A. Preliminary

Notation. We first introduce necessary notations as follows.

- $\mathbf{x}^{(k)} = [(x_1^{(k)})^T; (x_2^{(k)})^T; \cdots; (x_n^{(k)})^T] \in \mathbb{R}^{n\times d}$
- $\nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k)}) = [\nabla F_1(x_1^{(k)}; \xi_1^{(k)})^T; \cdots; \nabla F_n(x_n^{(k)}; \xi_n^{(k)})^T] \in \mathbb{R}^{n\times d}$
- $\nabla f(\mathbf{x}^{(k)}) = [\nabla f_1(x_1^{(k)})^T; \nabla f_2(x_2^{(k)})^T; \cdots; \nabla f_n(x_n^{(k)})^T] \in \mathbb{R}^{n\times d}$
- $f(\mathbf{x}^{(k)}) = \sum_{i=1}^n f(x_i^{(k)})$
- $\bar{\mathbf{x}}^{(k)} = [(\bar{x}^{(k)})^T; (\bar{x}^{(k)})^T; \cdots; (\bar{x}^{(k)})^T] \in \mathbb{R}^{n\times d}$ where $\bar{x}^{(k)} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^{(k)}$
- $\mathbf{x}^\star = [(x^\star)^T; (x^\star)^T; \cdots; (x^\star)^T] \in \mathbb{R}^{n\times d}$ where $x^\star$ is the global solution to problem (1).
- $W = [w_{ij}] \in \mathbb{R}^{n\times n}$ is the weight matrix.
- $\mathbb{1}_n = \mathrm{col}\{1, 1, \cdots, 1\} \in \mathbb{R}^n$.
- Given two matrices $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n\times d}$, we define inner product $\langle \mathbf{x}, \mathbf{y}\rangle = \mathrm{tr}(\mathbf{x}^T\mathbf{y})$, the Frobenius norm $\|\mathbf{x}\|^2 = \langle \mathbf{x}, \mathbf{x}\rangle$, and the $\|\mathbf{x}\|_2$ as $\mathbf{x}$'s ℓ_2 norm. Furthermore, for a positive semi-definite matrix $A \in \mathbb{R}^{n\times n}$, we define $\langle \mathbf{x}, \mathbf{y}\rangle_A = \mathrm{tr}(\mathbf{x}^T A\mathbf{y})$ and $\|\mathbf{x}\|_A^2 = \langle \mathbf{x}, \mathbf{x}\rangle_A$ for simplicity.
- Given $W \in \mathbb{R}^{n\times n}$, we let $\|W\|_2 = \sigma_{\max}(W)$ where $\sigma_{\max}(\cdot)$ denote the maximum sigular value.

GmSGD in matrix notation. For ease of analysis, we rewrite the recursion of GmSGD in Algorithm 1 with matrix notation:

$$\mathbf{m}^{(k+1)} = \beta\boldsymbol{m}^{(k)} + \nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k)}) \tag{14}$$

$$\mathbf{x}^{(k+1)} = W(\mathbf{x}^{(k)} - \gamma\mathbf{m}^{(k+1)}) \tag{15}$$

DecentLaM in matrix notation. We can also rewrite DecentLaM in Algorithm 2 with matrix notation:

$$\tilde{\mathbf{g}}^{(k)} = \frac{1}{\gamma}\mathbf{x}^{(k)} - \frac{1}{\gamma}W(\mathbf{x}^{(k)} - \gamma\nabla F(\mathbf{x}^{(k)}, \boldsymbol{\xi}^{(k)})) \tag{16}$$

$$\mathbf{m}^{(k+1)} = \beta\mathbf{m}^{(k)} + \tilde{\mathbf{g}}^{(k)} \tag{17}$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \gamma\mathbf{m}^{(k+1)} \tag{18}$$

Moreover, we define $\mathbf{g}^{(k)} = \mathbb{E}[\tilde{\mathbf{g}}^{(k)}] = \frac{1}{\gamma}\mathbf{x}^{(k)} - \frac{1}{\gamma}W(\mathbf{x}^{(k)} - \gamma\nabla f(\mathbf{x}^{(k)}))$

Smoothness. Since each $f_i(x)$ is assumed to be L-smooth in Assumption A.1, it holds that $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ is also L-smooth. As a result, the following inequality holds for any $x, y \in \mathbb{R}^d$:

$$f_i(x) - f_i(y) - \frac{L}{2}\|x - y\|^2 \le \langle \nabla f_i(y), x - y\rangle \tag{19}$$

Network weight matrix. Suppose a symmetric matrix $W \in \mathbb{R}^{n\times n}$ satisfies Assumption A.3, and λ_j denotes its j-th largest eigenvalue. It holds that $1 = \lambda_1 > \lambda_2 \ge \cdots \ge \lambda_n > -1$. As a result, it holds that

$$\|W\|_2 = 1, \quad \text{and} \quad \|I - W\|_2 \le 1 - \lambda_n. \tag{20}$$

If W satisfying Assumption A.3 is further assumed to be positive-definite, it holds that $1 = \lambda_1 > \lambda_2 \ge \cdots \ge \lambda_n > 0$.

Submultiplicativity of the Frobenius norm. Given matrices $W \in \mathbb{R}^{n\times n}$ and $\mathbf{y} \in \mathbb{R}^{n\times d}$, it holds that

$$\|W\mathbf{y}\| \le \|W\|_2\|\mathbf{y}\|. \tag{21}$$

To verify it, by letting y_j be the j-th column of $\mathbf{y}$, we have $\|W\mathbf{y}\|^2 = \sum_{j=1}^d \|Wy_j\|_2^2 \le \sum_{j=1}^d \|W\|_2^2\|y_j\|_2^2 = \|W\|_2^2\|\mathbf{y}\|^2$.

B. Reformulation of DmSGD and DecentLaM

B.1. Reformulation of DmSGD

In this section we show how DmSGD algorithm 1 can be rewritten as (6). To this end, we rewrite (15) as

$$\beta \mathbf{x}^{(k)} = W(\beta \mathbf{x}^{(k-1)} - \gamma \beta \mathbf{m}^{(k)}). \tag{22}$$

Subtracting (22) from (15), we have

$$\mathbf{x}^{(k+1)} - \beta \mathbf{x}^{(k)} = W\big(\mathbf{x}^{(k)} - \beta \mathbf{x}^{(k-1)} - \gamma(\mathbf{m}^{(k+1)} - \beta \mathbf{m}^{(k)})\big) \overset{(14)}{=} W\big(\mathbf{x}^{(k)} - \beta \mathbf{x}^{(k-1)} - \gamma \nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k)})\big) \tag{23}$$

which is equivalent to

$$\mathbf{x}^{(k+1)} = \underbrace{W\big(\mathbf{x}^{(k)} - \gamma \nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k)})\big)}_{\text{DSGD}} + \underbrace{\beta(\mathbf{x}^{(k)} - W\mathbf{x}^{(k-1)})}_{\text{momentum}}. \tag{24}$$

When a full-batch gradient is used, the above recursion becomes

$$\mathbf{x}^{(k+1)} = W\big(\mathbf{x}^{(k)} - \gamma \nabla f(\mathbf{x}^{(k)})\big) + \beta(\mathbf{x}^{(k)} - W\mathbf{x}^{(k-1)}) \tag{25}$$

which is essentially recursion (6) in the matrix notation.

B.2. Reformulation of DecentLaM

In this section we show the equivalence between Algorithm 2 and recursion (8). To this end, we rewrite (18) as

$$\beta \mathbf{x}^{(k)} = \beta \mathbf{x}^{(k-1)} - \gamma \beta \mathbf{m}^{(k)}. \tag{26}$$

Subtracting (26) from (18), we have

$$\begin{aligned} \mathbf{x}^{(k+1)} - \beta \mathbf{x}^{(k)} &= \mathbf{x}^{(k)} - \beta \mathbf{x}^{(k-1)} - \gamma(\mathbf{m}^{(k+1)} - \beta \mathbf{m}^{(k)}) \\ &\overset{(17)}{=} \mathbf{x}^{(k)} - \beta \mathbf{x}^{(k-1)} - \gamma \tilde{\boldsymbol{g}}^{(k)} \\ &\overset{(16)}{=} W(\mathbf{x}^{(k)} - \gamma \nabla F(\mathbf{x}^{(k)}, \boldsymbol{\xi}^{(k)})) - \beta \mathbf{x}^{(k-1)} \end{aligned} \tag{27}$$

which is equivalent to

$$\mathbf{x}^{(k+1)} = \underbrace{W(\mathbf{x}^{(k)} - \gamma \nabla F(\mathbf{x}^{(k)}, \boldsymbol{\xi}^{(k)}))}_{\text{DSGD}} + \underbrace{\beta(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})}_{\text{momentum}} \tag{28}$$

When a full-batch gradient is used, the above recursion becomes

$$\mathbf{x}^{(k+1)} = W\big(\mathbf{x}^{(k)} - \gamma \nabla f(\mathbf{x}^{(k)})\big) + \beta(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}) \tag{29}$$

which is essentially recursion (8) in the matrix notation.

C. Limiting Bias of Decentralized Algorithms

C.1. Limiting bias of DSGD

In this section we illustrate the stochastic bias and inconsistency bias in the DSGD algorithm. It is established in [57] that DSGD in the strongly-convex scenario will converge as follows:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\|x_i^{(k)} - x^\star\|^2 = O\Big(\underbrace{(1-\gamma\mu)^k}_{\text{convg. rate}} + \underbrace{\gamma\sigma^2}_{\text{sto. bias}} + \underbrace{\gamma^2 b^2}_{\text{inconsis. bias}}\Big). \tag{30}$$

where σ^2 is the variance of gradient noise, and b^2 is the data inconsistency (see the definition in Proposition 2). When learning rate is constant and iteration k goes to infinity, DSGD will converge with limiting bias. i.e.,

$$\text{Limiting bias} = \limsup_{k\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\|x_i^{(k)} - x^\star\|^2 = O\Big(\underbrace{\gamma\sigma^2}_{\text{sto. bias}} + \underbrace{\gamma^2 b^2}_{\text{inconsis. bias}} \Big) \tag{31}$$

As we discussed in Sec. 4, the limiting bias can be divided into two categories: stochastic bias and inconsistency bias. The stochastic bias is caused by the gradient noise. In the large-batch scenario in which the gradient noise σ^2 gets significantly reduced, the inconsistency bias will dominate the magnitude of DSGD's limiting bias.

C.2. Inconsistency bias of DmSGD (Proof of Proposition 2)

In this section we will prove Proposition 2. To achieve the inconsistency bias, we let $\mathbf{x}_{\mathrm{m}}$ be the fixed point of $\mathbf{x}^{(k)}$, i.e., $\mathbf{x}^{(k)} \to \mathbf{x}_{\mathrm{m}}$. From recursion (25), it is derived that $\mathbf{x}_{\mathrm{m}}$ satisfies

$$(1-\beta)(I-W)\mathbf{x}_{\mathrm{m}} = -\gamma W \nabla f(\mathbf{x}_{\mathrm{m}}). \tag{32}$$

Bound of $\|\mathbf{x}_{\mathrm{m}} - \bar{\mathbf{x}}_{\mathrm{m}}\|$. Letting $\bar{\mathbf{x}}_{\mathrm{m}} = \frac{1}{n}\mathbb{1}\mathbb{1}^T\mathbf{x}_{\mathrm{m}}$, it holds that $(I-W)\bar{\mathbf{x}}_{\mathrm{m}} = 0$ because $W\mathbb{1} = \mathbb{1}$ (see Assumption A.3). Substituting $(I-W)\bar{\mathbf{x}}_{\mathrm{m}} = 0$ into (32), we have

$$(1-\beta)(I-W)(\mathbf{x}_{\mathrm{m}} - \bar{\mathbf{x}}_{\mathrm{m}}) = -\gamma W \nabla f(\mathbf{x}_{\mathrm{m}}). \tag{33}$$

Since W is symmetric and satisfies $W\mathbb{1} = \mathbb{1}$ (see Assumption A.3), we can eigen-decompose it as

$$W = \underbrace{[\frac{1}{\sqrt{n}}\mathbb{1} \quad U_1]}_{U} \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & \Lambda_1 \end{bmatrix}}_{\Lambda} \underbrace{\begin{bmatrix} \frac{1}{\sqrt{n}}\mathbb{1}^T \\ U_1^T \end{bmatrix}}_{U^T} \tag{34}$$

where U is the orthonormal matrix, and $\Lambda_1 = \mathrm{diag}\{\lambda_2, \cdots, \lambda_n\}$ is a diagonal matrix. With (34), we have

$$\begin{aligned}
\|(I-W)(\mathbf{x}_{\mathrm{m}} - \bar{\mathbf{x}}_{\mathrm{m}})\|^2 &= \|U(I-\Lambda)U^T(\mathbf{x}_{\mathrm{m}} - \bar{\mathbf{x}}_{\mathrm{m}})\|^2 \\
&\overset{(a)}{=} \|(I-\Lambda)U^T(\mathbf{x}_{\mathrm{m}} - \bar{\mathbf{x}}_{\mathrm{m}})\|^2 \\
&\overset{(b)}{=} \|(I-\Lambda_1)U_1^T(\mathbf{x}_{\mathrm{m}} - \bar{\mathbf{x}}_{\mathrm{m}})\|^2 \\
&\geq (1-\lambda_2)^2\|U_1^T(\mathbf{x}_{\mathrm{m}} - \bar{\mathbf{x}}_{\mathrm{m}})\|^2 \\
&\overset{(c)}{=} (1-\lambda_2)^2\|U^T(\mathbf{x}_{\mathrm{m}} - \bar{\mathbf{x}}_{\mathrm{m}})\|^2 \\
&\overset{(d)}{=} (1-\lambda_2)^2\|\mathbf{x}_{\mathrm{m}} - \bar{\mathbf{x}}_{\mathrm{m}}\|^2
\end{aligned} \tag{35}$$

where (a) and (d) hold because U is orthonormal, and (b) and (c) hold because $\|U^T(\mathbf{x}_{\mathrm{m}} - \bar{\mathbf{x}}_{\mathrm{m}})\|^2 \overset{(34)}{=} \|\frac{1}{\sqrt{n}}\mathbb{1}^T(\mathbf{x}_{\mathrm{m}} - \bar{\mathbf{x}}_{\mathrm{m}})\|^2 + \|U_1^T(\mathbf{x}_{\mathrm{m}} - \bar{\mathbf{x}}_{\mathrm{m}})\|^2 = \|U_1^T(\mathbf{x}_{\mathrm{m}} - \bar{\mathbf{x}}_{\mathrm{m}})\|^2$. With (33) and (35), we have

$$\begin{aligned}
(1-\beta)(1-\lambda_2)\|\mathbf{x}_{\mathrm{m}} - \bar{\mathbf{x}}_{\mathrm{m}}\| &\leq (1-\beta)\|(I-W)(\mathbf{x}_{\mathrm{m}} - \bar{\mathbf{x}}_{\mathrm{m}})\| \\
&= \gamma\|W\nabla f(\mathbf{x}_{\mathrm{m}})\| \\
&\overset{(20)}{\leq} \gamma\|\nabla f(\mathbf{x}_{\mathrm{m}})\| \\
&\leq \gamma\|\nabla f(\mathbf{x}_{\mathrm{m}}) - \nabla f(\bar{\mathbf{x}}_{\mathrm{m}})\| + \gamma\|\nabla f(\bar{\mathbf{x}}_{\mathrm{m}}) - \nabla f(\mathbf{x}^\star)\| + \gamma\|\nabla f(\mathbf{x}^\star)\| \\
&\leq \gamma L\|\mathbf{x}_{\mathrm{m}} - \bar{\mathbf{x}}_{\mathrm{m}}\| + \sqrt{n}\gamma L\|\bar{x}_m - x^\star\| + \sqrt{n}\gamma b
\end{aligned} \tag{36}$$

where $x^\star$ is the global solution to problem (1), $\bar{x}_{\mathrm{m}} = \frac{1}{n}\mathbb{1}^T\mathbf{x}_{\mathrm{m}}$, and $b^2 = \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^\star)\|^2$.

