A. Preliminary

Notation. We first introduce necessary notations as follows.

- $\mathbf{x}^{(k)} = [(\boldsymbol{x}_1^{(k)})^T; (\boldsymbol{x}_2^{(k)})^T; \cdots; (\boldsymbol{x}_n^{(k)})^T] \in \mathbb{R}^{n\times d}$
- $\nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k+1)}) = [\nabla F_1(\boldsymbol{x}_1^{(k)}; \boldsymbol{\xi}_1^{(k+1)})^T; \cdots; \nabla F_n(\boldsymbol{x}_n^{(k)}; \boldsymbol{\xi}_n^{(k+1)})^T] \in \mathbb{R}^{n\times d}$
- $\nabla f(\mathbf{x}^{(k)}) = [\nabla f_1(\boldsymbol{x}_1^{(k)})^T; \nabla f_2(\boldsymbol{x}_2^{(k)})^T; \cdots; \nabla f_n(\boldsymbol{x}_n^{(k)})^T] \in \mathbb{R}^{n\times d}$
- $\bar{\mathbf{x}}^{(k)} = [(\bar{\boldsymbol{x}}^{(k)})^T; (\bar{\boldsymbol{x}}^{(k)})^T; \cdots; (\bar{\boldsymbol{x}}^{(k)})^T] \in \mathbb{R}^{n\times d}$ where $\bar{\boldsymbol{x}}^{(k)} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^{(k)}$
- $\mathbf{x}^\star = [(x^\star)^T; (x^\star)^T; \cdots; (x^\star)^T] \in \mathbb{R}^{n\times d}$ where $x^\star$ is the global solution to problem (1).
- $W = [w_{ij}] \in \mathbb{R}^{n\times n}$.
- $\mathbb{1}_n = \mathrm{col}\{1, 1, \cdots, 1\} \in \mathbb{R}^n$.
- Given two matrices $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n\times d}$, we define inner product $\langle \mathbf{x}, \mathbf{y}\rangle = \mathrm{tr}(\mathbf{x}^T\mathbf{y})$ and the Frobenius norm $\|\mathbf{x}\|_F^2 = \langle \mathbf{x}, \mathbf{x}\rangle$.
- Given $W \in \mathbb{R}^{n\times n}$, we let $\|W\|_2 = \sigma_{\max}(W)$ where $\sigma_{\max}(\cdot)$ denote the maximum sigular value.

Gossip-PGA in matrix notation. For ease of analysis, we rewrite the main recursion of Gossip-PGA in matrix notation:

$$\mathbf{x}^{(k+1)} = \begin{cases} W(\mathbf{x}^{(k)} - \gamma\nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k+1)})) & \text{If } \mathrm{mod}(k+1, H) \neq 0 \\ \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T(\mathbf{x}^{(k)} - \gamma\nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k+1)})) & \text{otherwise} \end{cases} \tag{10}$$

Gradient noise. We repeat the definition of filtration in Assumption 2 here for convenience.

$$\mathcal{F}^{(k)} = \big\{\{\boldsymbol{x}_i^{(k)}\}_{i=1}^n, \{\boldsymbol{\xi}_i^{(k)}\}_{i=1}^n, \cdots, \{\boldsymbol{x}_i^{(0)}\}_{i=1}^n, \{\boldsymbol{\xi}_i^{(0)}\}_{i=1}^n\big\} \tag{11}$$

- With Assumption 2, we can evaluate the magnitude of the averaged gradient noise:

$$\begin{aligned} &\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^n \nabla F_i(\boldsymbol{x}_i^{(k-1)}; \boldsymbol{\xi}_i^{(k)}) - \frac{1}{n}\sum_{i=1}^n \nabla f_i(\boldsymbol{x}_i^{(k-1)})\|^2 | \mathcal{F}^{(k-1)}] \\ &\qquad \overset{(a)}{=} \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[\|\nabla F_i(\boldsymbol{x}_i^{(k-1)}; \boldsymbol{\xi}_i^{(k)}) - \nabla f_i(\boldsymbol{x}_i^{(k-1)})\|^2 | \mathcal{F}^{(k-1)}] \overset{(5)}{\leq} \frac{\sigma^2}{n} \end{aligned} \tag{12}$$

  where (a) holds because $\boldsymbol{\xi}_i^{(k)}$ is independent for any i and $\mathbb{E}[\nabla F_i(\boldsymbol{x}_i^{(k-1)}; \boldsymbol{\xi}_i^{(k)}) - \nabla f_i(\boldsymbol{x}_i^{(k-1)}) | \mathcal{F}^{(k-1)}] = 0$.

- We define gradient noise as $\boldsymbol{s}_i^{(k)} = \nabla F_i(\boldsymbol{x}_i^{(k-1)}; \boldsymbol{\xi}_i^{(k)}) - \nabla f_i(\boldsymbol{x}_i^{(k-1)})$. For any $0 \leq j \leq k < \ell$, it holds that

$$\mathbb{E}\big[\big(\boldsymbol{s}_i^{(k)}\big)^T \boldsymbol{s}_i^{(\ell)} | \mathcal{F}^{(j)}\big] \overset{(a)}{=} \mathbb{E}_{\mathcal{F}^{(\ell-1)}/\mathcal{F}^{(j)}}\big[\mathbb{E}\big[\big(\boldsymbol{s}_i^{(k)}\big)^T \boldsymbol{s}_i^{(\ell)} | \mathcal{F}^{(\ell-1)}\big]\big] = \mathbb{E}_{\mathcal{F}^{(\ell-1)}/\mathcal{F}^{(j)}}\big[\big(\boldsymbol{s}_i^{(k)}\big)^T \mathbb{E}[\boldsymbol{s}_i^{(\ell)} | \mathcal{F}^{(\ell-1)}]\big] \overset{(4)}{=} 0 \tag{13}$$

  where $\mathcal{F}^{(\ell-1)}/\mathcal{F}^{(j)} := \big\{\{\boldsymbol{x}_i^{(j+1)}\}_{i=1}^n, \{\boldsymbol{\xi}_i^{(j+1)}\}_{i=1}^n, \cdots, \{\boldsymbol{x}_i^{(\ell-1)}\}_{i=1}^n, \{\boldsymbol{\xi}_i^{(\ell-1)}\}_{i=1}^n\big\}$ and (a) holds due to the law of total expectation.

- For any $0 \leq k < \ell$, it holds that

$$\mathbb{E}[\|\boldsymbol{s}_i^{(\ell)}\|^2 | \mathcal{F}^{(k)}] = \mathbb{E}_{\mathcal{F}^{(\ell-1)}/\mathcal{F}^{(k)}}\big[\mathbb{E}[\|\boldsymbol{s}_i^{(\ell)}\|^2 | \mathcal{F}^{\ell-1}]\big] \overset{(5)}{\leq} \mathbb{E}_{\mathcal{F}^{(\ell-1)}/\mathcal{F}^{(k)}}\big[\sigma^2\big] = \sigma^2 \tag{14}$$

Smoothness. Since each $f_i(x)$ is assumed to be L-smooth in Assumption 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 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$:

$$f_i(\boldsymbol{x}) - f_i(\boldsymbol{y}) - \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2 \leq \langle \nabla f_i(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle \tag{15}$$

Smoothness and convexity. If each $f_i(x)$ is further assumed to be convex (see Assumption 4), it holds that $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ is also convex. For this scenario, it holds for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ that:

$$\|\nabla f(\boldsymbol{x}) - \nabla f(x^\star)\|^2 \leq 2L\big(f(\boldsymbol{x}) - f(x^\star)\big) \tag{16}$$

$$f_i(\boldsymbol{x}) - f_i(\boldsymbol{y}) \leq \langle \nabla f_i(\boldsymbol{x}), \boldsymbol{x}-\boldsymbol{y}\rangle \tag{17}$$

Network weighting matrix. Suppose a weighting matrix $W \in \mathbb{R}^{n\times n}$ satisfies Assumption 3, it holds that

$$\|W - \frac{1}{n}\mathbb{1}\mathbb{1}^T\|_2 \leq \beta, \quad \|(W - \frac{1}{n}\mathbb{1}\mathbb{1}^T)^k\|_2 \leq \beta^k, \quad \forall\, k. \tag{18}$$

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

$$\|W\mathbf{y}\|_F \leq \|W\|_2\|\mathbf{y}\|_F. \tag{19}$$

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

B. Convergence analysis for convex scenario

B.1. Proof Outline for Theorem 1

The following lemma established in (Koloskova et al., 2020, Lemma 8) shows how $\mathbb{E}\|\bar{\boldsymbol{x}}^{(k)} - x^\star\|^2$ evolves with iterations.

