VII. APPENDIX

A Description of Data Sets

TABLE I
SUMMARY OF DATA SETS USED FOR KL-NMF

Type	Name	# of samples	# of features	# of nonzeros	Sparsity
Synthetic	Pois1	1,000	1,000	900,000	0.90
Synthetic	Pois2	3,000	3,000	900,000	0.10
Synthetic	Pois3	9,000	9,000	900,000	0.01
Real	NIPS	1,500	12,419	280,000	0.985
Real	WT	287	19,200	5,510,000	0.000
Real	KOS	3,430	6,906	950,000	0.960
Real	MITF	361	2,429	877,000	0.000

The 4 real data sets in the table are retrieved from https://archive.ics.uci.edu/ml/datasets/bag+of+words, https://www.microsoft.com/en-us/research/project and https://cbcl.mit.edu/cbcl. They have already been used in the previous papers such as [3], [30]. We preprocess the real data sets by removing few rows and columns having sums less than 20 for NIPS and KOS data sets.

For synthetic data, $V \in \mathbb{R}^{N\times M}$ generated from i.i.d. Poisson random variables, i.e. $V_{ij} \sim \text{Poisson}(-\log(1-\rho))$. Here ρ denotes sparsity or proportion of nonzero entries of V. This corresponds to the null signal case since in this case KL-NMF is the maximum likelihood estimation problem when $WH = 0$.

B Proofs

In what follows, we frequently use the fact that for $0 < \eta \le 1$, $\eta \le 1/\max(1,\nu)$ implies

C Equations

$$n\nu \le 1. \tag{21}$$

Using $\Delta_0 \le 1 - 1/\sqrt{2}$ which follows from (6), we often use

$$\frac{\sqrt{\Delta_0}}{1-\Delta_0} \le 1, \quad \frac{1}{1-\Delta_0} \le \sqrt{2}. \tag{22}$$

Proof of Lemma III.1. From the update rule in Algorithm 1, we have

$$\begin{aligned} x_{t+1} &= (1-\eta)x_t + \frac{\eta}{\|x_t\|^{p-2}}\left(\nabla f_{S_t}(x_t) - \alpha_t \nabla f_{S_t}(y_0) + \alpha_t \tilde{g}\right) \\ &= (1-\eta)x_t + \frac{\eta}{\|x_t\|^{p-2}}\nabla f(x_t) \\ &\quad + \frac{\eta}{\|x_t\|^{p-2}}\left[\nabla f_{S_t}(x_t) - \nabla f(x_t) - \alpha_t\left(\nabla f_{S_t}(y_0) - \nabla f(y_0)\right)\right]. \end{aligned} \tag{23}$$

Since $\nabla_i f$ is twice continuously differentiable on an open set containing $\partial\mathcal{B}_d$, using the Taylor theorem, we obtain

$$\nabla_i f(y_t) = \nabla_i f(x^*) + \nabla\nabla_i f(x^*)(y_t - x^*) + \frac{1}{2}(y_t - x^*)^T H_i(\hat{y}_t^i)(y_t - x^*) \tag{24}$$

where $\hat{y}_t^i \in \mathcal{N}(y_t, x^*) \triangleq \{z \,|\, z = \mu y_t + (1-\mu)x^*, 0 \le \mu \le 1\}$. Since f is scale invariant with the degree of p, by [3, Proposition 3], we have $c\nabla f(cx) = |c|^p \nabla f(x)$, leading to

$$\frac{\nabla f(x_t)^T z}{\|x_t\|^{p-1}} = \nabla f(x^*)^T z + (y_t - x^*)^T \nabla^2 f(x^*) z + \frac{1}{2}(y_t - x^*)^T \sum_{i=1}^{d} z_i H_i(\hat{y}_t^i)(y_t - x^*) \tag{25}$$

for any vector $z \in \mathbb{R}^d$. For $k = 1$, using $v_1 = x^*$, we have

$$\begin{aligned} &\nabla f(x^*)^T v_1 = \nabla f(x^*)^T x^* = \lambda^*, \\ &(y_t - x^*)^T \nabla^2 f(x^*) v_1 = (y_t - x^*)^T \nabla^2 f(x^*) x^* = \lambda_1 (y_t^T x^* - 1), \end{aligned}$$

which from (25) with $z = v_1$ results in

$$
\begin{aligned}
\frac{\nabla f(x_t)^T v_1}{\|x_t\|^{p-1}} &= \lambda^* - \lambda_1(1 - y_t^T x^*) + \frac{1}{2}(y_t - x^*)^T \sum_{i=1}^d v_{1i} H_i(\hat{y}_t^i)(y_t - x^*) \\
&= \lambda^* y_t^T x^* + (\lambda^* - \lambda_1)(1 - y_t^T x^*) + \frac{1}{2}(y_t - x^*)^T \sum_{i=1}^d v_{1i} H_i(\hat{y}_t^i)(y_t - x^*) \\
&= \lambda^* y_t^T x^* + \frac{1}{2}(y_t - x^*)^T \big[(\lambda^* - \lambda_1) I + \sum_{i=1}^d v_{1i} H_i(\hat{y}_t^i)\big](y_t - x^*) \\
&= \lambda^* y_t^T x^* + \frac{1}{2}(y_t - x^*)^T F_1(\hat{y}_t^1, \cdots, \hat{y}_t^d)(y_t - x^*)).
\end{aligned} \tag{26}
$$

For $2 \le k \le d$, from (25) with $z = v_k$, $(x^*)^T v_k = v_1^T v_k = 0$ and $\nabla f(x^*)^T v_k = \lambda^* v_1^T v_k = 0$, we have

$$
\frac{\nabla f(x_t)^T v_k}{\|x_t\|^{p-1}} = \lambda_k y_t^T x^* + \frac{1}{2}(y_t - x^*)^T F_k(\hat{y}_t^1, \cdots, \hat{y}_t^d)(y_t - x^*). \tag{27}
$$

Since ∇f_l is scale invariant with the degree of $p-1$ for each $l \in [n]$, we have

$$
\nabla f_l(x_t) = \|x_t\|^{p-1} \nabla f_l(y_t), \quad \alpha_t \nabla f_l(y_0) = \|x_t\|^{p-1}(y_t^T y_0)^{p-1} \nabla f_l(y_0),
$$

which leads to

$$
\frac{1}{\|x_t\|^{p-1}} \left(\nabla f_{S_t}(x_t) - \nabla f(x_t) - \alpha_t \left(\nabla f_{S_t}(y_0) - \nabla f(y_0)\right)\right) = \nabla g_{S_t}(y_t) - \nabla g_{S_t}\big((y_t^T y_0) y_0\big).
$$

Using the Taylor approximation of $\nabla_k g_{S_t}$ around $(y_t^T y_0) y_0$, we have

$$
\nabla_k g_{S_t}(y_t) - \nabla_k g_{S_t}((y_t^T y_0) y_0) = \nabla \nabla_k g_{S_t}\left(\bar{y}_t^k\right)^T \left(y_t - (y_t^T y_0) y_0\right)
$$

where $\bar{y}_t^k \in \mathcal{N}(y_t, (y_t^T y_0) y_0)$. This leads to

$$
\frac{1}{\|x_t\|^{p-2}} \left(\nabla f_{S_t}(x_t) - \nabla f(x_t) - \alpha_t \left(\nabla f_{S_t}(y_0) - \nabla f(y_0)\right)\right) = G_{S_t}(\bar{y}_t^1, \cdots, \bar{y}_t^d)\big(x_t - (x_t^T y_0) y_0\big). \tag{28}
$$