Bound of $\|\bar{x}_m - x^\star\|$. Left-multiplying $\frac{1}{n}\mathbb{1}^T$ to both sides of (32), we achieve $\frac{1}{n}\mathbb{1}^T\nabla f(\mathbf{x}_{\mathrm{m}}) = 0$. With this fact we have

$$\|\bar{x}_{\mathrm{m}} - x^\star\| = \|\bar{x}_{\mathrm{m}} - x^\star - \gamma\big(\frac{1}{n}\mathbb{1}^T\nabla F(\mathbf{x}_{\mathrm{m}}) - \frac{1}{n}\mathbb{1}^T\nabla F(\mathbf{x}^\star)\big)\|$$

$$
\begin{aligned}
&= \|\bar{x}_{\rm m} - x^\star - \gamma\big(\frac{1}{n}\mathbb{1}^T\nabla F(\bar{\mathbf{x}}_{\rm m}) - \frac{1}{n}\mathbb{1}^T\nabla F(\mathbf{x}^\star)\big)\| + \gamma\|\frac{1}{n}\mathbb{1}^T\nabla F(\mathbf{x}_{\rm m}) - \frac{1}{n}\mathbb{1}^T\nabla F(\bar{\mathbf{x}}_{\rm m})\| \\
&\overset{(a)}{\le} (1 - \frac{\gamma\mu}{2})\|\bar{x}_{\rm m} - x^\star\| + \frac{\gamma L}{\sqrt{n}}\|\mathbf{x}_{\rm m} - \bar{\mathbf{x}}_{\rm m}\| \qquad (37)
\end{aligned}
$$

where (a) holds because $\frac{1}{n}\sum_{i=1}^n f_i(x)$ is L-smooth and μ-strongly convex (see Assumptions A.1 and A.4). We thus have

$$
\sqrt{n}\|\bar{x}_{\rm m} - x^\star\| \le \frac{2L}{\mu}\|\mathbf{x}_{\rm m} - \bar{\mathbf{x}}_{\rm m}\|. \qquad (38)
$$

Proof of Proposition 2. Substituting (38) to (36), we achieve

$$
\begin{aligned}
(1-\beta)(1-\lambda_2)\|\mathbf{x}_{\rm m} - \bar{\mathbf{x}}_{\rm m}\| &\le \Big(\gamma L + \frac{2\gamma L^2}{\mu}\Big)\|\mathbf{x}_{\rm m} - \bar{\mathbf{x}}_{\rm m}\| + \sqrt{n}\gamma b \\
&\le \frac{3\gamma L^2}{\mu}\|\mathbf{x}_{\rm m} - \bar{\mathbf{x}}_{\rm m}\| + \sqrt{n}\gamma b \qquad (39)
\end{aligned}
$$

If $\gamma \le \frac{\mu(1-\beta)(1-\lambda)}{6L^2}$, the above inequality becomes

$$
\|\mathbf{x}_{\rm m} - \bar{\mathbf{x}}_{\rm m}\| \le \frac{2\sqrt{n}\gamma b}{(1-\beta)(1-\lambda_2)}. \qquad (40)
$$

With (38) and (40), we have

$$
\|\mathbf{x}_{\rm m} - \mathbf{x}^\star\| \le \|\mathbf{x}_{\rm m} - \bar{\mathbf{x}}_{\rm m}\| + \sqrt{n}\|\bar{x}_{\rm m} - x^\star\| \le (1 + \frac{2L}{\mu})\frac{2\sqrt{n}\gamma b}{(1-\beta)(1-\lambda_2)}, \qquad (41)
$$

which leads to

$$
\lim_{k\to\infty}\frac{1}{n}\sum_{i=1}^n\|x_i^{(k)} - x^\star\|^2 = \frac{1}{n}\|\mathbf{x}_{\rm m} - \mathbf{x}^\star\|^2 \overset{(41)}{=} O\Big(\frac{\gamma^2 b^2}{(1-\beta)^2}\Big). \qquad (42)
$$

This concludes the proof of Proposition 2.

C.3. Inconsistency bias of DecentLaM (Proof of Proposition 3)

We let $\mathbf{x}_{\rm L}$ be the fixed point of the DecentLaM iterate $\mathbf{x}^{(k)}$, i.e., $\mathbf{x}^{(k)} \to \mathbf{x}_{\rm L}$. From DecentLaM recursion (29), we have

$$
(I - W)\mathbf{x}_{\rm L} = -\gamma W\nabla f(\mathbf{x}_{\rm L}). \qquad (43)
$$

By following the arguments in (33)–(42), we can prove Proposition 3.

D. Fundamental Supporting Lemmas

In this section, we establish some key lemmas to facilitate the convergence analysis in Appendices E and F. This section assumes the weight matrix W to be positive-definite to simplify the derivations.

Define $\mathcal{F}(\mathbf{s}) = f(W^{\frac{1}{2}}\mathbf{s}) + \frac{1}{2\gamma}\|\mathbf{s}\|_{I-W}^2$ and let $\mathbf{x} = W^{\frac{1}{2}}\mathbf{s}$. It follows from the chain rule that:

$$
\nabla_{\mathbf{s}}\mathcal{F}(\mathbf{s}, \boldsymbol{\xi}) = W^{\frac{1}{2}}\nabla F(\mathbf{x}, \boldsymbol{\xi}) + \frac{1}{\gamma}(I - W)W^{-\frac{1}{2}}\mathbf{x} = W^{-\frac{1}{2}}\tilde{\mathbf{g}}. \qquad (44)
$$

where $\tilde{\mathbf{g}} = \frac{1}{\gamma}\mathbf{x} - \frac{1}{\gamma}W(\mathbf{x} - \gamma\nabla F(\mathbf{x}, \boldsymbol{\xi}))$. Substituting (44) and $\mathbf{x} = W^{\frac{1}{2}}\mathbf{s}$ into (17) and (18), we achieve

$$
\begin{aligned}
&\mathbf{m}^{(k+1)} = \beta\mathbf{m}^{(k)} + W^{\frac{1}{2}}\nabla_{\mathbf{s}}\mathcal{F}(\mathbf{s}^{(k)}, \boldsymbol{\xi}^{(k)}) \qquad (45) \\
&W^{\frac{1}{2}}\mathbf{s}^{(k+1)} = W^{\frac{1}{2}}\mathbf{s}^{(k)} - \gamma\mathbf{m}^{(k+1)}. \qquad (46)
\end{aligned}
$$

If we introduce $\mathbf{m}_{\mathbf{s}}^{(k+1)} = (1-\beta)W^{-\frac{1}{2}}\mathbf{m}^{(k+1)}$, the above recursions (45) and (46) are equivalent to

$$\mathbf{m}_{\mathbf{s}}^{(k+1)} = \beta\mathbf{m}_{\mathbf{s}}^{(k)} + (1-\beta)\nabla_{\mathbf{s}}\mathcal{F}(\mathbf{s}^{(k)}, \boldsymbol{\xi}^{(k)}) \tag{47}$$

$$\mathbf{s}^{(k+1)} = \mathbf{s}^{(k)} - \frac{\gamma}{1-\beta}\mathbf{m}_{\mathbf{s}}^{(k+1)} \tag{48}$$

The new recursions (47) and (48) imply that DecentLaM (16)–(18) can be interpreted as the standard momentum stochastic gradient descent method to solve the optimization problem $\min_{\mathbf{s}} \mathcal{F}(\mathbf{s})$. Such interpretation will be critical to establish the convergence analysis for DecentLaM. To proceed, we need to characterize the smoothness and strong convexity of $\mathcal{F}(\mathbf{s})$.

Lemma 1. *a. If f_i is L-smooth ($1 \le i \le n$), then $\mathcal{F}(\mathbf{s})$ is $L' \triangleq L + \frac{1}{\gamma}(1-\lambda_n)$-smooth with respect to* $\mathbf{s}$.
b. If f_i is μ-strongly convex ($1 \le i \le n$), then $\mathcal{F}(\mathbf{s})$ is $\mu' \triangleq \min\{\mu, \frac{1}{\gamma}\}$-strongly convex with respect to $\mathbf{s}$.

Proof. Since f_i ($1 \le i \le n$) are L-smooth, it holds that $f(\mathbf{x})$ is also L-smooth in terms of $\mathbf{x}$. With transformations $\mathbf{x} = W^{\frac{1}{2}}\mathbf{s}$ and $\mathbf{x}' = W^{\frac{1}{2}}\mathbf{s}'$, we have

$$\begin{aligned}
&\mathcal{F}(\mathbf{s}') - \mathcal{F}(\mathbf{s}) - \langle\nabla_{\mathbf{s}}\mathcal{F}(\mathbf{s}), \mathbf{s}'-\mathbf{s}\rangle \\
=&\Big(f(W^{\frac{1}{2}}\mathbf{s}') - f(W^{\frac{1}{2}}\mathbf{s}) - \langle W^{\frac{1}{2}}(\nabla_{\mathbf{x}}f(W^{\frac{1}{2}}\mathbf{s}) - \nabla_{\mathbf{x}'}f(W^{\frac{1}{2}}\mathbf{s}')), \mathbf{s}'-\mathbf{s}\rangle\Big) \\
&+ \frac{1}{\gamma}\left(\frac{1}{2}\|\mathbf{s}'\|_{I-W}^2 - \frac{1}{2}\|\mathbf{s}\|_{I-W}^2 - \langle (I-W)\mathbf{s}, \mathbf{s}'-\mathbf{s}\rangle\right) \\
=&\Big(f(W^{\frac{1}{2}}\mathbf{s}') - f(W^{\frac{1}{2}}\mathbf{s}) - \langle(\nabla_{\mathbf{x}}f(W^{\frac{1}{2}}\mathbf{s}) - \nabla_{\mathbf{x}'}f(W^{\frac{1}{2}}\mathbf{s}')), W^{\frac{1}{2}}\mathbf{s}' - W^{\frac{1}{2}}\mathbf{s}\rangle\Big) + \frac{1}{2\gamma}\|\mathbf{s}'-\mathbf{s}\|_{I-W}^2 \\
\le&\frac{L}{2}\|W^{\frac{1}{2}}\mathbf{s}' - W^{\frac{1}{2}}\mathbf{s}\|^2 + \frac{1}{2\gamma}\|\mathbf{s}'-\mathbf{s}\|_{I-W}^2 \\
\le&\frac{L'}{2}\|\mathbf{s}'-\mathbf{s}\|^2
\end{aligned}$$

which leads to the conclusion that $\mathcal{F}(\mathbf{s})$ is L'-smooth.

Similarly, since f_i ($1 \le i \le n$) is μ-strongly convex, it holds that $f(\mathbf{x})$ is also μ-strongly convex in terms of $\mathbf{x}$. We have

$$\begin{aligned}
&\mathcal{F}(\mathbf{s}') - \mathcal{F}(\mathbf{s}) - \langle\nabla_{\mathbf{s}}\mathcal{F}(\mathbf{s}), \mathbf{s}'-\mathbf{s}\rangle \\
=&\Big(f(W^{\frac{1}{2}}\mathbf{s}') - f(W^{\frac{1}{2}}\mathbf{s}) - \langle(\nabla_{\mathbf{x}}f(W^{\frac{1}{2}}\mathbf{s}) - \nabla_{\mathbf{x}'}f(W^{\frac{1}{2}}\mathbf{s}')), W^{\frac{1}{2}}\mathbf{s}' - W^{\frac{1}{2}}\mathbf{s}\rangle\Big) + \frac{1}{2\gamma}\|\mathbf{s}'-\mathbf{s}\|_{I-W}^2 \\
\ge&\frac{\mu}{2}\|W^{\frac{1}{2}}\mathbf{s}' - W^{\frac{1}{2}}\mathbf{s}\|^2 + \frac{1}{2\gamma}\|\mathbf{s}'-\mathbf{s}\|_{I-W}^2 \\
=&\frac{\mu}{2}\|\mathbf{s}'-\mathbf{s}\|_W^2 + \frac{1}{2\gamma}\|\mathbf{s}'-\mathbf{s}\|_{I-W}^2 \\
\ge&\frac{\mu'}{2}\|\mathbf{s}'-\mathbf{s}\|^2
\end{aligned}$$

□

We notice from (47) that, without loss of generality, if $\mathbf{m}^{(0)} = 0$, i.e. $\mathbf{m}_{\mathbf{s}}^{(0)} = 0$, then $\mathbf{m}_{\mathbf{s}}^{(k+1)}$ can be calculated by

$$\mathbf{m}_{\mathbf{s}}^{(k+1)} = (1-\beta)\sum_{i=0}^{k}\beta^{k-i}\nabla_{\mathbf{s}}\mathcal{F}(\mathbf{s}^{(i)}, \boldsymbol{\xi}^{(i)}). \tag{49}$$

For notation simplicity, we define $\tilde{\mathbf{g}}_{\mathbf{s}}^{(k)} = \nabla_{\mathbf{s}}\mathcal{F}(\mathbf{s}^{(k)}, \boldsymbol{\xi}^{(k)})$ and $\mathbf{g}_{\mathbf{s}}^{(k)} = \nabla_{\mathbf{s}}\mathcal{F}(\mathbf{s}^{(k)}) = \mathbb{E}_{\boldsymbol{\xi}^{(k)}}\nabla\mathcal{F}(\mathbf{s}^{(k)}, \boldsymbol{\xi}^{(k)})$. The following lemma establishes the variance of $\mathbf{m}_{\mathbf{s}}^{(k+1)}$.

