A. About the polar retraction

Given the polar decomposition of $x + \xi = QH$, where $Q \in \mathbb{R}^{d\times r}$ is orthogonal and $H \in \mathbb{R}^{r\times r}$ is positive definite. The polar retraction is the polar factor

$$\mathcal{R}_x(\xi) = Q = (x+\xi)(I_r + \xi^\top \xi)^{-1/2}, \tag{A.1}$$

which is also the orthogonal projection of $x + \xi$ onto $\mathrm{St}(d, r)$. The computation complexity is $\mathcal{O}(dr^2)$. (Liu et al., 2019, Append. E) showed that if $\|\xi\|_{\mathrm{F}} \le 1$ then $M = 1$ for polar retraction. The boundedness of ξ can be verified in our convergence analysis. Therefore, we have $M = 1$ in this paper.

B. More details on linear rate of consensus

The following results were provided in (Chen et al., 2021).

If there exists an integer $t \ge 0$ such that

$$\max_{i\in[n]} \| \sum_{j=1}^{n} (W_{ij}^t - 1/n)(x_j - \bar{x}) \|_{\mathrm{F}} \le \max_{i\in[n]} \sum_{j=1}^{n} |W_{ij}^t - 1/n| \|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F},\infty} \le \frac{1}{2} \|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F},\infty}, \tag{B.1}$$

then it suffices to show the sequence $\{\mathbf{x}_k\}$ of DRCS satisfying $\mathbf{x}_k \in \mathcal{N}$ with $t \ge \lceil \log_{\sigma_2}(\frac{1}{2\sqrt{n}}) \rceil$ steps of communication.

Denote the smallest eigenvalue of W^t by $\lambda_n(W^t)$, the constant L_t is given by

$$L_t = 1 - \lambda_n(W^t). \tag{B.2}$$

It is the Lipschitz constant of $\nabla \varphi^t(\mathbf{x})$. Since $L_t \in (0, 2]$, if $\lambda_n(W^t)$ is unknown, one can use $L_t = 2$. Define the second largest eigenvalue of W^t by $\lambda_2(W^t)$ and

$$\mu_t = 1 - \lambda_2(W^t).$$

The formal statement of Fact 3.1 is given as follows.

Fact B.1. *(Chen et al., 2021) Under Assumption 1, let the stepsize α satisfy $0 < \alpha \le \bar{\alpha} := \min\{\nu \frac{\Phi}{L_t}, 1, \frac{1}{M}\}$ and $t \ge \lceil \log_{\sigma_2}(\frac{1}{2\sqrt{n}}) \rceil$, where $\nu \in [0, 1]$, $\Phi = 2 - \delta_2^2$ and M is given in Lemma 2.3. The sequence $\{\mathbf{x}_k\}$ of (3.2) achieves consensus linearly if the initialization satisfies $\mathbf{x}_0 \in \mathcal{N}$ defined by (3.6). That is, we have $\mathbf{x}_k \in \mathcal{N}$ for all $k \ge 0$ and*

$$\begin{aligned} \|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|_F &\le \|\mathbf{x}_k - \alpha \mathrm{grad}\varphi^t(\mathbf{x}_k) - \bar{\mathbf{x}}_k\|_F \\ &\le \sqrt{1 - 2(1-\nu)\alpha\gamma_t} \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_F, \end{aligned} \tag{B.3}$$

where $\gamma_t = (1 - 4r\delta_1^2)(1 - \frac{\delta_2^2}{2})\mu_t \ge \frac{\mu_t}{2} \ge \frac{1-\sigma_2^t}{2}$.

If $\nu = 1/2$, we have $\alpha \le \bar{\alpha} := \min\{\frac{\Phi}{2L_t}, 1, 1/M\}$ and

$$\rho_t = \sqrt{1 - \gamma_t \alpha}.$$

Recall that M is the constant given in Lemma 2.3. We also have $M = \mathcal{O}(1)$ which is discussed in appendix A. If $\alpha = 1$ is admissible, then the rate is $\rho_t = \sqrt{\frac{1+\sigma_2^t}{2}}$ which is worse that the Euclidean rate σ_2^t. Moreover, it was shown in (Chen et al., 2021) in a smaller region, i.e., $\varphi^t(\mathbf{x}) = \mathcal{O}(\sigma_2^t)$ and $\|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}}^2 = \mathcal{O}(1)$, it follows asymptotically $\rho_t = \sigma_2^t$ with $\alpha = 1$. For simplicity, we will only discuss the convergence of our proposed algorithms using (B.3) with $\nu = 1/2$. Note that this may imply $\bar{\alpha} < 1$, but we find that $\alpha = 1$ always works for our proposed algorithms.

C. Proofs for Section 2

Denote $\mathcal{P}_{N_x\mathcal{M}}$ as the orthogonal projection onto the normal space $N_x\mathcal{M}$. One can rewrite the projection $\mathcal{P}_{\mathrm{T}_x\mathcal{M}}(y - x), \forall y \in \mathrm{St}(d, r)$ (Chen et al., 2021) as follows

$$\begin{aligned} \mathcal{P}_{\mathrm{T}_x\mathcal{M}}(y - x) &= y - x - \mathcal{P}_{N_x\mathcal{M}}(y - x) \\ &= y - x + \frac{1}{2} x (x - y)^\top (x - y). \end{aligned} \tag{P2}$$

This implies that

$$\mathcal{P}_{\mathrm{T}_x\mathcal{M}}(y-x) = y - x + \mathcal{O}(\|y-x\|_{\mathrm{F}}^2).$$

The relationship (P2) helps us to prove Lemma 2.4.

Proof of Lemma 2.4. Firstly, since $\nabla f(x)$ is $L-$Lipschitz in Euclidean space, one has

$$|f(y) - [f(x) + \langle \nabla f(x), y - x\rangle]| \leq \frac{L}{2}\|y-x\|_{\mathrm{F}}^2. \tag{C.1}$$

Since $\mathrm{grad} f(x) = \mathcal{P}_{\mathrm{T}_x\mathcal{M}}\nabla f(x)$, we have

$$\begin{aligned}\langle \mathrm{grad} f(x), y-x\rangle &= \langle \nabla f(x), \mathcal{P}_{\mathrm{T}_x\mathcal{M}}(y-x)\rangle \\ &\overset{\text{(P2)}}{=} \langle \nabla f(x), y-x\rangle + \left\langle \nabla f(x), \frac{1}{2}x(y-x)^\top(y-x)\right\rangle.\end{aligned}$$

Using

$$\left\langle \nabla f(x), \frac{1}{2}x(y-x)^\top(y-x)\right\rangle \leq \|\nabla f(x)\|_2 \cdot \|x\|_2 \cdot \frac{1}{2}\|x-y\|_{\mathrm{F}}^2 \leq \frac{1}{2}\|\nabla f(x)\|_2 \cdot \|x-y\|_{\mathrm{F}}^2$$

implies

$$|\langle \mathrm{grad} f(x), y-x\rangle - \langle \nabla f(x), y-x\rangle| \leq \frac{1}{2}\max_{x\in\mathrm{St}(d,r)}\|\nabla f(x)\|_2 \cdot \|y-x\|_{\mathrm{F}}^2, \tag{C.2}$$

where $\|\nabla f(x)\|_2$ represents the operator norm of $\nabla f(x)$. Since $\mathrm{St}(d,r)$ is a compact set and $\nabla f(x)$ is continuous, we denote $L_n = \max_{x\in\mathrm{St}(d,r)}\|\nabla f(x)\|_2$. Let $L_g = L_n + L$. Combining (C.1) with (C.2) yields

$$|f(y) - [f(x) + \langle \mathrm{grad} f(x), y-x\rangle]| \leq \frac{L_g}{2}\|y-x\|_{\mathrm{F}}^2. \tag{C.3}$$

Secondly, using $\mathrm{grad} f(x) = \nabla f(x) - \mathcal{P}_{N_x\mathcal{M}}\nabla f(x)$ and $\mathrm{grad} f(y) = \nabla f(y) - \mathcal{P}_{N_y\mathcal{M}}\nabla f(y)$ implies

$$\begin{aligned}&\|\mathrm{grad} f(x) - \mathrm{grad} f(y)\|_{\mathrm{F}} \\ \leq\ & \|\nabla f(x) - \nabla f(y)\|_{\mathrm{F}} + \|\mathcal{P}_{N_x\mathcal{M}}\nabla f(y) - \mathcal{P}_{N_y\mathcal{M}}\nabla f(y)\|_{\mathrm{F}} \\ =\ & \|\nabla f(x) - \nabla f(y)\|_{\mathrm{F}} + \frac{1}{2}\|x(x^\top\nabla f(y) + \nabla f(y)^\top x) - y(y^\top \nabla f(y) + \nabla f(y)^\top y)\|_{\mathrm{F}} \\ \leq\ & \|\nabla f(x) - \nabla f(y)\|_{\mathrm{F}} + 2L_n\|x-y\|_{\mathrm{F}} \\ \leq\ & (L+2L_n)\|x-y\|_{\mathrm{F}}.\end{aligned} \tag{C.4}$$

In (C.4) we used

$$\begin{aligned}&\|x(x^\top\nabla f(y) + \nabla f(y)^\top x) - y(y^\top\nabla f(y) + \nabla f(y)^\top y)\|_{\mathrm{F}} \\ \leq\ & \|x((x-y)^\top\nabla f(y) + \nabla f(y)^\top(x-y))\|_{\mathrm{F}} + \|(x-y)(y^\top\nabla f(y) + \nabla f(y)^\top y)\|_{\mathrm{F}} \\ \leq\ & 4L_n\|x-y\|_{\mathrm{F}}.\end{aligned}$$

The proof is completed. □

C.1. Comparison on different Lipschitz-type inequalities

Using Taylor's Theorem(Absil et al., 2009, Lemma 7.4.7), L_g' corresponds to the leading eigenvalue of Riemannian Hessian. According to (Absil et al., 2013), it follows for any $\eta \in \mathrm{T}_x\mathcal{M}$ that

$$\begin{aligned}\mathrm{Hess} f(x)[\eta] &= \mathcal{P}_{T_x\mathcal{M}}\left(D\mathrm{grad} h(x)[\eta]\right) \\ &= \mathcal{P}_{T_x\mathcal{M}}\nabla^2 f(x)\eta - \eta x^\top \mathcal{P}_{N_x}\nabla f(x) - x\frac{1}{2}\left(\eta^\top\mathcal{P}_{N_x}\nabla f(x_i) + (\mathcal{P}_{N_x}\nabla f(x))^\top\eta\right),\end{aligned} \tag{C.5}$$

where $\mathcal{P}_{N_x}$ is the orthogonal projection onto the normal space $N_x\mathcal{M}$. Since $x\frac{1}{2}\left(\eta^\top \mathcal{P}_{N_x}\nabla f(x_i) + (\mathcal{P}_{N_x}\nabla f(x))^\top \eta\right) \in N_x\mathcal{M}$, we have

$$\langle \eta, \mathrm{Hess} f(x)[\eta]\rangle = \langle \eta, \nabla^2 f(x)\eta\rangle - \langle \eta, \eta x^\top \mathcal{P}_{N_x}\nabla f(x)\rangle = \langle \eta, \nabla^2 f(x)\eta\rangle - \left\langle \eta, \eta\frac{1}{2}(x^\top \nabla f(x) + \nabla f(x)^\top x)\right\rangle, \tag{C.6}$$

where we use $\mathcal{P}_{T_x\mathcal{M}}\nabla^2 f(x)\eta = \nabla^2 f(x)\eta - \mathcal{P}_{N_x}\nabla^2 f(x)\eta$. Therefore, we get

$$L_g' \le \lambda_{\max}(\nabla^2 f(x)) + \max_{x\in \mathrm{St}(d,r)} \|\nabla f(x)\|_2 = L + L_n. \tag{C.7}$$

The restricted inequality proposed in (Boumal et al., 2019) is related to the pull back function $g(\xi) := f(\mathcal{R}_x(\xi))$, whose Lipschitz constant $\tilde{L}_g$ relies on the retraction. Specifically, $\tilde{L}_g = M_0^2 L + 2ML_n$, where M_0 is a constant related to the retraction, M and L_n are the same constants in Lemma 2.3.

C.2. Technical lemmas

Lemma C.1. *(Chen et al., 2021) For any $\mathbf{x} \in \mathrm{St}(d,r)^n$, let $\hat{x} = \frac{1}{n}\sum_{i=1}^n x_i$ be the Euclidean mean and denote $\hat{\mathbf{x}} = \mathbf{1}_n \otimes \hat{x}$. Similarly, let $\bar{\mathbf{x}} = \mathbf{1}_n \otimes \bar{x}$, where $\bar{x}$ is the IAM defined in (IAM). Moreover, if $\|\mathbf{x} - \bar{\mathbf{x}}\|_F^2 \le n/2$, one has*

$$\|\bar{x} - \hat{x}\|_F \le \frac{2\sqrt{r}\|\mathbf{x} - \bar{\mathbf{x}}\|_F^2}{n}. \tag{P1}$$

The following lemma will be useful to bound the Euclidean distance between two average points $\bar{x}_k$ and $\bar{x}_{k+1}$.