Lemma 1 (DESCENT LEMMA (Koloskova et al., 2020))**.** *When Assumptions 1–4 hold and step-size $\gamma < \frac{1}{4L}$, it holds for $k = 1, 2, \cdots$ that*

$$\mathbb{E}\|\bar{\boldsymbol{x}}^{(k)} - x^\star\|^2 \leq \mathbb{E}\|\bar{\boldsymbol{x}}^{(k-1)} - x^\star\|^2 - \gamma\big(\mathbb{E}f(\bar{\boldsymbol{x}}^{(k-1)}) - f(x^\star)\big) + \frac{3L\gamma}{2n}\mathbb{E}\|\mathbf{x}^{(k-1)} - \bar{\mathbf{x}}^{(k-1)}\|_F^2 + \frac{\gamma^2\sigma^2}{n}, \tag{20}$$

where $\bar{\mathbf{x}}^{(k)} = [(\bar{\boldsymbol{x}}^{(k)})^T; \cdots; (\bar{\boldsymbol{x}}^{(k)})^T] \in \mathbb{R}^{n\times d}$.

Remark 7. *It is worth noting that Lemma 1 also holds for the standard Gossip SGD algorithm. The periodic global averaging step does not affect this descent lemma.*

Next we establish the consensus lemmas in which Gossip-PGA is fundamentally different from Gossip SGD. Note that Gossip-PGA takes global average every H iterations. For any $k = 0, 1, \cdots$, we define

$$\tau(k) = \max\{\ell : \ell \leq k \text{ and } \mathrm{mod}(\ell, H) = 0\} \tag{21}$$

as the most recent iteration when global average is conducted. In Gossip-PGA, it holds that $\bar{\boldsymbol{x}}^{\tau(k)} = \boldsymbol{x}_i^{\tau(k)}$ for any $i \in [n]$. This is different from Gossip SGD in which $\bar{\boldsymbol{x}}^{(k)} = \boldsymbol{x}_i^{(k)}$ can only happen when $k = 0$.

For Gossip-PGA, the real challenge is to investigate how the periodic global averaging helps reduce consensus error and hence accelerate the convergence rate. In fact, there are two forces in Gossip-PGA that drive local model parameters to reach consensus: the gossip communication and the periodic global averaging. Each of these two forces is possible to dominate the consensus controlling in different scenarios:

Scenario I. Global averaging is more critical to guarantee consensus on large or sparse network, or when global averaging is conducted frequently.

Scenario II. Gossip communication is more critical to guarantee consensus on small or dense network, or when global averaging is conducted infrequently.

Ignoring either of the above scenario will lead to incomplete or even incorrect conclusions, as shown in Remark 5. In the following, we will establish a specific consensus lemma for each scenario and then unify them into one that precisely characterize how the consensus distance evolves with iterations in Gossip-PGA.

Lemma 2 (CONSENSUS LEMMA: GLOBAL AVERAGING DOMINATING)**.** *Under Assumptions 1–4, it holds for* $k = \tau(k), \tau(k)+1, \cdots, \tau(k)+H-1$ *that*

$$\mathbb{E}\|\mathbf{x}^{(k+1)} - \bar{\mathbf{x}}^{(k+1)}\|_F^2 \leq 6H\gamma^2\beta^2L^2 \sum_{\ell=\tau(k)}^{k} \beta^{k-\ell}\mathbb{E}\|\mathbf{x}^{(\ell)}-\bar{\mathbf{x}}^{(\ell)}\|_F^2 + 12nH\gamma^2\beta^2L \sum_{\ell=\tau(k)}^{k} \beta^{k-\ell}\mathbb{E}(f(\bar{\boldsymbol{x}}^{(\ell)}) - f(x^\star)) + 2n\gamma^2\beta^2C_\beta(3b^2+\sigma^2) \tag{22}$$

where $b^2 = \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^\star)\|^2$ *implies data heterogeneity and* $C_\beta = \sum_{k=0}^{H-1}\beta^k = (1-\beta^H)/(1-\beta)$.

Lemma 3 (CONSENSUS LEMMA: GOSSIP DOMINATING)**.** *Under Assumptions 1–4, it holds for* $k = \tau(k), \tau(k)+1, \cdots, \tau(k)+H-1$ *that*

$$\mathbb{E}\|\mathbf{x}^{(k+1)} - \bar{\mathbf{x}}^{(k+1)}\|_F^2 \leq \frac{6\gamma^2\beta^2L^2}{1-\beta} \sum_{\ell=\tau(k)}^{k} \beta^{k-\ell}\mathbb{E}\|\mathbf{x}^{(\ell)}-\bar{\mathbf{x}}^{(\ell)}\|_F^2 + \frac{12n\gamma^2\beta^2L}{1-\beta} \sum_{\ell=\tau(k)}^{k} \beta^{k-\ell}\mathbb{E}(f(\bar{\boldsymbol{x}}^{(\ell)}) - f(x^\star)) + 2n\gamma^2\beta^2C_\beta(\frac{3b^2}{1-\beta}+\sigma^2) \tag{23}$$

Observing Lemmas 2 and 3, it is found that bounds (22) and (23) are in the same shape except for some critical coefficients. With the following relation:

$$\begin{cases} y \leq a_1 x + b \\ y \leq a_2 x + b \end{cases} \quad \Longrightarrow \quad y \leq \min\{a_1, a_2\}x + b, \tag{24}$$

we can unify Lemmas 2 and 3 into:

Lemma 4 (UNIFIED CONSENSUS LEMMA)**.** *Under Assumptions 1–4, it holds for* $k = \tau(k), \tau(k)+1, \cdots, \tau(k)+H-1$ *that*

$$\mathbb{E}\|\mathbf{x}^{(k+1)} - \bar{\mathbf{x}}^{(k+1)}\|_F^2 \leq 6D_\beta\gamma^2\beta^2L^2 \sum_{\ell=\tau(k)}^{k} \beta^{k-\ell}\mathbb{E}\|\mathbf{x}^{(\ell)}-\bar{\mathbf{x}}^{(\ell)}\|_F^2 + 12nD_\beta\gamma^2\beta^2L \sum_{\ell=\tau(k)}^{k} \beta^{k-\ell}\mathbb{E}(f(\bar{\boldsymbol{x}}^{(\ell)}) - f(x^\star)) + 2n\gamma^2\beta^2C_\beta(3D_\beta b^2+\sigma^2) \tag{25}$$

where $b^2 = \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^\star)\|^2$ *implies data heterogeneity,* $C_\beta = \sum_{k=0}^{H-1}\beta^k = (1-\beta^H)/(1-\beta)$*, and* $D_\beta = \min\{\frac{1}{1-\beta}, H\}$.

Remark 8. *This lemma reflects how the network topology and the global averaging period contribute to the consensus controlling. For scenario I where the network is large or sparse such that* $1/(1-\beta) > H$*, Lemma 4 indicates that the consensus error is mainly controlled by the global averaging period (i.e.,* $D_\beta = H$*). On the other hand, for scenario II where the network is small or dense such that* $1/(1-\beta) < H$*, Lemma 4 indicates that the consensus error is mainly controlled by gossip communication (i.e.,* $D_\beta = 1/(1-\beta)$*).*

Using Lemma 4, we derive the upper bound of the weighted running average of $\mathbb{E}\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|_F^2$:

Lemma 5 (RUNNING CONSENSUS LEMMA)**.** *Suppose Assumptions 1–4 hold and step-size* $\gamma < 1/(4L\beta D_\beta)$*, it holds for* $T > 0$ *that*

$$\frac{1}{T+1}\sum_{k=0}^{T}\mathbb{E}\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|_F^2 \leq \frac{2c_2D_\beta}{T+1}\sum_{k=0}^{T}\big(\mathbb{E}f(\bar{\boldsymbol{x}}^{(k)}) - f(x^\star)\big) + 2c_3 \tag{26}$$

where c_2 *and* c_3 *are constants defined as*

$$c_2 = 12n\beta^2D_\beta\gamma^2L, \tag{27}$$

$$c_3 = 2n\beta^2\gamma^2C_\beta(3D_\beta b^2 + \sigma^2) \tag{28}$$

With Lemmas 1 and 5, we can establish the final convergence Theorem 1, see the proof in Sec. B.5.