Using (23), (26), (27) and (28), we have

$$
\begin{aligned}
x_{t+1}^T v_k = &\left(1 - \eta + \eta(\lambda_k + (\lambda^* - \lambda_1)\mathbb{1}_{k=1})\right) x_t^T v_k + \frac{1}{2}\eta \|x_t\| (y_t - x^*)^T F_k(\hat{y}_t^1, \cdots, \hat{y}_t^d)(y_t - x^*) \\
&+ \eta \left(G_{S_t}(\bar{y}_t^1, \cdots, \bar{y}_t^d)\big(x_t - (x_t^T y_0) y_0\big)\right)^T v_k.
\end{aligned} \tag{29}
$$

$\square$

Proof of Lemma III.2. We prove by induction. Suppose that we have $\Delta_s \le \Delta_0$ for $s \le t < m$. Since $\Delta_0 \le 1 - 1/\sqrt{2}$, this implies that $y_t^T x^* \ge 1/\sqrt{2}$ and $y_0^T x^* \ge 1/\sqrt{2}$. Therefore, we have

$$
\begin{aligned}
y_t^T y_0 &= \left[(y_t^T x^*) x^* + y_t - (y_t^T x^*) x^*\right]^T \left[(y_0^T x^*) x^* + y_0 - (y_0^T x^*) x^*\right] \\
&= (y_t^T x^*)(y_0^T x^*) + (y_t - (y_t^T x^*) x^*)^T (y_0 - (y_0^T x^*) x^*) \\
&\ge (y_t^T x^*)(y_0^T x^*) - \|y_t - (y_t^T x^*) x^*\| \|y_0 - (y_0^T x^*) x^*\| \\
&\ge (y_t^T x^*)(y_0^T x^*) - \sqrt{1 - (y_t^T x^*)^2} \sqrt{1 - (y_0^T x^*)^2} \\
&\ge 0,
\end{aligned}
$$

which leads to

$$
\|x_t - (x_t^T y_0) y_0\|^2 = \|x_t\|^2 (1 - (y_k^T y_0)^2) \le 2\|x_t\|^2 (1 - y_t^T y_0) = \|x_t\|^2 \|y_t - y_0\|^2.
$$

By the triangular inequality, $(a+b)^2 \le 2(a^2 + b^2)$ and $\Delta_t \le \Delta_0$, we have

$$
\|y_t - y_0\|^2 \le 2(\|y_t - x^*\|^2 + \|y_0 - x^*\|^2) \le 4\|y_0 - x^*\|^2.
$$

From $y_0^T x^* \geq 0$, we further obtain

$$\|x_t - (x_t^T y_0) y_0\|^2 \leq 4\|x_t\|^2 \|y_0 - x^*\|^2 = 8\|x_t\|^2 (1 - y_0^T x^*) \tag{30}$$

$$\leq 8\|x_t\|^2 (1 - (y_0^T x^*)^2) = 8\|x_t\|^2 \sum_{k=2}^{d} (y_0^T v_k)^2. \tag{31}$$

Using Lemma III.1, the definitions of M and L, (30) and that $\Delta_t \leq \Delta_0$, we have

$$\begin{aligned} x_{t+1}^T v_1 &\geq (1 - \eta + \eta\lambda^*)\, x_t^T v_1 - \frac{1}{2}\eta M \|x_t\| \|y_t - x^*\|^2 - \eta\sqrt{L}\|x_t - (x_t^T y_0) y_0\| \\ &\geq (1 - \eta + \eta\lambda^*)\, x_t^T v_1 - \eta M (1 - y_t^T x^*)\|x_t\| - \eta\sqrt{8L(1 - y_0^T x^*)}\|x_t\| \\ &\geq \left[1 - \eta + \eta\left(\lambda^* - \frac{M\Delta_0}{1 - \Delta_0} - \frac{\sqrt{8L\Delta_0}}{1 - \Delta_0}\right)\right] y_0^T x^* \|x_t\|. \end{aligned} \tag{32}$$

By (22), (6) and that $L \leq L_0$, we have

$$\lambda^* - \frac{M\Delta_0}{1 - \Delta_0} - \frac{\sqrt{8L\Delta_0}}{1 - \Delta_0} \geq \lambda^* - \big(M + 4\sqrt{L}\big)\sqrt{\Delta_0} \geq \lambda^* - \frac{(\lambda^* - \bar{\lambda})\big(M + 4\sqrt{L}\big)}{2M + 4\sqrt{L_0}} \geq 0.$$

This leads to $x_{t+1}^T v_1 \geq 0$.

Now, we prove that $\Delta_{t+1} \leq \Delta_0$. Since $\{v_1, \ldots, v_d\}$ forms an orthogonal basis, we have $\|x_t\|^2 = \sum_{k=1}^d (x_t^T v_k)^2$. Since

$$\sum_{k=2}^{d} (1 - \eta + \eta\lambda_k)^2 (x_t^T v_k)^2 \leq (1 - \eta + \eta\bar{\lambda})^2 \sum_{k=2}^{d} (x_t^T v_k)^2 \tag{33}$$

$$\sum_{k=1}^{d} \left(1 - \eta + \eta(\lambda_k + (\lambda^* - \lambda_1)\mathbb{1}_{k=1})\right)^2 (x_t^T v_k)^2 \leq (1 - \eta + \eta\lambda^*)^2 \|x_t\|^2 \tag{34}$$

$$\begin{aligned} \sum_{k=2}^{d} \left[(y_t - x^*)^T F_k(\hat{y}_t^1, \cdots, \hat{y}_t^d)(y_t - x^*)\right]^2 &\leq \sum_{k=1}^{d} \left[(y_t - x^*)^T F_k(\hat{y}_t^1, \cdots, \hat{y}_t^d)(y_t - x^*)\right]^2 \\ &\leq M^2 \|y_t - x^*\|^4 \end{aligned} \tag{35}$$

$$\begin{aligned} \sum_{k=2}^{d} \left[v_k^T G_{S_t}(\bar{y}_t^1, \cdots, \bar{y}_t^d)\big(x_t - (x_t^T y_0) y_0\big)\right]^2 &\leq \sum_{k=1}^{d} \left[v_k^T G_{S_t}(\bar{y}_t^1, \cdots, \bar{y}_t^d)\big(x_t - (x_t^T y_0) y_0\big)\right]^2 \\ &\leq L \|x_t - (x_t^T y_0) y_0\|^2 \end{aligned} \tag{36}$$

where (36) follows from $\|\sum_{k=1}^d v_k v_k^T\| = 1$. By Lemma III.1 and the Cauchy-Schwarz inequality, we have

$$\sum_{k=2}^{d} (x_{t+1}^T v_k)^2 \leq \left[(1 - \eta + \eta\bar{\lambda})\sqrt{\sum_{k=2}^{d} (x_t^T v_k)^2} + \frac{1}{2}\eta M \|x_t\| \|y_t - x^*\|^2 + \eta\sqrt{L}\|x_t - (x_t^T y_0) y_0\|\right]^2 \tag{37}$$

$$\|x_{t+1}\|^2 = \sum_{k=1}^{d} (x_{t+1}^T v_k)^2 \leq \left[1 - \eta + \eta\lambda^* + \frac{1}{2}\eta M \|y_0 - x^*\|^2 + \eta\sqrt{L}\|y_t - (y_t^T y_0) y_0\|\right]^2 \|x_t\|^2. \tag{38}$$