Lemma 2. *Under assumption A.2., the momentum vector $\mathbf{m}_{\mathbf{s}}^{(k+1)}$ satisfies*

$$\mathbb{E}\left[\left\|\mathbf{m}_{\mathbf{s}}^{(k+1)} - (1-\beta)\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)}\right\|^2\right] \le \frac{1-\beta}{1+\beta}(1-\beta^{2(k+1)})n\sigma^2. \tag{50}$$

Proof. Since $\mathbf{m}_{\mathbf{s}}^{(k+1)} = (1-\beta)\sum\limits_{i=0}^{k}\beta^{k-i}\tilde{\mathbf{g}}_{\mathbf{s}}^{(i)}$, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{m}_{\mathbf{s}}^{(k+1)} - (1-\beta)\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)}\right\|^2\right] &= (1-\beta)^2\mathbb{E}\left[\left\|\sum_{i=0}^{k}\beta^{k-i}\left(\tilde{\mathbf{g}}_{\mathbf{s}}^{(i)} - \mathbf{g}_{\mathbf{s}}^{(i)}\right)\right\|^2\right] \\
&= (1-\beta)^2\mathbb{E}\left[\sum_{i=0}^{k}\sum_{j=0}^{k}\left\langle\beta^{k-i}\left(\tilde{\mathbf{g}}_{\mathbf{s}}^{(i)} - \mathbf{g}_{\mathbf{s}}^{(i)}\right), \beta^{k-j}\left(\tilde{\mathbf{g}}_{\mathbf{s}}^{(j)} - \mathbf{g}_{\mathbf{s}}^{(j)}\right)\right\rangle\right] \qquad (51) \\
&= (1-\beta)^2\sum_{i=0}^{k}\beta^{2(k-i)}\mathbb{E}_{\boldsymbol{\xi}^{(i)}}\left[\left\|\tilde{\mathbf{g}}_{\mathbf{s}}^{(i)} - \mathbf{g}_{\mathbf{s}}^{(i)}\right\|^2\right]
\end{aligned}$$

where the last equality holds because of the independence between $\boldsymbol{\xi}^{(0)}, \boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(k)}$ (see Assumption A.2). Since Assumption A.2 implies $\mathbb{E}\left[\|\nabla F(\mathbf{x},\boldsymbol{\xi}) - \nabla f(\mathbf{x})\|^2\right] \leq n\sigma^2$, we have

$$\mathbb{E}\left[\left\|\tilde{\mathbf{g}}_{\mathbf{s}}^{(i)} - \mathbf{g}_{\mathbf{s}}^{(i)}\right\|^2\right] = \mathbb{E}\left[\left\|\nabla_{\mathbf{s}}\mathcal{F}(\mathbf{s}^{(k)},\boldsymbol{\xi}^{(k)}) - \nabla_{\mathbf{s}}\mathcal{F}(\mathbf{s}^{(k)})\right\|^2\right] \overset{(44)}{=} \mathbb{E}\left[\left\|\nabla F(\mathbf{x}^{(k)},\boldsymbol{\xi}^{(k)}) - \nabla f(\mathbf{x}^{(k)})\right\|_W^2\right] \leq n\sigma^2.$$

Combining the above inequality with (51), we achieve the result. □

The next lemma examines the distance between the expectation of $\mathbf{m}_{\mathbf{s}}^{(k+1)}$ (scaled by $1/(1-\beta^{k+1})$) and the real descent gradient $\mathbf{g}_{\mathbf{s}}^{(k)}$.

Lemma 3. *Under assumption A.1, we have*

$$\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)} - \mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] \leq \sum_{i=0}^{k-1}a_{k,i}\mathbb{E}\left[\left\|\mathbf{s}^{(i+1)} - \mathbf{s}^{(i)}\right\|^2\right] \qquad (52)$$

where $a_{k,i} = \frac{(L')^2\beta^{k-i}}{1-\beta^{k+1}}\left(k-i+\frac{\beta}{1-\beta}\right)$.

Proof. We have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)} - \mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] \\
=&\left(\frac{1-\beta}{1-\beta^{k+1}}\right)^2\sum_{i,j=0}^{k}\mathbb{E}\left[\left\langle\beta^{k-i}\left(\mathbf{g}_{\mathbf{s}}^{(i)} - \mathbf{g}_{\mathbf{s}}^{(k)}\right), \beta^{k-j}\left(\mathbf{g}_{\mathbf{s}}^{(j)} - \mathbf{g}_{\mathbf{s}}^{(k)}\right)\right\rangle\right] \\
\overset{(a)}{\leq}&\left(\frac{1-\beta}{1-\beta^{k+1}}\right)^2\sum_{i,j=0}^{k}\beta^{2k-i-j}\left(\frac{1}{2}\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(i)} - \mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] + \frac{1}{2}\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(j)} - \mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right]\right) \\
=&\left(\frac{1-\beta}{1-\beta^{k+1}}\right)^2\sum_{j=0}^{k}\left(\sum_{i=0}^{k}\beta^{2k-i-j}\right)\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(i)} - \mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] \\
=&\left(\frac{1-\beta}{1-\beta^{k+1}}\right)^2\sum_{i=0}^{k}\frac{\beta^{k-i}\left(1-\beta^{k+1}\right)}{1-\beta}\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(j)} - \mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] \\
=&\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(i)} - \mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right]
\end{aligned}$$

where we have used the Cauchy-Schwarz inequality in (a).

By applying the triangle inequality and Lemma 1, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)}-\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right]\\
\leq&\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}(k-i)\sum_{j=i}^{k-1}\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(j+1)}-\mathbf{g}_{\mathbf{s}}^{(j)}\right\|^2\right] \qquad .\\
\leq&\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}(k-i)\sum_{j=i}^{k-1}(L')^2\mathbb{E}\left[\left\|\mathbf{s}^{(j+1)}-\mathbf{s}^{(j)}\right\|^2\right]\\
=&\frac{1-\beta}{1-\beta^{k+1}}\sum_{j=0}^{k-1}\left(\sum_{i=0}^{j}\beta^{k-i}(k-i)\right)(L')^2\mathbb{E}\left[\left\|\mathbf{s}^{(j+1)}-\mathbf{s}^{(j)}\right\|^2\right].
\end{aligned}$$

Furthermore, it can be shown that

$$\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{j}\beta^{k-i}(k-i)(L')^2\leq\frac{(L')^2\beta^{k-j}}{1-\beta^{k+1}}(k-j+\frac{\beta}{1-\beta})\triangleq a_{k,j}.$$

which completes the proof. □

Next we introduce a key Lyapunov function, which is inspired by [31]:

$$\mathcal{L}^k=\mathcal{F}\left(\mathbf{t}^{(k)}\right)-\mathcal{F}^{\star}+\sum_{i=0}^{k-1}c_i\|\mathbf{s}^{(k-i)}-\mathbf{s}^{(k-1-i)}\|^2 \tag{53}$$

where c_i are positive constants to be specified later, $\mathcal{F}^{\star}=\min\mathcal{F}(\mathbf{s})$, and $\mathbf{t}^{(k)}$ is an auxiliary sequence defined as

$$\mathbf{t}^{(k)}=\begin{cases}\mathbf{s}^{(0)} & k=0,\\ \frac{1}{1-\beta}\mathbf{s}^{(k)}-\frac{\beta}{1-\beta}\mathbf{s}^{(k-1)} & k\geq 1.\end{cases} \tag{54}$$

The introduction of $\mathbf{t}^{(k)}$ is inspired from [56]. It is delicately designed to enjoy the following property:

Lemma 4. $\mathbf{t}^{(k)}$ *defined in* (54) *satisfies*

$$\mathbf{t}^{(k+1)}=\mathbf{t}^{(k)}-\frac{\gamma}{1-\beta}\tilde{\mathbf{g}}_{\mathbf{s}}^{(k)}. \tag{55}$$

Proof. We prove this by direct calculation. When $k=1$,

$$\mathbf{t}^{(1)}-\mathbf{t}^{(0)}=\frac{1}{1-\beta}\mathbf{s}^{(1)}-\frac{\beta}{1-\beta}\mathbf{s}^{(0)}-\mathbf{s}^{(0)}=\frac{1}{1-\beta}\left(\mathbf{s}^{(1)}-\mathbf{s}^{(0)}\right)=-\frac{\gamma}{1-\beta}\tilde{\mathbf{g}}_{\mathbf{s}}^{(0)}.$$

For $k\geq 1$, we have

$$\begin{aligned}
\mathbf{t}^{k+1}-\mathbf{t}^{k}&=\frac{1}{1-\beta}\left(\mathbf{s}^{(k+1)}-\mathbf{s}^{(k)}\right)-\frac{\beta}{1-\beta}\left(\mathbf{s}^{(k)}-\mathbf{s}^{(k-1)}\right)\\
&=\frac{1}{1-\beta}\left(-\frac{\gamma}{1-\beta}\mathbf{m}_{\mathbf{s}}^{(k+1)}\right)-\frac{\beta}{1-\beta}\left(-\frac{\gamma}{1-\beta}\mathbf{m}_{\mathbf{s}}^{(k)}\right)\\
&=-\frac{\gamma}{1-\beta}\left(\frac{1}{1-\beta}\mathbf{m}_{\mathbf{s}}^{(k+1)}-\frac{\beta}{1-\beta}\mathbf{m}_{\mathbf{s}}^{(k)}\right)\\
&=-\frac{\gamma}{1-\beta}\tilde{\mathbf{g}}_{\mathbf{s}}^{(k)}
\end{aligned}$$

□

Finally, we establish the following descent lemma in terms of $\mathbb{E}\left[\mathcal{F}\left(\mathbf{t}^{(k)}\right)\right]$, which plays a critical role in both proofs of the strongly convex scenario and non-convex scenario.

Lemma 5. *Under Assumption A.1-A.3, it holds that*

$$\begin{aligned}\mathbb{E}\left[\mathcal{F}\left(\mathbf{t}^{(k+1)}\right)\right] \leq & \mathbb{E}\left[\mathcal{F}\left(\mathbf{t}^{(k)}\right)\right] + \left(-\frac{\gamma}{1-\beta} + \frac{2\beta^2-\beta+3}{2(1-\beta)}L'\frac{\gamma^2}{(1-\beta)^2}\right)\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] \\ & + \frac{\beta^2+\beta+1}{2(1+\beta)}L'\frac{\gamma^2}{(1-\beta)^2}n\sigma^2 + \frac{\left(1-\beta^{k+1}\right)^2 L'\gamma^2}{(1-\beta)^3}\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)} - \mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right]\end{aligned} \tag{56}$$

Proof. Since $\mathcal{F}(\mathbf{s})$ is L'-smooth, it holds that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}^{(k)}}\left[\mathcal{F}\left(\mathbf{t}^{(k+1)}\right)\right] & \leq \mathcal{F}\left(\mathbf{t}^{(k)}\right) + \mathbb{E}_{\boldsymbol{\xi}^{(k)}}\left[\left\langle\nabla\mathcal{F}\left(\mathbf{t}^{(k)}\right), \mathbf{t}^{(k+1)}-\mathbf{t}^{(k)}\right\rangle\right] + \frac{L'}{2}\mathbb{E}_{\boldsymbol{\xi}^{(k)}}\left[\left\|\mathbf{t}^{(k+1)}-\mathbf{t}^{(k)}\right\|^2\right] \\ & \overset{(55)}{=} \mathcal{F}\left(\mathbf{t}^{(k)}\right) + \mathbb{E}_{\boldsymbol{\xi}^{(k)}}\left[\left\langle\nabla\mathcal{F}\left(\mathbf{t}^{(k)}\right), -\frac{\gamma}{1-\beta}\tilde{\mathbf{g}}_{\mathbf{s}}^{(k)}\right\rangle\right] + \frac{L'\gamma^2}{2(1-\beta)^2}\mathbb{E}_{\boldsymbol{\xi}^{(k)}}\left[\left\|\tilde{\mathbf{g}}_{\mathbf{s}}^{(k)}\right\|^2\right]\end{aligned} \tag{57}$$