Lemma C.2. *(Chen et al., 2021) Suppose $\mathbf{x}, \mathbf{y} \in \mathcal{N}_1$, where $\mathcal{N}_1$ is defined in (3.4). Then we have*

$$\|\bar{x} - \bar{y}\|_F \le \frac{1}{1 - 2\delta_1^2}\|\hat{x} - \hat{y}\|_F,$$

where $\bar{x}$ and $\bar{y}$ are the IAM of $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, respectively.

We also need the following bounds for $\mathrm{grad}\varphi^t(\mathbf{x})$.

Lemma C.3. *(Chen et al., 2021) For any $\mathbf{x} \in \mathrm{St}(d,r)^n$, it follows that*

$$\|\sum_{i=1}^n \mathrm{grad}\varphi^t(x_i)\|_F \le L_t\|\mathbf{x} - \bar{\mathbf{x}}\|_F^2 \tag{C.8}$$

and

$$\|\mathrm{grad}\varphi^t(\mathbf{x})\|_F \le L_t\|\mathbf{x} - \bar{\mathbf{x}}\|_F, \tag{C.9}$$

where L_t is the constant given in (B.2). Moreover, suppose $\mathbf{x} \in \mathcal{N}_2$, where $\mathcal{N}_2$ is defined by (3.5). We then have

$$\max_{i\in[n]} \|\mathrm{grad}\varphi_i^t(\mathbf{x})\|_F \le 2\delta_2. \tag{C.10}$$

Applying Lemma C.2 to the update rule of our algorithms gives the following lemma.

Lemma C.4. *If $\mathbf{x}_k \in \mathcal{N}, \mathbf{x}_{k+1} \in \mathcal{N}_1$ and $x_{i,k+1} = \mathcal{R}_{x_{i,k}}(-\alpha \mathrm{grad}\varphi_i^t(\mathbf{x}_k) + \beta u_{i,k})$, where $u_{i,k} \in \mathrm{T}_{x_{i,k}}\mathcal{M}$, $0 \le \alpha \le \frac{1}{M}$, $0 \le \beta$. Let $\mathbf{u}_k^\top = (u_{1,k}^\top \ \ldots \ u_{n,k}^\top)$ and $\hat{u}_k = \frac{1}{n}\sum_{i=1}^n u_{i,k}$. It follows that*

$$\|\bar{x}_k - \bar{x}_{k+1}\|_F \le \frac{1}{1 - 2\delta_1^2}\left(\frac{2L_t^2\alpha + L_t\alpha}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_F^2 + \beta\|\hat{u}_k\|_F + \frac{2M\beta^2}{n}\|\mathbf{u}_k\|_F^2\right).$$

Proof. From Lemma 2.3 and Lemma C.3, we have

$$\begin{aligned}
&\|\hat{x}_k - \hat{x}_{k+1}\|_{\mathrm{F}} \\
\leq\ & \|\hat{x}_k + \frac{1}{n}\sum_{i=1}^n(-\alpha \mathrm{grad}\varphi_i^t(\mathbf{x}_k) + \beta u_{i,k}) - \hat{x}_{k+1}\|_{\mathrm{F}} + \|\frac{1}{n}\sum_{i=1}^n(-\alpha \mathrm{grad}\varphi_i^t(\mathbf{x}_k) + \beta u_{i,k})\|_{\mathrm{F}} \\
\stackrel{(P1)}{\leq}\ & \frac{M}{n}\sum_{i=1}^n \|\alpha \mathrm{grad}\varphi_i^t(\mathbf{x}_k) + \beta u_{i,k}\|_{\mathrm{F}}^2 + \alpha\|\frac{1}{n}\sum_{i=1}^n \mathrm{grad}\varphi_i^t(\mathbf{x}_k)\|_{\mathrm{F}} + \beta\|\hat{u}_k\|_{\mathrm{F}} \\
\leq\ & \frac{2M\alpha^2}{n}\|\mathrm{grad}\varphi^t(\mathbf{x}_k)\|_{\mathrm{F}}^2 + \frac{2M\beta^2}{n}\|\mathbf{u}_k\|_{\mathrm{F}}^2 + \alpha\|\frac{1}{n}\sum_{i=1}^n \mathrm{grad}\varphi_i^t(\mathbf{x}_k)\|_{\mathrm{F}} + \beta\|\hat{u}_k\|_{\mathrm{F}} \\
\stackrel{(C.9)(C.8)}{\leq}\ & \frac{2L_t^2M\alpha^2 + L_t\alpha}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \frac{2M\beta^2}{n}\|\mathbf{u}_k\|_{\mathrm{F}}^2 + \beta\|\hat{u}_k\|_{\mathrm{F}}.
\end{aligned}$$

Therefore, it follows from Lemma C.2 that

$$\|\bar{x}_k - \bar{x}_{k+1}\|_{\mathrm{F}} \leq \frac{1}{1-2\delta_1^2}\|\hat{x}_k - \hat{x}_{k+1}\|_{\mathrm{F}} \leq \frac{1}{1-2\delta_1^2}\left(\frac{2L_t^2\alpha + L_t\alpha}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \beta\|\hat{u}_k\|_{\mathrm{F}} + \frac{2M\beta^2}{n}\|\mathbf{u}_k\|_{\mathrm{F}}^2\right),$$

where we use the fact that $\alpha \leq \frac{1}{M}$. □

D. Proofs for Section 4

We use the notations

$$\mathbf{v}_k = [v_{1,k}^\top\ \ldots\ v_{n,k}^\top]^\top, \quad \hat{v}_k = \frac{1}{n}\sum_{i=1}^n v_{i,k},$$

$$g_{i,k} = \mathrm{grad}f_i(x_{i,k}) \quad \textit{and} \quad \hat{g}_k = \frac{1}{n}\sum_{i=1}^n g_{i,k}.$$

The following lemma is useful to show $\mathbf{x}_k \in \mathcal{N}$ for all k.

Lemma D.1. *(Chen et al., 2021, Lemma 11) Given any $\mathbf{x} \in \mathcal{N}_2$, where $\mathcal{N}_2$ is defined in (3.5), if $t \geq \lceil \log_{\sigma_2}(\frac{1}{2\sqrt{n}})\rceil$, we have*

$$\max_{i\in[n]}\|\sum_{j=1}^n(W_{ij}^t - 1/n)x_j\|_F \leq \frac{\delta_2}{2}. \tag{D.1}$$

Lemma D.2. *Under the same conditions of Fact B.1, if $\mathbf{x}_k \in \mathcal{N}$ and*

$$x_{i,k+1} = \mathcal{R}_{x_{i,k}}(-\alpha \mathrm{grad}\varphi_i^t(\mathbf{x}_k) - \beta v_{i,k}), \quad \forall i \in [n],$$

where $v_{i,k} \in \mathrm{T}_{x_{i,k}}\mathcal{M}$, the following holds

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|_F \leq \rho_t\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_F + \beta_k\|\mathbf{v}_k\|_F.$$

Proof. By the definition of IAM, we have

$$\begin{aligned}
&\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|_{\mathrm{F}}^2 \leq \|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 \\
=& \sum_{i=1}^n \|\mathcal{R}_{x_{i,k}}\left(-\alpha \mathrm{grad}\varphi_i^t(\mathbf{x}_k) - \beta_k v_{i,k}\right) - \bar{x}_k\|_{\mathrm{F}}^2 \\
\stackrel{(2.4)}{\leq}& \sum_{i=1}^n \|x_{i,k} - \alpha \mathrm{grad}\varphi_i^t(\mathbf{x}_k) - \beta_k v_{i,k} - \bar{x}_k\|_{\mathrm{F}}^2.
\end{aligned} \tag{D.2}$$

Let $\mathbf{v}_k = [v_{1,k}^\top\ \ldots\ v_{n,k}^\top]^\top$. Then, we get

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|_{\mathrm{F}} &\leq \|\mathbf{x}_k - \alpha \mathrm{grad}\varphi^t(\mathbf{x}_k) - \beta_k\mathbf{v}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}} \\
&\leq \|\mathbf{x}_k - \alpha \mathrm{grad}\varphi^t(\mathbf{x}_k) - \bar{\mathbf{x}}_k\|_{\mathrm{F}} + \beta_k\|\mathbf{v}_k\|_{\mathrm{F}}.
\end{aligned} \tag{D.3}$$

By combining inequality (B.3) of Fact B.1, we get

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|_{\mathrm{F}} \leq \rho_t \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}} + \beta_k \|\mathbf{v}_k\|_{\mathrm{F}}. \tag{D.4}$$

The proof is completed.

□

Proof of Lemma 4.1 . We prove that $\mathbf{x}_k \in \mathcal{N}$ for all $k \geq 0$ by induction. Suppose $\mathbf{x}_k \in \mathcal{N}$, let us show $\mathbf{x}_{k+1} \in \mathcal{N}$. Note $\|\mathbf{v}_k\|_{\mathrm{F}} \leq \sqrt{n} D$. Using Lemma D.2 yields

$$\begin{aligned} \|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|_{\mathrm{F}} &\leq \rho_t \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}} + \beta_k \sqrt{n} D \\ &\leq \rho_t \sqrt{n} \delta_1 + \beta_k \sqrt{n} D \\ &\leq \sqrt{n} \delta_1, \end{aligned} \tag{D.5}$$

where the last inequality follows from $\beta_k \leq \frac{1-\rho_t}{D}\delta_1$. Hence $\mathbf{x}_{k+1} \in \mathcal{N}_1$. Secondly, let us verify $\mathbf{x}_{k+1} \in \mathcal{N}_2$. It follows from $\beta_k \leq \frac{\alpha\delta_1}{5D} \leq \frac{\alpha}{2D}$ and $\alpha \leq 1$ that

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_{\mathrm{F},\infty} \overset{(2.4)}{\leq} \max_{i \in [n]} \|\alpha \mathrm{grad} \varphi^t(x_{k,i})\|_{\mathrm{F}} + \beta_k D \overset{(C.10)}{\leq} 2\alpha\delta_2 + \frac{\alpha}{2} \leq 1 - \delta_1^2.$$

Then, we can use Lemma C.4 to get

$$\begin{aligned} \|\bar{x}_k - \bar{x}_{k+1}\|_{\mathrm{F}} &\leq \frac{1}{1 - 2\delta_1^2} \left(\frac{2L_t^2\alpha + L_t\alpha}{n} \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \beta \|\hat{v}_k\|_{\mathrm{F}} + \frac{2M\beta_k^2}{n} \|\mathbf{v}_k\|_{\mathrm{F}}^2 \right) \\ &\leq \frac{1}{1 - 2\delta_1^2} \left[(2L_t^2\alpha + L_t\alpha)\delta_1^2 + \beta_k D + 2M\beta_k^2 D^2 \right]. \end{aligned}$$

Furthermore, since $L_t \leq 2$, $\beta_k \leq \frac{\alpha\delta_1}{5D}$, $\alpha \leq 1/M$, we get

$$\|\bar{x}_k - \bar{x}_{k+1}\|_{\mathrm{F}} \leq \frac{1}{1 - 2\delta_1^2} \left(\frac{252}{25}\alpha\delta_1^2 + \frac{\alpha\delta_1}{5} \right) \leq \frac{1}{1 - 2\delta_1^2} \left(\frac{252}{625r}\alpha\delta_2^2 + \frac{1}{25\sqrt{r}}\alpha\delta_2 \right), \tag{D.6}$$

where the last inequality follows from $\delta_1 \leq \frac{1}{5\sqrt{r}}\delta_2$. Then, one has

$$\begin{aligned} &\|x_{i,k+1} - \bar{x}_{k+1}\|_{\mathrm{F}} \\ \leq &\|x_{i,k+1} - \bar{x}_k\|_{\mathrm{F}} + \|\bar{x}_k - \bar{x}_{k+1}\|_{\mathrm{F}} \\ \overset{(2.4)}{\leq} &\|x_{i,k} - \alpha \mathrm{grad} \varphi_i^t(\mathbf{x}_k) - \beta_k v_{i,k} - \bar{x}_k\|_{\mathrm{F}} + \|\bar{x}_k - \bar{x}_{k+1}\|_{\mathrm{F}} \\ \leq &\|x_{i,k} - \alpha \mathrm{grad} \varphi_i^t(\mathbf{x}_k) - \bar{x}_k\|_{\mathrm{F}} + \frac{1}{5}\alpha\delta_1 + \|\bar{x}_k - \bar{x}_{k+1}\|_{\mathrm{F}}. \end{aligned} \tag{D.7}$$