B.2. Proof of Lemma 1.

This lemma was first established in (Koloskova et al., 2020, Lemma 8). We made slight improvement to tight constants appeared in step-size ranges and upper bound (20). For readers' convenience, we repeat arguments here.

Recall the algorithm in (10). By taking the average on both sides, we reach that

$$\bar{\boldsymbol{x}}^{(k)} - x^\star = \bar{\boldsymbol{x}}^{(k-1)} - x^\star - \frac{\gamma}{n}\sum_{i=1}^n \nabla F_i(\boldsymbol{x}_i^{(k-1)}; \boldsymbol{\xi}_i^{(k)}), \quad \forall k = 1, 2, \cdots \tag{29}$$

By taking expectation over the square of both sides of the above recursion conditioned on $\mathcal{F}^{(k-1)}$, we have

$$\begin{aligned}
&\mathbb{E}[\|\bar{\boldsymbol{x}}^{(k)} - x^\star\|^2 | \mathcal{F}^{(k-1)}] \\
&\overset{(4)}{=} \|\bar{\boldsymbol{x}}^{(k-1)} - x^\star - \frac{\gamma}{n}\sum_{i=1}^n \nabla f_i(\boldsymbol{x}_i^{(k-1)})\|^2 + \gamma^2 \mathbb{E}[\|\frac{1}{n}\sum_{i=1}^n \nabla F_i(\boldsymbol{x}_i^{(k-1)}; \boldsymbol{\xi}_i^{(k)}) - \frac{1}{n}\sum_{i=1}^n \nabla f_i(\boldsymbol{x}_i^{(k-1)})\|^2 | \mathcal{F}^{(k-1)}] \\
&\overset{(12)}{\leq} \|\bar{\boldsymbol{x}}^{(k-1)} - x^\star - \frac{\gamma}{n}\sum_{i=1}^n \nabla f_i(\boldsymbol{x}_i^{(k-1)})\|^2 + \frac{\gamma^2\sigma^2}{n}
\end{aligned} \tag{30}$$

Note that the first term can be expanded as follows.

$$\begin{aligned}
&\|\bar{\boldsymbol{x}}^{(k-1)} - x^\star - \frac{\gamma}{n}\sum_{i=1}^n \nabla f_i(\boldsymbol{x}_i^{(k-1)})\|^2 \\
&= \|\bar{\boldsymbol{x}}^{(k-1)} - x^\star - \frac{\gamma}{n}\sum_{i=1}^n [\nabla f_i(\boldsymbol{x}_i^{(k-1)}) - \nabla f_i(x^\star)]\|^2 \\
&= \|\bar{\boldsymbol{x}}^{(k-1)} - x^\star\|^2 \underbrace{- \frac{2\gamma}{n}\sum_{i=1}^n \langle \bar{\boldsymbol{x}}^{(k-1)} - x^\star, \nabla f_i(\boldsymbol{x}_i^{(k-1)}) - \nabla f_i(x^\star)\rangle}_{(A)} + \underbrace{\gamma^2 \|\frac{1}{n}\sum_{i=1}^n [\nabla f_i(\boldsymbol{x}_i^{(k-1)}) - \nabla f_i(x^\star)]\|^2}_{(B)}
\end{aligned} \tag{31}$$

We now bound the term (A):

$$\begin{aligned}
&\frac{2\gamma}{n}\sum_{i=1}^n \langle \bar{\boldsymbol{x}}^{(k-1)} - x^\star, \nabla f_i(\boldsymbol{x}_i^{(k-1)}) - \nabla f_i(x^\star)\rangle \\
&= \frac{2\gamma}{n}\sum_{i=1}^n \langle \bar{\boldsymbol{x}}^{(k-1)} - x^\star, \nabla f_i(\boldsymbol{x}_i^{(k-1)})\rangle \\
&= \frac{2\gamma}{n}\sum_{i=1}^n \langle \bar{\boldsymbol{x}}^{(k-1)} - \boldsymbol{x}_i^{(k-1)}, \nabla f_i(\boldsymbol{x}_i^{(k-1)})\rangle + \frac{2\gamma}{n}\sum_{i=1}^n \langle \boldsymbol{x}_i^{(k-1)} - x^\star, \nabla f_i(\boldsymbol{x}_i^{(k-1)})\rangle \\
&\overset{(a)}{\geq} \frac{2\gamma}{n}\sum_{i=1}^n \Big(f_i(\bar{\boldsymbol{x}}^{(k-1)}) - f_i(\boldsymbol{x}_i^{(k-1)}) - \frac{L}{2}\|\bar{\boldsymbol{x}}^{(k-1)} - \boldsymbol{x}_i^{(k-1)}\|^2 \Big) + \frac{2\gamma}{n}\sum_{i=1}^n \Big(f_i(\boldsymbol{x}_i^{(k-1)}) - f_i(x^\star) \Big) \\
&= \frac{2\gamma}{n}\sum_{i=1}^n \Big(f_i(\bar{\boldsymbol{x}}^{(k-1)}) - f_i(x^\star) \Big) - \frac{\gamma L}{n}\|\bar{\mathbf{x}}^{(k-1)} - \mathbf{x}^{(k-1)}\|_F^2 \\
&= 2\gamma\big(f(\bar{\boldsymbol{x}}^{(k-1)}) - f(x^\star) \big) - \frac{\gamma L}{n}\|\bar{\mathbf{x}}^{(k-1)} - \mathbf{x}^{(k-1)}\|_F^2
\end{aligned} \tag{32}$$

where (a) holds because of the inequality (15) and (17). We next bound term (B) in (31):

$$\gamma^2 \|\frac{1}{n}\sum_{i=1}^n [\nabla f_i(\boldsymbol{x}_i^{(k-1)}) - \nabla f_i(x^\star)]\|^2$$

$$
\begin{aligned}
&= \gamma^2 \| \frac{1}{n} \sum_{i=1}^{n} [\nabla f_i(\boldsymbol{x}_i^{(k-1)}) - \nabla f_i(\bar{\boldsymbol{x}}^{(k-1)}) + \nabla f_i(\bar{\boldsymbol{x}}^{(k-1)}) - \nabla f_i(x^\star)] \|^2 \\
&\overset{(3)}{\leq} \frac{2\gamma^2 L^2}{n} \|\mathbf{x}^{(k-1)} - \bar{\mathbf{x}}^{(k-1)}\|_F^2 + 2\gamma^2 \|\nabla f(\bar{\boldsymbol{x}}^{(k-1)}) - \nabla f(x^\star)\|^2 \\
&\overset{(16)}{\leq} \frac{2\gamma^2 L^2}{n} \|\mathbf{x}^{(k-1)} - \bar{\mathbf{x}}^{(k-1)}\|_F^2 + 4L\gamma^2 \big(f(\bar{\boldsymbol{x}}^{(k-1)}) - f(x^\star)\big).
\end{aligned} \tag{33}
$$

Substituting (33) and (32) into (31), we have

$$
\begin{aligned}
&\|\bar{\boldsymbol{x}}^{(k-1)} - x^\star - \frac{\gamma}{n} \sum_{i=1}^{n} \nabla f_i(\boldsymbol{x}_i^{(k-1)})\|^2 \\
&\leq \|\bar{\boldsymbol{x}}^{(k-1)} - x^\star\|^2 - 2\gamma(1 - 2L\gamma)\big(f(\bar{\boldsymbol{x}}^{(k-1)}) - f(x^\star)\big) + \Big(\frac{\gamma L}{n} + \frac{2\gamma^2 L^2}{n}\Big) \|\bar{\mathbf{x}}^{(k-1)} - \mathbf{x}^{(k-1)}\|_F^2 \\
&\leq \|\bar{\boldsymbol{x}}^{(k-1)} - x^\star\|^2 - \gamma\big(f(\bar{\boldsymbol{x}}^{(k-1)}) - f(x^\star)\big) + \frac{3\gamma L}{2n} \|\bar{\mathbf{x}}^{(k-1)} - \mathbf{x}^{(k-1)}\|_F^2
\end{aligned} \tag{34}
$$

where the last inequality holds when $\gamma < \frac{1}{4L}$. Substituting the above inequality into (30) and taking expectation over the filtration, we reach the result in (20).