Note that $\mathbf{t}^{(k)}$ is determined by previous k random samples $\boldsymbol{\xi}^0, \ldots, \boldsymbol{\xi}^{(k-1)}$ which are independent of $\boldsymbol{\xi}^{(k)}$ and $\mathbb{E}_{\boldsymbol{\xi}^{(k)}}\left[\tilde{\mathbf{g}}_{\mathbf{s}}^{(k)}\right] = \mathbf{g}_{\mathbf{s}}^{(k)}$. Taking expectations over all historical $\boldsymbol{\xi}$'s, we reach

$$\begin{aligned}& \mathbb{E}\left[\left\langle\nabla\mathcal{F}\left(\mathbf{t}^{(k)}\right), -\frac{\gamma}{1-\beta}\tilde{\mathbf{g}}_{\mathbf{s}}^{(k)}\right\rangle\right] \\ & = \mathbb{E}\left[\left\langle\nabla\mathcal{F}\left(\mathbf{t}^{k}\right) - \mathbf{g}_{\mathbf{s}}^{(k)}, -\frac{\gamma}{1-\beta}\mathbf{g}_{\mathbf{s}}^{(k)}\right\rangle\right] - \frac{\gamma}{1-\beta}\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] \\ & \overset{(a)}{\leq} \frac{\gamma}{1-\beta}\frac{\rho_0}{2}(L')^2\mathbb{E}\left[\left\|\mathbf{t}^{(k)}-\mathbf{s}^{(k)}\right\|^2\right] + \frac{\gamma}{1-\beta}\frac{1}{2\rho_0}\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] - \frac{\gamma}{1-\beta}\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] \\ & \overset{(b)}{\leq} \frac{(1-\beta)L'}{4}\mathbb{E}\left[\left\|\mathbf{t}^{(k)}-\mathbf{s}^{(k)}\right\|^2\right] + \frac{\gamma^2 L'}{(1-\beta)^3}\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] - \frac{\gamma}{1-\beta}\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right]\end{aligned}$$

where (a) holds because of the Cauchy's inequality $\langle x, y\rangle \leq \frac{a}{2}\|x\|^2 + \frac{1}{2a}\|y\|^2$ for any vector x, y and positive constant a, and (b) holds by letting $\rho_0 = \frac{(1-\beta)^2}{2L'\gamma}$. Combining (57) and the above inequality, we arrive at

$$\mathbb{E}\left[\mathcal{F}\left(\mathbf{t}^{(k+1)}\right)\right] \leq \mathbb{E}\left[\mathcal{F}\left(\mathbf{t}^{(k)}\right)\right] + \frac{(1-\beta)L'}{4}\mathbb{E}\left[\left\|\mathbf{t}^{k}-\mathbf{s}^{k}\right\|^2\right] + \frac{\gamma}{1-\beta}\left(\frac{L'\gamma}{(1-\beta)^2}-1\right)\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] + \frac{L'\gamma^2}{2(1-\beta)^2}\mathbb{E}\left[\left\|\tilde{\mathbf{g}}_{\mathbf{s}}^{(k)}\right\|^2\right]$$

Substituting (48) into (55), we achieve

$$\mathbf{t}^{(k)} - \mathbf{s}^{(k)} = -\frac{\gamma\beta}{(1-\beta)^2}\mathbf{m}_{\mathbf{s}}^{(k)} \tag{58}$$

Consequently, we have

$$\mathbb{E}\left[\mathcal{F}\left(\mathbf{t}^{(k+1)}\right)\right] \leq \mathbb{E}\left[\mathcal{F}\left(\mathbf{t}^{(k)}\right)\right] + \frac{L'\gamma^2\beta^2}{4(1-\beta)^3}\mathbb{E}\left[\left\|\mathbf{m}_{\mathbf{s}}^{(k)}\right\|^2\right] + \frac{\gamma}{1-\beta}\left(\frac{L'\gamma}{(1-\beta)^2}-1\right)\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] + \frac{L'\gamma^2}{2(1-\beta)^2}\mathbb{E}\left[\left\|\tilde{\mathbf{g}}_{\mathbf{s}}^{(k)}\right\|^2\right] \tag{59}$$

On the other hand, we know form Lemma 2 that

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{m}_{\mathbf{s}}^{(k)}\right\|^2\right] & \leq 2\mathbb{E}\left[\left\|\mathbf{m}_{\mathbf{s}}^{(k)} - (1-\beta)\sum_{i=0}^{k-1}\beta^{k-1-i}\mathbf{g}_{\mathbf{s}}^{(i)}\right\|^2\right] + 2\mathbb{E}\left[\left\|(1-\beta)\sum_{i=0}^{k-1}\beta^{k-1-i}\mathbf{g}_{\mathbf{s}}^{(i)}\right\|^2\right] \\ & \leq 2\frac{1-\beta}{1+\beta}n\sigma^2 + 2\mathbb{E}\left[\left\|(1-\beta)\sum_{i=0}^{k-1}\beta^{k-1-i}\mathbf{g}_{\mathbf{s}}^{(i)}\right\|^2\right] \\ \mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^k}\sum_{i=0}^{k-1}\beta^{k-1-i}\mathbf{g}_{\mathbf{s}}^{(i)}\right\|^2\right] & \leq 2\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] + 2\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^k}\sum_{i=0}^{k-1}\beta^{k-1-i}\mathbf{g}_{\mathbf{s}}^{(i)} - \mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] \\ \mathbb{E}\left[\left\|\tilde{\mathbf{g}}_{\mathbf{s}}^{(k)}\right\|^2\right] & \leq n\sigma^2 + \mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right].\end{aligned} \tag{60}$$

Substituting these inequalities into (59), we arrive at

$$\begin{aligned}\mathbb{E}\left[\mathcal{F}\left(\mathbf{t}^{(k+1)}\right)\right] \leq & \mathbb{E}\left[\mathcal{F}\left(\mathbf{t}^{(k)}\right)\right] + \left(\frac{\gamma}{1-\beta}\left(\frac{L'\gamma}{(1-\beta)^2}-1\right)+\frac{L'\gamma^2\beta^2}{(1-\beta)^3}\left(1-\beta^k\right)^2+\frac{L'\gamma^2}{2(1-\beta)^2}\right)\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] \\ & + \left(\frac{L'\gamma^2\beta^2}{2(1-\beta)^3}\frac{1-\beta}{1+\beta}n\sigma^2+\frac{L'\gamma^2}{2(1-\beta)^2}n\sigma^2\right) \\ & + \frac{L'\gamma^2\beta^2}{(1-\beta)^3}\left(1-\beta^k\right)^2\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^k}\sum_{i=0}^{k-1}\beta^{k-1-i}\mathbf{g}_{\mathbf{s}}^{(i)}-\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right].\end{aligned}$$

Substituting

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)}-\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\beta\sum_{i=0}^{k-1}\beta^{k-1-i}\mathbf{g}_{\mathbf{s}}^{(i)}-\frac{1-\beta^k}{1-\beta^{k+1}}\beta\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] \\ &= \beta^2\left(\frac{1-\beta^k}{1-\beta^{k+1}}\right)^2\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^k}\sum_{i=0}^{k-1}\beta^{k-1-i}\mathbf{g}_{\mathbf{s}}^{(i)}-\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right]\end{aligned}$$

into the last inequality produces

$$\begin{aligned}\mathbb{E}\left[\mathcal{F}\left(\mathbf{t}^{(k+1)}\right)\right] \leq & \mathbb{E}\left[\mathcal{F}\left(\mathbf{t}^{(k)}\right)\right] + \left(-\frac{\gamma}{1-\beta}+\frac{L'\gamma^2(-\beta+3)}{2(1-\beta)^3}+\frac{L'\gamma^2\beta^2}{(1-\beta)^3}\left(1-\beta^k\right)^2\right)\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] \\ & + \left(\frac{L'\gamma^2}{2(1-\beta)^2}\frac{\beta^2}{1+\beta}n\sigma^2+\frac{L'\gamma^2}{2(1-\beta)^2}n\sigma^2\right) \\ & + \frac{L'\gamma^2}{(1-\beta)^3}\left(1-\beta^{k+1}\right)^2\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-1-i}\mathbf{g}_{\mathbf{s}}^{(i)}-\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right].\end{aligned} \tag{61}$$

Finally, using $1-\beta^k<1$ leads to the conclusion. □

E. Convergence analysis for strongly-convex scenario

Proposition 4. *Under Assumptions A.1-A.4 and W is positive-definite, there exists positive constants c_i for (53) such that for all $\gamma=\mathcal{O}\left(\min\left\{\frac{(1-\beta)^2}{2\sqrt{3}L'},\frac{(1-\beta)^2}{6\sqrt{\beta}L'\left(3-\beta+2\beta^2+8\left(1+\frac{13\mu'/L'}{2(1+6\mu'/L')}\right)\right)}\right\}\right)$ and $k\geq\lfloor\frac{\log(0.5)}{\log(\beta)}\rfloor$, it holds that*

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}^{k+1}-\mathcal{L}^k\right] \leq & -\frac{\gamma\mu}{(1+\frac{6\mu'}{L'})(1-\beta)}\mathbb{E}\left[\mathcal{L}^k\right]+\left(\frac{1+\beta+\beta^2}{2(1+\beta)}L'+\frac{1-\beta}{1+\beta}2c_0\right)\frac{\gamma^2}{(1-\beta)^2}n\sigma^2 \\ & + \frac{\beta^2+\frac{L'\gamma}{2}\frac{\beta^2}{(1-\beta)^2}}{\left(1+\frac{6\mu'}{L'}\right)(1+\beta)}\frac{2\mu'\gamma^2n\sigma^2}{(1-\beta)^2}.\end{aligned} \tag{62}$$

where $\mathcal{L}^k$ is defined in (53).

Proof. We first derive a lower bound for the gradient norm. Following the strong convexity of $\mathcal{F}$, we have

$$\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right]=\mathbb{E}\left[\left\|\nabla\mathcal{F}\left(\mathbf{s}^{(k)}\right)\right\|^2\right]\geq 2\mu'\mathbb{E}\left[\mathcal{F}\left(\mathbf{s}^{(k)}\right)-\mathcal{F}^\star\right]. \tag{63}$$

Since $\mathcal{F}$ is L-smooth, we have

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{F}\left(\mathbf{t}^{(k)}\right)\right] \leq & \mathbb{E}\left[\mathcal{F}\left(\mathbf{s}^{(k)}\right)\right] + \mathbb{E}\left[\left\langle \mathbf{g}_{\mathbf{s}}^{(k)}, \mathbf{t}^{(k)} - \mathbf{s}^{(k)} \right\rangle\right] + \frac{L'}{2}\mathbb{E}\left[\left\| \mathbf{t}^{(k)} - \mathbf{s}^{(k)} \right\|^2\right] \\
\overset{(58)}{=} & \mathbb{E}\left[\mathcal{F}\left(\mathbf{s}^{(k)}\right)\right] + \mathbb{E}\left[\left\langle \mathbf{g}_{\mathbf{s}}^{(k)} - \frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)} + \frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)}, -\frac{\gamma\beta}{(1-\beta)^2}\mathbf{m}_{\mathbf{s}}^{(k)} \right\rangle\right] \\
& + \frac{L'}{2}\mathbb{E}\left[\left\| \frac{\gamma\beta}{(1-\beta)^2}\mathbf{m}_{\mathbf{s}}^{(k)} \right\|^2\right] \\
\overset{(a)}{\leq} & \mathbb{E}\left[\mathcal{F}\left(\mathbf{s}^{(k)}\right)\right] + \frac{\gamma}{2(1-\beta)}\rho\mathbb{E}\left[\left\| \mathbf{g}_{\mathbf{s}}^{(k)} - \frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)} \right\|^2\right] + \frac{\gamma}{2(1-\beta)\rho}\mathbb{E}\left[\left\| \frac{\beta}{1-\beta}\mathbf{m}_{\mathbf{s}}^{(k)} \right\|^2\right] \\
& + \mathbb{E}\left[\left\langle \frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)}, -\frac{\gamma\beta}{(1-\beta)^2}\mathbf{m}_{\mathbf{s}}^{(k)} \right\rangle\right] + \frac{L'}{2}\mathbb{E}\left[\left\| \frac{\gamma\beta}{(1-\beta)^2}\mathbf{m}_{\mathbf{s}}^{(k)} \right\|^2\right] \\
\overset{(b)}{\leq} & \mathbb{E}\left[\mathcal{F}\left(\mathbf{s}^{(k)}\right)\right] + \frac{\gamma}{2(1-\beta)}\rho\mathbb{E}\left[\left\| \mathbf{g}_{\mathbf{s}}^{(k)} - \frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)} \right\|^2\right] \\
& + \left(\frac{\gamma}{2(1-\beta)\rho}\left(\frac{\beta}{1-\beta}\right)^2 + \frac{L'\gamma^2}{2(1-\beta)^2}\left(\frac{\beta}{1-\beta}\right)^2\right)\mathbb{E}\left[\left\| \mathbf{m}_{\mathbf{s}}^{(k)} \right\|^2\right] \\
& + \frac{\gamma\beta}{(1-\beta)^2}\left(\frac{\rho_1}{2}\mathbb{E}\left[\left\| \frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)} \right\|^2\right] + \frac{1}{2\rho_1}\mathbb{E}\left[\left\| \mathbf{m}_{\mathbf{s}}^{(k)} \right\|^2\right]\right) \\
\overset{(c)}{=} & \mathbb{E}\left[\mathcal{F}\left(\mathbf{s}^{(k)}\right)\right] + \frac{\gamma}{2(1-\beta)^2}\mathbb{E}\left[\left\| \mathbf{g}_{\mathbf{s}}^{(k)} - \frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)} \right\|^2\right] \\
& + \left(\frac{\gamma\beta^2}{(1-\beta)^2} + \frac{L'\gamma^2}{2}\frac{\beta^2}{(1-\beta)^4}\right)\mathbb{E}\left[\left\| \mathbf{m}_{\mathbf{s}}^{(k)} \right\|^2\right] \\
& + \frac{\gamma}{2(1-\beta)^2}\mathbb{E}\left[\left\| \frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)} \right\|^2\right],
\end{aligned}
$$

where (a) and (b) hold because of the Cauchy's inequality, and (c) holds by letting $\rho = \frac{1}{1-\beta}$, $\rho_1 = \frac{1}{\beta}$, respectively. Combining the above inequality with (63), we have

$$
\begin{aligned}
\mathbb{E}\left[\left\| \mathbf{g}_{\mathbf{s}}^{(k)} \right\|^2\right] \geq 2\mu' \Bigg(& \mathbb{E}\left[\mathcal{F}\left(\mathbf{t}^{(k)}\right)\right] - \mathcal{F}^{\star} - \frac{\gamma}{2(1-\beta)^2}\mathbb{E}\left[\left\| \mathbf{g}_{\mathbf{s}}^{(k)} - \frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)} \right\|^2\right] \\
& - \left(\frac{\gamma\beta^2}{(1-\beta)^2} + \frac{L'\gamma^2}{2}\frac{\beta^2}{(1-\beta)^4}\right)\mathbb{E}\left[\left\| \mathbf{m}_{\mathbf{s}}^{(k)} \right\|^2\right] \\
& - \frac{\gamma}{2(1-\beta)^2}\mathbb{E}\left[\left\| \frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)} \right\|^2\right] \Bigg).
\end{aligned}
\tag{64}
$$