Now, we proceed by using the same lines in the proof of (Chen et al., 2021, Lemma 13) as follows

$$\mathrm{grad} \varphi_i^t(\mathbf{x}) = x_i - \sum_{j=1}^n W_{ij} x_j - \frac{1}{2} x_i \sum_{j=1}^n W_{ij}^t (x_i - x_j)^\top (x_i - x_j), \tag{D.8}$$

and

$$\begin{aligned} &\|x_{i,k} - \alpha \mathrm{grad} \varphi_i^t(\mathbf{x}_k) - \bar{x}_k\|_{\mathrm{F}} \\ \overset{(D.8)}{=} &\|(1-\alpha)(x_{i,k} - \bar{x}_k) + \alpha(\hat{x}_k - \bar{x}_k) + \alpha \sum_{j=1}^n W_{ij}^t (x_{j,k} - \hat{x}_k) + \frac{\alpha}{2} x_{i,k} \sum_{j=1}^n W_{ij}^t (x_{i,k} - x_{j,k})^\top (x_{i,k} - x_{j,k})\|_{\mathrm{F}} \\ \leq &(1-\alpha)\delta_2 + \alpha\|\hat{x}_k - \bar{x}_k\|_{\mathrm{F}} + \alpha\| \sum_{j=1}^n (W_{ij}^t - \frac{1}{n}) x_{j,k}\|_{\mathrm{F}} + \frac{1}{2}\|\alpha \sum_{j=1}^n W_{ij}^t (x_{i,k} - x_{j,k})^\top (x_{i,k} - x_{j,k})\|_{\mathrm{F}} && \text{(D.9)} \\ \leq &(1-\alpha)\delta_2 + 2\alpha\delta_{1,t}^2\sqrt{r} + \alpha\| \sum_{j=1}^n (W_{ij}^t - \frac{1}{n}) x_{j,k}\|_{\mathrm{F}} + 2\alpha\delta_2^2 && \text{(D.10)} \\ \leq &(1 - \frac{\alpha}{2})\delta_2 + 2\alpha\delta_1^2\sqrt{r} + 2\alpha\delta_2^2, && \text{(D.11)} \end{aligned}$$

where (D.9) follows from $\alpha \in [0,1]$, (D.10) holds by Lemma C.1 and (D.11) follows from Lemma D.1. Combining this with (D.7) implies

$$
\begin{aligned}
&\|x_{i,k+1} - \bar{x}_{k+1}\|_{\mathrm{F}} \\
\le &(1 - \frac{\alpha}{2})\delta_2 + 2\alpha\delta_1^2\sqrt{r} + 2\alpha\delta_2^2 + \frac{1}{5}\alpha\delta_1 + \|\bar{x}_k - \bar{x}_{k+1}\|_{\mathrm{F}} \\
\overset{(D.6)}{\le} &(1 - \frac{\alpha}{2})\delta_2 + 2\alpha\delta_1^2\sqrt{r} + 2\alpha\delta_2^2 + \frac{1}{5}\alpha\delta_1 + \frac{1}{1-2\delta_1^2}\left(\frac{252}{625r}\alpha\delta_2^2 + \frac{1}{25\sqrt{r}}\alpha\delta_2\right).
\end{aligned} \tag{D.12}
$$

Therefore, substituting the conditions (3.6) on δ_1, δ_2 into (D.12) yields

$$\|x_{i,k+1} - \bar{x}_{k+1}\|_{\mathrm{F}} \le \delta_2.$$

The proof of the first statement is completed. Finally, it follows from (D.5) that

$$
\begin{aligned}
\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|_{\mathrm{F}} &\le \rho_t \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}} + \beta_k\sqrt{n}D \\
&\le \rho_t^{k+1}\|\mathbf{x}_0 - \bar{\mathbf{x}}_0\|_{\mathrm{F}} + \sqrt{n}D\sum_{l=0}^{k}\rho_t^{k-l}\beta_l.
\end{aligned} \tag{D.13}
$$

□

An immediate result of Lemma 4.1 is that the rate of consensus $\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 = \mathcal{O}(\beta_k^2)$ if $\beta_k = \mathcal{O}(\frac{1}{k^p})$. The proof is similar as (Liu et al., 2017, Proposition 8), we provide it for completeness.

Lemma D.3. *Under Assumptions 1 to 4, for Algorithm 1, if* $\mathbf{x}_0 \in \mathcal{N}$, $0 < \alpha \le \min\{\frac{\Phi}{2L_t}, 1, \frac{1}{M}\}$, $t \ge \lceil \log_{\sigma_2}(\frac{1}{2\sqrt{n}})\rceil$ *and*

$$\beta_k = \min\{\frac{\alpha\delta_1}{5D}\cdot\frac{1}{(k+1)^p}, \frac{1-\rho_t}{D}\delta_1\}, \quad p \in (0,1], \tag{D.14}$$

then there exists a constant $C > 0$ *such that* $\frac{1}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_F^2 \le CD^2\beta_k^2$ *for any* $k \ge 0$*, where* C *is independent of* D *and* n*.*

Proof of Lemma D.3. The proof relies on Lemma 4.1. Let $a_k := \frac{\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}}{\sqrt{n}\beta_k}$.

It follows from (D.13) that

$$a_{k+1} \le \rho_t a_k + D\cdot\frac{\beta_k}{\beta_{k+1}} \le \rho_t^{k+1-K}a_K + D\sum_{l=K}^{k}\rho_t^{k-l}\frac{\beta_l}{\beta_{l+1}}. \tag{D.15}$$

Recall that $\beta_k = \mathcal{O}(1/D)$ and $\frac{1}{n}\|\mathbf{x}_0 - \bar{\mathbf{x}}_0\|_{\mathrm{F}}^2 \le \delta_1^2$, it follows that $a_0 \le \delta_1/\beta_0 = \mathcal{O}(D)$. Since $\lim_{k\to\infty}\frac{\beta_{k+1}}{\beta_k} = 1$, there exists sufficiently large K such that

$$\frac{\beta_k}{\beta_{k+1}} \le 2, \quad \forall k \ge K.$$

For $0 \le k \le K$, there exists some $C' > 0$ such that

$$a_k^2 \le C'D^2,$$

where C' is independent of D and n. For $k \ge K$, using (D.15) gives $a_k^2 \le CD^2$, where $C = 2C' + \frac{8}{(1-\rho_t)^2}$. Hence, we get $\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2/n \le CD^2\beta_k^2$ for all $k \ge 0$, where $C = \mathcal{O}(\frac{1}{(1-\rho_t)^2})$. □

Lemma D.4. *Under Assumptions 1 to 4, suppose* $\mathbf{x}_k \in \mathcal{N}$, $t \ge \lceil \log_{\sigma_2}(\frac{1}{2\sqrt{n}})\rceil$, $0 < \alpha \le \min\{\frac{\Phi}{2L_t}, 1, \frac{1}{M}\}$*. If* $x_{i,k+1} = \mathcal{R}_{x_{i,k}}(-\alpha\mathrm{grad}\varphi^t(x_{i,k}) - \beta_k v_{i,k})$, $0 < \beta_k \le \min\{\frac{1}{5L_g}, \frac{\alpha\delta_1}{5D}\}$ *and* $\beta_k \ge \beta_{k+1}$*, where* $v_{i,k}$ *satisfies Assumption 3 and* L_g *is given in Lemma 2.4. It follows that*

$$
\begin{aligned}
\mathbb{E}_k f(\bar{x}_{k+1}) \le f(\bar{x}_k) - \frac{\beta_k}{4}\|\hat{g}_k\|_F^2 - \frac{\beta_k}{4}\|\mathrm{grad} f(\bar{x}_k)\|_F^2 \\
+ \frac{3L_g\Xi^2}{2n}\beta_k^2 + (\frac{CD^2L_G^2}{2} + \mathcal{T}_1 D^4)\beta_k^3 + \mathcal{T}_2 L_g D^4\beta_k^4,
\end{aligned} \tag{D.16}
$$

where L_G *is given in Lemma 2.4,* C *is given in Lemma D.3,* $\mathcal{T}_1 = 2(4\sqrt{r} + 6\alpha)^2C^2 + 8M^2$ *and* $\mathcal{T}_2 = 201\alpha^2C^2 + 9M^2$.

Note the variance term is in the order of $\mathcal{O}(\frac{\Xi^2}{n}\beta_k^2)$, since the gradient batch size is n.

Proof of Lemma D.4. Denote the conditional expectation $\mathbb{E}_{i,k} v_{i,k} := \mathbb{E}[v_{i,k}|x_{i,k}]$ and $\mathbb{E}_k := \mathbb{E}[\cdot|\mathbf{x}_k]$. By invoking Lemma 2.4, we have

$$\begin{aligned}
&\mathbb{E}_k f(\bar{x}_{k+1}) \leq f(\bar{x}_k) + \langle \mathrm{grad} f(\bar{x}_k), \mathbb{E}_k \bar{x}_{k+1} - \bar{x}_k \rangle + \frac{L_g}{2}\mathbb{E}_k \|\bar{x}_{k+1} - \bar{x}_k\|_{\mathrm{F}}^2 \\
&= f(\bar{x}_k) - \langle \mathrm{grad} f(\bar{x}_k), \beta_k \hat{g}_k \rangle + \langle \mathrm{grad} f(\bar{x}_k), \mathbb{E}_k[\bar{x}_{k+1} - \bar{x}_k + \beta_k \hat{v}_k] \rangle + \frac{L_g}{2}\mathbb{E}_k \|\bar{x}_{k+1} - \bar{x}_k\|_{\mathrm{F}}^2 \\
&= f(\bar{x}_k) - \frac{\beta_k}{2}\|\mathrm{grad} f(\bar{x}_k)\|_{\mathrm{F}}^2 - \frac{\beta_k}{2}\|\hat{g}_k\|_{\mathrm{F}}^2 + \frac{\beta_k}{2}\|\mathrm{grad} f(\bar{x}_k) - \hat{g}_k\|_{\mathrm{F}}^2 \\
&\quad + \langle \mathrm{grad} f(\bar{x}_k), \mathbb{E}_k \bar{x}_{k+1} - \bar{x}_k + \beta_k \hat{g}_k \rangle + \frac{L_g}{2}\mathbb{E}_k \|\bar{x}_{k+1} - \bar{x}_k\|_{\mathrm{F}}^2,
\end{aligned} \tag{D.17}$$

where $\hat{v}_k = \frac{1}{n}\sum_{i=1}^n v_{i,k}$ and we use $\mathbb{E}_k \hat{v}_k = \hat{g}_k$ in the first equation.

Note that for $\beta_k > 0$, we have

$$\langle \mathrm{grad} f(\bar{x}_k), \mathbb{E}_k \bar{x}_{k+1} - \bar{x}_k + \beta_k \hat{g}_k \rangle \leq \frac{\beta_k}{4}\|\mathrm{grad} f(\bar{x}_k)\|_{\mathrm{F}}^2 + \frac{1}{\beta_k}\|\mathbb{E}_k \bar{x}_{k+1} - \bar{x}_k + \beta_k \hat{g}_k\|_{\mathrm{F}}^2.$$

Plugging this into (D.17) yields

$$\begin{aligned}
&\mathbb{E}_k f(\bar{x}_{k+1}) \\
&\leq f(\bar{x}_k) - \frac{\beta_k}{2}\|\hat{g}_k\|_{\mathrm{F}}^2 - \frac{\beta_k}{4}\|\mathrm{grad} f(\bar{x}_k)\|_{\mathrm{F}}^2 + \frac{\beta_k}{2}\underbrace{\|\mathrm{grad} f(\bar{x}_k) - \hat{g}_k\|_{\mathrm{F}}^2}_{:=a_1} + \frac{1}{\beta_k}\underbrace{\|\mathbb{E}_k[\bar{x}_{k+1} - \bar{x}_k + \beta_k \hat{v}_k]\|_{\mathrm{F}}^2}_{:=a_2} \\
&\quad + \frac{L_g}{2}\underbrace{\mathbb{E}_k \|\bar{x}_{k+1} - \bar{x}_k\|_{\mathrm{F}}^2}_{:=a_3}.
\end{aligned} \tag{D.18}$$

Using Lemma 2.4 implies

$$a_1 \leq \frac{1}{n}\sum_{i=1}^n \|\mathrm{grad} f(x_{i,k}) - \mathrm{grad} f(\bar{x}_k)\|_{\mathrm{F}}^2 \overset{(2.6)}{\leq} \frac{L_G^2}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2.$$

Secondly, we use the following inequality to derive the upper bound of a_2. From Lemma 4.1, we have $\mathbf{x}_{k+1} \in \mathcal{N}$. One has

$$\begin{aligned}
&\|\bar{x}_{k+1} - \bar{x}_k + \beta_k \hat{v}_k\|_{\mathrm{F}} \\
&\leq \|\bar{x}_k - \hat{x}_k\|_{\mathrm{F}} + \|\bar{x}_{k+1} - \hat{x}_{k+1}\|_{\mathrm{F}} + \|\hat{x}_k - \beta_k \hat{v}_k - \hat{x}_{k+1}\|_{\mathrm{F}} \\
&\overset{(P1)}{\leq} \frac{2\sqrt{r}}{n}(\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|_{\mathrm{F}}^2) + \|\hat{x}_k - \beta_k \hat{v}_k - \hat{x}_{k+1}\|_{\mathrm{F}} \\
&\leq \frac{4\sqrt{r}}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \|\hat{x}_k - \beta_k \hat{v}_k - \hat{x}_{k+1}\|_{\mathrm{F}},
\end{aligned} \tag{D.19}$$

where we use $\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 \geq \|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|_{\mathrm{F}}^2$ in the last inequality.