B.3. Proofs of Lemma 2 and 3.

Note the gossip averaging is conducted when $k = \tau(k), \tau(k)+1, \cdots, \tau(k)+H-1$, i.e.,

$$
\mathbf{x}^{(k+1)} = W(\mathbf{x}^{(k)} - \gamma \nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k+1)})). \tag{35}
$$

Since $\bar{\mathbf{x}}^{(k+1)} = \frac{1}{n}\mathbb{1}\mathbb{1}^T \mathbf{x}^{(k+1)}$, it holds that

$$
\bar{\mathbf{x}}^{(k+1)} = \frac{1}{n}\mathbb{1}\mathbb{1}^T (\mathbf{x}^{(k)} - \gamma \nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k+1)})). \tag{36}
$$

With the above two recursions, we have

$$
\mathbf{x}^{(k+1)} - \bar{\mathbf{x}}^{(k+1)} = (W - \frac{1}{n}\mathbb{1}\mathbb{1}^T)(\mathbf{x}^{(k)} - \gamma \nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k+1)})) \tag{37}
$$

In the following we will derive two upper bounds for $\mathbb{E}\|\mathbf{x}^{(k+1)} - \bar{\mathbf{x}}^{(k+1)}\|^2$.

Bound in Lemma 2. With (37), we have

$$
\begin{aligned}
\mathbf{x}^{(k+1)} - \bar{\mathbf{x}}^{(k+1)} &= (W - \frac{1}{n}\mathbb{1}\mathbb{1}^T)(\mathbf{x}^{(k)} - \gamma \nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k+1)})) \\
&= (W - \frac{1}{n}\mathbb{1}\mathbb{1}^T)(\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)} - \gamma \nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k+1)})) \\
&= (W - \frac{1}{n}\mathbb{1}\mathbb{1}^T)^{k+1-\tau(k)}(\mathbf{x}^{(\tau(k))} - \bar{\mathbf{x}}^{(\tau(k))}) - \gamma \sum_{\ell=\tau(k)}^{k} (W - \frac{1}{n}\mathbb{1}\mathbb{1}^T)^{k+1-\ell} \nabla F(\mathbf{x}^{(\ell)}; \boldsymbol{\xi}^{(\ell+1)}) \\
&= -\gamma \sum_{\ell=\tau(k)}^{k} (W - \frac{1}{n}\mathbb{1}\mathbb{1}^T)^{k+1-\ell} \nabla F(\mathbf{x}^{(\ell)}; \boldsymbol{\xi}^{(\ell+1)})
\end{aligned} \tag{38}
$$

where the last equality holds because $\mathbf{x}^{(\tau(k))} = \bar{\mathbf{x}}^{(\tau(k))}$ after the global averaging at iteration $\tau(k)$. With the above inequality, we have

$$
\begin{aligned}
&\mathbb{E}[\|\mathbf{x}^{(k+1)} - \bar{\mathbf{x}}^{(k+1)}\|_F^2 | \mathcal{F}^{(\tau(k))}] \\
&= \gamma^2 \mathbb{E}[\| \sum_{\ell=\tau(k)}^{k} (W - \frac{1}{n}\mathbb{1}\mathbb{1}^T)^{k+1-\ell} \nabla F(\mathbf{x}^{(\ell)}; \boldsymbol{\xi}^{(\ell+1)})\|_F^2 | \mathcal{F}^{(\tau(k))}]
\end{aligned}
$$

$$
\begin{aligned}
&\leq 2\gamma^2\mathbb{E}[\|\sum_{\ell=\tau(k)}^{k}(W-\frac{1}{n}\mathbb{1}\mathbb{1}^T)^{k+1-\ell}\nabla f(\mathbf{x}^{(\ell)})\|_F^2|\mathcal{F}^{(\tau(k))}] \\
&\quad +2\gamma^2\mathbb{E}\Big[\|\sum_{\ell=\tau(k)}^{k}(W-\frac{1}{n}\mathbb{1}\mathbb{1}^T)^{k+1-\ell}[\nabla F(\mathbf{x}^{(\ell)};\boldsymbol{\xi}^{(\ell+1)})-\nabla f(\mathbf{x}^{\ell})]\|_F^2|\mathcal{F}^{(\tau(k))}\Big] \\
&\overset{(13)}{=} 2\gamma^2\mathbb{E}[\|\sum_{\ell=\tau(k)}^{k}(W-\frac{1}{n}\mathbb{1}\mathbb{1}^T)^{k+1-\ell}\nabla f(\mathbf{x}^{(\ell)})\|_F^2|\mathcal{F}^{(\tau(k))}] \\
&\quad +2\gamma^2\sum_{\ell=\tau(k)}^{k}\mathbb{E}\Big[\|(W-\frac{1}{n}\mathbb{1}\mathbb{1}^T)^{k+1-\ell}[\nabla F(\mathbf{x}^{(\ell)};\boldsymbol{\xi}^{(\ell+1)})-\nabla f(\mathbf{x}^{\ell})]\|_F^2|\mathcal{F}^{(\tau(k))}\Big] \\
&\overset{(a)}{\leq} 2\gamma^2(k+1-\tau(k))\sum_{\ell=\tau(k)}^{k}\beta^{2(k+1-\ell)}\mathbb{E}[\|\nabla f(\mathbf{x}^{(\ell)})\|_F^2|\mathcal{F}^{(\tau(k))}]+2n\gamma^2\sigma^2\sum_{\ell=\tau(k)}^{k}\beta^{2(k+1-\ell)} \\
&\leq 2\gamma^2 H\sum_{\ell=\tau(k)}^{k}\beta^{2(k+1-\ell)}\mathbb{E}[\|\nabla f(\mathbf{x}^{(\ell)})\|_F^2|\mathcal{F}^{(\tau(k))}]+2n\gamma^2\beta^2\sigma^2C_\beta \qquad (39)
\end{aligned}
$$

where inequality (a) holds because of (14), (18) and (19), and the last inequality holds because $\sum_{\ell=\tau(k)}^{k}\beta^{2(k+1-\ell)} \leq \sum_{\ell=1}^{H}\beta^{2\ell}=\beta^2(1-\beta^{2H})/(1-\beta^2)\leq\beta^2(1-\beta^H)/(1-\beta)=\beta^2C_\beta$ where we define $C_\beta=\sum_{\ell=0}^{H-1}\beta^\ell=(1-\beta^H)/(1-\beta)$. Note that

$$
\begin{aligned}
\|\nabla f(\mathbf{x}^{(\ell)})\|_F^2 &= \|\nabla f(\mathbf{x}^{(\ell)})-\nabla f(\bar{\mathbf{x}}^{(\ell)})+\nabla f(\bar{\mathbf{x}}^{(\ell)})-\nabla f(\mathbf{x}^\star)+\nabla f(\mathbf{x}^\star)\|_F^2 \\
&\leq 3\|\nabla f(\mathbf{x}^{(\ell)})-\nabla f(\bar{\mathbf{x}}^{(\ell)})\|_F^2+3\|\nabla f(\bar{\mathbf{x}}^{(\ell)})-\nabla f(\mathbf{x}^\star)\|_F^2+3\|\nabla f(\mathbf{x}^\star)\|_F^2 \\
&\leq 3L^2\|\mathbf{x}^{(\ell)}-\bar{\mathbf{x}}^{(\ell)}\|_F^2+6nL(f(\bar{\boldsymbol{x}}^{(\ell)})-f(x^\star))+3nb^2 \qquad (40)
\end{aligned}
$$

