A. Proof of Lemma 1

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

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

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

It is easy to verify that

$$\langle \mathbf{w}'_{t+1}, \mathbf{v}_i\rangle = a_i + z_i, \tag{15}$$

where

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

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

$$\|\mathbf{w}'_{t+1}\|^2 = \sum_{i=1}^d \langle \mathbf{v}_i, \mathbf{w}'_{t+1}\rangle^2 = \sum_{i=1}^d (a_i+z_i)^2. \tag{16}$$

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

$$\mathbb{E}\left[\langle \mathbf{w}_{t+1}, \mathbf{v}_1\rangle^2\right] = \mathbb{E}\left[\left\langle \frac{\mathbf{w}'_{t+1}}{\|\mathbf{w}'_{t+1}\|}, \mathbf{v}_1\right\rangle^2\right] = \mathbb{E}\left[\frac{\langle \mathbf{w}'_{t+1}, \mathbf{v}_1\rangle^2}{\|\mathbf{w}'_{t+1}\|^2}\right] = \mathbb{E}\left[\frac{(a_1+z_1)^2}{\sum_{i=1}^d (a_i+z_i)^2}\right]. \tag{17}$$

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:

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

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 $f(\mathbb{E}[x], \mathbb{E}[y]) = \frac{\mathbb{E}[(a_1+z_1)^2]}{\mathbb{E}[\sum_{i=1}^d (a_i+z_i)^2]}$. Specifically, we have that $\mathbb{E}_{x,y}[f(x,y)]$ can be lower bounded by

$$\begin{aligned}
&\mathbb{E}_{x,y}\left[f(\mathbb{E}[x],\mathbb{E}[y]) + \nabla f(\mathbb{E}[x],\mathbb{E}[y])^\top\left(\begin{pmatrix}x\\y\end{pmatrix} - \begin{pmatrix}\mathbb{E}[x]\\\mathbb{E}[y]\end{pmatrix}\right) - \max_{x,y}\|\nabla^2 f(x,y)\| \max_{x,y}\left\|\begin{pmatrix}x\\y\end{pmatrix} - \begin{pmatrix}\mathbb{E}[x]\\\mathbb{E}[y]\end{pmatrix}\right\|^2\right]\\
&\quad = f(\mathbb{E}[x],\mathbb{E}[y]) - \max_{x,y}\|\nabla^2 f(x,y)\| \max_{x,y}\left\|\begin{pmatrix}x-\mathbb{E}[x]\\y-\mathbb{E}[y]\end{pmatrix}\right\|^2,
\end{aligned} \tag{18}$$

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

$$\nabla^2 f(x,y) = \frac{1}{(x+y)^3}\begin{pmatrix}-2y & x-y\\ x-y & 2x\end{pmatrix}.$$

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

$$\max_{x,y}\|\nabla^2 f(x,y)\| \le \max_x \frac{4}{x^2} = \max_{z_1}\frac{4}{(a_1+z_1)^2}.$$

Now, recall that $a_1 \geq \frac{1}{2}$ by the Lemma's assumptions, and in contrast $|z_1| \leq \eta\left|\mathbf{v}_i^\top(\mathbf{x}\mathbf{x}^\top - A)(\mathbf{w}_t - \tilde{\mathbf{w}})\right| \leq \eta\|\mathbf{v}_i\|\|\mathbf{x}\mathbf{x}^\top - A\|\|\mathbf{w}_t - \tilde{\mathbf{w}}\| \leq c\eta$, so for η 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 $\max_{x,y}[(x-\mathbb{E}[x])^2+(y-\mathbb{E}[y])^2]$ 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

$$\max_{x,y}\left((x-\mathbb{E}[x])^2+(y-\mathbb{E}[y])^2\right) \;=\; \max_{z_1\ldots z_d} 4\left((a_1z_1)^2 + \left(\sum_{i=2}^d a_iz_i\right)^2\right)$$

By definition of a_i, z_i, and recalling that $\|\mathbf{w}_t\|$,$\|\mathbf{v}_1\|$,ηs_i and $\|\mathbf{x}\mathbf{x}^\top - A\|$ are all bounded by constants, this expression equals