For the second term, since $v_{i,k} \in \mathrm{T}_{x_{i,k}}\mathcal{M}$ we have

$$\begin{aligned}
&\|\hat{x}_k - \beta_k \hat{v}_k - \hat{x}_{k+1}\|_{\mathrm{F}} \leq \frac{1}{n}\sum_{i=1}^n \|x_{i,k} - \alpha \mathrm{grad}\varphi_i^t(\mathbf{x}_k) - \beta_k v_{i,k} - x_{i,k+1}\|_{\mathrm{F}} + \frac{\alpha}{n}\|\sum_{i=1}^n \mathrm{grad}\varphi_i^t(\mathbf{x}_k)\|_{\mathrm{F}} \\
&\overset{(P1)}{\leq} \frac{M}{n}\sum_{i=1}^n \|\alpha \mathrm{grad}\varphi_i^t(\mathbf{x}_k) + \beta_k v_{i,k}\|_{\mathrm{F}}^2 + \frac{\alpha}{n}\|\sum_{i=1}^n \mathrm{grad}\varphi_i^t(\mathbf{x}_k)\|_{\mathrm{F}} \\
&\overset{(C.8)}{\leq} \frac{2M\alpha^2}{n}\|\mathrm{grad}\varphi^t(\mathbf{x}_k)\|_{\mathrm{F}}^2 + \frac{2M\beta_k^2}{n}\|\mathbf{v}_k\|_{\mathrm{F}}^2 + \frac{L_t\alpha}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 \\
&\overset{(C.9)}{\leq} \frac{2L_t^2 M\alpha^2 + L_t\alpha}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \frac{2M\beta_k^2}{n}\|\mathbf{v}_k\|_{\mathrm{F}}^2 \\
&\leq \frac{10\alpha}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \frac{2M\beta_k^2}{n}\|\mathbf{v}_k\|_{\mathrm{F}}^2,
\end{aligned} \tag{D.20}$$

where we use $\alpha \leq \frac{1}{M}$ and $L_t \leq 2$ in the last inequality. Plugging (D.20) into (D.19) yields

$$\|\bar{x}_{k+1} - \bar{x}_k + \beta_k \hat{v}_k\|_{\mathrm{F}}^2 \leq 2(\frac{4\sqrt{r}+10\alpha}{n})^2 \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^4 + 2(\frac{2M\beta_k^2}{n})^2 \|\mathbf{v}_k\|_{\mathrm{F}}^4. \tag{D.21}$$

Then, using Jensen's inequality and $\|\mathbf{v}_k\|_{\mathrm{F}}^2 \leq nD^2$ implies

$$a_2 \leq \mathbb{E}_k[\|\bar{x}_{k+1} - \bar{x}_k + \beta_k \hat{v}_k\|_{\mathrm{F}}^2] \leq 2(\frac{4\sqrt{r}+10\alpha}{n})^2 \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^4 + 8M^2\beta_k^4 D^4.$$

Thirdly, invoking Lemma C.4 yields

$$\|\bar{x}_k - \bar{x}_{k+1}\|_{\mathrm{F}} \leq \frac{1}{1-2\delta_1^2}\left[\frac{10\alpha}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + 2M\beta_k^2 D^2 + \beta_k \|\hat{v}_k\|_{\mathrm{F}}\right].$$

Hence, it follows that

$$\begin{aligned}
a_3 &\leq \frac{2}{(1-2\delta_1^2)^2}\left[\frac{10\alpha}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + 2M\beta_k^2 D^2\right]^2 + \frac{2}{(1-2\delta_1^2)^2}\beta_k^2 \mathbb{E}_k \|\hat{v}_k\|_{\mathrm{F}}^2 \\
&= \frac{2}{(1-2\delta_1^2)^2}\left[\frac{10\alpha}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + 2M\beta_k^2 D^2\right]^2 + \frac{2}{(1-2\delta_1^2)^2}\beta_k^2 \mathbb{E}_k \|\hat{v}_k - \hat{g}_k\|_{\mathrm{F}}^2 + \frac{2}{(1-2\delta_1^2)^2}\beta_k^2 \|\hat{g}_k\|_{\mathrm{F}}^2 \\
&\overset{(i)}{=} \frac{2}{(1-2\delta_1^2)^2}\left[\frac{10\alpha}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + 2M\beta_k^2 D^2\right]^2 + \frac{2}{(1-2\delta_1^2)^2 n^2}\beta_k^2 \sum_{i=1}^n \mathbb{E}_k \|v_{i,k} - g_{i,k}\|_{\mathrm{F}}^2 + \frac{2}{(1-2\delta_1^2)^2}\beta_k^2 \|\hat{g}_k\|_{\mathrm{F}}^2 \\
&\overset{(ii)}{\leq} \frac{4}{(1-2\delta_1^2)^2}\left[\frac{100\alpha^2}{n^2}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^4 + 4M^2\beta_k^4 D^4\right] + \frac{2}{(1-2\delta_1^2)^2 n}\beta_k^2 \Xi^2 + \frac{2}{(1-2\delta_1^2)^2}\beta_k^2 \|\hat{g}_k\|_{\mathrm{F}}^2,
\end{aligned}$$

where (i) and (ii) hold by the independence of $v_{i,k}$ and bounded variance of Assumption 3, respectively. Therefore, by combining a_1, a_2, a_3 with (D.18) implies that

$$\begin{aligned}
&\mathbb{E}_k f(\bar{x}_{k+1}) \leq f(\bar{x}_k) - \frac{\beta_k}{2}\|\hat{g}_k\|_{\mathrm{F}}^2 - \frac{\beta_k}{4}\|\mathrm{grad} f(\bar{x}_k)\|_{\mathrm{F}}^2 + \frac{\beta_k}{2}a_1 + \frac{1}{\beta_k}a_2 + \frac{L_g}{2}a_3 \\
&\leq f(\bar{x}_k) - (\frac{\beta_k}{2} - \frac{L_g\beta_k^2}{(1-2\delta_1^2)^2})\|\hat{g}_k\|_{\mathrm{F}}^2 - \frac{\beta_k}{4}\|\mathrm{grad} f(\bar{x}_k)\|_{\mathrm{F}}^2 + \frac{\beta_k L_G^2}{2n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \frac{2}{\beta_k}(\frac{4\sqrt{r}+10\alpha}{n})^2\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^4 \\
&\quad + 8M^2\beta_k^3 D^4 + \frac{2L_g}{(1-2\delta_1^2)^2}\left[\frac{100\alpha^2}{n^2}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^4 + 4M^2\beta_k^4 D^4\right] + \frac{L_g}{(1-2\delta_1^2)^2 n}\beta_k^2 \Xi^2.
\end{aligned}$$

By Lemma D.3, we have $\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 \leq nCD^2\beta_k^2$. It follows that

$$\begin{aligned}
&\mathbb{E}_k f(\bar{x}_{k+1}) \\
&\leq f(\bar{x}_k) - (\frac{\beta_k}{2} - \frac{L_g\beta_k^2}{(1-2\delta_1^2)^2})\|\hat{g}_k\|_{\mathrm{F}}^2 - \frac{\beta_k}{4}\|\mathrm{grad} f(\bar{x}_k)\|_{\mathrm{F}}^2 \\
&\quad + \frac{L_g \Xi^2}{(1-2\delta_1^2)^2 n}\beta_k^2 + \left[\frac{CD^2L_G^2}{2} + \left(2(4\sqrt{r}+10\alpha)^2 C^2 + 8M^2\right) D^4\right]\beta_k^3 \\
&\quad + \frac{2L_g}{(1-2\delta_1^2)^2}\left[100\alpha^2 C^2 D^4 + 4M^2 D^4\right]\beta_k^4 \\
&\leq f(\bar{x}_k) - \frac{\beta_k}{4}\|\hat{g}_k\|_{\mathrm{F}}^2 - \frac{\beta_k}{4}\|\mathrm{grad} f(\bar{x}_k)\|_{\mathrm{F}}^2 \\
&\quad + \frac{3L_g\Xi^2}{2n}\beta_k^2 + \left[\frac{CD^2L_G^2}{2} + \left(2(4\sqrt{r}+10\alpha)^2 C^2 + 8M^2\right) D^4\right]\beta_k^3 \\
&\quad + \left(201\alpha^2 C^2 D^4 + 9M^2 D^4\right) L_g \beta_k^4,
\end{aligned}$$

where we use $\frac{1}{(1-2\delta_1^2)^2} \leq 1.002$ and $\beta_k \leq \frac{1}{5L_g}$ in the last inequality. The proof is completed. □

Proof of Theorem 4.2. Using (D.16) implies

$$\begin{aligned}&\mathbb{E}_k f(\bar{x}_{k+1})\\ &\le f(\bar{x}_k) - \frac{\beta_k}{4}\|\mathrm{grad} f(\bar{x}_k)\|_{\mathrm{F}}^2 + \frac{3L_g\Xi^2}{2n}\beta_k^2 + (\frac{CD^2L_g^2}{2} + \mathcal{T}_1 D^4)\beta_k^3 + \mathcal{T}_2 L_g D^4 \beta_k^4,\end{aligned} \tag{D.22}$$

Taking the expectation on all k and telescoping the right hand side give us for any $K > 0$

$$\sum_{k=0}^{K}\frac{\beta_k}{4}\mathbb{E}\|\mathrm{grad} f(\bar{x}_k)\|_{\mathrm{F}}^2 \le f(\bar{x}_0) - f^* + \frac{3L_g\Xi^2}{2n}\sum_{k=0}^{K}\beta_k^2 + (\frac{CD^2L_g^2}{2} + \mathcal{T}_1 D^4)\sum_{k=0}^{K}\beta_k^3 + \mathcal{T}_2 L_g D^4\sum_{k=0}^{K}\beta_k^4,$$

where $f^* = \min_{x\in\mathrm{St}(d,r)} f(x)$. Dividing both sides by $\sum_{k=0}^{K}\frac{\beta_k}{4}$ yields

$$\min_{k=0,\ldots,K}\mathbb{E}\|\mathrm{grad} f(\bar{x}_k)\|_{\mathrm{F}}^2 \le \frac{f(\bar{x}_0) - f^* + \frac{3L_g\Xi^2}{2n}\sum_{k=0}^{K}\beta_k^2 + (\frac{CD^2L_g^2}{2} + \mathcal{T}_1 D^4)\sum_{k=0}^{K}\beta_k^3 + \mathcal{T}_2 L_g D^4\sum_{k=0}^{K}\beta_k^4}{\sum_{k=0}^{K}\frac{\beta_k}{4}}.$$

Let $\tilde{\beta} = \min\{1/L_g, \frac{1-\rho_t}{D}\}$. Noticing that $\beta_k = \mathcal{O}(\min\{\frac{1-\rho_t}{D}, \frac{1}{L_G}\}\cdot\frac{1}{k})$, $\frac{\sum_{k=0}^{K}\beta_k^2}{\sum_{k=0}^{K}\beta_k} = \mathcal{O}(\tilde{\beta}\frac{\ln(K+1))}{\sqrt{K+1}})$, $\frac{\sum_{k=0}^{K}\beta_k^3}{\sum_{k=0}^{K}\beta_k} = \mathcal{O}(\frac{\tilde{\beta}^2}{\sqrt{K+1}})$ and $\frac{\sum_{k=0}^{K}\beta_k^4}{\sum_{k=0}^{K}\beta_k} = \mathcal{O}(\frac{\tilde{\beta}^3}{\sqrt{K+1}})$. The proof is completed. □

The following corollary follows (Lian et al., 2017), in which the convergence results of constant stepsize β_k is given.