where the last inequality holds because of (3) and (16). Notation b^2 is defined as $b^2=\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x^\star)\|^2$. Substituting (40) into (39), it holds for $k=\tau(k),\tau(k)+1,\cdots,\tau(k)+H-1$ that

$$
\begin{aligned}
&\mathbb{E}[\|\mathbf{x}^{(k+1)}-\bar{\mathbf{x}}^{(k+1)}\|_F^2|\mathcal{F}^{(\tau(k))}] \\
&\leq 6H\gamma^2L^2\sum_{\ell=\tau(k)}^{k}\beta^{2(k+1-\ell)}\mathbb{E}[\|\mathbf{x}^{(\ell)}-\bar{\mathbf{x}}^{(\ell)}\|_F^2|\mathcal{F}^{(\tau(k))}] \\
&\quad +12nH\gamma^2L\sum_{\ell=\tau(k)}^{k}\beta^{2(k+1-\ell)}\mathbb{E}[f(\bar{\boldsymbol{x}}^{(\ell)})-f(x^\star)|\mathcal{F}^{(\tau(k))}]+2n\gamma^2\beta^2C_\beta(3Hb^2+\sigma^2) \qquad (41)
\end{aligned}
$$

By taking expectations over the filtration $\mathcal{F}^{(\tau(k))}$, we have

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{x}^{(k+1)}-\bar{\mathbf{x}}^{(k+1)}\|_F^2 \\
&\leq 6H\beta^2\gamma^2L^2\sum_{\ell=\tau(k)}^{k}\beta^{2(k-\ell)}\mathbb{E}\|\mathbf{x}^{(\ell)}-\bar{\mathbf{x}}^{(\ell)}\|_F^2+12nH\beta^2\gamma^2L\sum_{\ell=\tau(k)}^{k}\beta^{2(k-\ell)}\mathbb{E}(f(\bar{\boldsymbol{x}}^{(\ell)})-f(x^\star)) \\
&\quad +2n\gamma^2\beta^2C_\beta(3Hb^2+\sigma^2) \\
&\leq 6H\beta^2\gamma^2L^2\sum_{\ell=\tau(k)}^{k}\beta^{k-\ell}\,\mathbb{E}\|\mathbf{x}^{(\ell)}-\bar{\mathbf{x}}^{(\ell)}\|_F^2+12nH\beta^2\gamma^2L\sum_{\ell=\tau(k)}^{k}\beta^{k-\ell}\,\mathbb{E}(f(\bar{\boldsymbol{x}}^{(\ell)})-f(x^\star)) \\
&\quad +2n\gamma^2\beta^2C_\beta(3Hb^2+\sigma^2) \qquad (42)
\end{aligned}
$$

Bound in Lemma 3. With (37), it holds for $k=\tau(k),\tau(k)+1,\cdots,\tau(k)+H-1$ that

$$\mathbb{E}[\|\mathbf{x}^{(k+1)}-\bar{\mathbf{x}}^{(k+1)}\|_F^2|\mathcal{F}^{(k)}]$$

$$
\begin{aligned}
&= \mathbb{E}[\|(W-\frac{1}{n}\mathbb{1}\mathbb{1}^T)\Big(\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}-\gamma\nabla F(\mathbf{x}^{(k)};\boldsymbol{\xi}^{(k+1)})\Big)\|_F^2|\mathcal{F}^{(k)}] \\
&\overset{(4)}{=} \|(W-\frac{1}{n}\mathbb{1}\mathbb{1}^T)\Big(\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}-\gamma\nabla f(\mathbf{x}^{(k)})\Big)\|_F^2+\gamma^2\mathbb{E}[\|(W-\frac{1}{n}\mathbb{1}\mathbb{1}^T)\big(\nabla F(\mathbf{x}^{(k)};\boldsymbol{\xi}^{(k+1)})-\nabla f(\mathbf{x}^{(k)})\big)\|_F^2|\mathcal{F}^{(k)}] \\
&\leq \|(W-\frac{1}{n}\mathbb{1}\mathbb{1}^T)\Big(\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}-\gamma\nabla f(\mathbf{x}^{(k)})\Big)\|_F^2+n\gamma^2\beta^2\sigma^2 \qquad (43)
\end{aligned}
$$

where the last inequality holds because of (5) and (18). We now bound the first term:

$$
\begin{aligned}
&\|(W-\frac{1}{n}\mathbb{1}\mathbb{1}^T)\Big(\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}-\gamma\nabla f(\mathbf{x}^{(k)})\Big)\|_F^2 \\
&\overset{(a)}{\leq} \frac{1}{t}\|(W-\frac{1}{n}\mathbb{1}\mathbb{1}^T)\big(\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}\big)\|_F^2+\frac{\gamma^2}{1-t}\|(W-\frac{1}{n}\mathbb{1}\mathbb{1}^T)\nabla f(\mathbf{x}^{(k)})\|_F^2 \\
&\overset{(b)}{=} \beta\|\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}\|_F^2+\frac{\beta^2\gamma^2}{1-\beta}\|\nabla f(\mathbf{x}^{(k)})\|_F^2 \\
&= \beta\|\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}\|_F^2+\frac{\beta^2\gamma^2}{1-\beta}\|\nabla f(\mathbf{x}^{(k)})-\nabla f(\bar{\mathbf{x}}^{(k)})+\nabla f(\bar{\mathbf{x}}^{(k)})-\nabla f(\mathbf{x}^\star)+\nabla f(\mathbf{x}^\star)\|_F^2 \\
&\overset{(c)}{\leq} \beta\|\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}\|_F^2+\frac{3\beta^2\gamma^2L^2}{1-\beta}\|\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}\|_F^2+\frac{6n\beta^2\gamma^2L}{1-\beta}\big(f(\bar{\boldsymbol{x}}^{(k)})-f(x^\star)\big)+\frac{3n\beta^2\gamma^2b^2}{1-\beta} \qquad (44)
\end{aligned}
$$

where (a) holds because of the Jensen's inequality for any $t\in(0,1)$, (b) holds by setting $t=\beta$, and (c) holds because of (3) and (16). Quantity $b^2=\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(x^\star)\|^2$ in the last inequality. Substituting (44) into (43), we have

$$
\begin{aligned}
&\mathbb{E}[\|\mathbf{x}^{(k+1)}-\bar{\mathbf{x}}^{(k+1)}\|_F^2|\mathcal{F}^{(k)}] \\
&\leq \beta\|\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}\|_F^2+\frac{3\beta^2\gamma^2L^2}{1-\beta}\|\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}\|_F^2+\frac{6n\beta^2\gamma^2L}{1-\beta}\big(f(\bar{\boldsymbol{x}}^{(k)})-f(x^\star)\big)+n\gamma^2\beta^2\sigma^2+\frac{3n\beta^2\gamma^2b^2}{1-\beta} \\
&= \beta^{k+1-\tau(k)}\|\mathbf{x}^{(\tau(k))}-\bar{\mathbf{x}}^{(\tau(k))}\|_F^2+\frac{3\beta^2\gamma^2L^2}{1-\beta}\sum_{\ell=\tau(k)}^{k}\beta^{k-\ell}\|\mathbf{x}^{(\ell)}-\bar{\mathbf{x}}^{(\ell)}\|_F^2+\frac{6n\beta^2\gamma^2L}{1-\beta}\sum_{\ell=\tau(k)}^{k}\beta^{k-\ell}\left(f(\bar{\boldsymbol{x}}^{(\ell)})-f(x^\star)\right) \\
&\quad +n\gamma^2\beta^2C_\beta(\frac{3b^2}{1-\beta}+\sigma^2) \\
&= \frac{3\beta^2\gamma^2L^2}{1-\beta}\sum_{\ell=\tau(k)}^{k}\beta^{k-\ell}\|\mathbf{x}^{(\ell)}-\bar{\mathbf{x}}^{(\ell)}\|_F^2+\frac{6n\beta^2\gamma^2L}{1-\beta}\sum_{\ell=\tau(k)}^{k}\beta^{k-\ell}\left(f(\bar{\boldsymbol{x}}^{(\ell)})-f(x^\star)\right)+n\gamma^2\beta^2C_\beta(\frac{3b^2}{1-\beta}+\sigma^2) \qquad (45)
\end{aligned}
$$