Following (60), we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{m}_{\mathbf{s}}^{(k)}\right\|^2\right] \\
\leq & 2\frac{1-\beta}{1+\beta}n\sigma^2 + 2\left(1-\beta^k\right)^2\left(2\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] + 2\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^k}\sum_{i=0}^{k-1}\beta^{k-1-i}\mathbf{g}_{\mathbf{s}}^{(i)} - \mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right]\right) \\
= & 2\frac{1-\beta}{1+\beta}n\sigma^2 + 4\left(1-\beta^k\right)^2\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] + \frac{4}{\beta^2}(1-\beta^{k+1})^2\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)} - \mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right]
\end{aligned}$$

and

$$\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)}\right\|^2\right] \leq 2\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] + 2\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)} - \mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right].$$

Substituting the above two inequalities into (64) and rearranging terms, we have

$$\begin{aligned}
&\left[1 + 2\mu'\left(\left(\frac{\gamma\beta^2}{(1-\beta)^2} + \frac{L'\gamma^2}{2}\frac{\beta^2}{(1-\beta)^4}\right)4\left(1-\beta^k\right)^2 + \frac{\gamma}{(1-\beta)^2}\right)\right]\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] \\
\geq & 2\mu'\left(\mathbb{E}\left[\mathcal{F}\left(\mathbf{t}^{(k)}\right)\right] - \mathcal{F}^\star - \frac{\gamma}{2(1-\beta)^2}\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)} - \frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)}\right\|^2\right]\right. \\
&- \left(\frac{\gamma\beta^2}{(1-\beta)^2} + \frac{L'\gamma^2}{2}\frac{\beta^2}{(1-\beta)^4}\right) \\
&\times\left(2\frac{1-\beta}{1+\beta}n\sigma^2 + \frac{4}{\beta^2}\left(1-\beta^{k+1}\right)^2\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)} - \mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right]\right) \\
&\left.- \frac{\gamma}{(1-\beta)^2}\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)} - \mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right]\right) \qquad (65) \\
= & 2\mu'\left(\mathbb{E}\left[\mathcal{F}\left(\mathbf{t}^{(k)}\right)\right] - \mathcal{F}^\star\right. \\
&- \left(\frac{\gamma\beta^2}{(1-\beta)^2} + \frac{L'\gamma^2}{2}\frac{\beta^2}{(1-\beta)^4}\right)2\frac{1-\beta}{1+\beta}n\sigma^2 \\
&- \left(\frac{3\gamma}{2(1-\beta)^2} + \left(\frac{\gamma}{(1-\beta)^2} + \frac{L'\gamma^2}{2(1-\beta)^4}\right)4(1-\beta^{k+1})^2\right) \\
&\left.\times\,\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)} - \mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right]\right).
\end{aligned}$$

When $\gamma \leq \frac{(1-\beta)^2}{2L'}$, it holds that

$$1 + 2\mu'\left(\left(\frac{\gamma\beta^2}{(1-\beta)^2} + \frac{L'\gamma^2}{2}\frac{\beta^2}{(1-\beta)^4}\right)4\left(1-\beta^k\right)^2 + \frac{\gamma}{(1-\beta)^2}\right) \leq 1 + 6\frac{\mu'}{L'}$$

which leads to

$$\begin{aligned}
&\left(1+6\frac{\mu'}{L'}\right)\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right]\\
\geq& 2\mu'\Bigg(\mathbb{E}\left[\mathcal{F}\left(\mathbf{t}^{(k)}\right)\right]-\mathcal{F}^\star\\
&-\left(\frac{\gamma\beta^2}{(1-\beta)^2}+\frac{L'\gamma^2}{2}\frac{\beta^2}{(1-\beta)^4}\right)2\frac{1-\beta}{1+\beta}n\sigma^2\\
&-\left(\frac{3\gamma}{2(1-\beta)^2}+\left(\frac{\gamma}{(1-\beta)^2}+\frac{L'\gamma^2}{2(1-\beta)^4}\right)4(1-\beta^{k+1})^2\right)\\
&\times\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)}-\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right]\Bigg).
\end{aligned}\tag{66}$$

On the other hand, by Lemma 5, we have

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}^{k+1}\right]\leq&\mathbb{E}\left[\mathcal{L}^{k}\right]+\left(-\frac{\gamma}{1-\beta}+\frac{2\beta^2-\beta+3}{2(1-\beta)}L'\frac{\gamma^2}{(1-\beta)^2}\right)\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right]\\
&+\frac{\beta^2+\beta+1}{2(1+\beta)}\frac{L'\gamma^2}{(1-\beta)^2}n\sigma^2+\frac{\left(1-\beta^{k+1}\right)^2L'\gamma^2}{(1-\beta)^3}\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)}-\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right]\\
&+\sum_{i=0}^{k-1}(c_{i+1}-c_i)\|\mathbf{s}^{(k-i)}-\mathbf{s}^{(k-1-i)}\|^2+c_0\|\mathbf{s}^{(k+1)}-\mathbf{s}^{(k)}\|^2.
\end{aligned}\tag{67}$$

Note that $c_0\|\mathbf{s}^{(k+1)}-\mathbf{s}^{(k)}\|^2$ can be bounded by

$$\begin{aligned}
&c_0\mathbb{E}\left[\left\|\mathbf{s}^{(k+1)}-\mathbf{s}^{(k)}\right\|^2\right]=c_0\frac{\gamma^2}{(1-\beta)^2}\mathbb{E}\left[\left\|\mathbf{m}_{\mathbf{s}}^{(k+1)}\right\|^2\right]\\
\overset{(60)}{\leq}& c_0\frac{\gamma^2}{(1-\beta)^2}\left(2\frac{1-\beta}{1+\beta}n\sigma^2+4\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right]\right)+4c_0\frac{\gamma^2}{(1-\beta)^2}\left(1-\beta^{k+1}\right)^2\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)}-\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right].
\end{aligned}$$

Combining the above inequality with (67), we obtain

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}^{k+1}-\mathcal{L}^{k}\right]\leq&\left(-\frac{\gamma}{1-\beta}+\frac{3-\beta+2\beta^2}{2(1-\beta)}\frac{L'\gamma^2}{(1-\beta)^2}+4c_0\frac{\gamma^2}{(1-\beta)^2}\right)\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right]\\
&+\left(\frac{\beta^2+\beta+1}{2(1+\beta)}\frac{L'\gamma^2}{(1-\beta)^2}n\sigma^2+2c_0\frac{\gamma^2}{1-\beta^2}n\sigma^2\right)\\
&+\sum_{i=0}^{k-1}\left(c_{i+1}-c_i\right)\mathbb{E}\left[\left\|\mathbf{s}^{(k-i)}-\mathbf{s}^{(k-i-1)}\right\|^2\right]\\
&+\left(4c_0\frac{\gamma^2}{(1-\beta)^2}+\frac{L'\gamma^2}{(1-\beta)^3}\right)\left(1-\beta^{k+1}\right)^2\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)}-\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right].
\end{aligned}\tag{68}$$

Suppose γ is sufficiently small such that

$$-\frac{\gamma}{1-\beta}+\frac{3-\beta+2\beta^2}{2(1-\beta)}\frac{L'\gamma^2}{(1-\beta)^2}+4c_0\frac{\gamma^2}{(1-\beta)^2}\leq-\frac{\gamma}{2(1-\beta)},\tag{69}$$

it follows that

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}^{k+1}-\mathcal{L}^{k}\right]\leq&-\frac{\gamma}{2(1-\beta)}\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^{2}\right]\\&+\left(\frac{\beta^{2}+\beta+1}{2(1+\beta)}\frac{L'\gamma^{2}}{(1-\beta)^{2}}n\sigma^{2}+2c_{0}\frac{\gamma^{2}}{1-\beta^{2}}n\sigma^{2}\right)\\&+\sum_{i=0}^{k-1}\left(c_{i+1}-c_{i}\right)\mathbb{E}\left[\left\|\mathbf{s}^{(k-i)}-\mathbf{s}^{(k-i-1)}\right\|^{2}\right]\\&+\left(4c_{0}\frac{\gamma^{2}}{(1-\beta)^{2}}+\frac{L'\gamma^{2}}{(1-\beta)^{3}}\right)\left(1-\beta^{k+1}\right)^{2}\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=1}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)}-\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^{2}\right].\end{aligned}\tag{70}$$

Substituting (66) into (70), we obtain

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}^{k+1}-\mathcal{L}^{k}\right]\leq&P_{1}\mathbb{E}\left[\mathcal{F}\left(\mathbf{t}^{(k)}\right)-\mathcal{F}^{\star}\right]+P_{2}\\&+P_{3}\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)}-\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^{2}\right]+\sum_{i=0}^{k-1}\left(c_{i+1}-c_{i}\right)\mathbb{E}\left[\left\|\mathbf{s}^{(k-i)}-\mathbf{s}^{(k-1-i)}\right\|^{2}\right]\end{aligned}\tag{71}$$

with

$$\begin{aligned}P_{1}=&-\frac{\gamma\mu'}{(1-\beta)(1+\frac{6\mu'}{L'})},\\P_{2}=&\frac{\beta^{2}+\beta+1}{2(1+\beta)}\frac{L'\gamma^{2}}{(1-\beta)^{2}}n\sigma^{2}+2c_{0}\frac{\gamma^{2}}{1-\beta^{2}}n\sigma^{2}+\frac{\gamma\mu'\left(\frac{\gamma\beta^{2}}{1-\beta^{2}}+\frac{L'\gamma^{2}\beta^{2}}{2(1-\beta)^{3}(1+\beta)}\right)2n\sigma^{2}}{(1-\beta)(1+\frac{6\mu'}{L'})},\\P_{3}=&4c_{0}\frac{\gamma^{2}}{(1-\beta)^{2}}+\frac{L'\gamma^{2}}{(1-\beta)^{3}}+\frac{\gamma\mu'\left(\frac{3\gamma}{2(1-\beta)^{2}}+4\left(\frac{\gamma}{(1-\beta)^{2}}+\frac{L'\gamma^{2}}{2(1-\beta)^{4}}\right)\right)}{(1-\beta)(1+\frac{6\mu'}{L'})}.\end{aligned}\tag{72}$$

By Lemma 3 we know that

$$\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)}-\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^{2}\right]\leq\sum_{i=0}^{k-1}a_{k,k-1-i}\mathbb{E}\left[\left\|\mathbf{s}^{(k-i)}-\mathbf{s}^{(k-1-i)}\right\|^{2}\right]$$

with $a_{k,k-i-1}=\frac{(L')^{2}\beta^{i+1}}{1-\beta^{k+1}}\left(i+1+\frac{\beta}{1-\beta}\right)$. Substituting the above inequality into (71), we obtain

$$\mathbb{E}\left[\mathcal{L}^{k+1}-\mathcal{L}^{k}\right]\leq P_{1}\mathbb{E}\left[\mathcal{F}\left(\mathbf{t}^{(k)}\right)-\mathcal{F}^{\star}\right]+P_{2}+\sum_{i=0}^{k-1}\left(c_{i+1}-c_{i}+P_{3}a_{k,k-i-1}\right)\mathbb{E}\left[\left\|\mathbf{s}^{(k-i)}-\mathbf{s}^{(k-1-i)}\right\|^{2}\right].\tag{73}$$

If there exists a positive sequence $\{c_i\}$ such that

$$c_{i+1}-c_{i}+P_{3}a_{k,k-i-1}\leq P_{1}c_{i},\tag{74}$$

then inequality (73) becomes

$$\mathbb{E}\left[\mathcal{L}^{k+1}-\mathcal{L}^{k}\right]\leq P_{1}\mathbb{E}\left[\mathcal{L}^{k}\right]+P_{2}\tag{75}$$

which is equivalent to the result in (62). To construct $\{c_i\}$, we notice that $\frac{1}{1-\beta^{k+1}}\leq 2$ when $k\geq\lfloor\frac{\log(0.5)}{\log(\beta)}\rfloor$. This implies

$$c_{i+1}-c_{i}+P_{3}a_{k,k-i-1}\leq c_{i+1}-c_{i}+P_{3}2(L')^{2}\beta^{i+1}\left(i+1+\frac{\beta}{1-\beta}\right),\quad k\geq\lfloor\frac{\log(0.5)}{\log(\beta)}\rfloor.$$

If we construct $\{c_i\}$ that satisfy

$$c_{i+1} - c_i + P_3 2(L')^2 \beta^{i+1} \left(i + 1 + \frac{\beta}{1-\beta} \right) = P_1 c_i, \tag{76}$$

then (74) holds. Multiplying $1/(1+P_1)^{i+1}$ to both sides of the above inequality, we have

$$\frac{c_i}{(1+P_1)^i} - \frac{c_{i+1}}{(1+P_1)^{i+1}} = \frac{P_3 2(L')^2 \beta^{i+1}}{(1+P_1)^{i+1}} \left(i + 1 + \frac{\beta}{1-\beta} \right). \tag{77}$$

Summing both sides of the above equality from $i = 0$ to $i = \infty$, we conclude that as long as c_0 satisfies

$$\begin{aligned} c_0 \geq & 2(L')^2 P_3 \sum_{i=0}^{\infty} \frac{\beta^{i+1}}{(1+P_1)^{i+1}} \left(i + 1 + \frac{\beta}{1-\beta} \right) \\ \overset{(a)}{=} & 2(L')^2 P_3 \left(\frac{\frac{\beta}{1+P_1}}{\left(1 - \frac{\beta}{1+P_1}\right)^2} + \frac{\frac{\beta}{1+P_1}}{1 - \frac{\beta}{1+P_1}} \frac{\beta}{1-\beta} \right), \end{aligned} \tag{78}$$

and c_i $(i \geq 1)$ is constructed following recursion (77), we will achieve a positive sequence $\{c_i\}$ that satisfies (74). Note that we used $\gamma \leq \frac{(1-\beta)^2}{2L'} < \frac{(1-\beta)^2\left(1+6\frac{\mu'}{L'}\right)}{\mu'(1+\sqrt{\beta})}$ to result in $\beta < \sqrt{\beta} < 1 + P_1$ in (a) so that the series is summable.