Corollary D.5. *Under Assumptions 1 to 4, suppose* $\mathbf{x}_k \in \mathcal{N}$, $t \ge \lceil \log_{\sigma_2}(\frac{1}{2\sqrt{n}})\rceil$, $0 < \alpha \le \bar{\alpha}$. *If constant stepsize* $\beta_k \equiv \beta = \frac{1}{2L_G + \Xi\sqrt{(K+1)/n}}$, *where*

$$K + 1 \ge \max\{\frac{n}{\Xi^2}(\max\{3L_G, \frac{5D}{\alpha\delta_1}, \frac{D\delta_1}{1-\rho_t}\})^2, \frac{n^3}{\Xi^6}\left(\frac{CD^2L_g^2 + (2\mathcal{T}_1 + \mathcal{T}_2)D^4}{2(f(\bar{x}_0) - f^*) + 3L_G}\right)^2\},$$

if follows that

$$\min_{k=0,\ldots,K}\mathbb{E}\|\mathrm{grad} f(\bar{x}_k)\|_F^2 \le \frac{8L_G(f(\bar{x}_0) - f^*)}{K+1} + \frac{8(f(\bar{x}_0) - f^* + \frac{3L_G}{2})\Xi}{\sqrt{n(K+1)}}.$$

Proof. Since $K + 1 \ge \frac{n}{\Xi^2}(\max\{3L_G, \frac{5D}{\alpha\delta_1}, \frac{D\delta_1}{1-\rho_t}\})^2$, we have

$$\beta_k \le \min\{\frac{1}{5L_G}, \frac{\alpha\delta_1}{5D}, \frac{1-\rho_t}{D}\delta_1\}$$

for all $k = 0, 1, \ldots, K$. Therefore, it follows that $\mathbf{x}_k \in \mathcal{N}$ for $k = 0, 1, \ldots, K$. Using Theorem 4.2, we have

$$\begin{aligned}&\min_{k=0,\ldots,K}\mathbb{E}\|\mathrm{grad} f(\bar{x}_k)\|_{\mathrm{F}}^2\\ &\le \frac{4(f(\bar{x}_0) - f^*)}{(K+1)\beta} + \frac{6L_g\beta\Xi^2}{n} + (2CD^2L_g^2 + 4\mathcal{T}_1 D^4)\beta^2 + 4\mathcal{T}_2 L_g D^4\beta^3\\ &\le \frac{8L_G(f(\bar{x}_0) - f^*)}{K+1} + \frac{4(f(\bar{x}_0) - f^*)\Xi}{\sqrt{n(K+1)}} + \frac{6L_G\Xi^2}{2nL_G + \Xi\sqrt{n(K+1)}} + \frac{2CD^2L_g^2 + (4\mathcal{T}_1 + 2\mathcal{T}_2)D^4}{(2L_G + \Xi\sqrt{(K+1)/n})^2} &\text{(D.23)}\\ &\le \frac{8L_G(f(\bar{x}_0) - f^*)}{K+1} + \frac{4(f(\bar{x}_0) - f^* + \frac{3L_G}{2})\Xi}{\sqrt{n(K+1)}} + \frac{2nCD^2L_g^2 + (4\mathcal{T}_1 + 2\mathcal{T}_2)nD^4}{\Xi^2(K+1)}, &\text{(D.24)}\end{aligned}$$

where we use $\beta \le \frac{1}{2L_G} \le \frac{1}{2L_g}$ in (D.23).

When

$$K + 1 \ge \frac{n^3}{\Xi^6}\left(\frac{CD^2L_g^2 + (2\mathcal{T}_1 + \mathcal{T}_2)D^4}{2(f(\bar{x}_0) - f^*) + 3L_G}\right)^2,$$

the second term in (D.24) is greater than the third term, we get

$$\begin{aligned}&\min_{k=0,\dots,K} \mathbb{E}\|\mathrm{grad} f(\bar{x}_k)\|_{\mathrm{F}}^2 \\ &\leq \frac{8L_G(f(\bar{x}_0)-f^*)}{K+1} + \frac{8(f(\bar{x}_0)-f^*+\frac{3L_G}{2})\Xi}{\sqrt{n(K+1)}},\end{aligned}$$

which completes the proof. □

E. Proofs for Section 5

In this section, we use the following notations

$$\mathbf{G}_k := \begin{bmatrix}\mathrm{grad} f_1(x_{1,k}) \\ \vdots \\ \mathrm{grad} f_n(x_{n,k})\end{bmatrix}, \ \mathbf{y}_k = \begin{bmatrix} y_{1,k} \\ \vdots \\ y_{n,k}\end{bmatrix}, \ \hat{y}_k := \frac{1}{n}\sum_{i=1}^n y_{i,k},$$

$$\hat{g}_k := \frac{1}{n}\sum_{i=1}^n \mathrm{grad} f_i(x_{i,k}), \quad \hat{\mathbf{G}}_k := (\mathbf{1}_n \otimes I_n)\hat{g}_k.$$

Proof of Lemma 5.1. We prove it by induction. Let $\hat{g}_{-1} = \hat{y}_0$, one has $\|y_{i,0}\|_{\mathrm{F}} \leq D$ and

$$\|y_{i,0} - \hat{g}_{-1}\|_{\mathrm{F}} \leq \|y_{i,0}\|_{\mathrm{F}} + \|\hat{g}_{-1}\|_{\mathrm{F}} \leq D + \frac{1}{n}\sum_{j=1}^n \|y_{j,0}\|_{\mathrm{F}} \leq 2D$$

for all $i \in [n]$ by Assumption 2. Suppose for some $k \geq 0$, it follows that $\|y_{i,k}\|_{\mathrm{F}} \leq 2D + L_G$ and $\|y_{i,k} - \hat{g}_{k-1}\|_{\mathrm{F}} \leq 2D + L_G$. We note that the bound of v_i becomes $2D + L_G$ here since $\|v_{i,k}\|_{\mathrm{F}} = \|\mathcal{P}_{T_{x_{i,k}}\mathcal{M}} y_{i,k}\|_{\mathrm{F}} \leq \|y_{i,k}\|_{\mathrm{F}}$. Following the same argument in the proof of Lemma 4.1, we get $\mathbf{x}_{k+1} \in \mathcal{N}$ since $0 < \alpha \leq \min\{\frac{\Phi}{2L_t}, 1, \frac{1}{M}\}$ and $0 \leq \beta \leq \min\{\frac{1-\rho_t}{L_G+2D}\delta_1, \frac{\alpha\delta_1}{5(L_G+2D)}\}$.

Then, we have

$$\begin{aligned}\|y_{i,k+1} - \hat{g}_k\|_{\mathrm{F}} &= \|\sum_{j=1}^n W_{i,j}^t y_{j,k} - \hat{g}_k + \mathrm{grad} f(x_{i,k+1}) - \mathrm{grad} f(x_{i,k})\|_{\mathrm{F}} \\ &= \|\sum_{j=1}^n (W_{i,j}^t - \frac{1}{n})(y_{j,k} - \hat{g}_{k-1}) + \mathrm{grad} f(x_{i,k+1}) - \mathrm{grad} f(x_{i,k})\|_{\mathrm{F}} \\ &\overset{(2.6)}{\leq} \sigma_2^t \sqrt{n}\|y_{j,k} - \hat{g}_{k-1}\|_{\mathrm{F}} + L_G\|x_{i,k+1} - x_{i,k}\|_{\mathrm{F}} \\ &\overset{(2.4)}{\leq} \sigma_2^t \sqrt{n}\|y_{j,k} - \hat{g}_{k-1}\|_{\mathrm{F}} + L_G(\alpha\|\mathrm{grad}\varphi_i^t(\mathbf{x}_k)\|_{\mathrm{F}} + \beta\|y_{i,k}\|_{\mathrm{F}}) \\ &\overset{(C.10)}{\leq} \frac{1}{2}\|y_{j,k} - \hat{g}_{k-1}\|_{\mathrm{F}} + 2L_G\alpha\delta_2 + L_G\beta\|y_{i,k}\|_{\mathrm{F}} \\ &\leq \frac{1}{2}(2D + L_G) + 2\delta_2 L_G + \frac{L_G}{5}\delta_1\alpha \\ &\overset{(3.6)}{\leq} D + L_G.\end{aligned}$$

Hence, $\|y_{i,k+1}\|_{\mathrm{F}} \leq \|y_{i,k+1} - \hat{g}_k\|_{\mathrm{F}} + \|\hat{g}_k\|_{\mathrm{F}} \leq L_G + 2D$, where we use $\|\hat{g}_k\|_{\mathrm{F}} \leq D$. Therefore, we get $\|y_{i,k}\|_{\mathrm{F}} \leq L_g + 2D$ for all i, k and $\mathbf{x}_k \in \mathcal{N}$.

Using the same argument of Lemma D.3, there exists some $C_1 = \mathcal{O}(\frac{1}{(1-\rho_t)^2})$ that is independent of L_G and D such that

$$\frac{1}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 \leq C_1(L_G + 2D)^2\beta^2, k \geq 0. \tag{E.1}$$

The proof is completed. □

Next, we present the relations between the consensus error and the gradient tracking error.

Lemma E.1. *Under the same conditions of Lemma 5.1, one has the following error bounds for any $k \geq 0$:*

1. *Successive gradient error:*
$$\|\mathbf{G}_{k+1} - \mathbf{G}_k\|_F \leq 2\alpha L_G \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_F + \beta L_G \|\mathbf{y}_k\|_F. \tag{E.2}$$
2. *Successive tracking error:*
$$\|\mathbf{y}_{k+1} - \hat{\mathbf{G}}_{k+1}\|_F \leq \sigma_2^t \|\mathbf{y}_k - \hat{\mathbf{G}}_k\|_F + \|\mathbf{G}_{k+1} - \mathbf{G}_k\|_F. \tag{E.3}$$
3. *Successive consensus error: for $\rho_t = \sqrt{1 - \gamma_t \alpha} \in (0, 1)$,*
$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|_F \leq \rho_t \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_F + \beta \|\mathbf{y}_k\|_F. \tag{E.4}$$
4. *Associating $\mathbf{y}_k, \hat{\mathbf{G}}_k$ with above items:*
$$\|\mathbf{y}_k\|_F \leq \|\mathbf{y}_k - \hat{\mathbf{G}}_k\|_F + \|\hat{\mathbf{G}}_k\|_F. \tag{E.5}$$

Proof of Lemma E.1. By Lemma 5.1, we know $\mathbf{x}_k \in \mathcal{N}$ for all $k \geq 0$.

1. Using Lemma 2.4 yields
$$\|\mathbf{G}_{k+1} - \mathbf{G}_k\|_\mathrm{F} \leq L_G \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_\mathrm{F}.$$
By Lemma 2.3, it follows that
$$\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_\mathrm{F} \leq \alpha \|\mathrm{grad}\varphi^t(\mathbf{x}_k)\|_\mathrm{F} + \beta \|\mathbf{v}_k\|_\mathrm{F} \overset{(C.9)}{\leq} 2\alpha \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_\mathrm{F} + \beta \|\mathbf{y}_k\|_\mathrm{F},$$
where we use $\|\mathbf{v}_k\|_\mathrm{F} \leq \|\mathbf{y}_k\|_\mathrm{F}$. Hence, the inequality (E.2) is proved.
2. Denote $J = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$. Note that
$$\begin{aligned}\mathbf{y}_{k+1} - \hat{\mathbf{G}}_{k+1} &= ((I_n - J) \otimes I_n)\mathbf{y}_{k+1} \\ &= ((I_n - J) \otimes I_n)\left[(W^t \otimes I_n)\mathbf{y}_k + \mathbf{G}_{k+1} - \mathbf{G}_k\right] \\ &= ((W^t - J) \otimes I_n)\mathbf{y}_k + ((I_n - J) \otimes I_n)(\mathbf{G}_{k+1} - \mathbf{G}_k)\end{aligned}$$
where we use $((I_n - J) \otimes I_n)(W^t \otimes I_n) = (W^t - J) \otimes I_n$. It follows that
$$\|\mathbf{y}_{k+1} - \hat{\mathbf{G}}_{k+1}\|_\mathrm{F} \leq \sigma_2^t \|\mathbf{y}_k - \hat{\mathbf{G}}_k\|_\mathrm{F} + \|\mathbf{G}_{k+1} - \mathbf{G}_k\|_\mathrm{F}$$
3. Note that $\|\mathbf{v}_k\|_\mathrm{F} \leq \|\mathbf{y}_k\|_\mathrm{F}$. Then the desired result follows the same line as that of Lemma D.2.
4. This follows from the triangle inequality.