By taking expectation over the filtration $\mathcal{F}^{(k)}$, we have

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{x}^{(k+1)}-\bar{\mathbf{x}}^{(k+1)}\|_F^2 \\
&\leq \frac{3\beta^2\gamma^2L^2}{1-\beta}\sum_{\ell=\tau(k)}^{k}\beta^{k-\ell}\,\mathbb{E}\|\mathbf{x}^{(\ell)}-\bar{\mathbf{x}}^{(\ell)}\|_F^2+\frac{6n\beta^2\gamma^2L}{1-\beta}\sum_{\ell=\tau(k)}^{k}\beta^{k-\ell}\,\mathbb{E}(f(\bar{\boldsymbol{x}}^{(\ell)})-f(x^\star))+n\gamma^2\beta^2C_\beta(\frac{3b^2}{1-\beta}+\sigma^2) \\
&< \frac{6\beta^2\gamma^2L^2}{1-\beta}\sum_{\ell=\tau(k)}^{k}\beta^{k-\ell}\,\mathbb{E}\|\mathbf{x}^{(\ell)}-\bar{\mathbf{x}}^{(\ell)}\|_F^2+\frac{12n\beta^2\gamma^2L}{1-\beta}\sum_{\ell=\tau(k)}^{k}\beta^{k-\ell}\,\mathbb{E}(f(\bar{\boldsymbol{x}}^{(\ell)})-f(x^\star))+2n\gamma^2\beta^2C_\beta(\frac{3b^2}{1-\beta}+\sigma^2) \qquad (46)
\end{aligned}
$$

B.4. Proof of Lemma 5.

To simplify the notation, we define

$$
\begin{aligned}
&A^{(k)}=\mathbb{E}\|\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}\|_F^2, \quad B^{(k)}=\mathbb{E}f(\bar{\boldsymbol{x}}^{(k)})-f(x^\star), \\
&c_1=6D_\beta\beta^2\gamma^2L^2, \qquad c_2=12nD_\beta\beta^2\gamma^2L, \qquad c_3=2n\beta^2\gamma^2C_\beta(3D_\beta b^2+\sigma^2). \qquad (47)
\end{aligned}
$$

Using these notations, we rewrite (22) for any $k=0,1,2,\cdots$ that

$$\begin{cases} A^{(k)} \leq c_1 \sum_{\ell=\tau(k)}^{k-1} \beta^{k-1-\ell} A^{(\ell)} + c_2 \sum_{\ell=\tau(k)}^{k-1} \beta^{k-1-\ell} B^{(\ell)} + c_3 & \text{if } k > \tau(k) \\ A^{(k)} = 0 & \text{if } k = \tau(k) \end{cases} \tag{48}$$

We next define

$$\Gamma_T := \{k | 0 \leq k \leq T \text{ and } \mathrm{mod}(k,H)=0\}, \quad \Gamma_T^c := \{k | 0 \leq k \leq T \text{ and } \mathrm{mod}(k,H)\neq 0\}. \tag{49}$$

By taking the running average over both sides in (48) and recalling $A^{(\tau(k))}=0$, it holds that

$$\sum_{k=0}^{T} A^{(k)} \leq c_1 \sum_{k\in\Gamma_T^c} \sum_{\ell=\tau(k)}^{k-1} \beta^{k-1-\ell} A^{(\ell)} + c_2 \sum_{k\in\Gamma_T^c} \sum_{\ell=\tau(k)}^{k-1} \beta^{k-1-\ell} B^{(\ell)} + c_3(T+1) \tag{50}$$

We further define

$$\Theta_T := \{\ell | 0 \leq \ell \leq T-1 \text{ and } \mathrm{mod}(\ell+1,H)=0\}, \quad \Theta_T^c := \{\ell | 0 \leq \ell \leq T-1 \text{ and } \mathrm{mod}(\ell+1,H)\neq 0\}. \tag{51}$$

With these notations, we have

$$\begin{aligned} \sum_{k=0}^{T} A^{(k)} &\leq c_1 \sum_{k\in\Gamma_T^c} \sum_{\ell=\tau(k)}^{k-1} \beta^{k-1-\ell} A^{(\ell)} + c_2 \sum_{k\in\Gamma_T^c} \sum_{\ell=\tau(k)}^{k-1} \beta^{k-1-\ell} B^{(\ell)} + c_3(T+1) \\ &= c_1 \sum_{\ell\in\Theta_T^c} A^{(\ell)} \Big(\sum_{k=\ell+1}^{\tau(\ell)+H-1} \beta^{k-1-\ell} \Big) + c_2 \sum_{\ell\in\Theta_T^c} B^{(\ell)} \Big(\sum_{k=\ell+1}^{\tau(\ell)+H-1} \beta^{k-1-\ell} \Big) + c_3(T+1) \\ &\overset{(a)}{\leq} c_1 C_\beta \sum_{\ell\in\Theta_T^c} A^{(\ell)} + c_2 C_\beta \sum_{\ell\in\Theta_T^c} B^{(\ell)} + c_3(T+1) \\ &\overset{(b)}{\leq} c_1 C_\beta \sum_{k=0}^{T} A^{(k)} + c_2 C_\beta \sum_{k=0}^{T} B^{(k)} + c_3(T+1) \\ &\overset{(c)}{\leq} c_1 D_\beta \sum_{k=0}^{T} A^{(k)} + c_2 D_\beta \sum_{k=0}^{T} B^{(k)} + c_3(T+1) \end{aligned} \tag{52}$$

where (a) holds because $\sum_{k=\ell+1}^{\tau(\ell)+H-1} \beta^{k-1-\ell} \leq \sum_{k=0}^{H-1} \beta^k = C_\beta$, (b) holds because $A^{(k)} \geq 0$ and $B^{(k)} \geq 0$, and (c) holds because $C_\beta = (1-\beta^H)/(1-\beta) \leq \min\{\frac{1}{1-\beta}, H\} = D_\beta$. If step-size γ is sufficiently small such that $1 - c_1 D_\beta \geq \frac{1}{2}$, it holds that

$$\sum_{k=0}^{T} A^{(k)} \leq 2c_2 D_\beta \sum_{k=0}^{T} B^{(k)} + 2c_3(T+1). \tag{53}$$

To guarantee $1 - c_1 D_\beta \geq \frac{1}{2}$, it is enough to let $\gamma \leq 1/(4L\beta D_\beta)$.

B.5. Proof of Theorem 1

Following the notation in (47), we further define $F^{(k)} := \mathbb{E}\|\bar{\boldsymbol{x}}^{(k)} - x^\star\|_F^2$. With these notations, the inequality (20) becomes

$$B^{(k)} \leq \frac{F^{(k)}}{\gamma} - \frac{F^{(k+1)}}{\gamma} + \frac{\gamma\sigma^2}{n} + \frac{3L}{2n} A^{(k)} \tag{54}$$

Taking weighted running average over the above inequality to get

$$\frac{1}{T+1} \sum_{k=0}^{T} B^{(k)} \leq \frac{F^{(0)}}{(T+1)\gamma} + \frac{3L}{2n(T+1)} \sum_{k=0}^{T} A^{(k)} + \frac{\gamma\sigma^2}{n}$$

$$\overset{(53)}{\leq} \frac{F^{(0)}}{(T+1)\gamma} + \frac{6Lc_2D_\beta}{n(T+1)}\sum_{k=0}^{T} B^{(k)} + \frac{3Lc_3}{n} + \frac{\gamma\sigma^2}{n}$$
$$\leq \frac{F^{(0)}}{(T+1)\gamma} + \frac{1}{2(T+1)}\sum_{k=0}^{T} B^{(k)} + \frac{3Lc_3}{n} + \frac{\gamma\sigma^2}{n} \tag{55}$$

where the last inequality holds when $\gamma \leq 1/(12L\beta D_\beta)$. Substituting c_3 into the above inequality, we have

$$\frac{1}{T+1}\sum_{k=0}^{T} B^{(k)} \leq \frac{2F^{(0)}}{(T+1)\gamma} + \frac{2\gamma\sigma^2}{n} + 12L\beta^2\gamma^2C_\beta\sigma^2 + 36L\beta^2\gamma^2C_\beta D_\beta b^2. \tag{56}$$