First, we consider the case when (7) holds. From $\Delta_t \leq \Delta_0 \leq 1$, we have $0 \leq y_t^T x^* \leq 1$ and $\sum_{k=2}^d (y_t^T v_k)^2 = 1 - (y_t^T x^*)^2 \leq 1 - (y_0^T x^*)^2 = \sum_{k=2}^d (y_0^T v_k)^2$, resulting in

$$\|y_t - x^*\|^2 \leq 2\sqrt{1 - y_t^T x^*}\sqrt{1 - (y_t^T x^*)^2} \leq 2\sqrt{\Delta_t}\sqrt{\sum_{k=2}^{d} (y_0^T v_k)^2} \leq 2\sqrt{\Delta_0}\sqrt{\sum_{k=2}^{d} (y_0^T v_k)^2}. \tag{39}$$

Plugging (31) and (39) into (37), we have

$$\sum_{k=2}^{d} (x_{t+1}^T v_k)^2 \leq \left[1 - \eta + \eta\left(\bar{\lambda} + M\sqrt{\Delta_0} + 2\sqrt{2L}\right)\right]^2 \|x_t\|^2 \sum_{k=2}^{d} (y_0^T v_k)^2. \tag{40}$$

Combining (32) and (40), we have

$$\frac{\sum_{k=2}^d (x_{t+1}^T v_k)^2}{(x_{t+1}^T v_1)^2} \leq \left[\frac{1 - \eta + \eta\big(\bar{\lambda} + M\sqrt{\Delta_0} + 2\sqrt{2L}\big)}{1 - \eta + \eta\left(\lambda^* - M\Delta_0/(1 - \Delta_0) - 2\sqrt{2L\Delta_0}/(1 - \Delta_0)\right)}\right]^2 \frac{\sum_{k=2}^d (y_0^T v_k)^2}{(y_0^T v_1)^2}. \tag{41}$$

Using (22) and (7), we have

$$\lambda^* - \frac{M\Delta_0}{1-\Delta_0} - \frac{2\sqrt{2L\Delta_0}}{1-\Delta_0} - \left(\bar{\lambda} + M\sqrt{\Delta_0} + 2\sqrt{2L}\right) \geq (\lambda^* - \bar{\lambda}) - 2M\sqrt{\Delta_0} - 4\sqrt{2L} \geq 0.$$

Therefore, from (41), we finally have

$$\frac{1-(y_{t+1}^T x^*)^2}{(y_{t+1}^T x^*)^2} = \frac{\sum_{k=2}^d (y_{t+1}^T v_k)^2}{(y_{t+1}^T v_1)^2} = \frac{\sum_{k=2}^d (x_{t+1}^T v_k)^2}{(x_{t+1}^T v_1)^2} \leq \frac{\sum_{k=2}^d (y_0^T v_k)^2}{(y_0^T v_1)^2} = \frac{1-(y_0^T x^*)^2}{(y_0^T x^*)^2},$$

which leads to $\Delta_{t+1} = 1 - y_{t+1}^T x^* \leq 1 - y_0^T x^* = \Delta_0$.

Next, we derive $\Delta_{t+1} \leq \Delta_0$ from (8). From (30) and (38), we have

$$\|x_{t+1}\|^2 \leq \left[1 - \eta + \eta\Big(\lambda^* + \frac{1}{2}M\|y_0 - x^*\|^2 + 2\sqrt{L}\|y_0 - x^*\|\Big)\right]^2 \|x_t\|^2.$$

Using induction, this leads to

$$\|x_{t+1}\|^2 \leq \left[1 - \eta + \eta\Big(\lambda^* + \frac{1}{2}M\|y_0 - x^*\|^2 + 2\sqrt{L}\|y_0 - x^*\|\Big)\right]^{2(t+1)} \|x_0\|^2. \tag{42}$$

On the other hand, from (23), (28), (30) and the definition of L, we have

$$\begin{aligned}
x_{t+1}^T y_0 &= (1-\eta)x_t^T y_0 + \frac{\eta \nabla f(x_t)^T y_0}{\|x_t\|^{p-2}} + \eta y_0^T G_{S_t}(\bar{y}_t^1, \cdots, \bar{y}_t^d)(x_t - (x_t^T y_0) y_0) \\
&\geq (1-\eta)x_t^T y_0 + \frac{\eta \nabla f(x_t)^T y_0}{\|x_t\|^{p-2}} - 2\eta\sqrt{L}\|y_0 - x^*\|\|x_t\|.
\end{aligned}$$

Using $z = y_0$ in (25) and using $\nabla f(x^*) = \lambda^* x^*$ and the definition of M, we have

$$\begin{aligned}
\frac{\nabla f(x_t)^T y_0}{\|x_t\|^{p-1}} &= \nabla f(x^*)^T y_0 + (y_t - x^*)^T \nabla^2 f(x^*) y_0 + \frac{1}{2}(y_t - x^*)^T \sum_{i=1}^d y_{0i} H_i(\hat{y}_t^i)(y_t - x^*) \\
&= \lambda^* y_t^T y_0 + (y_t - x^*)^T \left(\nabla^2 f(x^*) - \lambda^* I\right) y_0 - \frac{1}{2}M\|y_t - x^*\|^2 \\
&\geq \lambda^* y_t^T y_0 - (\lambda^* + \sigma)\|y_t - x^*\| - \frac{1}{2}M\|y_t - x^*\|^2 \\
&\geq \lambda^* y_t^T y_0 - (\lambda^* + \sigma)\|y_0 - x^*\| - \frac{1}{2}M\|y_0 - x^*\|^2
\end{aligned}$$

where the last inequality follows from

$$\|y_t - x^*\|^2 = 2(1 - y_t^T x^*) \leq 2(1 - y_0^T x^*) = \|y_0 - x^*\|^2.$$

This results in

$$\begin{aligned}
x_{t+1}^T y_0 &\geq (1 - \eta + \eta\lambda^*)x_t^T y_0 - \eta\big(\lambda^* + \sigma + \frac{1}{2}M\|y_0 - x^*\| + 2\sqrt{L}\big)\|y_0 - x^*\|\|x_t\| \\
&= (1 - \eta + \eta\lambda^*)x_t^T y_0 - \eta\theta_1\sqrt{2\Delta_0}\|x_t\|.
\end{aligned}$$