$$\begin{aligned}
&4\eta^2\left(\left((1+\eta s_1)\langle\mathbf{w}_t,\mathbf{v}_1\rangle\mathbf{v}_1^\top(\mathbf{x}\mathbf{x}^\top - A)(\mathbf{w}_t-\tilde{\mathbf{w}})\right)^2 + \left(\sum_{i=2}^d(1+\eta s_i)\langle\mathbf{w}_t,\mathbf{v}_i\rangle\mathbf{v}_i^\top(\mathbf{x}\mathbf{x}^\top - A)(\mathbf{w}_t-\tilde{\mathbf{w}})\right)^2\right)\\
&\quad\leq c\eta^2\left(\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2 + \left(\left\|\sum_{i=2}^d(1+\eta s_i)\langle\mathbf{v}_i,\mathbf{w}_t\rangle\mathbf{v}_i\right\|\|\mathbf{w}_t-\tilde{\mathbf{w}}\|\right)^2\right)\\
&\quad= c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2\left(1+\left\|\sum_{i=2}^d(1+\eta s_i)\mathbf{v}_i\mathbf{v}_i^\top\mathbf{w}_t\right\|^2\right)\\
&\quad\leq c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2\left(1+\|\sum_{i=2}^d(1+\eta s_i)\mathbf{v}_i\mathbf{v}_i^\top\|^2\right)\\
&\quad\leq c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2,
\end{aligned}$$

where in the last inequality we used the fact that $\mathbf{v}_2\ldots\mathbf{v}_d$ 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

$$f(\mathbb{E}[x],\mathbb{E}[y]) - c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2 = \frac{a_1^2+z_1^2}{\sum_{i=1}^d(a_i^2+z_i^2)} - c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2 \tag{19}$$

By definition of z_i and the fact that $\mathbf{v}_1,\ldots,\mathbf{v}_d$ are orthonormal (hence $\sum_i \mathbf{v}_i\mathbf{v}_i^\top$ is the identity matrix), we have

$$\begin{aligned}
\sum_{i=1}^d z_i^2 &= \eta^2(\mathbf{w}_t-\tilde{\mathbf{w}})^\top(\mathbf{x}\mathbf{x}^\top - A)\left(\sum_{i=1}^d\mathbf{v}_i\mathbf{v}_i^\top\right)(\mathbf{x}\mathbf{x}^\top - A)(\mathbf{w}_t-\tilde{\mathbf{w}})\\
&= \eta^2(\mathbf{w}_t-\tilde{\mathbf{w}})^\top(\mathbf{x}\mathbf{x}^\top - A)(\mathbf{x}\mathbf{x}^\top - A)(\mathbf{w}_t-\tilde{\mathbf{w}})\\
&= \eta^2\|(\mathbf{x}\mathbf{x}^\top - A)(\mathbf{w}_t-\tilde{\mathbf{w}})\|^2 \;\leq\; c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2,
\end{aligned}$$

so we can lower bound Eq. (19) by

$$\frac{a_1^2}{\sum_{i=1}^d a_i^2 + c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2} - c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2. \tag{20}$$

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

$$
\begin{aligned}
&\frac{(1+\eta s_1)^2\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2}{(1+\eta s_1)^2\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2+\sum_{i=2}^d(1+\eta s_i)^2\langle \mathbf{v}_i,\mathbf{w}_t\rangle^2+c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2}\\
&\geq \frac{\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2}{\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2+\left(\frac{1+\eta s_2}{1+\eta s_1}\right)^2\sum_{i=2}^d\langle \mathbf{v}_i,\mathbf{w}_t\rangle^2+c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2}\\
&= \frac{\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2}{\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2+\left(\frac{1+\eta s_2}{1+\eta s_1}\right)^2(1-\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2)+c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2}\\
&= \frac{\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2}{1-\left(1-\left(\frac{1+\eta s_2}{1+\eta s_1}\right)^2\right)(1-\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2)+c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2}\\
&\geq \langle \mathbf{w}_t,\mathbf{v}_1\rangle^2\left(1+\left(1-\left(\frac{1+\eta s_2}{1+\eta s_1}\right)^2\right)\left(1-\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2\right)-c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2\right),
\end{aligned}
$$

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 $\langle \mathbf{w}_t,\mathbf{v}_1\rangle\leq 1$ and $\frac{1+\eta s_2}{1+\eta s_1}\leq 1$). This can be further lower bounded by

$$
\begin{aligned}
&\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2\left(1+\left(1-\left(\frac{1+\eta s_2}{1+\eta s_1}\right)\right)\left(1-\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2\right)-c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2\right)\\
&= \langle \mathbf{w}_t,\mathbf{v}_1\rangle^2\left(1+\frac{\eta(s_1-s_2)}{1+\eta s_1}\left(1-\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2\right)-c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2\right)\\
&\geq \langle \mathbf{w}_t,\mathbf{v}_1\rangle^2\left(1+\frac{\eta\lambda}{2}\left(1-\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2\right)-c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2\right),
\end{aligned}
$$

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 η is sufficiently small).