The way to choose step-size γ is adapted from Lemma 15 in (Koloskova et al., 2020). For simplicity, we let

$$r_0 = 2F^{(0)}, \quad r_1 = \frac{2\sigma^2}{n}, \quad r_2 = 12L\beta^2C_\beta\sigma^2 + 36L\beta^2C_\beta D_\beta b^2, \tag{57}$$

and inequality (56) becomes

$$\frac{1}{T+1}\sum_{k=0}^{T} B^{(k)} \leq \frac{r_0}{(T+1)\gamma} + r_1\gamma + r_2\gamma^2. \tag{58}$$

Now we let $\gamma = \min\big\{\frac{1}{12\beta LD_\beta}, \Big(\frac{r_0}{r_1(T+1)}\Big)^{\frac{1}{2}}, \Big(\frac{r_0}{r_2(T+1)}\Big)^{\frac{1}{3}}\big\}$:

- If $\Big(\frac{r_0}{r_2(T+1)}\Big)^{\frac{1}{3}}$ is the smallest, we let $\gamma = \Big(\frac{r_0}{r_2(T+1)}\Big)^{\frac{1}{3}}$. With $\Big(\frac{r_0}{r_2(T+1)}\Big)^{\frac{1}{3}} \leq \Big(\frac{r_0}{r_1(T+1)}\Big)^{\frac{1}{2}}$, (58) becomes

$$\frac{1}{T+1}\sum_{k=0}^{T} B^{(k)} \leq 2r_2^{\frac{1}{3}}\Big(\frac{r_0}{T+1}\Big)^{\frac{2}{3}} + r_1\Big(\frac{r_0}{r_2(T+1)}\Big)^{\frac{1}{3}} \leq 2r_2^{\frac{1}{3}}\Big(\frac{r_0}{T+1}\Big)^{\frac{2}{3}} + \Big(\frac{r_0r_1}{T+1}\Big)^{\frac{1}{2}}. \tag{59}$$

- If $\Big(\frac{r_0}{r_1(T+1)}\Big)^{\frac{1}{2}}$ is the smallest, we let $\gamma = \Big(\frac{r_0}{r_1(T+1)}\Big)^{\frac{1}{2}}$. With $\Big(\frac{r_0}{r_1(T+1)}\Big)^{\frac{1}{2}} \leq \Big(\frac{r_0}{r_2(T+1)}\Big)^{\frac{1}{3}}$, (58) becomes

$$\frac{1}{T+1}\sum_{k=0}^{T} B^{(k)} \leq 2\Big(\frac{r_0r_1}{T+1}\Big)^{\frac{1}{2}} + \frac{r_0r_2}{r_1(T+1)} \leq 2\Big(\frac{r_0r_1}{T+1}\Big)^{\frac{1}{2}} + r_2^{\frac{1}{3}}\Big(\frac{r_0}{T+1}\Big)^{\frac{2}{3}}. \tag{60}$$

- If $\frac{1}{12\beta LD_\beta} \leq \Big(\frac{r_0}{r_1(T+1)}\Big)^{\frac{1}{2}}$ and $\frac{1}{12\beta LD_\beta} \leq \Big(\frac{r_0}{r_2(T+1)}\Big)^{\frac{1}{3}}$, we let $\gamma = \frac{1}{12\beta LD_\beta}$ and (58) becomes

$$\frac{1}{T+1}\sum_{k=0}^{T} B^{(k)} \leq \frac{12\beta LD_\beta r_0}{T+1} + \Big(\frac{r_0r_1}{T+1}\Big)^{\frac{1}{2}} + r_2^{\frac{1}{3}}\Big(\frac{r_0}{T+1}\Big)^{\frac{2}{3}}. \tag{61}$$

Combining (59), (60) and (61), we have

$$\frac{1}{T+1}\sum_{k=0}^{T} B^{(k)} \leq \frac{12r_0LD_\beta\beta}{T+1} + 2\Big(\frac{r_0r_1}{T+1}\Big)^{\frac{1}{2}} + 2r_2^{\frac{1}{3}}\Big(\frac{r_0}{T+1}\Big)^{\frac{2}{3}}. \tag{62}$$

Substituting constants r_0, r_1, and r_2, we have the final result:

$$\frac{1}{T+1}\sum_{k=0}^{T} B^{(k)} = O\Big(\frac{\sigma}{\sqrt{nT}} + \frac{(C_\beta)^{\frac{1}{3}}\beta^{\frac{2}{3}}\sigma^{\frac{2}{3}}}{T^{\frac{2}{3}}} + \frac{(C_\beta)^{\frac{1}{3}}(D_\beta)^{\frac{1}{3}}\beta^{\frac{2}{3}}b^{\frac{2}{3}}}{T^{\frac{2}{3}}} + \frac{\beta D_\beta}{T}\Big). \tag{63}$$

C. Convergence analysis for non-convex scenario

C.1. Proof Outline for Theorem 2.

The proof outline for Theorem 2 is similar to that for Theorem 1. The descent lemma (Koloskova et al., 2020, Lemma 10) was established follows.

Lemma 6 (DESCENT LEMMA (Koloskova et al., 2020)). *Under Assumption 1–3 and step-size $\gamma < \frac{1}{4L}$, it holds for $k = 1, 2, \cdots$ that*

$$\mathbb{E}f(\bar{\boldsymbol{x}}^{(k)}) \leq \mathbb{E}f(\bar{\boldsymbol{x}}^{(k-1)}) - \frac{\gamma}{4}\mathbb{E}\|\nabla f(\bar{\boldsymbol{x}}^{(k-1)})\|^2 + \frac{\gamma^2 L\sigma^2}{2n} + \frac{3\gamma L^2}{4n}\mathbb{E}\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|_F^2. \tag{64}$$

The consensus distance is examined in the following two lemmas. Similar to Lemma 4, we use $D_\beta = \min\{H, 1/(1-\beta)\}$.

Lemma 7 (UNIFIED CONSENSUS LEMMA). *Under Assumptions 1–3 and 5, it holds for $k = \tau(k), \tau(k)+1, \cdots, \tau(k)+H-1$ that*

$$\begin{aligned}\mathbb{E}\|\mathbf{x}^{(k+1)} - \bar{\mathbf{x}}^{(k+1)}\|_F^2 \leq & 6D_\beta\beta^2\gamma^2L^2\sum_{\ell=\tau(k)}^{k}\beta^{2(k+1-\ell)}\mathbb{E}\|\mathbf{x}^{(\ell)} - \bar{\mathbf{x}}^{(\ell)}\|_F^2 \\ & + 6nD_\beta\beta^2\gamma^2\sum_{\ell=\tau(k)}^{k}\beta^{2(k+1-\ell)}\mathbb{E}\|\nabla f(\bar{\boldsymbol{x}}^{(\ell)})\|^2 + 2n\beta^2\gamma^2C_\beta(3H\hat{b}^2+\sigma^2)\end{aligned} \tag{65}$$

where $D_\beta = \min\{H, \frac{1}{1-\beta}\}$.

Lemma 8 (RUNNING CONSENSUS LEMMA). *When Assumptions 1–3 and 5 hold and step-size $\gamma < \frac{1}{4L\beta D_\beta}$, it holds for any $T > 0$ that*

$$\frac{1}{T+1}\sum_{k=0}^{T}\mathbb{E}\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|_F^2 \leq \frac{2c_2D_\beta}{T+1}\sum_{k=0}^{T}\mathbb{E}\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\|^2 + 2c_3 \tag{66}$$

where c_2 and c_3 are constants defined as

$$c_2 = 6nD_\beta\beta^2\gamma^2, \tag{67}$$

$$c_3 = 2n\beta^2\gamma^2C_\beta(3D_\beta\hat{b}^2+\sigma^2). \tag{68}$$

With Lemmas 6 and 8, we can establish the convergence rate in Theorem 2.