Next we simplify the expression in (78). With $\gamma \leq \frac{(1-\beta)^2}{2L'}$, we have $\beta < \sqrt{\beta} < 1 + P_1$ and $\frac{\beta}{1+P_1} \leq \sqrt{\beta}$. These facts lead to

$$\begin{aligned} & 2(L')^2 P_3 \left(\frac{\frac{\beta}{1+P_1}}{\left(1 - \frac{\beta}{1+P_1}\right)^2} + \frac{\frac{\beta}{1+P_1}}{1 - \frac{\beta}{1+P_1}} \frac{\beta}{1-\beta} \right) \\ \leq & 2(L')^2 P_3 \left(\frac{\sqrt{\beta}}{\left(1-\sqrt{\beta}\right)^2} + \frac{\sqrt{\beta}}{1-\sqrt{\beta}} \frac{\beta}{1-\beta} \right) \\ = & 2(L')^2 \left(4c_0 \frac{\gamma^2}{(1-\beta)^2} + \frac{L'\gamma^2}{(1-\beta)^3} + \frac{\gamma\mu'\left(\frac{3\gamma}{2(1-\beta)^2} + \left(\frac{\gamma}{(1-\beta)^2} + \frac{L'\gamma^2}{2(1-\beta)^4}\right)4\right)}{(1-\beta)(1+\frac{6\mu'}{L'})} \right) \\ & \times \left(\frac{\sqrt{\beta}}{\left(1-\sqrt{\beta}\right)^2} + \frac{\sqrt{\beta}}{1-\sqrt{\beta}} \frac{\beta}{1-\beta} \right) \\ \leq & 2(L')^2 \left(4c_0 \frac{\gamma^2}{(1-\beta)^2} + \frac{L'\gamma^2}{(1-\beta)^3} + \frac{\mu' \frac{13\gamma^2}{2(1-\beta)^3}}{1 + \frac{6\mu'}{L'}} \right) \left(\frac{\sqrt{\beta}}{\left(1-\sqrt{\beta}\right)^2} + \frac{\sqrt{\beta}}{1-\sqrt{\beta}} \frac{\beta}{1-\beta} \right). \end{aligned} \tag{79}$$

To guarantee (78), it suffices to let

$$c_0 \geq 2(L')^2 \left(4c_0 \frac{\gamma^2}{(1-\beta)^2} + \frac{L'\gamma^2}{(1-\beta)^3} + \frac{\mu' \frac{13\gamma^2}{2(1-\beta)^3}}{1 + \frac{6\mu'}{L'}} \right) \left(\frac{\sqrt{\beta}}{\left(1-\sqrt{\beta}\right)^2} + \frac{\sqrt{\beta}}{1-\sqrt{\beta}} \frac{\beta}{1-\beta} \right). \tag{80}$$

When $\gamma \leq \frac{(1-\beta)^2}{2\sqrt{3}L'}$, it holds that

$$1 - 8\frac{\gamma^2}{(1-\beta)^2}(L')^2 \left(\frac{\sqrt{\beta}}{(1-\sqrt{\beta})^2} + \frac{\sqrt{\beta}}{1-\sqrt{\beta}} \frac{\beta}{1-\beta} \right) \geq \frac{1}{2}.$$

Withe the help of the above inequality, if c_0 is constructed as

$$c_0 \geq 4(L')^2 \left(\frac{L'\gamma^2}{(1-\beta)^3} + \frac{\mu' \frac{13\gamma^2}{2(1-\beta)^3}}{1 + \frac{6\mu'}{L'}} \right) \left(\frac{\sqrt{\beta}}{\left(1-\sqrt{\beta}\right)^2} + \frac{\sqrt{\beta}}{1-\sqrt{\beta}} \frac{\beta}{1-\beta} \right), \tag{81}$$

and and c_i $(i \geq 1)$ is constructed following recursion (77), we will achieve a positive sequence $\{c_i\}$ that satisfies (74).

Finally we come back to examine the condition on γ that satisfies (69). To guarantee

$$\begin{aligned}
&-\frac{\gamma}{1-\beta}+\frac{3-\beta+2\beta^2}{2(1-\beta)}\frac{L'\gamma^2}{(1-\beta)^2}+4c_0\frac{\gamma^2}{(1-\beta)^2}\\
=&-\frac{\gamma}{1-\beta}+\frac{3-\beta+2\beta^2}{2(1-\beta)}\frac{L'\gamma^2}{(1-\beta)^2}+4(L')^2\left(\frac{L'\gamma^2}{(1-\beta)^3}+\frac{\mu'\frac{13\gamma^2}{2(1-\beta)^3}}{1+\frac{6\mu'}{L'}}\right)\left(\frac{\sqrt{\beta}}{\left(1-\sqrt{\beta}\right)^2}+\frac{\sqrt{\beta}}{1-\sqrt{\beta}}\frac{\beta}{1-\beta}\right)\\
\leq&-\frac{\gamma}{2(1-\beta)},
\end{aligned}$$

it is enough to require

$$\begin{aligned}
&\frac{L'\gamma}{(1-\beta)^2}\left(3-\beta+2\beta^2+8\left(1+\frac{13\mu'/L'}{2(1+6\mu'/L')}\right)\right)\left(\sqrt{\beta}(1+\sqrt{\beta})^2+\beta^{\frac{3}{2}}(1+\sqrt{\beta})\right)\leq 1\\
\Longleftarrow&\frac{L'\gamma}{(1-\beta)^2}\left(3-\beta+2\beta^2+8\left(1+\frac{13\mu'/L'}{2(1+6\mu'/L')}\right)\right)6\sqrt{\beta}\leq 1\\
\Longleftrightarrow&\gamma\leq\frac{(1-\beta)^2}{6\sqrt{\beta}L'\left(3-\beta+2\beta^2+8\left(1+\frac{13\mu'/L'}{2(1+6\mu'/L')}\right)\right)}.
\end{aligned}$$

In summary, if $\gamma \leq \mathcal{O}\big(\min\big\{\frac{(1-\beta)^2}{2\sqrt{3}L'}, \frac{(1-\beta)^2}{6\sqrt{\beta}L'\left(3-\beta+2\beta^2+8\left(1+\frac{13\mu'/L'}{2(1+6\mu'/L')}\right)\right)}\big\}\big) = \mathcal{O}\left(\frac{(1-\beta)^2}{L'}\right)$, it holds for $k \geq \lfloor\frac{\log(0.5)}{\log(\beta)}\rfloor$ that

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}^{k+1}-\mathcal{L}^k\right]\leq&-\frac{\gamma\mu}{(1+\frac{6\mu'}{L'})(1-\beta)}\mathbb{E}\left[L^k\right]+\left(\frac{1+\beta+\beta^2}{2(1+\beta)}L'+\frac{1-\beta}{1+\beta}2c_0\right)\frac{\gamma^2}{(1-\beta)^2}n\sigma^2\\
&+\frac{\beta^2+\frac{L'\gamma}{2}\frac{\beta^2}{(1-\beta)^2}}{\left(1+\frac{6\mu'}{L'}\right)(1+\beta)}\frac{2\mu'\gamma^2n\sigma^2}{(1-\beta)^2}.
\end{aligned}\tag{82}$$

□

E.1. Proof of Theorem 1

With Proposition 4, we are able to prove Theorem 1.

Proof of Theorem 1. From Proposition (4), we know for all $k \geq k_0 \triangleq \lfloor\frac{\log(0.5)}{\log(\beta)}\rfloor$, it holds that

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}^{k+1}-\mathcal{L}^k\right]\leq&-\frac{\gamma\mu'}{(1+\frac{6\mu'}{L'})(1-\beta)}\mathbb{E}\left[\mathcal{L}^k\right]+\left(\frac{1+\beta+\beta^2}{2(1+\beta)}L'+\frac{1-\beta}{1+\beta}2c_0\right)\frac{\gamma^2}{(1-\beta)^2}n\sigma^2\\
&+\frac{\beta^2+\frac{L'\gamma}{2}\frac{\beta^2}{(1-\beta)^2}}{\left(1+\frac{6\mu'}{L'}\right)(1+\beta)}\frac{2\mu'\gamma^2n\sigma^2}{(1-\beta)^2}
\end{aligned}$$

The above inequality can be rearranged as

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}^{k+1}\right]\leq&\left(1-\frac{\gamma\mu'}{(1+\frac{6\mu'}{L'})(1-\beta)}\right)\mathbb{E}\left[\mathcal{L}^k\right]+\left(\frac{1+\beta+\beta^2}{2(1+\beta)}L'+\frac{1-\beta}{1+\beta}2c_0\right)\frac{\gamma^2}{(1-\beta)^2}n\sigma^2\\
&+\frac{\beta^2+\frac{L'\gamma}{2}\frac{\beta^2}{(1-\beta)^2}}{\left(1+\frac{6\mu'}{L'}\right)(1+\beta)}\frac{2\mu'\gamma^2n\sigma^2}{(1-\beta)^2}\\
\leq&\left(1-\frac{\gamma\mu'}{(1+\frac{6\mu'}{L'})(1-\beta)}\right)\mathbb{E}\left[\mathcal{L}^k\right]+\left(\frac{1+\beta+\beta^2}{2(1+\beta)}L'+\frac{1-\beta}{1+\beta}2c_0\right)\frac{\gamma^2}{(1-\beta)^2}n\sigma^2\\
&+\frac{\frac{5\beta^2}{2}}{1+\frac{6\mu'}{L'}}\frac{1}{1+\beta}\frac{\mu'\gamma^2n\sigma^2}{(1-\beta)^2}
\end{aligned}$$

in which we used $\gamma \le \frac{(1-\beta)^2}{2L'}$ in the second inequality. Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{L}^{k+1}\right] - \left(1+\frac{6\mu'}{L'}\right)\left(\left(\frac{1+\beta+\beta^2}{2(1+\beta)}\frac{L'}{\mu'} + \frac{1-\beta}{1+\beta}\frac{2c_0}{\mu'}\right)\frac{\gamma}{1-\beta}n\sigma^2 + \frac{\frac{5\beta^2}{2}}{1+\frac{6\mu'}{L'}}\frac{1}{1+\beta}\frac{\mu'\gamma n\sigma^2}{1-\beta}\right)\\
\le&\left(1-\frac{\gamma\mu'}{(1+\frac{6\mu'}{L'})(1-\beta)}\right)\times\\
&\left(\mathbb{E}\left[\mathcal{L}^{k}\right] - \left(1+\frac{6\mu'}{L'}\right)\left(\left(\frac{1+\beta+\beta^2}{2(1+\beta)}\frac{L'}{\mu'} + \frac{1-\beta}{1+\beta}\frac{2c_0}{\mu'}\right)\frac{\gamma}{1-\beta}n\sigma^2 + \frac{\frac{5\beta^2}{2}}{1+\frac{6\mu'}{L'}}\frac{1}{1+\beta}\frac{\mu'\gamma n\sigma^2}{1-\beta}\right)\right),
\end{aligned}$$

which directly yields

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}^{k}\right] \le& \left(1+\frac{6\mu'}{L'}\right)\left(\left(\frac{1+\beta+\beta^2}{2(1+\beta)}\frac{L'}{\mu'} + \frac{1-\beta}{1+\beta}\frac{2c_0}{\mu'}\right)\frac{\gamma}{1-\beta}n\sigma^2 + \frac{\frac{5\beta^2}{2}}{1+\frac{6\mu'}{L'}}\frac{1}{1+\beta}\frac{\mu'\gamma n\sigma^2}{1-\beta}\right)\\
&+\left(1-\frac{\gamma\mu'}{(1+\frac{6\mu'}{L'})(1-\beta)}\right)^{k-k_0}\times\\
&\left(\mathbb{E}\left[\mathcal{L}^{k_0}\right] - \left(1+\frac{6\mu'}{L'}\right)\left(\left(\frac{1+\beta+\beta^2}{2(1+\beta)}\frac{L'}{\mu'} + \frac{1-\beta}{1+\beta}\frac{2c_0}{\mu'}\right)\frac{\gamma}{1-\beta}n\sigma^2 + \frac{\frac{5\beta^2}{2}}{1+\frac{6\mu'}{L'}}\frac{1}{1+\beta}\frac{\mu'\gamma n\sigma^2}{1-\beta}\right)\right)\\
&\le \left(1+\frac{6\mu'}{L'}\right)\left(\left(\frac{1+\beta+\beta^2}{2(1+\beta)}\frac{L'}{\mu'} + \frac{1-\beta}{1+\beta}\frac{2c_0}{\mu'}\right)\frac{\gamma}{1-\beta}n\sigma^2 + \frac{\frac{5\beta^2}{2}}{1+\frac{6\mu'}{L'}}\frac{1}{1+\beta}\frac{\mu'\gamma n\sigma^2}{1-\beta}\right)\\
&+\left(1-\frac{\gamma\mu'}{(1+\frac{6\mu'}{L'})(1-\beta)}\right)^{k-k_0}\mathbb{E}\left[\mathcal{L}^{k_0}\right].
\end{aligned}$$

Since c_i are delicately chosen to be positive, we achieve

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{F}\left(\mathbf{t}^{(k)}\right)-\mathcal{F}^\star\right]\\
\le&\left(1+\frac{6\mu'}{L'}\right)\left(\left(\frac{1+\beta+\beta^2}{2(1+\beta)}\frac{L'}{\mu'} + \frac{1-\beta}{1+\beta}\frac{2c_0}{\mu'}\right)\frac{\gamma}{1-\beta}n\sigma^2 + \frac{\frac{5\beta^2}{2}}{1+\frac{6\mu'}{L'}}\frac{1}{1+\beta}\frac{\mu'\gamma n\sigma^2}{1-\beta}\right)\\
&+\left(1-\frac{\gamma\mu'}{(1+\frac{6\mu'}{L'})(1-\beta)}\right)^{k-k_0}\mathbb{E}\left[\mathcal{L}^{k_0}\right]\\
=&\mathcal{O}\left(\mathbb{E}\left[\mathcal{L}^{k_0}\right]\left(1-\frac{\mu'\gamma}{1-\beta}\right)^{k}+\frac{\gamma}{1-\beta}n\sigma^2\right)
\end{aligned} \tag{83}$$

where we use the fact in (81) that $c_0 = \mathcal{O}(\frac{1}{1-\beta})$ when $\gamma = \mathcal{O}(\frac{(1-\beta)^2}{L'})$.