Combining with (42), we obtain

$$\begin{aligned}
x_{t+1}^T y_0 &\geq (1 - \eta + \eta\lambda^*)x_t^T y_0 \\
&\quad - \eta\theta_1\sqrt{2\Delta_0}\left[1 - \eta + \eta\Big(\lambda^* + \frac{1}{2}M\|y_0 - x^*\|^2 + 2\sqrt{L}\|y_0 - x^*\|\Big)\right]^t \|x_0\| \\
&\geq (1 - \eta + \eta\lambda^*)x_t^T y_0 - \eta\theta_1\sqrt{2\Delta_0}\left[1 - \eta + \eta\lambda^* + \eta\theta_1\sqrt{2\Delta_0}\right]^t \|x_0\|.
\end{aligned}$$

By recursion, we further have

$$\begin{aligned}
x_{t+1}^T y_0 &\geq (1 - \eta + \eta\lambda^*)^{t+1}\|x_0\| \qquad (43) \\
&\quad - \eta\theta_1\sqrt{2\Delta_0}\sum_{i=1}^{t+1}\left(1 - \eta + \eta\lambda^*\right)^{i-1}\left[1 - \eta + \eta\lambda^* + \eta\theta_1\sqrt{2\Delta_0}\right]^{t+1-i}\|x_0\| \\
&= \left[2(1 - \eta + \eta\lambda^*)^{t+1} - \left(1 - \eta + \eta\lambda^* + \eta\theta_1\sqrt{2\Delta_0}\right)^{t+1}\right]\|x_0\|. \qquad (44)
\end{aligned}$$

Also, by the definition of ν_1 and requirement (8) that $\eta\nu_1 \le 1$ which yields

$$\bar{x} = \frac{\eta\theta_1\sqrt{2\Delta_0}}{1-\eta+\eta\lambda^*} \le \frac{1}{2m},$$

it is easy to establish that

$$(1+x)^t \le \exp(xt) \le 2xt + 1 \tag{45}$$

for any $0 \le x \le 1/2t$. Since $t < m$, by considering $x = \bar{x}$, we obtain from (44) inequality

$$x_{t+1}^T y_0 \ge (1-\eta+\eta\lambda^*)^{t+1}(1-2m\bar{x}) \ge 0.$$

Since $\|x_t - (x_t^T y_0)y_0\|^2 = \|x_t\|^2 - (x_t^T y_0)^2$, using (42), (44) and elementary algebraic manipulations, we have

$$\|x_t - (x_t^T y_0)y_0\|^2 \le 4(1-\eta+\eta\lambda^*)^{2t}\left[\left(1+\frac{\eta\theta_1\sqrt{2\Delta_0}}{1-\eta+\eta\lambda^*}\right)^t - 1\right]\|x_0\|^2.$$

By (8), (9a) and (21), we have $\eta(1-\lambda^*+2\theta_1 m\sqrt{2\Delta_0}) \le 1$ or

$$\frac{\eta\theta_1 m\sqrt{2\Delta_0}}{1-\eta+\eta\lambda^*} \le \frac{1}{2}.$$

Since

$$\frac{\eta\theta_1 t\sqrt{2\Delta_0}}{1-\eta+\eta\lambda^*} \le \frac{\eta\theta_1 m\sqrt{2\Delta_0}}{1-\eta+\eta\lambda^*} \le \frac{1}{2} < 1,$$

using (45), we obtain

$$\|x_t - (x_t^T y_0)y_0\|^2 \le 8\eta\theta_1(1-\eta+\eta\lambda^*)^{2t-1}\sqrt{2\Delta_0}\,t\|x_0\|^2. \tag{46}$$

Plugging (39) and (46) into the square root of (37) and then apply recursion, we have

$$\begin{aligned}
\sqrt{\sum_{k=2}^d (x_{t+1}^T v_k)^2} &\le \left[1-\eta+\eta(\bar{\lambda}+M\sqrt{\Delta_0})\right]\sqrt{\sum_{k=2}^d (x_t^T v_k)^2} \\
&\quad + \eta\sqrt{\frac{8\eta L\theta_1\sqrt{2\Delta_0}}{1-\eta+\eta\lambda^*}}(1-\eta+\eta\lambda^*)^t t\|x_0\| \\
&\le \left[1-\eta+\eta(\bar{\lambda}+M\sqrt{\Delta_0})\right]^{t+1}\sqrt{\sum_{k=2}^d (x_0^T v_k)^2} \\
&\quad + \eta\sqrt{\frac{8\eta L\theta_1\sqrt{2\Delta_0}}{1-\eta+\eta\lambda^*}}\sum_{i=1}^t i\,(1-\eta+\eta\lambda^*)^i\left[1-\eta+\eta(\bar{\lambda}+M\sqrt{\Delta_0})\right]^{t-i}\|x_0\|.
\end{aligned} \tag{47}$$

For a positive integer t and a non-negative real number $r \ge 0$ such that $rt \le 1$, we have

$$(1+r)^t - 1 = r\left((1+r)^{t-1} + (1+r)^{t-2} + \cdots + 1\right) \ge rt$$

and (45) with $x = r$, which results in

$$\begin{aligned}
\sum_{i=1}^t (1+r)^i i &= \frac{1+r}{r^2}\left(t(1+r)^{t+1} - (t+1)(1+r)^t + 1\right) \\
&\le \frac{1+r}{r^2}\left(t(1+r)^{t+1} - t(1+r)^t - rt\right) \\
&= \frac{(1+r)t}{r}\left((1+r)^t - 1\right) \\
&\le 2(1+r)t^2.
\end{aligned} \tag{48}$$

By (8), (9b) and (21), we have

$$\frac{1-\eta+\eta\lambda^*}{1-\eta+\eta(\bar{\lambda}+M\sqrt{\Delta_0})} - 1 \le \frac{1}{m}.$$

Also, by (6), we have $\lambda^* - \bar{\lambda} - M\sqrt{\Delta_0} \ge 0$, leading to

$$\frac{1-\eta+\eta\lambda^*}{1-\eta+\eta(\bar{\lambda}+M\sqrt{\Delta_0})} - 1 = \frac{\eta(\lambda^*-\bar{\lambda}-M\sqrt{\Delta_0})}{1-\eta+\eta(\bar{\lambda}+M\sqrt{\Delta_0})} \ge 0.$$

Therefore, using (48), we have

$$\begin{aligned}
&\sum_{i=1}^{t} i\,(1-\eta+\eta\lambda^*)^i\big[1-\eta+\eta(\bar{\lambda}+M\sqrt{\Delta_0})\big]^{t-i} \\
&= \Big[1-\eta+\eta(\bar{\lambda}+M\sqrt{\Delta_0})\Big]^t \sum_{i=1}^{t} i\,\left[\frac{1-\eta+\eta\lambda^*}{1-\eta+\eta(\bar{\lambda}+M\sqrt{\Delta_0})}\right]^i \\
&\le 2(1-\eta+\eta\lambda^*)t^2\big[1-\eta+\eta(\bar{\lambda}+M\sqrt{\Delta_0})\big]^{t-1}.
\end{aligned} \tag{49}$$

Plugging (49) into (47), we obtain

$$\begin{aligned}
\sqrt{\sum_{k=2}^{d}(x_{t+1}^T v_k)^2} \le \big[1-\eta+\eta(\bar{\lambda}+M\sqrt{\Delta_0})\big]^{t+1}\sqrt{\sum_{k=2}^{d}(x_0^T v_k)^2} \\
+ 2\eta\sqrt{8(1-\eta+\eta\lambda^*)\eta L\theta_1\sqrt{2\Delta_0}}\, t^2\big[1-\eta+\eta(\bar{\lambda}+M\sqrt{\Delta_0})\big]^{t-1}\|x_0\|.
\end{aligned} \tag{50}$$