□

To show Theorem 5.2, we firstly show a descent lemma. Note that an extra $\|\hat{\mathbf{G}}_k\|_\mathrm{F}^2 = n\|\hat{g}_k\|_\mathrm{F}^2$ appears in (E.5), what is we aim at bounding in the optimization problem (1.1). By combining with the following lemmas, we can quickly obtain the final convergence result.

Lemma E.2. *Under the same conditions of Lemma 5.1, it follows that*
$$f(\bar{x}_{k+1}) \leq f(\bar{x}_k) - (\beta - 4L_G\beta^2)\|\hat{g}_k\|_F^2 + \mathcal{G}_0 \frac{L_G}{n} \|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|_F^2 + \mathcal{G}_1 \frac{L_G}{n} \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_F^2 + \mathcal{G}_2 \frac{L_G}{n} \beta^2 \|\mathbf{y}_k\|_F^2, \tag{E.6}$$
where $\mathcal{G}_0 = \frac{4r(L_G+2D)^2C_1}{L_G^2}$, $\mathcal{G}_1 = 1 + \mathcal{G}_0 + \frac{2D\alpha + 8MD\alpha^2}{L_G} + 13C_1\delta_1^2\alpha^4$, $\mathcal{G}_2 = \frac{2MD}{L_G} + \frac{\delta_1^2}{2} + 5$ and C_1 is given in Lemma 5.1.

Since $D = \max_{x\in \mathrm{St}(d,r)} \|\nabla f(x)\|_{\mathrm{F}} \leq \sqrt{r}\cdot \max_{x\in \mathrm{St}(d,r)} \|\nabla f(x)\|_2 = \sqrt{r} L_n$. By the choice of α, the constants in Lemma E.2 are given by $\mathcal{G}_0 = \mathcal{O}(r^2 C_1)$, $\mathcal{G}_1 = \mathcal{O}(r^2 C_1)$ and $\mathcal{G}_2 = \mathcal{O}(M)$.

Proof of Lemma E.2. It follows from Lemma 2.4 that

$$\|\hat{g}_k - \mathrm{grad} f(\bar{x}_k)\|_{\mathrm{F}}^2 \leq \frac{1}{n}\sum_{i=1}^n \|\mathrm{grad} f_i(x_{i,k}) - \mathrm{grad} f(\bar{x}_k)\|_{\mathrm{F}}^2 \leq \frac{L_G^2}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2. \tag{E.7}$$

By invoking Lemma 2.4 and noting $L_g \leq L_G$, we also have

$$\begin{aligned} f(\bar{x}_{k+1}) &\leq f(\bar{x}_k) + \langle \mathrm{grad} f(\bar{x}_k), \bar{x}_{k+1} - \bar{x}_k\rangle + \frac{L_g}{2}\|\bar{x}_{k+1} - \bar{x}_k\|_{\mathrm{F}}^2 \\ &\leq f(\bar{x}_k) + \langle \hat{g}_k, \bar{x}_{k+1} - \bar{x}_k\rangle + \langle \mathrm{grad} f(\bar{x}_k) - \hat{g}_k, \bar{x}_{k+1} - \bar{x}_k\rangle + \frac{L_G}{2}\|\bar{x}_{k+1} - \bar{x}_k\|_{\mathrm{F}}^2 \\ &\leq f(\bar{x}_k) + \langle \hat{g}_k, \bar{x}_{k+1} - \bar{x}_k\rangle + \frac{1}{L_G}\|\mathrm{grad} f(\bar{x}_k) - \hat{g}_k\|_{\mathrm{F}}^2 + \frac{3L_G}{4}\|\bar{x}_{k+1} - \bar{x}_k\|_{\mathrm{F}}^2 \\ &\overset{(E.7)}{\leq} f(\bar{x}_k) + \langle \hat{g}_k, \bar{x}_{k+1} - \bar{x}_k\rangle + \frac{L_G}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \frac{3L_G}{4}\|\bar{x}_{k+1} - \bar{x}_k\|_{\mathrm{F}}^2 \\ &= f(\bar{x}_k) + \langle \hat{g}_k, \hat{x}_{k+1} - \hat{x}_k\rangle + \langle \hat{g}_k, \bar{x}_{k+1} - \hat{x}_{k+1} + \hat{x}_k - \bar{x}_k\rangle + \frac{L_G}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \frac{3L_G}{4}\|\bar{x}_{k+1} - \bar{x}_k\|_{\mathrm{F}}^2. \end{aligned} \tag{E.8}$$

Note that for $\beta > 0$, we have

$$\langle \hat{g}_k, \bar{x}_{k+1} - \hat{x}_{k+1} + \hat{x}_k - \bar{x}_k\rangle \leq \frac{\beta^2 L_G}{2}\|\hat{g}_k\|_{\mathrm{F}}^2 + \frac{1}{\beta^2 L_G}\|\hat{x}_k - \bar{x}_k\|_{\mathrm{F}}^2 + \frac{1}{\beta^2 L_G}\|\bar{x}_{k+1} - \hat{x}_{k+1}\|_{\mathrm{F}}^2.$$

Plugging this into (E.8) yields

$$\begin{aligned} &f(\bar{x}_{k+1}) \\ &\leq f(\bar{x}_k) + \underbrace{\langle \hat{g}_k, \hat{x}_{k+1} - \hat{x}_k\rangle}_{:=b_1} + \frac{\beta^2 L_G}{2}\|\hat{g}_k\|_{\mathrm{F}}^2 + \underbrace{\frac{1}{\beta^2 L_G}(\|\hat{x}_k - \bar{x}_k\|_{\mathrm{F}}^2 + \|\bar{x}_{k+1} - \hat{x}_{k+1}\|_{\mathrm{F}}^2)}_{:=b_2} \\ &\quad + \frac{L_G}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \underbrace{\frac{3L_G}{4}\|\bar{x}_{k+1} - \bar{x}_k\|_{\mathrm{F}}^2}_{:=b_3}. \end{aligned} \tag{E.9}$$

Firstly, we have

$$\begin{aligned} b_1 &= \langle \hat{g}_k, \hat{x}_{k+1} - \hat{x}_k - \beta \hat{g}_k + \beta \hat{g}_k\rangle \\ &= -\beta\|\hat{g}_k\|_{\mathrm{F}}^2 + \left\langle \hat{g}_k, \frac{1}{n}\sum_{i=1}^n [x_{i,k+1} - (x_{i,k} - \beta v_{i,k} - \alpha \mathrm{grad}\varphi_i^t(\mathbf{x}_k))]\right\rangle \\ &\quad + \left\langle \hat{g}_k, \frac{1}{n}\sum_{i=1}^n [\beta(y_{i,k} - v_{i,k}) - \alpha \mathrm{grad}\varphi_i^t(\mathbf{x}_k)]\right\rangle. \end{aligned} \tag{E.10}$$

Since $y_{i,k} - v_{i,k} \in N_{x_{i,k}}\mathcal{M}$, it follows that

$$\begin{aligned}
&\left\langle \hat{g}_k, \frac{\beta}{n}\sum_{i=1}^{n}(y_{i,k} - v_{i,k}) - \alpha \mathrm{grad}\varphi_i^t(\mathbf{x}_k) \right\rangle \\
\overset{(C.8)}{\leq}\ & \frac{\beta}{n}\sum_{i=1}^{n} \langle \hat{g}_k - \mathrm{grad} f_i(x_{i,k}), y_{i,k} - v_{i,k} \rangle + \frac{2\alpha}{n}\|\hat{g}_k\|_{\mathrm{F}} \cdot \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 \\
\leq\ & \frac{1}{4nL_G}\sum_{i=1}^{n}\|\hat{g}_k - \mathrm{grad} f_i(x_{i,k})\|_{\mathrm{F}}^2 + \frac{\beta^2 L_G}{n}\sum_{i=1}^{n}\|\mathcal{P}_{N_{x_{i,k}}} y_{i,k}\|_{\mathrm{F}}^2 + \frac{2\alpha D}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 \\
\leq\ & \frac{1}{4n^2 L_G}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathrm{grad} f_j(x_{j,k}) - \mathrm{grad} f_i(x_{i,k})\|_{\mathrm{F}}^2 + \frac{\beta^2 L_G}{n}\|\mathbf{y}_k\|_{\mathrm{F}}^2 + \frac{2\alpha D}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 \\
\leq\ & \frac{L_G + 2\alpha D}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \frac{\beta^2 L_G}{n}\|\mathbf{y}_k\|_{\mathrm{F}}^2,
\end{aligned}$$

where we use Lemma 2.4 in the last inequality. This, together with (E.10) and (C.8) implies

$$\begin{aligned}
b_1 &\leq -\beta\|\hat{g}_k\|_{\mathrm{F}}^2 + \left\langle \hat{g}_k, \frac{1}{n}\sum_{i=1}^{n}[x_{i,k+1} - (x_{i,k} - \beta v_{i,k} - \alpha \mathrm{grad}\varphi_i^t(\mathbf{x}_k))] \right\rangle + \frac{L_G + 2D\alpha}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \frac{\beta^2 L_G}{n}\|\mathbf{y}_k\|_{\mathrm{F}}^2 \\
&\leq -\beta\|\hat{g}_k\|_{\mathrm{F}}^2 + \frac{D}{n}\sum_{i=1}^{n}\|x_{i,k} - \alpha \mathrm{grad}\varphi_i^t(\mathbf{x}_k) - \beta v_{i,k} - x_{i,k+1}\|_{\mathrm{F}} + \frac{L_G + 2D\alpha}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \frac{\beta^2 L_G}{n}\|\mathbf{y}_k\|_{\mathrm{F}}^2 \\
&\overset{(P1)}{\leq} -\beta\|\hat{g}_k\|_{\mathrm{F}}^2 + \frac{MD}{n}\sum_{i=1}^{n}\|\alpha \mathrm{grad}\varphi_i^t(\mathbf{x}_k) + \beta v_{i,k}\|_{\mathrm{F}}^2 + \frac{L_G + 2D\alpha}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \frac{\beta^2 L_G}{n}\|\mathbf{y}_k\|_{\mathrm{F}}^2 \\
&\overset{(C.8)}{\leq} -\beta\|\hat{g}_k\|_{\mathrm{F}}^2 + \frac{2MD\alpha^2}{n}\|\mathrm{grad}\varphi^t(\mathbf{x}_k)\|_{\mathrm{F}}^2 + \frac{2MD\beta^2}{n}\|\mathbf{y}_k\|_{\mathrm{F}}^2 + \frac{L_G + 2D\alpha}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \frac{\beta^2 L_G}{n}\|\mathbf{y}_k\|_{\mathrm{F}}^2 \\
&\overset{(C.9)}{\leq} -\beta\|\hat{g}_k\|_{\mathrm{F}}^2 + \frac{8MD\alpha^2 + 2D\alpha + L_G}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \frac{(2MD + L_G)\beta^2}{n}\|\mathbf{y}_k\|_{\mathrm{F}}^2,
\end{aligned} \tag{E.11}$$

where we use $\|\hat{g}_k\|_{\mathrm{F}} \leq D$.