Now we let $\mathbf{x}_{\mathrm{L}}$ be the fixed point of the DecentLaM recursion (25) when the full-batch gradient $\nabla f(\mathbf{x})$ can be accessed per iteration. It is known from (43) that

$$(I-W)W^{-\frac{1}{2}}\mathbf{x}_{\mathrm{L}} + \gamma W^{\frac{1}{2}}\nabla f(\mathbf{x}_L) = 0. \tag{84}$$

Moreover, it is derived from Appendix C.3 that

$$\frac{1}{n}\mathbb{E}\left[\|\mathbf{x}_{\mathrm{L}} - \mathbf{x}^\star\|^2\right] = \mathcal{O}(\gamma^2 b^2) \tag{85}$$

Next we let $\mathbf{s}_{\mathrm{L}} = \arg\min_{\mathbf{s}} \mathcal{F}(\mathbf{s})$. With the definition of $\mathcal{F}(\mathbf{s})$, $\mathbf{s}_{\mathrm{L}}$ satisfies

$$W^{\frac{1}{2}}\nabla f(W^{\frac{1}{2}}\mathbf{s}_{\mathrm{L}}) + \frac{1}{\gamma}(I-W)\mathbf{s}_{\mathrm{L}} = 0. \tag{86}$$

Since $\mathcal{F}(\mathbf{s})$ is strongly convex, there is only one root of $\nabla\mathcal{F}(\mathbf{s})$, it is derived from (84) and (86) that

$$\mathbf{x}_{\mathrm{L}} = W^{\frac{1}{2}}\mathbf{s}_{\mathrm{L}}. \tag{87}$$

Now we are ready to establish the convergence rate. Since $\mathbf{t}^{(k)} = \frac{1}{1-\beta}\mathbf{s}^{(k)} - \frac{\beta}{1-\beta}\mathbf{s}^{(k-1)}$, we have

$$\mathbf{s}^{(k)} = (1-\beta)\sum_{i=1}^{k}\beta^{k-i}\mathbf{t}^{(i)} + \beta^{k}\mathbf{t}^{(0)}$$

Since $(1-\beta)\sum_{i=1}^{k}\beta^{k-i} + \beta^{k} = 1$ and $\mathcal{F}(\cdot)$ is a convex function, it holds from the Jensen's inequality that

$$\begin{aligned}\mathbb{E}\left[\mathcal{F}(\mathbf{s}^{(k)}) - \mathcal{F}^{\star}\right] \leq & (1-\beta)\sum_{i=1}^{k}\beta^{k-i}\mathbb{E}\left[\mathcal{F}(\mathbf{t}^{(i)}) - \mathcal{F}^{\star}\right] + \beta^{k}\mathbb{E}\left[\mathcal{F}(\mathbf{t}^{(0)}) - \mathcal{F}^{\star}\right] \\ = & \mathcal{O}\left(\mathbb{E}\left[\mathcal{L}^{k_0}\right]\left(1 - \frac{\mu'\gamma}{1-\beta}\right)^{k} + \frac{\gamma}{1-\beta}n\sigma^{2}\right).\end{aligned} \tag{88}$$

where the last equality holds because of the inequality (83) and $\beta < 1 - \frac{\gamma\mu'}{1-\beta}$ when $\gamma \leq \frac{(1-\beta)^2}{L'}$. Therefore, following the strong convexity of $\mathcal{F}$ and (88), we have

$$\begin{aligned}\frac{1}{n}\mathbb{E}\left[\left\|\mathbf{x}^{(k)} - \mathbf{x}_{\mathrm{L}}\right\|^{2}\right] & \overset{(87)}{=} \frac{1}{n}\mathbb{E}\left[\left\|\mathbf{s}^{(k)} - \mathbf{s}_{\mathrm{L}}\right\|_{W}^{2}\right] \\ & \leq \frac{2}{n}\mathbb{E}\left[\mu'\left(\mathcal{F}(\mathbf{s}^{(k)}) - \mathcal{F}^{\star}\right)\right] = \mathcal{O}\left(\left(1 - \frac{\mu'\gamma}{1-\beta}\right)^{k} + \frac{\gamma}{1-\beta}\sigma^{2}\right).\end{aligned} \tag{89}$$

Combining (85) and (89) and note $\mu' = \mu$ because $\gamma \leq \frac{1}{\mu}$, we conclude

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|x_i^{(k)} - x^{\star}\right\|^{2}\right] & = \frac{1}{n}\mathbb{E}\left[\left\|\mathbf{x}^{(k)} - \mathbf{x}^{\star}\right\|^{2}\right] \\ & \leq \frac{2}{n}\mathbb{E}\left[\left\|\mathbf{x}^{(k)} - \mathbf{x}_{\mathrm{L}}\right\|^{2}\right] + \frac{2}{n}\mathbb{E}\left[\|\mathbf{x}_{\mathrm{L}} - \mathbf{x}^{\star}\|^{2}\right] \\ & = \mathcal{O}\left((1 - \frac{\gamma\mu}{1-\beta})^{k} + \frac{\gamma\sigma^{2}}{1-\beta} + \gamma^{2}b^{2}\right)\end{aligned} \tag{90}$$

□

E.2. Proof of Corollary 1

Proof of Corollary 1. In fact, by choosing $\gamma = \frac{\mu(1-\beta)}{k}\log(\frac{k\mu}{\sigma^2})$, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|x_i^{(k)} - x^{\star}\right\|^{2}\right] & = \mathcal{O}\left((1 - \frac{\gamma\mu}{1-\beta})^{k} + \frac{\gamma\sigma^{2}}{1-\beta} + \gamma^{2}b^{2}\right) \\ & \leq \mathcal{O}\left(e^{-\frac{k\gamma\mu}{1-\beta}} + \frac{\gamma\sigma^{2}}{1-\beta} + \gamma^{2}b^{2}\right) \\ & \leq \mathcal{O}\left(\frac{\sigma^{2}}{k\mu} + \frac{\mu\sigma^{2}}{k}\log(\frac{k\mu}{\sigma^{2}}) + \frac{\mu^{2}(1-\beta)^{2}}{k^{2}}2\log(\frac{k\mu}{\sigma^{2}})b^{2}\right) = \tilde{\mathcal{O}}(\frac{1}{k}).\end{aligned}$$

□

F. Convergence analysis for non-convex scenario

Proposition 5. *Under Assumptions A.1-A.3 and W is positive-definite, there exists positive constants c_i for (53) such that for all $\gamma < \frac{(1-\beta)^2}{2\sqrt{2}L'\sqrt{\beta+\beta^2}}$, it holds that*

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}^{k+1}-\mathcal{L}^{k}\right] \le & \left(-\frac{\gamma}{1-\beta}+\frac{3-\beta+\beta^2}{2(1-\beta)^2}L'\gamma^2+4c_0\frac{\gamma^2}{(1-\beta)^2}\right)\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] \\ & + \left(\frac{(\beta^2+\beta+1)\gamma^2}{2(1+\beta)(1-\beta)^2}L'n\sigma^2+2c_0\frac{\gamma^2}{1-\beta^2}n\sigma^2\right).\end{aligned} \tag{91}$$

Proof. Following arguments (67)–(68) (note that these arguments do not need convexity of $\mathcal{F}$), we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}^{k+1}-\mathcal{L}^{k}\right] \le & \left(-\frac{\gamma}{1-\beta}+\frac{3-\beta+2\beta^2}{2(1-\beta)}\frac{L'\gamma^2}{(1-\beta)^2}+4c_0\frac{\gamma^2}{(1-\beta)^2}\right)\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] \\ & + \left(\frac{\beta^2+\beta+1}{2(1+\beta)}L'+2c_0\frac{1-\beta}{1+\beta}\right)\frac{\gamma^2 n\sigma^2}{(1-\beta)^2} \\ & \underbrace{+\sum_{i=0}^{k-1}(c_{i+1}-c_i)\,\mathbb{E}\left[\left\|\mathbf{s}^{(k-i)}-\mathbf{s}^{(k-i-1)}\right\|^2\right]}_{A} \\ & \underbrace{+\left(4c_0\frac{\gamma^2}{(1-\beta)^2}+\frac{L'\gamma^2}{(1-\beta)^3}\right)\left(1-\beta^{k+1}\right)^2\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)}-\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right]}_{B}.\end{aligned} \tag{92}$$

Next we show that the sum of terms A and B can be non-positive by choosing appropriate positive c_i. By Lemma 3 we know

$$\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^{k+1}}\sum_{i=0}^{k}\beta^{k-i}\mathbf{g}_{\mathbf{s}}^{(i)}-\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] \le \sum_{i=0}^{k-1}a_{k,k-1-i}\mathbb{E}\left[\left\|\mathbf{s}^{(k-i)}-\mathbf{s}^{(k-1-i)}\right\|^2\right]$$

with $a_{k,k-i-1}=\frac{(L')^2\beta^{i+1}}{1-\beta^{k+1}}\left(i+1+\frac{\beta}{1-\beta}\right)$. To achieve $A+B\le 0$, it suffices to let

$$c_{i+1}\le c_i-\left(4c_0\frac{\gamma^2}{(1-\beta)^2}+\frac{L'\gamma^2}{(1-\beta)^3}\right)\left(1-\beta^{k+1}\right)^2 a_{k,k-1-i}. \tag{93}$$

Since $(1-\beta^{k+1})<1$, we can require for all $i\ge 0$ that

$$c_{i+1}\le c_i-\left(4c_0\frac{\gamma^2}{(1-\beta)^2}+\frac{L'\gamma^2}{(1-\beta)^3}\right)(L')^2\beta^{i+1}(i+1+\frac{\beta}{1-\beta}). \tag{94}$$

In order to construct positive $\{c_i\}$ to satisfy the above relation, we can choose

$$\begin{aligned}c_0 &= \left(4c_0\frac{\gamma^2}{(1-\beta)^2}+\frac{L'\gamma^2}{(1-\beta)^3}\right)\sum_{i=0}^{\infty}\beta^{i+1}\left(i+1+\frac{\beta}{1-\beta}\right)(L')^2 \\ &= \left(4c_0\frac{\gamma^2}{(1-\beta)^2}+\frac{L'\gamma^2}{(1-\beta)^3}\right)\frac{\beta+\beta^2}{(1-\beta)^2}(L')^2\end{aligned} \tag{95}$$

This indeed yields

$$c_0=\frac{\frac{\beta+\beta^2}{(1-\beta)^5}\gamma^2(L')^3}{1-4\gamma^2\frac{\beta+\beta^2}{(1-\beta)^4}(L')^2} \tag{96}$$

which is positive when $\gamma < \frac{(1-\beta)^2}{2L'\sqrt{\beta+\beta^2}}$. As a result, if c_0 is constructed as in (96), and c_i is constructed as in recursion (94) (replace "$\le$" with "$=$"), then all c_i's are positive and the inequality (93) holds, which leads to $A+B\le 0$. Substituting $A+B\le 0$ into (92), we achieve the result. □

F.1. Proof of Theorem 2

With Proposition 5, we are ready to establish Theorem 2.

Proof of Theorem 2. From Proposition 5, we know that

$$\mathbb{E}\left[\mathcal{L}^{k+1} - \mathcal{L}^{k}\right] \leq -Q_1 \mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] + Q_2 \tag{97}$$

with

$$Q_1 = \frac{\gamma}{1-\beta} - \frac{3-\beta+\beta^2}{2(1-\beta)^2} L'\gamma^2 - 4c_0 \frac{\gamma^2}{(1-\beta)^2} \tag{98}$$

$$Q_2 = \frac{(\beta^2+\beta+1)\gamma^2}{2(1+\beta)(1-\beta)^2} L' n\sigma^2 + 2c_0 \frac{\gamma^2}{1-\beta^2} n\sigma^2. \tag{99}$$

This immediately yields that

$$Q_1 \sum_{k=0}^{T-1} \mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] \leq \mathbb{E}\left[\mathcal{L}^0 - \mathcal{L}^T\right] + TQ_2 \leq \mathcal{L}^0 + TQ_2$$

and hence

$$\frac{1}{T} \sum_{k=0}^{T-1} \mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] \leq \frac{\mathcal{L}^0}{TQ_1} + \frac{Q_2}{Q_1}. \tag{100}$$

Next we bound Q_1 and Q_2 properly. Note that

$$\begin{aligned} Q_1 \geq \frac{\gamma}{2(1-\beta)} &\iff \frac{\gamma}{1-\beta} - \frac{3-\beta+\beta^2}{2(1-\beta)^2} L'\gamma^2 - 4c_0 \frac{\gamma^2}{(1-\beta)^2} \geq \frac{\gamma}{2(1-\beta)} \\ &\iff 1-\beta \geq (3-\beta+\beta^2) L'\gamma + 8c_0\gamma. \end{aligned} \tag{101}$$

With the expression in (96), when $\gamma \leq \frac{(1-\beta)^2}{2\sqrt{2}L'\sqrt{\beta+\beta^2}}$, we have

$$c_0 \leq 2\frac{\beta+\beta^2}{(1-\beta)^5}\gamma^2 (L')^3 \leq \frac{L'}{4(1-\beta)}.$$

According to (101), to achieve $Q_1 \geq \frac{\gamma}{2(1-\beta)}$, it suffices to let

$$\begin{aligned} &1-\beta \geq (3-\beta+\beta^2)L'\gamma + \frac{2}{1-\beta}L'\gamma \\ \iff &\gamma \leq \frac{(1-\beta)^2}{\left(2+(1-\beta)(3-\beta+2\beta^2)\right)L'}. \end{aligned} \tag{102}$$

As to Q_2, note that

$$\begin{aligned} Q_2 &= \frac{(\beta^2+\beta+1)\gamma^2}{2(1+\beta)(1-\beta)^2} L' n\sigma^2 + 2c_0 \frac{\gamma^2}{1-\beta^2} n\sigma^2 \\ &\leq \frac{(\beta^2+\beta+1)\gamma^2}{2(1+\beta)(1-\beta)^2} L' n\sigma^2 + \frac{L'}{2(1-\beta)} \frac{\gamma^2}{1-\beta^2} n\sigma^2 \end{aligned} \tag{103}$$