On the other hand, from (32) and

$$\begin{aligned}
(1-y_t^T v_1)\|x_t\| &= \frac{1-y_t^T v_1}{y_t^T v_1}x_t^T v_1 \le \frac{1-y_0^T v_1}{y_0^T v_1}x_t^T v_1 = \frac{\Delta_0}{1-\Delta_0}x_t^T v_1, \\
\sqrt{1-y_0^T v_1}\|x_t\| &= \frac{\sqrt{1-y_0^T v_1}}{y_t^T v_1}x_t^T v_1 \le \frac{\sqrt{1-y_0^T v_1}}{y_0^T v_1}x_t^T v_1 = \frac{\sqrt{1-\Delta_0}}{1-\Delta_0}x_t^T v_1,
\end{aligned}$$

we have

$$x_{t+1}^T v_1 \ge \left[1-\eta+\eta\left(\lambda^*-\frac{M\Delta_0}{1-\Delta_0}-\frac{2\sqrt{2L\Delta_0}}{1-\Delta_0}\right)\right]^{t+1} x_0^T v_1. \tag{51}$$

Combining (50) and (51), we have

$$\begin{aligned}
\frac{\sqrt{\sum_{k=2}^{d}(x_{t+1}^T v_k)^2}}{x_{t+1}^T v_1} \le \left[\frac{1-\eta+\eta(\bar{\lambda}+M\sqrt{\Delta_0})}{1-\eta+\eta\left[\lambda^*-(M\Delta_0+2\sqrt{2L\Delta_0})/(1-\Delta_0)\right]}\right]^{t+1}\frac{\sqrt{\sum_{k=2}^{d}(x_0^T v_k)^2}}{x_0^T v_1} \\
+\frac{2\eta t^2\sqrt{8(1-\eta+\eta\lambda^*)\eta L\theta_1\sqrt{2\Delta_0}}\,\big[1-\eta+\eta(\bar{\lambda}+M\sqrt{\Delta_0})\big]^{t-1}}{\big(1-\eta+\eta\left[\lambda^*-(M\Delta_0+2\sqrt{2L\Delta_0})/(1-\Delta_0)\right]\big)^{t+1} y_0^T v_1}.
\end{aligned} \tag{52}$$

Since $0<\eta\le 1$ and $\bar{\lambda}<\lambda^*$, we have

$$\frac{\bar{\lambda}}{\lambda^*} \le \frac{1-\eta+\eta\bar{\lambda}}{1-\eta+\eta\lambda^*} \le \frac{1-\eta+\eta(\bar{\lambda}+M\sqrt{\Delta_0})}{1-\eta+\eta\lambda^*}. \tag{53}$$

Let

$$\gamma = \frac{\lambda^*-\bar{\lambda}-M\sqrt{\Delta_0}-(M\Delta_0+2\sqrt{2L\Delta_0})/(1-\Delta_0)}{1-\eta+\eta\left[\lambda^*-(M\Delta_0+2\sqrt{2L\Delta_0})/(1-\Delta_0)\right]}. \tag{54}$$

By (22) and $\theta_2\ge 0$ due to (6), we have

$$\begin{aligned}
\frac{1}{1-\eta+\eta(\bar{\lambda}+M\sqrt{\Delta_0})} &= \frac{\gamma}{1-\eta\gamma}\left[\frac{1}{\lambda^*-\bar{\lambda}-M\sqrt{\Delta_0}-(M\Delta_0+2\sqrt{2L\Delta_0})/(1-\Delta_0)}\right] \\
&\le \frac{\gamma}{\theta_2(1-\eta\gamma)}.
\end{aligned} \tag{55}$$

Using (53), (55) and that $y_0^T v_1 \ge 1/\sqrt{2}$, we have

$$\frac{2\eta t^2\sqrt{8(1-\eta+\eta\lambda^*)\eta L\theta_1\sqrt{2\Delta_0}}}{y_0^T v_1(1-\eta+\eta(\bar{\lambda}+M\sqrt{\Delta_0}))^2} \le 8\sqrt{2}\sqrt{\frac{\lambda^*}{\bar{\lambda}}}\sqrt{\frac{\eta L\theta_1\sqrt{\Delta_0}}{1-\eta+\eta(\bar{\lambda}+M\sqrt{\Delta_0})}}\frac{\eta\gamma t^2}{\theta_2(1-\eta\gamma)}.$$

By (8), (9c) and (21), we have

$$\eta\left(\frac{128L\theta_1\lambda^* m^2}{\theta_2^2\bar{\lambda}\Delta_0\sqrt{\Delta_0}}+1-(\bar{\lambda}+M\sqrt{\Delta_0})\right)\le 1$$

or

$$\frac{\eta L\theta_1\sqrt{\Delta_0}}{1-\eta+\eta\big(\bar{\lambda}+M\sqrt{\Delta_0}\big)} \le \frac{\theta_2^2\bar{\lambda}\Delta_0^2}{128\lambda^* m^2},$$

which results in

$$\frac{2\eta t^2\sqrt{8(1-\eta+\eta\lambda^*)\eta L\theta_1\sqrt{2\Delta_0}}}{y_0^T v_1(1-\eta+\eta(\bar{\lambda}+M\sqrt{\Delta_0}))^2} \le \frac{\eta\gamma t^2\Delta_0}{(1-\eta\gamma)m} \le \frac{\eta\gamma t^2}{(1-\eta\gamma)m}\frac{\sum_{k=2}^d (x_0^T v_k)^2}{(x_0^T v_1)^2}. \tag{56}$$

The last inequality follows from

$$\Delta_0 = 1 - y_0^T x^* \le 1-(y_0^T x^*)^2 \le \frac{\sum_{k=2}^d (y_0^T v_k)^2}{(y_0^T v_1)^2} = \frac{\sum_{k=2}^d (x_0^T v_k)^2}{(x_0^T v_1)^2}.$$

Plugging (54) and (56) into (52), we have

$$\frac{\sqrt{\sum_{k=2}^d (x_{t+1}^T v_k)^2}}{x_{t+1}^T v_1} \le (1-\eta\gamma)^{t+1}\left[1+\frac{\eta\gamma t^2}{(1-\eta\gamma)m}\right]\frac{\sqrt{\sum_{k=2}^d (x_0^T v_k)^2}}{x_0^T v_1}.$$