Secondly, we use the following inequality to derive the upper bound of b_2. From Lemma 5.1, we have $\mathbf{x}_{k+1} \in \mathcal{N}$. One has

$$\begin{aligned}
&\|\bar{x}_k - \hat{x}_k\|_{\mathrm{F}}^2 + \|\bar{x}_{k+1} - \hat{x}_{k+1}\|_{\mathrm{F}}^2 \\
\overset{(P1)}{\leq}\ & \frac{4r}{n^2}(\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^4 + \|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|_{\mathrm{F}}^4).
\end{aligned} \tag{E.12}$$

We then obtain

$$b_2 \leq \frac{4r}{n^2\beta^2 L_G}(\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^4 + \|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|_{\mathrm{F}}^4).$$

Thirdly, invoking Lemma C.4 and $\alpha \leq 1/M$ yields

$$\|\bar{x}_k - \bar{x}_{k+1}\|_{\mathrm{F}} \leq \frac{1}{1 - 2\delta_1^2}\left[\frac{10\alpha}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \frac{2M\beta^2}{n}\|\mathbf{y}_k\|_{\mathrm{F}}^2 + \beta\|\hat{v}_k\|_{\mathrm{F}}\right].$$

Then, it follows from $\beta\|\mathbf{y}_k\|_{\mathrm{F}} \le \frac{\alpha\delta_1}{5}$ that

$$\begin{aligned}
b_3 &\le \frac{3L_G}{4}\left(\frac{2}{(1-2\delta_1^2)^2}\left[\frac{10\alpha}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \frac{2M\beta^2}{n}\|\mathbf{y}_k\|_{\mathrm{F}}^2\right]^2 + \frac{2}{(1-2\delta_1^2)^2}\beta^2\|\hat{v}_k\|_{\mathrm{F}}^2\right)\\
&\le \frac{3L_G}{(1-2\delta_1^2)^2}\left[\frac{100\alpha^2}{n^2}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^4 + \frac{(M\alpha\delta_1\beta)^2}{10n}\|\mathbf{y}_k\|_{\mathrm{F}}^2\right] + \frac{3L_G}{(1-2\delta_1^2)^2}\beta^2(\|\hat{y}_k\|_{\mathrm{F}}^2 + \|\hat{v}_k - \hat{y}_k\|_{\mathrm{F}}^2)\\
&\le \frac{3L_G}{(1-2\delta_1^2)^2}\left[\frac{100\alpha^2}{n^2}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^4 + \frac{(M\alpha\delta_1\beta)^2}{10n}\|\mathbf{y}_k\|_{\mathrm{F}}^2\right] + \frac{3L_G}{(1-2\delta_1^2)^2}\beta^2(\|\hat{g}_k\|_{\mathrm{F}}^2 + \frac{1}{n}\|\mathbf{y}_k\|_{\mathrm{F}}^2),
\end{aligned}$$

where we use $\hat{y}_k = \hat{g}_k$ and $\|\hat{v}_k - \hat{y}_k\|_{\mathrm{F}}^2 \le \frac{1}{n}\|\mathcal{P}_{N_{x_{i,k}}} y_{i,k}\|_{\mathrm{F}}^2 \le \frac{1}{n}\|\mathbf{y}_k\|_{\mathrm{F}}^2$. It follows from (5.1) that

$$\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 \le C_1(L_G + 2D)^2\beta^2 \le \frac{C_1\alpha^2\delta_1^2}{25},$$

where we use $\beta \le \frac{\alpha\delta_1}{5(L_G+2D)}$. Therefore, we get

$$b_2 \le \frac{4r(L_G+2D)^2C_1}{nL_G}(\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|_{\mathrm{F}}^2). \tag{E.13}$$

and

$$\begin{aligned}
b_3 &\le \frac{3L_G}{(1-2\delta_1^2)^2}\left[\frac{4C_1\delta_1^2\alpha^4}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \frac{\delta_1^2}{10n}\beta^2\|\mathbf{y}_k\|_{\mathrm{F}}^2\right] + \frac{3L_G}{(1-2\delta_1^2)^2}\beta^2(\|\hat{g}_k\|_{\mathrm{F}}^2 + \frac{1}{n}\|\mathbf{y}_k\|_{\mathrm{F}}^2)\\
&\le \frac{13L_GC_1\delta_1^2\alpha^4}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \frac{7}{2}L_G\beta^2\|\hat{g}_k\|_{\mathrm{F}}^2 + \frac{\frac{\delta_1^2}{2}+4}{n}L_G\beta^2\|\mathbf{y}_k\|_{\mathrm{F}}^2,
\end{aligned} \tag{E.14}$$

where we use $\alpha \le \frac{1}{M}$ and $\frac{1}{(1-2\delta_1^2)^2} \le 1.002$. Therefore, by combining the upper bound of b_1, b_2, b_3 with (E.9) implies

$$\begin{aligned}
f(\bar{x}_{k+1}) &\le f(\bar{x}_k) + b_1 + \frac{\beta^2L_G}{2}\|\hat{g}_k\|_{\mathrm{F}}^2 + b_2 + \frac{L_G}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + b_3\\
&\le f(\bar{x}_k) - (\beta - 4L_G\beta^2)\|\hat{g}_k\|_{\mathrm{F}}^2 + \frac{L_G + \frac{4r(L_G+2D)^2C_1}{L_G} + 2D\alpha + 8MD\alpha^2 + 13L_GC_1\delta_1^2\alpha^4}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2\\
&\quad + \frac{4r(L_G+2D)^2C_1}{nL_G}\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|_{\mathrm{F}}^2 + \frac{2MD + (\frac{\delta_1^2}{2}+5)L_G}{n}\beta^2\|\mathbf{y}_k\|_{\mathrm{F}}^2.
\end{aligned}$$

The proof is completed.

□

To proceed, we need the following recursive lemma, which is helpful to combine Lemma E.1 and Lemma E.2. It is a little different from the original one in (Xu et al., 2015). We only change $\sqrt{\sum_{l=0}^k u_i^2}$ and $\sqrt{\sum_{l=0}^k w_i^2}$ to be $\sum_{l=0}^k u_i^2$ and $\sum_{l=0}^k w_i^2$.

Lemma E.3. *(Xu et al., 2015, Lemma 2) Let $\{u_k\}_{k\ge0}$ and $\{w_k\}_{k\ge0}$ be two positive scalar sequences such that for all $k \ge 0$*

$$u_{k+1} \le \eta u_k + w_k,$$

where $\eta \in (0,1)$ is the decaying factor. Let $\Gamma(k) = \sum_{l=0}^k u_i^2$ and $\Omega(k) = \sum_{l=0}^k w_i^2$. Then we have

$$\Gamma(k) \le c_0\Omega(k) + c_1,$$

where $c_0 = \frac{2}{(1-\eta)^2}$ and $c_1 = \frac{2}{1-\eta^2}u_0^2$.

Proof of Theorem 5.2. Applying Lemma E.3 to (E.4) yields

$$\frac{1}{n}\sum_{k=0}^{K}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 \le \tilde{C}_0 \cdot \frac{\beta^2}{n}\sum_{k=0}^{K}\|\mathbf{y}_k\|_{\mathrm{F}}^2 + \tilde{C}_1, \tag{E.15}$$

where $\tilde{C}_0 = \frac{2}{(1-\rho_t)^2}$ and $\tilde{C}_1 = \frac{2}{1-\rho_t^2}\frac{1}{n}\|\mathbf{x}_0 - \bar{\mathbf{x}}_0\|_{\mathrm{F}}^2$.

It follows from Lemma E.2 that

$$
\begin{aligned}
& f(\bar{x}_{K+1}) \\
\leq\ & f(\bar{x}_0) - (\beta - 4L_G\beta^2)\sum_{k=0}^{K}\|\hat{g}_k\|_{\mathrm{F}}^2 + \frac{\mathcal{G}_1 L_G}{n}\sum_{k=0}^{K}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \frac{\mathcal{G}_0 L_G}{n}\sum_{k=1}^{K+1}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \frac{\mathcal{G}_2 L_G}{n}\beta^2\sum_{k=0}^{K}\|\mathbf{y}_k\|_{\mathrm{F}}^2 \\
\overset{\text{(E.15)}}{\leq}\ & f(\bar{x}_0) - (\beta - 4L_G\beta^2)\sum_{k=0}^{K}\|\hat{g}_k\|_{\mathrm{F}}^2 + (\mathcal{G}_1\tilde{C}_0 + \mathcal{G}_0\tilde{C}_0 + \mathcal{G}_2)\frac{L_G\beta^2}{n}\sum_{k=0}^{K}\|\mathbf{y}_k\|_{\mathrm{F}}^2 + \frac{\mathcal{G}_0\tilde{C}_0 L_G\beta^2\|\mathbf{y}_{K+1}\|_{\mathrm{F}}^2}{n} + \tilde{C}_1(\mathcal{G}_1 + \mathcal{G}_0)L_G \\
\leq\ & f(\bar{x}_0) - \frac{\beta}{2}\sum_{k=0}^{K}\|\hat{g}_k\|_{\mathrm{F}}^2 + \mathcal{G}_3\frac{L_G\beta^2}{n}\sum_{k=0}^{K}\|\mathbf{y}_k\|_{\mathrm{F}}^2 + \mathcal{G}_4 L_G,
\end{aligned}
\tag{E.16}
$$

where we use $\beta \leq \min\{\frac{1}{8L_G}, \frac{\alpha\delta_1}{5(L_G+2D)}\}$, $\beta^2\|\mathbf{y}_{K+1}\|_{\mathrm{F}}^2 \leq n(L_G+2D)^2\beta^2 \leq \frac{\delta_1^2\alpha^2 n}{25}$ and $\mathcal{G}_3 := \mathcal{G}_1\tilde{C}_0 + \mathcal{G}_0\tilde{C}_0 + \mathcal{G}_2$ and $\mathcal{G}_4 := \frac{\mathcal{G}_0\tilde{C}_0\delta_1^2\alpha^2}{25} + \tilde{C}_1(\mathcal{G}_1 + 4rC_1)$ in the last inequality.

We are going to associate $\|\hat{g}_k\|_{\mathrm{F}}^2$ with $\|\mathbf{y}_k\|_{\mathrm{F}}^2$. By (E.5), we get

$$
-\sum_{k=0}^{K}\|\hat{g}_k\|_{\mathrm{F}}^2 = -\frac{1}{n}\sum_{k=0}^{K}\|\hat{\mathbf{G}}_k\|_{\mathrm{F}}^2 \leq \frac{1}{n}\sum_{k=0}^{K}\|\mathbf{y}_k - \hat{\mathbf{G}}_k\|_{\mathrm{F}}^2 - \frac{1}{2n}\sum_{k=0}^{K}\|\mathbf{y}_k\|_{\mathrm{F}}^2
\tag{E.17}
$$

Again, applying Lemma E.3 to (E.3) yields

$$
\begin{aligned}
\frac{1}{n}\sum_{k=0}^{K}\|\mathbf{y}_k - \hat{\mathbf{G}}_k\|_{\mathrm{F}}^2 &\leq \tilde{C}_2\frac{1}{n}\sum_{k=0}^{K}\|\mathbf{G}_{k+1} - \mathbf{G}_k\|_{\mathrm{F}}^2 + \tilde{C}_3 \\
&\overset{\text{(E.2)}}{\leq} \tilde{C}_2\frac{1}{n}\sum_{k=0}^{K}(8\alpha^2 L_G^2\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 + 2\beta^2 L_G^2\|\mathbf{y}_k\|_{\mathrm{F}}^2) + \tilde{C}_3 \\
&\overset{\text{(E.15)}}{\leq} (8\alpha^2\tilde{C}_0\tilde{C}_2 + 2\tilde{C}_2)L_G^2\beta^2\frac{1}{n}\sum_{k=0}^{K}\|\mathbf{y}_k\|_{\mathrm{F}}^2 + 8\alpha^2\tilde{C}_1\tilde{C}_2 L_G^2 + \tilde{C}_3 \\
&\leq (8\tilde{C}_0 + \frac{1}{2}\tilde{C}_2)\alpha\delta_1 L_G\beta\frac{1}{n}\sum_{k=0}^{K}\|\mathbf{y}_k\|_{\mathrm{F}}^2 + 8\alpha^2\tilde{C}_1\tilde{C}_2 L_G^2 + \tilde{C}_3,
\end{aligned}
$$

where $\tilde{C}_2 = \frac{2}{(1-\sigma_2^t)^2}$ and $\tilde{C}_3 = \frac{2}{1-\sigma_2^{2t}} \cdot \frac{1}{n}\|\mathbf{y}_0 - \hat{\mathbf{G}}_0\|_{\mathrm{F}}^2$. The last line is due to $\beta \leq \frac{\alpha\delta_1}{5L_G}$ and $\alpha^2\tilde{C}_2 \leq \tilde{C}_2 \leq \frac{2}{(1-\frac{1}{2\sqrt{n}})^2} \leq 5$. Plugging this into (E.17) implies

$$
-\sum_{k=0}^{K}\|\hat{g}_k\|_{\mathrm{F}}^2 \leq \left[(8\tilde{C}_0 + \frac{1}{2}\tilde{C}_2)\alpha\delta_1 L_G\beta - \frac{1}{2}\right]\frac{1}{n}\sum_{k=0}^{K}\|\mathbf{y}_k\|_{\mathrm{F}}^2 + 8\alpha^2\tilde{C}_1\tilde{C}_2 L_G^2 + \tilde{C}_3.
\tag{E.18}
$$

Hence, it follows from equation (E.16) that

$$
\begin{aligned}
& f(\bar{x}_{K+1}) \\
\overset{\text{(E.18)}}{\leq}\ & f(\bar{x}_0) - \frac{\beta}{2}\left(\frac{1}{2} - \left[2\mathcal{G}_3 + (8\tilde{C}_0 + \frac{1}{2}\tilde{C}_2)\alpha\delta_1\right]L_G\beta\right)\frac{1}{n}\sum_{k=0}^{K}\|\mathbf{y}_k\|_{\mathrm{F}}^2 + \frac{\beta}{2}\left(8\alpha^2\tilde{C}_1\tilde{C}_2 L_G^2 + \tilde{C}_3\right) + \mathcal{G}_4 L_G \\
\leq\ & f(\bar{x}_0) - \frac{\beta}{8}\frac{1}{n}\sum_{k=0}^{K}\|\mathbf{y}_k\|_{\mathrm{F}}^2 + \frac{\beta}{2}\left(8\alpha^2\tilde{C}_1\tilde{C}_2 L_G^2 + \tilde{C}_3\right) + \mathcal{G}_4 L_G
\end{aligned}
\tag{E.19}
$$

where the last inequality is due to $\beta \leq \dfrac{1}{4L_G(2\mathcal{G}_3 + (8\tilde{C}_0 + \frac{1}{2}\tilde{C}_2)\alpha\delta_1)}$.