Plugging this lower bound on the first term in Eq. (20), and recalling that $\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2$ is assumed to be at least $1/4$, we get the following lower bound on Eq. (20):

$$
\begin{aligned}
&\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2\left(1+\frac{\eta\lambda}{2}\left(1-\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2\right)-c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2\right)-c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2\\
&\geq\ \langle \mathbf{w}_t,\mathbf{v}_1\rangle^2\left(1+\frac{\eta\lambda}{2}\left(1-\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2\right)-c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2\right).
\end{aligned}
$$

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

$$
\mathbb{E}[\langle \mathbf{w}_{t+1},\mathbf{v}_1\rangle^2]\geq\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2\left(1+\frac{\eta\lambda}{2}\left(1-\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2\right)-c\eta^2\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2\right). \tag{21}
$$

We now get rid of the $\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2$ term, by noting that since $(x+y)^2\leq 2(x^2+y^2)$ and $\|\mathbf{w}_t\|=\|\mathbf{v}_1\|=1$,

$$
\begin{aligned}
\|\mathbf{w}_t-\tilde{\mathbf{w}}\|^2&\leq\left(\|\mathbf{w}_t-\mathbf{v}_1\|+\|\tilde{\mathbf{w}}-\mathbf{v}_1\|\right)^2\leq 2\left(\|\mathbf{w}_t-\mathbf{v}_1\|^2+\|\tilde{\mathbf{w}}-\mathbf{v}_1\|^2\right)\\
&=2\left(2-2\langle \mathbf{w}_t,\mathbf{v}_1\rangle+2-2\langle \tilde{\mathbf{w}},\mathbf{v}_1\rangle\right).
\end{aligned}
$$

Since we assume that $\langle \mathbf{w}_t,\mathbf{v}_1\rangle,\langle \tilde{\mathbf{w}},\mathbf{v}_1\rangle$ are both positive, and they are also at most 1, this is at most

$$
2\left(2-2\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2+2-2\langle \tilde{\mathbf{w}},\mathbf{v}_1\rangle^2\right)=4\left(1-\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2\right)+4\left(1-\langle \tilde{\mathbf{w}},\mathbf{v}_1\rangle^2\right).
$$

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

$$
\mathbb{E}[\langle \mathbf{w}_{t+1},\mathbf{v}_1\rangle^2]\geq\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2\left(1+\left(\frac{\eta\lambda}{2}-c\eta^2\right)\left(1-\langle \mathbf{w}_t,\mathbf{v}_1\rangle^2\right)-c\eta^2\left(1-\langle \tilde{\mathbf{w}},\mathbf{v}_1\rangle^2\right)\right),
$$

and since we can assume $\frac{\eta\lambda}{2} - c\eta^2 \geq \frac{\eta\lambda}{4}$ by picking η sufficiently smaller than λ, this can be simplified to

$$\mathbb{E}[\langle \mathbf{w}_{t+1}, \mathbf{v}_1\rangle^2] \geq \langle \mathbf{w}_t, \mathbf{v}_1\rangle^2 \left(1 + \frac{\eta\lambda}{4}\left(1 - \langle \mathbf{w}_t, \mathbf{v}_1\rangle^2\right) - c\eta^2\left(1 - \langle \tilde{\mathbf{w}}, \mathbf{v}_1\rangle^2\right)\right).$$

The final stage of the proof consists of converting the bound above to a bound on $\mathbb{E}[1 - \langle \mathbf{w}_{t+1}, \mathbf{v}_1\rangle^2]$ in terms of $1 - \langle \mathbf{w}_t, \mathbf{v}_1\rangle^2$. To simplify the notation, let $b = \left(1 - \langle \mathbf{w}_t, \mathbf{v}_1\rangle^2\right)$ and $\tilde{b} = c\eta^2\left(1 - \langle \tilde{\mathbf{w}}, \mathbf{v}_1\rangle^2\right)$, so the bound above implies

$$\begin{aligned}
\mathbb{E}[1 - \langle \mathbf{w}_{t+1}, \mathbf{v}_1\rangle^2] \;&\leq\; 1 - (1-b)\left(1 + \frac{\eta\lambda}{4}b - \tilde{b}\right) \\
&= 1 - (1-b) - \frac{\eta\lambda}{4}b(1-b) + (1-b)\tilde{b} \\
&= b - \frac{\eta\lambda}{4}b(1-b) - b\tilde{b} + \tilde{b} \\
&= b\left(1 - \frac{\eta\lambda}{4}(1-b) - \tilde{b}\right) + \tilde{b}.
\end{aligned}$$