Using $1+nx \le (1+x)^n$ for $x \ge 0$ and the fact that $\gamma \ge 0$ by (6), we have

$$\begin{aligned}(1-\eta\gamma)^{t+1}\left[1+\frac{\eta\gamma t^2}{(1-\eta\gamma)m}\right] &= 1-\left[\left(1+\frac{\eta\gamma}{1-\eta\gamma}\right)^{t+1} - 1 - \frac{\eta\gamma t^2}{(1-\eta\gamma)m}\right](1-\eta\gamma)^{t+1} \\ &\le 1-\left(t+1-\frac{t^2}{m}\right)\eta\gamma(1-\eta\gamma)^t,\end{aligned}$$

which yields

$$\frac{\sqrt{\sum_{k=2}^d (x_{t+1}^T v_k)^2}}{x_{t+1}^T v_1} \le \frac{\sqrt{\sum_{k=2}^d (x_0^T v_k)^2}}{x_0^T v_1}$$

due to $t < m$. We obtain

$$\frac{1-(y_{t+1}^T x^*)^2}{(y_{t+1}^T x^*)^2} = \frac{\sum_{k=2}^d (x_{t+1}^T v_k)^2}{(x_{t+1}^T v_1)^2} \le \frac{\sum_{k=2}^d (x_0^T v_k)^2}{(x_0^T v_1)^2} = \frac{1-(y_0^T x^*)^2}{(y_0^T x^*)^2}$$

and we finally have $\Delta_{t+1} = 1 - y_{t+1}^T x^* \le 1 - y_0^T x^* = \Delta_0$. □

Proof of Lemma III.3. By Lemma III.1, we have

$$\begin{aligned} x_{t+1}^T v_k = {} & (1-\eta+\eta(\lambda_k+(\lambda^*-\lambda_1)\mathbb{1}_{k=1}))\, x_t^T v_k + \frac{1}{2}\eta\|x_t\|(y_t-x^*)^T F_k(\hat{y}_t^1,\cdots,\hat{y}_t^d)(y_t-x^*) \\ & + \eta\big(G_{S_t}(\bar{y}_t^1,\cdots,\bar{y}_t^d)(x_t-(x_t^T y_0)y_0)\big)^T v_k. \end{aligned}$$

Since S_t is sampled uniformly at random, we have $E[f_{S_t}(y)] = f(y)$ for all $y \in \mathbb{R}^d$, which leads to

$$E[G_{S_t}(\bar{y}_t^1,\cdots,\bar{y}_t^d)] = E[E[G_{S_t}(\bar{y}_t^1,\cdots,\bar{y}_t^d)|\bar{y}_t^1,\bar{y}_t^2,\cdots,\bar{y}_t^d]] = 0.$$

Therefore,

$$\begin{aligned} E[(x_{t+1}^T v_1)^2\,|\,x_t] = {} & \big[(1-\eta+\eta\lambda^*)x_t^T v_1 + \frac{1}{2}\eta\|x_t\|(y_t-x^*)^T F_1(\hat{y}_t^1,\cdots,\hat{y}_t^d)(y_t-x^*)\big]^2 \\ & + \eta^2(x_t-(x_t^T y_0)y_0)^T E[G_{S_t}(\bar{y}_t^1,\cdots,\bar{y}_t^d)^T v_1 v_1^T G_{S_t}(\bar{y}_t^1,\cdots,\bar{y}_t^d)](x_t-(x_t^T y_0)y_0). \end{aligned} \tag{57}$$

In the same way, for $2 \le k \le d$, we have

$$\begin{aligned} E[(x_{t+1}^T v_k)^2\,|\,x_t] = {} & \big[(1-\eta+\eta\lambda_k)x_t^T v_k + \frac{1}{2}\eta\|x_t\|(y_t-x^*)^T F_k(\hat{y}_t^1,\cdots,\hat{y}_t^d)(y_t-x^*)\big]^2 \\ & + \eta^2(x_t-(x_t^T y_0)y_0)^T E[G_{S_t}(\bar{y}_t^1,\cdots,\bar{y}_t^d)^T v_k v_k^T G_{S_t}(\bar{y}_t^1,\cdots,\bar{y}_t^d)](x_t-(x_t^T y_0)y_0). \end{aligned} \tag{58}$$

Using the definition of M and $\|\sum_{k=1}^d v_k v_k^T\| = 1$, we have

$$\eta^2(x_t-(x_t^T y_0)y_0)^T \sum_{k=1}^d E[\|G_{S_t}(\bar{y}_t^1,\cdots,\bar{y}_t^d)^T v_k\|^2](x_t-(x_t^T y_0)y_0) \le \eta^2 K\|x_t-(x_t^T y_0)y_0\|^2. \tag{59}$$

Using (57), (58), (34), (35), (59) and the Cauchy-Schwarz inequality for the cross term as

$$\begin{aligned}\frac{1}{2}\eta\|x_t\|\sum_{k=1}^{K}(1-\eta+\eta(\lambda_k+(\lambda^*-\lambda_1)\mathbb{1}_{k=1}))x_k^T v_k(y_t-x^*)^T F_1(\hat{y}_t^1,\cdots,\hat{y}_t^d)(y_t-x^*)\\ \leq \frac{1}{2}\eta M(1-\eta+\eta\lambda^*)\|x_t\|\|y_k-x^*\|^2,\end{aligned} \tag{60}$$

we have

$$\begin{aligned}E[\|x_{t+1}\|^2|x_t] \leq (1-\eta+\eta\lambda^*)^2\|x_t\|^2 + \frac{1}{2}\eta M(1-\eta+\eta\lambda^*)\|x_t\|\|y_k-x^*\|^2\\ +\frac{1}{4}\eta^2M^2\|x_t\|^2\|y_k-x^*\|^4+\eta^2K\|x_t-(x_t^Ty_0)y_0\|^2.\end{aligned} \tag{61}$$

Using $\|x_t-(x_t^Ty_0)y_0\|^2 \leq \|x_t\|^2$ in (61), we obtain

$$\begin{aligned}E[\|x_{t+1}\|^2|x_t] &\leq \Big[\big(1-\eta+\eta\lambda^*+\frac{1}{2}\eta M\|y_t-x^*\|^2\big)^2+\eta^2K\Big]\|x_t\|^2\\ &=\Big[\big(1-\eta+\eta\lambda^*+\eta M(1-y_t^Tx^*)\big)^2+\eta^2K\Big]\|x_t\|^2,\end{aligned} \tag{62}$$

which establishes the first statement.

In the same way, using (58), (33), (35), (59) and the Cauchy-Schwarz inequality similarly to (60), we have

$$\begin{aligned}E\Big[\sum_{k=2}^{d}(x_{t+1}^Tv_k)^2|x_t\Big] \leq \Bigg[(1-\eta+\eta\bar{\lambda})\sqrt{\sum_{k=2}^{d}(x_t^Tv_k)^2}+\frac{1}{2}\eta M\|x_t\|\|y_t-x^*\|^2\Bigg]^2\\ +\eta^2K\|x_t-(x_t^Ty_0)y_0\|^2.\end{aligned} \tag{63}$$

By Lemma III.2, we have $\Delta_t \leq \Delta_0 \leq 1-1/\sqrt{2}$ and thus $y_t^Tx^* \geq 1/\sqrt{2}$ and $y_0^Tx^* \geq 1/\sqrt{2}$. Since $y_t^Tx^* \geq 0$, using (39), we have

$$\frac{1}{2}\eta M\|x_t\|\|y_t-x^*\|^2 \leq \eta M\sqrt{\Delta_t}\sqrt{\sum_{k=2}^{d}(x_t^Tv_k)^2}.$$

As a result of (31) which we can use since $\Delta_t \leq \Delta_0$, we obtain

$$E\Big[\sum_{k=2}^{d}(x_{t+1}^Tv_k)^2|x_t\Big] \leq \Big[1-\eta+\eta\bar{\lambda}+\eta M\sqrt{\Delta_t}\Big]^2\sum_{k=2}^{d}(x_t^Tv_k)^2+8\eta^2K\|x_t\|^2\sum_{k=2}^{d}(y_0^Tv_k)^2, \tag{64}$$

which shows the second statement in the lemma.