Then, we get

$$\frac{\beta}{8}\sum_{k=0}^{K}\|\hat{g}_k\|_{\mathrm{F}}^2 \leq \frac{\beta}{8}\cdot\frac{1}{n}\sum_{k=0}^{K}\|\mathbf{y}_k\|_{\mathrm{F}}^2 \leq f(\bar{x}_0) - f^* + \tilde{C}_4 + \mathcal{G}_4 L_G, \tag{E.20}$$

where $\tilde{C}_4 = (8\alpha^2\tilde{C}_1\tilde{C}_2L_G^2 + \tilde{C}_3)\frac{\beta}{2} = \mathcal{O}(\frac{r\delta_1^2 L_G}{(1-\sigma_2^t)^2})$ and $f^* = \min_{x\in\mathrm{St}(d,r)} f(x)$. This implies

$$\min_{k=0,\ldots,K}\|\hat{g}_k\|_{\mathrm{F}}^2 = \min_{k=0,\ldots,K}\|\hat{y}_k\|_{\mathrm{F}}^2 \leq \min_{k=0,\ldots,K}\frac{1}{n}\|\mathbf{y}_k\|_{\mathrm{F}}^2 \leq \frac{8(f(\bar{x}_0) - f^* + \tilde{C}_4 + \mathcal{G}_4 L_G)}{\beta\cdot K}. \tag{E.21}$$

It then follows from (E.15) that

$$\min_{k=0,\ldots,K}\frac{1}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 \leq \frac{8\beta(f(\bar{x}_0) - f^* + \tilde{C}_4 + \mathcal{G}_4 L_G)\tilde{C}_0 + \tilde{C}_1}{K}.$$

Finally, noticing $\beta \leq \frac{\alpha\delta_1}{5L_G}$ and

$$\|\mathrm{grad} f(\bar{x}_k)\|_{\mathrm{F}}^2 \leq 2\|\hat{g}_k\|_{\mathrm{F}}^2 + 2\|\mathrm{grad} f(\bar{x}_k) - \hat{g}_k\|_{\mathrm{F}}^2 \leq 2\|\hat{g}_k\|_{\mathrm{F}}^2 + \frac{2L_G^2}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2.$$

We finally have

$$\min_{k=0,\ldots,K}\|\mathrm{grad} f(\bar{x}_k)\|_{\mathrm{F}}^2 \leq \frac{(16 + \alpha^2\delta_1^2\tilde{C}_0)(f(\bar{x}_0) - f^* + \tilde{C}_4 + \mathcal{G}_4 L_G) + \tilde{C}_1 L_G}{\beta\cdot K}.$$

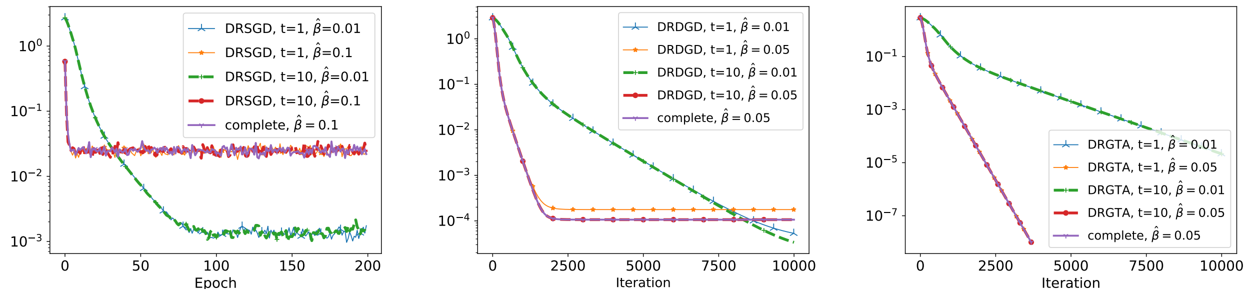
The proof is completed. □

F. Supplementary numerical results

We report the numerical results on different networks and data size in this section.

F.1. Synthetic data

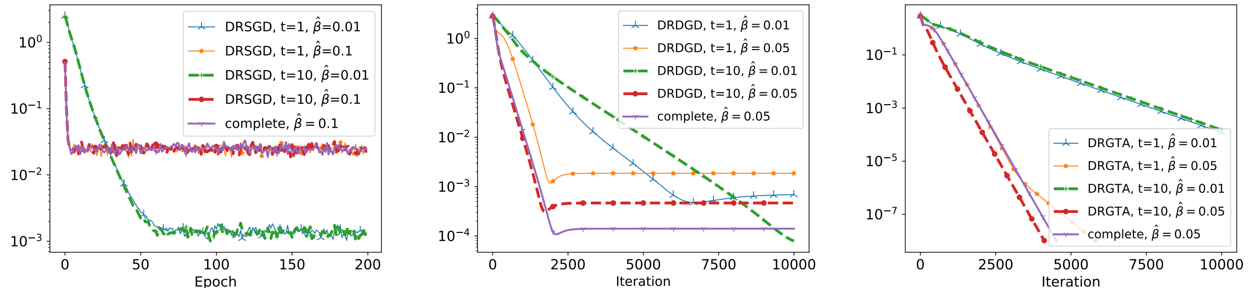
Figure 3 shows the results on the same data set as that of Figure 1. However, the network is an Erdös-Rényi model $\mathsf{ER}(n,p)$, which means the probability of each edge is included in the graph with probability p. The Metropolis constant matrix is associated with the graph. Since the $\mathsf{ER}(32,0.3)$ is more well-connected than the ring graph, we see that the results for different $t \in \{1, 10, \infty\}$ are almost the same except for DRDGD with $\hat{\beta} = 0.05$. Moreover, the solutions accuracy and convergence rate of DRDGD and DRGTA are better than those shown in Figure 1.



(a) DRSGD (b) DRDGD (c) DRGTA

Figure 3. Synthetic data, agents number $n = 32$, eigengap $\Delta = 0.8$, Graph: $\mathsf{ER}(32, 0.3)$.

In Figure 4, we show the results when the initial point does not satisfy $\mathbf{x}_0 \in \mathcal{N}$. Specifically, we randomly generate $x_{1,0},\ldots,x_{n,0}$ on $\mathcal{M}$, and the other settings are the same as Figure 1. Surprisingly, we find that the proposed algorithms still converge. As suggested by (Markdahl et al., 2020; Chen et al., 2021), the consensus algorithm can achieve global consensus with random initialization when $r \leq \frac{2}{3}d - 1$. The iteration in DRSGD and DRGTA is a perturbation of the consensus iteration. It will be interesting to study it further.



(a) DRSGD (b) DRDGD (c) DRGTA

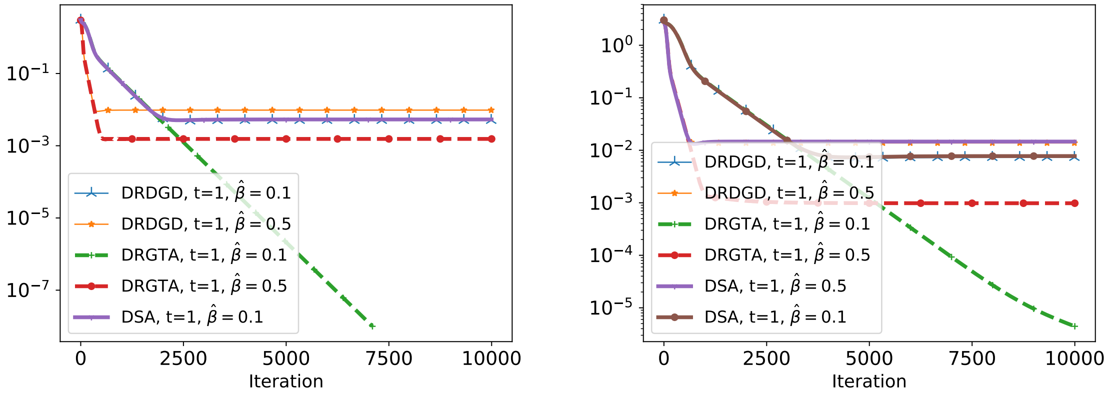
Figure 4. Synthetic data, agents number $n = 32$, eigengap $\Delta = 0.8$, Graph: Ring.

F.2. Real-world data

We compare our algorithms with a recently proposed algorithm decentralized Sanger's algorithm (DSA) (Gang & Bajwa, 2021), which is a Euclidean-type algorithm. To solve the eigenvector problem (6.1), DSA is shown to converge linearly to a neighborhood of the optimal solution. The computation of DSA iteration is cheaper than DRDGD since there is no retraction step. For simplicity, we fix $t = 1$ and $r = 5$ in this section.

We provide some numerical results on the MNIST dataset(LeCun). The graph is still the ring and W is the Metropolis constant weight matrix. The data set is evenly partitioned into n subsets. The stepsizes of DRDGD and DRGTA are set to $\beta = \frac{\hat{\beta}}{60000}$.

The results for MNIST data set with $n = 20, 40$ are shown in Figure 5. We see that the convergence rate of DSA and DRDGD are almost the same and DRGTA with $\hat{\beta} = 0.1$ can achieve the most accurate solution. When n becomes larger, the convergence rate of all algorithms is slower. Although the computation of DSA is cheaper than DRDGD, we find that when $\hat{\beta} = 0.5, n = 20$, DSA does not converge, which is not shown in the Figure 5 (a). This is probably because DSA is not a feasible method and needs carefully tuned stepsize.

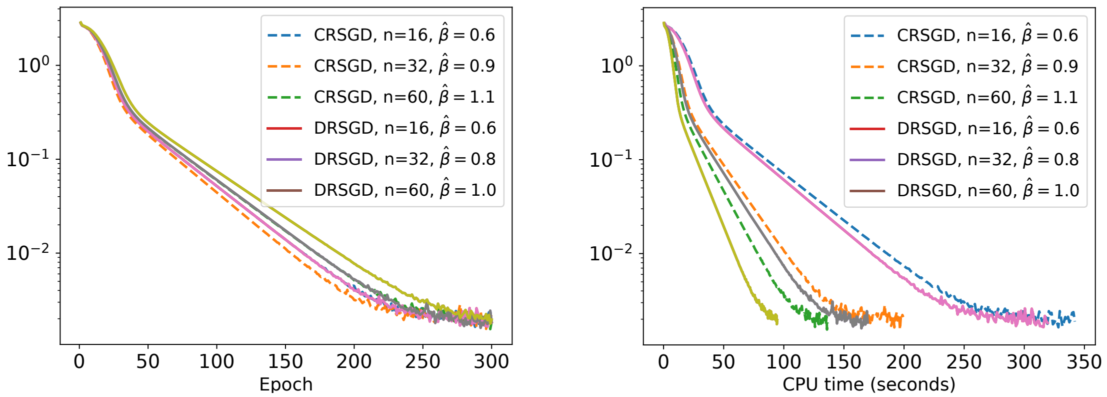


(a) MNIST, $n = 20$, ring graph (b) MNIST, $n = 40$, ring graph

Figure 5. Numerical results of DRDGD, DRGTA, DSA on MNIST data set.

Finally, we demonstrate that DRSGD is indeed faster than the centralized Riemannian stochastic gradient descent(CRSGD). We implemented the standard parameter server-based synchronous CRSGD using mpi4py. One node will serve as the parameter server in our implementation. In Figure 6, we show the comparison results between DRSGD and CRSGD. The settings of DRSGD are the same as those in Figure 2. And the stepsize of CRSGD is given by $\beta = \frac{\sqrt{n}}{10000\sqrt{300}}\hat{\beta}$. For different number of nodes, we tuned the stepsize $\hat{\beta}$ to get the best one.

We see that convergence rate of DRSGD and CRSGD are the same w.r.t the epoch number in 6 (a). This means that they can both achieve linear speedup. In Figure 6 (b), DRSGD is faster than CRSGD in CPU time. This is because DRSGD needs fewer number of communications between nodes than CRSGD.



(a) Epoch

(b) time

Figure 6. Comparison results of different number of nodes on MNIST. Ring graph associated with Metropolis constant weight matrix, $t = 1$, $\beta = \frac{\sqrt{n}}{10000\sqrt{300}}\hat{\beta}$.