Therefore, we have

$$\frac{Q_2}{Q_1} \leq \frac{(\beta^2+\beta+1)\gamma}{1-\beta^2} L' n\sigma^2 + \frac{L'\gamma}{1-\beta^2} n\sigma^2 = \mathcal{O}(\frac{L'\gamma n\sigma^2}{1-\beta}). \tag{104}$$

To summarize, as long as $\gamma < \min\{\frac{(1-\beta)^2}{2\sqrt{2}L'\sqrt{\beta+\beta^2}}, \frac{(1-\beta)^2}{(2+(1-\beta)(3-\beta+2\beta^2))L'}\} = \mathcal{O}\left(\frac{(1-\beta)^2}{L'}\right)$, it holds that

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] = \mathcal{O}(\frac{(1-\beta)\mathcal{L}^0}{\gamma T} + \frac{L'\gamma n\sigma^2}{1-\beta}). \tag{105}$$

On the other hand, recall the definition of $\mathbf{g}_{\mathbf{s}}^{(k)}$, we have

$$W^{\frac{1}{2}}\mathbf{g}_{\mathbf{s}}^{(k)} = W\nabla f(\mathbf{x}^{(k)}) + \frac{1}{\gamma}(I-W)\mathbf{x}^{(k)}. \tag{106}$$

where $\mathbf{x}^{(k)} = W^{\frac{1}{2}}\mathbf{s}^{(k)}$. Since $W\mathbb{1} = \mathbb{1}$ and W is positive-definite, it holds that $W^{\frac{1}{2}}$ is also positive-definite and $W^{\frac{1}{2}}\mathbb{1} = \mathbb{1}$. Taking global average over both sides of (106), we reach

$$\frac{1}{n}\mathbb{1}^T\mathbf{g}_{\mathbf{s}}^{(k)} = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i^{(k)}). \tag{107}$$

Substituting $(I-W)\bar{\mathbf{x}}^{(k)} = 0$ into (106), we have

$$\frac{1}{\gamma}(I-W)(\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}) = W\nabla f(\mathbf{x}^{(k)}) - W^{\frac{1}{2}}\mathbf{g}_{\mathbf{s}}^{(k)}, \tag{108}$$

which implies that

$$\begin{aligned}\frac{1-\lambda_2}{\gamma}\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\| &\le \|W\nabla f(\mathbf{x}^{(k)})\| + \|W^{\frac{1}{2}}\mathbf{g}_{\mathbf{s}}^{(k)}\| \\ &\le \|\nabla f(\mathbf{x}^{(k)})\| + \|\mathbf{g}_{\mathbf{s}}^{(k)}\|,\end{aligned} \tag{109}$$

where the first inequality holds by following arguments in (35). By taking expectation over the square of both sides, we achieve

$$\frac{(1-\lambda_2)^2}{\gamma^2}\mathbb{E}\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|^2 \le 2\mathbb{E}\|\nabla f(\mathbf{x}^{(k)})\|^2 + 2\mathbb{E}\|\mathbf{g}_{\mathbf{s}}^{(k)}\|^2. \tag{110}$$

Note that

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{(k)})\right\|^2\right] &= \sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i(x_i^{(k)})\right\|^2\right] \\ &= \sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i(x_i^{(k)}) - \nabla f_i(\bar{x}^{(k)}) + \nabla f_i(\bar{x}^{(k)}) - \nabla f(\bar{x}^{(k)}) + \nabla f(\bar{x}^{(k)})\right\|^2\right] \\ &\le 3L^2\mathbb{E}\left[\left\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\right\|^2\right] + 3\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i(\bar{x}^{(k)}) - \nabla f(\bar{x}^{(k)})\right\|^2\right] + 3n\mathbb{E}\left[\left\|\nabla f(\bar{x}^{(k)})\right\|^2\right] \\ &\le 3L^2\mathbb{E}\left[\left\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\right\|^2\right] + 3n\hat{b}^2 + 3n\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\bar{x}^{(k)})\right\|^2\right] \\ &\le 3L^2\mathbb{E}\left[\left\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\right\|^2\right] + 3n\hat{b}^2 + 3n\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i^{(k)}) + \frac{1}{n}\sum_{i=1}^{n}\left(\nabla f_i(\bar{x}^{(k)}) - \nabla f_i(x_i^{(k)})\right)\right\|^2\right] \\ &\le 9L^2\mathbb{E}\left[\left\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\right\|^2\right] + 3n\hat{b}^2 + 6n\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i^{(k)})\right\|^2\right]\end{aligned} \tag{111}$$

When $\gamma \le \frac{1-\lambda_2}{6L}$, i.e., $\frac{(1-\lambda_2)^2}{\gamma^2} - 18L^2 \ge \frac{(1-\lambda_2)^2}{2\gamma^2}$, we combine (110) and (111) to achieve

$$\mathbb{E}\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|^2 \le \frac{24n\gamma^2}{(1-\lambda_2)^2}\mathbb{E}\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i^{(k)})\|^2 + \frac{4\gamma^2}{(1-\lambda_2)^2}\mathbb{E}\|\mathbf{g}_{\mathbf{s}}^{(k)}\|^2 + \frac{12n\gamma^2\hat{b}^2}{(1-\lambda_2)^2}. \tag{112}$$

Recalling (107), we have

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i^{(k)})\right\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{n}\mathbb{1}^T\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] \le \frac{1}{n}\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] \tag{113}$$

Substituting (113) into (112), we achieve

$$\mathbb{E}\left[\left\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\right\|^2\right] \le \frac{28\gamma^2}{(1-\lambda_2)^2}\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] + \frac{12n\gamma^2\hat{b}^2}{(1-\lambda_2)^2}. \tag{114}$$

Therefore, when $\gamma \le \min\left\{\frac{(1-\beta)^2}{2\sqrt{2}L'\sqrt{\beta+\beta^2}}, \frac{(1-\beta)^2}{(2+(1-\beta)(3-\beta+2\beta^2))L'}, \frac{1-\lambda_2}{6L}\right\} = \mathcal{O}\left(\min\left\{\frac{(1-\beta)^2}{L'}, \frac{1-\lambda_2}{L}\right\}\right)$, combining (114) with (113) and (105), we have

$$\begin{aligned}
&\frac{1}{T}\sum_{k=0}^{T-1}\left(\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i^{(k)})\right\|^2\right] + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2\right]\right)\\
\le&\frac{1}{T}\sum_{k=0}^{T-1}\left(\frac{1}{n}\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] + \frac{1}{n}\left(\frac{28\gamma^2}{(1-\lambda_2)^2}\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] + \frac{12n\gamma^2\hat{b}^2}{(1-\lambda_2)^2}\right)\right)\\
=&\mathcal{O}\left(\frac{1}{n}\mathbb{E}\left[\left\|\mathbf{g}_{\mathbf{s}}^{(k)}\right\|^2\right] + \gamma^2\hat{b}^2\right)\\
=&\mathcal{O}\left(\frac{1-\beta}{\gamma T} + \frac{\gamma\sigma^2}{1-\beta} + \gamma^2\hat{b}^2\right).
\end{aligned} \tag{115}$$

□

F.2. Proof of Corollary 2

Proof of Corollary 2. If we let $\gamma = \mathcal{O}(\frac{1-\beta}{\sqrt{T}})$, it then holds that

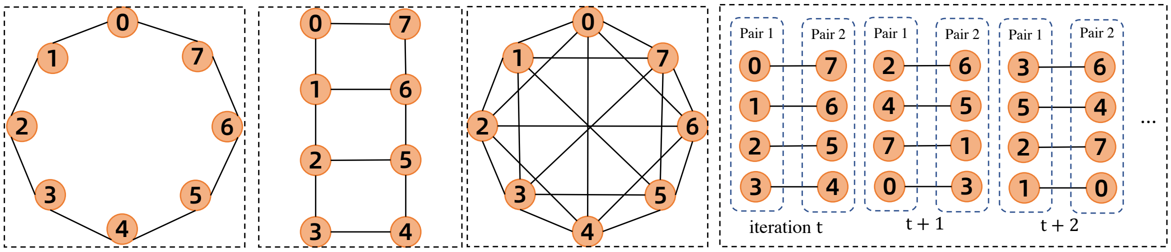
$$\begin{aligned}
&\frac{1}{T}\sum_{k=0}^{T-1}\left(\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i^{(k)})\right\|^2\right] + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2\right]\right)\\
=&\mathcal{O}\left(\frac{1-\beta}{\gamma T} + \frac{\gamma\sigma^2}{1-\beta} + \gamma^2\hat{b}^2\right)\\
=&\mathcal{O}\left(\frac{1}{\sqrt{T}} + \frac{\sigma^2}{\sqrt{T}} + \frac{(1-\beta)^2\hat{b}^2}{T}\right)\\
=&\mathcal{O}\left(\frac{1}{\sqrt{T}}\right).
\end{aligned} \tag{116}$$

□

G. More Experimental Details

G.1. Experimental setting for Table 1

Cifar-10 dataset contains 50,000 training samples and 10,000 validating samples. We follow the SOTA training scheme and train totally 200 epochs. The learning rate is linearly scaled and gradually warmed up form a relatively small value (e.g. 0.1) in the first 5 epochs. We decay the learning rate by a factor of 10 at 100, 150 epochs. To eliminate the effect of topology with different size in decentralized algorithms, we used 8 workers (i.e. 8 GPUs) in all decentralized training and changed the batch size of every single GPU respectively. For ImageNet experiments, the training setting is introduced in Sec. 7 in details.



a) Ring b) Mesh c) symmetric exponential d) bipartite random match

Figure 6. Different topologies with 8 nodes.

G.2. Experimental setting for linear regression (i.e., Figs. 2 and 3)

In this experiment, we consider a linear regression problem:

$$\min_{x\in\mathbb{R}^d} \quad \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{where} \quad f_i(x) = \frac{1}{2}\|A_i x - b_i\|^2. \tag{117}$$

In the above problem, we set $n = 8$ and all computing nodes are organized into the mesh topolology, see Fig. 6. The weight matrix is generated from the Metropolis-Hastings rule [42, Table 14.1] so that it satisfies Assumption A.3. Quantities $A_i \in \mathbb{R}^{50\times 30}$ and $b_i \in \mathbb{R}^{50}$ are local data held in node i. Each A_i is generated from the standard Gaussian distribution $\mathcal{N}(0,1)$, and $b_i = A_x x^o + s$ in which $x^o \in \mathbb{R}^{30}$ is a predefined solution, and s is a white noise with magnitude 0.01. For DSGD, DmSGD, and DecentLaM, we set learning rate $\gamma = 0.001$ and $\beta = 0.8$. To evaluate the inconsistency bias, we let each node i access the accurate gradient $\nabla f_i(x) = A_i^T(A_i x - b_i)$ rather than the stochastic gradient descent. The y-axis indicates the relative error $\frac{1}{n}\sum_{i=1}^{n}\|x_i^{(k)} - x^\star\|^2/\|x^\star\|^2$ in which $x^\star$ is the optimal solution to problem (117).

G.3. Network topologes used in Table 6

We empirically investigate a series of undirected deterministic and time-varying topologies. We generate the weight matrix W according to the Metropolis-Hastings rule [42, Table 14.1] so that it is satisfies Assumption A.3. A positive-definite W is not required in any of our experiments. We organize all computing nodes into the following topologies with BlueFog [1].

- **Ring**. All nodes forms a logical ring topology. Every node communicate with its direct neighbors (i.e. 2 peers).
- **Mesh**. All nodes forms a logical mesh topology. Every node communicate with its direct neighbors. It is a multi-peer topology.
- **Symmetric Exponential Graph [1, 4]**. The node with odd rank i communicates with even ranks $i + 2^0 - 1, i + 2^1 - 1, ..., i + 2^{\lfloor \log_2(n-1) \rfloor}$ by sending a message and waiting for a response.
- **Bipartite Random Match**. All nodes are evenly divided into two non-overlapping groups randomly per iteration. Communication is only allowed within each pair of the nodes. We keep the same random seed in all nodes to avoid deadlocks.

We also visually illustrate the aforementioned topologies with 8 nodes in Fig. 6.

G.4. Training time comparison

The end-to-end training speed varies across different models and network bandwidth conditions. For the sake of brevity, we compare the runtime of PmSGD, DmSGD and DecentLaM in training ResNet-50 (ImageNet) with different batch sizes and network bandwidths. The speed-up of DmSGD/DecentLaM over PmSGD is consistent for other backbones and tasks. In Fig. 7, DecentLaM and DmSGD have equivalent runtime because they are based on the same partial averaging operation. However, they can achieve 1.2∼1.9×speed-up compared to PmSGD, which is consistent with [4].

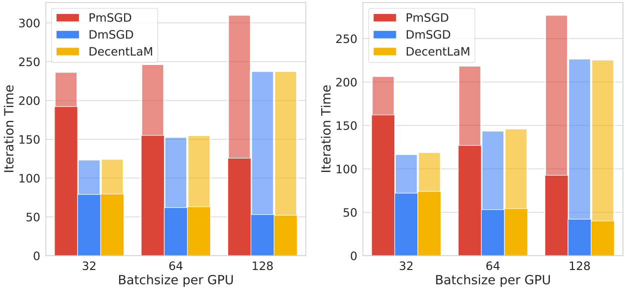


Figure 7. Runtime comparison on ResNet-50 with different batch sizes and network bandwidth (Left: 10Gbps; Right: 25Gbps). Each column indicates the averaged iteration runtime of 500 iterations. The thick part highlights the communication overhead.

G.5. Performance with different topologies.

We now examine how DecentLaM is robust to different topologies. To this end, we first organize all computing nodes into ring, mesh, symmetric exponential, or the bipartite random match topology, see Appendix G.3 for details of these topologies. Next we test the performance of DecentLaM on ResNet-50 with these topologies and the results are in Table 6. It is observed that DecentLaM has a consistent performance with different topologies. The ring topology is sparser than symmetric exponential topology. Interestingly, it is observed to have a better accuracy in the 32K batch-size setting. It is conjectured that the ring topology can help escape from shallow local minimums when batch-size is large. We leave the justification as the future work.

TOPOLOGY	Batch Size	
	16K	32K
Ring	76.65	76.34
Mesh	76.54	76.47
Symmetric Exponential	76.73	76.22
BIPARTITE RANDOM MATCH	76.53	76.11

Table 6. DecentLaM has consistent performance with different network topologies on ImageNet (ResNet-50).