Lastly, from (57), we have

$$E[(x_{t+1}^Tv_1)^2|x_t] \geq \Big[(1-\eta+\eta\lambda^*)x_t^Tv_1+\frac{1}{2}\eta\|x_t\|(y_t-x^*)^TF_1(\hat{y}_t^1,\cdots,\hat{y}_t^d)(y_t-x^*)\Big]^2$$

By (11) and (21), we have $\eta(1-\lambda^*+M\Delta_0\sqrt{2}) \leq 1$. Since $1/(1-\Delta_0) \leq \sqrt{2}$ by (6), we further have

$$\eta\left(\frac{M\Delta_0}{1-\Delta_0}+1-\lambda^*\right) \leq 1.$$

Due to $\Delta_t \leq \Delta_0$, this implies that

$$\begin{aligned}(1-\eta+\eta\lambda^*)x_t^Tv_1-\frac{1}{2}\eta M\|x_t\|\|y_t-x^*\|^2 &= \Big[(1-\eta+\eta\lambda^*)(1-\Delta_t)-\eta M\Delta_t\Big]\|x_t\|\\ &=\Big[1-\eta\Big(\frac{M\Delta_t}{1-\Delta_t}+1-\lambda^*\Big)\Big](1-\Delta_t)\|x_t\|\\ &\geq \Big[1-\eta\Big(\frac{M\Delta_0}{1-\Delta_0}+1-\lambda^*\Big)\Big](1-\Delta_t)\|x_t\|\\ &\geq 0.\end{aligned}$$

Since $(a+b)^2 \geq (a-c)^2$ holds if $a \geq c$ and $|b| \leq c$, we finally have

$$\begin{aligned}
E[(x_{t+1}^T v_1)^2 | x_t] &\geq \left[(1-\eta+\eta\lambda^*)x_t^T v_1 - \frac{1}{2}\eta M \|x_t\| \|y_t - x^*\|^2\right]^2 \\
&= \left[1-\eta+\eta\lambda^* - \eta M \left(\frac{1-y_t^T x^*}{y_t^T x^*}\right)\right]^2 (x_t^T v_1)^2 \\
&= \left[\alpha(\eta) - \frac{\eta M \Delta_t}{1-\Delta_t}\right]^2 (x_t^T v_1)^2.
\end{aligned}$$

□

Proof of Lemma III.4. By Lemma III.2, we have $\Delta_t \leq \Delta_0$. Repeatedly applying Lemma III.3, we have

$$\begin{aligned}
E[\|x_t\|^2 | x_0] = E[E[\|x_t\|^2 | x_{t-1}] | x_0] &\leq \left[\big(\alpha(\eta) + \eta M \Delta_0\big)^2 + \eta^2 K\right] E[\|x_{t-1}\|^2 | x_0] \\
&\leq \left[\big(\alpha(\eta) + \eta M \Delta_0\big)^2 + \eta^2 K\right]^t \|x_0\|^2.
\end{aligned} \tag{65}$$

Using (65), we have

$$\begin{aligned}
E\Big[\|x_t\|^2 \sum_{k=2}^d (y_0^T v_k)^2\Big] &= E\Big[E\Big[\|x_t\|^2 \sum_{k=2}^d (y_0^T v_k)^2 | x_0\Big]\Big] = E\Big[E\big[\|x_t\|^2 | x_0\big] \sum_{k=2}^d (y_0^T v_k)^2\Big] \\
&= E\Big[\big[(\alpha(\eta) + \eta M \Delta_0)^2 + \eta^2 K\big]^t \|x_0\|^2 \sum_{k=2}^d (y_0^T v_k)^2\Big] \\
&= \big[(\alpha(\eta) + \eta M \Delta_0)^2 + \eta^2 K\big]^t E\Big[\sum_{k=2}^d (x_0^T v_k)^2\Big].
\end{aligned} \tag{66}$$

Using Lemma III.3 and that $\Delta_t \leq \Delta_0$, we have

$$E\Big[\sum_{k=2}^d (x_t^T v_k)^2\Big] \leq \big(\beta(\eta) + \eta M \sqrt{\Delta_0}\big)^2 E\Big[\sum_{k=2}^d (x_{t-1}^T v_k)^2\Big] + 8\eta^2 K E\Big[\|x_{t-1}\|^2 \sum_{k=2}^d (y_0^T v_k)^2\Big]. \tag{67}$$

By induction on (67) using (66), we have

$$\begin{aligned}
E\Big[\sum_{k=2}^d (x_t^T v_k)^2\Big] &\leq \big(\beta(\eta) + \eta M \sqrt{\Delta_0}\big)^2 E\Big[\sum_{k=2}^d (x_{t-1}^T v_k)^2\Big] \\
&\qquad + 8\eta^2 K \big[(\alpha(\eta) + \eta M \Delta_0)^2 + \eta^2 K\big]^{t-1} E\Big[\sum_{k=2}^d (x_0^T v_k)^2\Big] \\
&\leq E\Big[\sum_{k=2}^d (x_0^T v_k)^2\Big] \Big[\big(\beta(\eta) + \eta M \sqrt{\Delta_0}\big)^{2t} \\
&\qquad + 8\eta^2 K \sum_{s=1}^t \big(\alpha(\eta) + \eta M \sqrt{\Delta_0}\big)^{2(t-s)} \big[(\alpha(\eta) + \eta M \sqrt{\Delta_0})^2 + \eta^2 K\big]^{s-1}\Big] \\
&\leq E\Big[\sum_{k=2}^d (x_0^T v_k)^2\Big] \Bigg[\big(\beta(\eta) + \eta M \sqrt{\Delta_0}\big)^{2t} \\
&\qquad + 8\big(\alpha(\eta) + \eta M \sqrt{\Delta_0}\big)^{2t} \left[\left(1 + \frac{\eta^2 K}{(\alpha(\eta) + \eta M \sqrt{\Delta_0})^2}\right)^t - 1\right]\Bigg] \\
&\leq E\Big[\sum_{k=2}^d (x_0^T v_k)^2\Big] \Bigg[\big(\beta(\eta) + \eta M \sqrt{\Delta_0}\big)^{2t} \\
&\qquad + 8\big(\alpha(\eta) + \eta M \sqrt{\Delta_0}\big)^{2t} \left[\exp\left(\frac{\eta^2 K t}{(\alpha(\eta) + \eta M \sqrt{\Delta_0})^2}\right) - 1\right]\Bigg].
\end{aligned}$$

By (12) and (21), we have $\eta\big(1 - \lambda^* - M\sqrt{\Delta_0} + \sqrt{Km}\big) \leq 1$, which leads to

$$0 \leq \frac{\eta^2 K t}{\big(\alpha(\eta) + \eta M \sqrt{\Delta_0}\big)^2} \leq 1.$$

