, v_1 \rangle^2] \leq (1 - \langle w_t, v_1 \rangle^2) \left( 1 - \frac{n \lambda}{4} \langle w_t, v_1 \rangle^2 - c \eta^2 (1 - \langle \tilde{w}, v_1 \rangle^2) \right) + c \eta^2 (1 - \langle \tilde{w}, v_1 \rangle^2).
\]

Since we assume \( \langle w_t, v_1 \rangle \geq \frac{1}{2} \), we can upper bound this by
\[
(1 - \langle w_t, v_1 \rangle^2) \left( 1 - \frac{n \lambda}{4} \right) + c \eta^2 (1 - \langle \tilde{w}, v_1 \rangle^2)
\]
as required. Note that to get this bound, we assumed at several places that \( \eta \) is smaller than either a constant, or a constant factor times \( \lambda \) (which is at most 1). Hence, the bound holds by assuming \( \eta \leq c \lambda \) for a sufficiently small numerical \( c \).

**B. Implementing Epochs in \( \mathcal{O}(d_s(m + n)) \) Runtime**

As discussed in remark 2, the runtime of each iteration in our algorithm (as presented in our pseudo-code) is \( \mathcal{O}(d) \), and the total runtime of each epoch is \( \mathcal{O}(dm + d_n) \), where \( d_s \) is the average sparsity (number of non-zero entries) in the data points \( x_i \). Here, we explain how the total epoch runtime can be improved (at least in terms of the theoretical analysis) to \( \mathcal{O}(d_s(m + n)) \). For ease of exposition, we reproduce the pseudo-code together with line numbers below:

1: **Parameters:** Step size \( \eta \), epoch length \( m \)
2: **Input:** Data matrix \( X = (x_1, \ldots, x_n) \); Initial unit vector \( \tilde{w}_0 \)
3: for \( s = 1, 2, \ldots \) do
4: \hspace{1em} \( \tilde{u} = \frac{1}{n} \sum_{i=1}^{n} x_i \langle x_i^\top \tilde{w}_{s-1} \rangle \)
5: \hspace{1em} \( w_0 = \tilde{w}_{s-1} \)
6: for \( t = 1, 2, \ldots, m \) do
7: \hspace{2em} Pick \( i_t \in \{1, \ldots, n\} \) uniformly at random
8: \hspace{2em} \( w_t = w_{t-1} + \eta (x_{i_t} (x_{i_t}^\top w_{t-1} - x_{i_t}^\top \tilde{w}_{s-1}) + \tilde{u}) \)
9: \hspace{2em} \( w_{\tilde{t}} = \frac{1}{\|w_t\|} w_t \)
10: end for
11: \( \tilde{w}_s = w_{\tilde{t}} \)
12: end for

First, we can assume without loss of generality that \( d \leq d_s n \). Otherwise, the number of non-zeros in the \( n \times d \) data matrix \( X \) is smaller than \( d \), so the matrix must contain some all-zeros columns. But then, we can simply drop those columns (the value of the largest singular vectors in the corresponding entries will be zero anyway), hence reducing the effective dimension \( d \) to be at most \( d_s n \). Therefore, given a vector \( w_{s-1} \), we can implement line (4) in \( \mathcal{O}(d + d_s n) \leq \mathcal{O}(d_s n) \) time, by initializing the \( d \)-dimensional vector \( \tilde{u} \) to be 0, and iteratively adding to it the sparse (on-average) vector \( x_{i_t} (x_{i_t}^\top \tilde{w}_{s-1}) \). Similarly, we can implement lines (5),(11) in \( \mathcal{O}(d) \leq \mathcal{O}(d_s n) \) time.
It remains to show that we can implement each iteration in lines (8) and (9) in $O(d_s)$ time. To do so, instead of explicitly storing $w_t, w'_t$, we only store $\tilde{u}$, an auxiliary vector $g$, and auxiliary scalars $\alpha, \beta, \gamma, \delta, \zeta$, such that

- At the end of line (8), $w'_t$ is stored as $\alpha g + \beta \tilde{u}$
- At the end of line (9), $w_t$ is stored as $\alpha g + \beta \tilde{u}$

- It holds that $\gamma = \| \alpha g \|^2$, $\delta = \langle \alpha g, \tilde{u} \rangle$, $\zeta = \| \tilde{u} \|^2$. This ensures that $\gamma + 2\delta + \zeta$ expresses $\| \alpha g + \beta \tilde{u} \|^2$.

Before the beginning of the epoch (line (5)), we initialize $g = \tilde{w}_{s-1}$, $\alpha = 1$, $\beta = 0$ and compute $\gamma = \| \alpha g \|^2$, $\delta = \langle \alpha g, \tilde{u} \rangle$, $\zeta = \| \tilde{u} \|^2$, all in time $O(d) \leq O(d_s n)$. This ensures that $w_0 = \alpha g + \beta \tilde{u}$. Line (8) can be implemented in $O(d_s)$ time as follows:

- Compute the sparse (on-average) update vector $\Delta g := \eta x_i, (x_i^T w_{t-1} - x_i^T \tilde{w}_{s-1})$
- Update $g := g + \Delta g / \alpha$; $\beta := \beta + \eta$; $\gamma := \gamma + 2\alpha \langle g, \Delta g \rangle + \| \Delta g \|^2$; $\delta := \delta + \langle \Delta g, \tilde{u} \rangle$. This implements line (8), and ensures that $w'_t$ is represented as $\alpha g + \beta \tilde{u}$, and its squared norm equals $\gamma + 2\delta + \zeta$.

To implement line (9), we simply divide $\alpha, \beta$ by $\sqrt{\gamma + 2\delta + \zeta}$ (which equals the norm of $w'_t$), and recompute $\gamma, \delta$ accordingly. After this step, $w_t$ is represented by $\alpha g + \beta \tilde{u}$ as required.