Plugging back the definitions of $\tilde{b}, b$, we get that

$$\mathbb{E}[1 - \langle \mathbf{w}_{t+1}, \mathbf{v}_1\rangle^2] \;\leq\; \left(1 - \langle \mathbf{w}_t, \mathbf{v}_1\rangle^2\right)\left(1 - \frac{\eta\lambda}{4}\langle \mathbf{w}_t, \mathbf{v}_1\rangle^2 - c\eta^2\left(1 - \langle \tilde{\mathbf{w}}, \mathbf{v}_1\rangle^2\right)\right) + c\eta^2\left(1 - \langle \tilde{\mathbf{w}}, \mathbf{v}_1\rangle^2\right).$$

Since we assume $\langle \mathbf{w}_t, \mathbf{v}_1\rangle \geq \frac{1}{2}$, we can upper bound this by

$$\left(1 - \langle \mathbf{w}_t, \mathbf{v}_1\rangle^2\right)\left(1 - \frac{\eta\lambda}{16}\right) + c\eta^2\left(1 - \langle \tilde{\mathbf{w}}, \mathbf{v}_1\rangle^2\right)$$

as required. Note that to get this bound, we assumed at several places that η is smaller than either a constant, or a constant factor times λ (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 $\mathcal{O}(dm + d_s n)$, where d_s is the average sparsity (number of non-zero entries) in the data points $\mathbf{x}_i$. 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 η, epoch length m
2: **Input:** Data matrix $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$; Initial unit vector $\tilde{\mathbf{w}}_0$
3: **for** $s = 1, 2, \ldots$ **do**
4:     $\tilde{\mathbf{u}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\left(\mathbf{x}_i^\top \tilde{\mathbf{w}}_{s-1}\right)$
5:     $\mathbf{w}_0 = \tilde{\mathbf{w}}_{s-1}$
6:     **for** $t = 1, 2, \ldots, m$ **do**
7:         Pick $i_t \in \{1, \ldots, n\}$ uniformly at random
8:         $\mathbf{w}'_t = \mathbf{w}_{t-1} + \eta\left(\mathbf{x}_{i_t}\left(\mathbf{x}_{i_t}^\top \mathbf{w}_{t-1} - \mathbf{x}_{i_t}^\top \tilde{\mathbf{w}}_{s-1}\right) + \tilde{\mathbf{u}}\right)$
9:         $\mathbf{w}_t = \frac{1}{\|\mathbf{w}'_t\|}\mathbf{w}'_t$
10:     **end for**
11:     $\tilde{\mathbf{w}}_s = \mathbf{w}_m$
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 $\tilde{\mathbf{w}}_{s-1}$, 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 $\tilde{\mathbf{u}}$ to be 0, and iteratively adding to it the sparse (on-average) vector $\mathbf{x}_i\left(\mathbf{x}_i^\top \tilde{\mathbf{w}}_{s-1}\right)$. 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 $\mathcal{O}(d_s)$ time. To do so, instead of explicitly storing $\mathbf{w}_t, \mathbf{w}'_t$, we only store $\tilde{\mathbf{u}}$, an auxiliary vector $\mathbf{g}$, and auxiliary scalars $\alpha, \beta, \gamma, \delta, \zeta$, such that

- At the end of line (8), $\mathbf{w}'_t$ is stored as $\alpha\mathbf{g} + \beta\tilde{\mathbf{u}}$
- At the end of line (9), $\mathbf{w}_t$ is stored as $\alpha\mathbf{g} + \beta\tilde{\mathbf{u}}$
- It holds that $\gamma = \|\alpha\mathbf{g}\|^2$, $\delta = \langle \alpha\mathbf{g}, \tilde{\mathbf{u}} \rangle$, $\zeta = \|\tilde{\mathbf{u}}\|^2$. This ensures that $\gamma + 2\delta + \zeta$ expresses $\|\alpha\mathbf{g} + \beta\tilde{\mathbf{u}}\|^2$.

Before the beginning of the epoch (line (5)), we initialize $\mathbf{g} = \tilde{\mathbf{w}}_{s-1}$, $\alpha = 1, \beta = 0$ and compute $\gamma = \|\alpha\mathbf{g}\|^2$, $\delta = \langle \alpha\mathbf{g}, \tilde{\mathbf{u}} \rangle$, $\zeta = \|\tilde{\mathbf{u}}\|^2$, all in time $\mathcal{O}(d) \leq \mathcal{O}(d_s n)$. This ensures that $\mathbf{w}_0 = \alpha\mathbf{g} + \beta\mathbf{u}$. Line (8) can be implemented in $\mathcal{O}(d_s)$ time as follows:

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

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