Using $\exp(x)-1\le 2x$ for $x\in[0,1]$, we have

$$E\Big[\sum_{k=2}^{d}(x_t^Tv_k)^2\Big]\le E\Big[\sum_{k=2}^{d}(x_0^Tv_k)^2\Big]\left[\big(\beta(\eta)+\eta M\sqrt{\Delta_0}\big)^{2t}+16\eta^2Kt\big(\alpha(\eta)+\eta M\sqrt{\Delta_0}\big)^{2(t-1)}\right].$$

On the other hand, using $\Delta_t\le\Delta_0$ and Lemma III.3, we have

$$E[(x_t^Tv_1)^2]=E[E[(x_t^Tv_1)^2|x_{t-1}]]\ge\left[\alpha(\eta)-\frac{\eta M\Delta_0}{1-\Delta_0}\right]^2E[(x_{t-1}^Tv_1)^2]. \tag{68}$$

By induction on (68) using $\Delta_t\le\Delta_0$, we finally have

$$E[(x_t^Tv_1)^2]\ge\left[\alpha(\eta)-\frac{\eta M\Delta_0}{1-\Delta_0}\right]^{2t}E[(x_0^Tv_1)^2].$$

□

Proof of Lemma III.5. By (13) and (14a), we have (12). Also, (13), (14b) and the fact that $\sqrt{2\Delta_0}\le 1$ which holds from (6) imply (11). Therefore, by Lemma III.4, we have

$$\delta_m\le\left[\left(\frac{\beta(\eta)+\eta M\sqrt{\Delta_0}}{\alpha(\eta)-\eta M\Delta_0/(1-\Delta_0)}\right)^{2m}+\frac{16\eta^2Km\big[\alpha(\eta)+\eta M\sqrt{\Delta_0}\big]^{2(m-1)}}{[\alpha(\eta)-\eta M\Delta_0/(1-\Delta_0)]^{2m}}\right]\delta_0 \tag{69}$$

where

$$\delta_t=\frac{E[\sum_{k=2}^{d}(x_t^Tv_k)^2]}{E[(x_t^Tv_1)^2]}.$$

By (22) which follows from (6) and the fact that $(1+x)^m\le\exp(mx)$ for all $x\in\mathbb{R}$, we have

$$\begin{aligned}\left(\frac{\beta(\eta)+\eta M\sqrt{\Delta_0}}{\alpha(\eta)-\eta M\Delta_0/(1-\Delta_0)}\right)^{2m}&\le\left(1-\frac{\eta(\lambda^*-\bar\lambda-2M\sqrt{\Delta_0})}{1-\eta+\eta(\lambda^*-M\sqrt{\Delta_0})}\right)^{2m}\\&\le\exp\left(-\frac{2\eta m(\lambda^*-\bar\lambda-2M\sqrt{\Delta_0})}{1-\eta+\eta(\lambda^*-M\sqrt{\Delta_0})}\right).\end{aligned}$$

Since (13), (14b) and (21) imply

$$\frac{2\eta m(\lambda^*-\bar\lambda-2M\sqrt{\Delta_0})}{1-\eta+\eta(\lambda^*-M\sqrt{\Delta_0})}\le 1,$$

using the fact that $\exp(-x)\le 1-x/2$ for $0\le x\le 1$, we have

$$\exp\left(-\frac{2\eta m(\lambda^*-\bar\lambda-2M\sqrt{\Delta_0})}{1-\eta+\eta(\lambda^*-M\sqrt{\Delta_0})}\right)\le 1-\frac{\eta m(\lambda^*-\bar\lambda-2M\sqrt{\Delta_0})}{1-\eta+\eta(\lambda^*-M\sqrt{\Delta_0})}=1-2\rho. \tag{70}$$

On the other hand, by (22) and the fact that $(1+x)^n\le\exp(nx)$, we have

$$\begin{aligned}\frac{16\eta^2Km\big[\alpha(\eta)+\eta M\sqrt{\Delta_0}\big]^{2(m-1)}}{[\alpha(\eta)-\eta M\Delta_0/(1-\Delta_0)]^{2m}}&\le\frac{16\eta^2Km}{(\alpha(\eta)+\eta M\sqrt{\Delta_0})^2}\left(1+\frac{2\eta M\sqrt{\Delta_0}}{\alpha(\eta)-\eta M\sqrt{\Delta_0}}\right)^{2m}\\&\le\frac{16\eta^2Km}{(\alpha(\eta)+\eta M\sqrt{\Delta_0})^2}\exp\left(\frac{4\eta mM\sqrt{\Delta_0}}{\alpha(\eta)-\eta M\sqrt{\Delta_0}}\right).\end{aligned} \tag{71}$$

By (13), (14a) and (21), we have

$$\eta\left(1-\lambda^*-M\sqrt{\Delta_0}+\frac{64K}{\lambda^*-\bar\lambda-2M\sqrt{\Delta_0}}\right)\le 1,$$

which leads to

$$\frac{\rho}{2}-\frac{16\eta^2Km}{(\alpha(\eta)+\eta M\sqrt{\Delta_0})^2}\ge\frac{\eta m(\lambda^*-\bar\lambda-2M\sqrt{\Delta_0})}{4\big(1-\eta+\eta(\lambda^*+M\sqrt{\Delta_0})\big)}-\frac{16\eta^2Km}{(1-\eta+\eta(\lambda^*+M\sqrt{\Delta_0}))^2}\ge 0. \tag{72}$$

In a similar way, by (13), (14b) and (21), we have

$$\eta\left(1-\lambda^*+M\sqrt{\Delta_0}+\frac{4mM\sqrt{\Delta_0}}{\log 2}\right)\le 1,$$

which results in

$$\exp\left(\frac{4\eta m M\sqrt{\Delta_0}}{\alpha(\eta)-\eta M\sqrt{\Delta_0}}\right) \leq 2. \tag{73}$$

Using (70), (71), (72) and (73) in (69), we finally have

$$\frac{E[\sum_{k=2}^{d}(x_m^T v_k)^2]}{E[(x_m^T v_1)^2]} \leq (1-\rho)\cdot\frac{E[\sum_{k=2}^{d}(x_0^T v_k)^2]}{E[(x_0^T v_1)^2]}.$$

$\square$

Proof of Theorem III.6. Since η, s and $x_0 = \tilde{x}_0$ satisfy (6), (7) (or (8)) and (13), by Lemmas III.2 and III.5, we have

$$\tilde{\Delta}_1 = \Delta_m \leq \Delta_0 = \tilde{\Delta}_0, \quad \tilde{\delta}_1 = \delta_m \leq (1-\rho)\delta_0 = (1-\rho)\tilde{\delta}_0.$$

By repeatedly applying the same argument, we have $\tilde{\delta}_\tau \leq (1-\rho)^\tau\tilde{\delta}_0$. Since $\tau \geq (1/\rho)\log(\tilde{\delta}_0/\epsilon)$, we finally obtain

$$\tilde{\delta}_\tau \leq (1-\rho)^\tau\tilde{\delta}_0 \leq \exp(-\tau\rho)\tilde{\delta}_0 \leq \epsilon.$$

This completes the proof. $\square$