C.2. Proof of Lemma 6.

This lemma was first established in (Koloskova et al., 2020, Lemma 10). We made slight improvement to tight constants appeared in step-size ranges and upper bound (64). For readers' convenience, we repeat arguments here. Recall that

$$\bar{\boldsymbol{x}}^{(k+1)} = \bar{\boldsymbol{x}}^{(k)} - \frac{\gamma}{n}\sum_{i=1}^{n}\nabla F_i(\boldsymbol{x}_i^{(k)}; \boldsymbol{\xi}_i^{(k+1)}). \tag{69}$$

Since $f(x)$ is L-smooth, it holds that

$$\begin{aligned}\mathbb{E}[f(\bar{\boldsymbol{x}}^{(k+1)})|\mathcal{F}^k] &\overset{(15)}{\leq} f(\bar{\boldsymbol{x}}^{(k)}) - \mathbb{E}\big[\langle\nabla f(\bar{\boldsymbol{x}}^{(k)}), \frac{\gamma}{n}\sum_{i=1}^{n}\nabla F_i(\boldsymbol{x}_i^{(k)}; \boldsymbol{\xi}_i^{(k+1)})\rangle|\mathcal{F}^k\big] + \frac{\gamma^2L}{2}\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\boldsymbol{x}_i^{(k)}; \boldsymbol{\xi}_i^{(k+1)})\|^2|\mathcal{F}^k] \\ &\overset{(4)}{=} f(\bar{\boldsymbol{x}}^{(k)}) - \langle\nabla f(\bar{\boldsymbol{x}}^{(k)}), \frac{\gamma}{n}\sum_{i=1}^{n}\nabla f_i(\boldsymbol{x}_i^{(k)})\rangle + \frac{\gamma^2L}{2}\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\boldsymbol{x}_i^{(k)}; \boldsymbol{\xi}_i^{(k+1)})\|^2|\mathcal{F}^k] \\ &\overset{(a)}{\leq} f(\bar{\boldsymbol{x}}^{(k)}) - \langle\nabla f(\bar{\boldsymbol{x}}^{(k)}), \frac{\gamma}{n}\sum_{i=1}^{n}\nabla f_i(\boldsymbol{x}_i^{(k)})\rangle + \frac{\gamma^2L\sigma^2}{2n} + \frac{\gamma^2L}{2}\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\boldsymbol{x}_i^{(k)})\|^2\end{aligned} \tag{70}$$

where (a) holds because

$$\begin{aligned}&\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^n\nabla F_i(\boldsymbol{x}_i^{(k)};\boldsymbol{\xi}_i^{(k+1)})-\nabla f_i(\boldsymbol{x}_i^{(k)})+\nabla f_i(\boldsymbol{x}_i^{(k)})\|^2|\mathcal{F}^k]\\ \overset{(4)}{=}&\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^n\nabla F_i(\boldsymbol{x}_i^{(k)};\boldsymbol{\xi}_i^{(k+1)})-\nabla f_i(\boldsymbol{x}_i^{(k)})\|^2|\mathcal{F}^k]+\|\frac{1}{n}\sum_{i=1}^n\nabla f_i(\boldsymbol{x}_i^{(k)})\|^2\overset{(12)}{\leq}\frac{\sigma^2}{n}+\|\frac{1}{n}\sum_{i=1}^n\nabla f_i(\boldsymbol{x}_i^{(k)})\|^2.\end{aligned}\tag{71}$$

Note that

$$\begin{aligned}-\langle\nabla f(\bar{\boldsymbol{x}}^{(k)}),\frac{\gamma}{n}\sum_{i=1}^n\nabla f_i(\boldsymbol{x}_i^{(k)})\rangle&=-\langle\nabla f(\bar{\boldsymbol{x}}^{(k)}),\frac{\gamma}{n}\sum_{i=1}^n[\nabla f_i(\boldsymbol{x}_i^{(k)})-\nabla f_i(\bar{\boldsymbol{x}}^{(k)})+\nabla f_i(\bar{\boldsymbol{x}}^{(k)})]\rangle\\&\leq-\gamma\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\|^2+\frac{\gamma}{2}\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\|^2+\frac{\gamma}{2n}\sum_{i=1}^n\|\nabla f_i(\boldsymbol{x}_i^{(k)})-\nabla f_i(\bar{\boldsymbol{x}}^{(k)})\|^2\\&\leq-\frac{\gamma}{2}\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\|^2+\frac{\gamma L^2}{2n}\|\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}\|_F^2\end{aligned}\tag{72}$$

and

$$\|\frac{1}{n}\sum_{i=1}^n\nabla f_i(\boldsymbol{x}_i^{(k)})\|^2\leq\frac{2L^2}{n}\|\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}\|_F^2+2\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\|^2\tag{73}$$

Substituting (72) and (73) into (70), taking expectations over $\mathcal{F}^{(k)}$ and using the fact that $\gamma<\frac{1}{4L}$, we reach the result in (64).

C.3. Proof of Lemma 7.

Similar to the proof of Lemmas 2 and 3, we will derive two bounds for $\mathbb{E}\|\mathbf{x}^{(k+1)}-\bar{\mathbf{x}}^{(k+1)}\|_F^2$:

Bound 1. Following (35)-(39), it holds for $k=\tau(k),\tau(k)+1,\cdots,\tau(k)+H-1$ that

$$\mathbb{E}[\|\mathbf{x}^{(k+1)}-\bar{\mathbf{x}}^{(k+1)}\|_F^2|\mathcal{F}^{(\tau(k))}]\leq2\gamma^2H\sum_{\ell=\tau(k)}^k\beta^{2(k+1-\ell)}\mathbb{E}[\|\nabla f(\mathbf{x}^{(\ell)})\|_F^2|\mathcal{F}^{(\tau(k))}]+2n\gamma^2\beta^2\sigma^2C_\beta\tag{74}$$

Note that

$$\begin{aligned}\|\nabla f(\mathbf{x}^{(k)})\|_F^2&=\sum_{i=1}^n\|\nabla f_i(\boldsymbol{x}_i^{(k)})\|^2\\&=\sum_{i=1}^n\|\nabla f_i(\boldsymbol{x}_i^{(k)})-\nabla f_i(\bar{\boldsymbol{x}}^{(k)})+\nabla f_i(\bar{\boldsymbol{x}}^{(k)})-\nabla f(\bar{\boldsymbol{x}}^k)+\nabla f(\bar{\boldsymbol{x}}^k)\|^2\\&\leq3L^2\|\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}\|_F^2+3n\hat{b}^2+3n\|\nabla f(\bar{\boldsymbol{x}}^k)\|^2.\end{aligned}\tag{75}$$

where the last inequality holds because of Assumption 5. Substituting (75) into (74) and taking expectation on $\mathcal{F}^{(\tau(k))}$, we get

$$\begin{aligned}&\mathbb{E}\|\mathbf{x}^{(k+1)}-\bar{\mathbf{x}}^{(k+1)}\|_F^2\\&\leq6H\beta^2\gamma^2L^2\sum_{\ell=\tau(k)}^k\beta^{2(k-\ell)}\mathbb{E}\|\mathbf{x}^{(\ell)}-\bar{\mathbf{x}}^{(\ell)}\|_F^2+6nH\beta^2\gamma^2\sum_{\ell=\tau(k)}^k\beta^{2(k-\ell)}\|\nabla f(\bar{\boldsymbol{x}}^{(\ell)})\|^2+2n\gamma^2\beta^2C_\beta(3H\hat{b}^2+\sigma^2)\\&\leq6H\gamma^2L^2\beta^2\sum_{\ell=\tau(k)}^k\beta^{k-\ell}\|\mathbf{x}^{(\ell)}-\bar{\mathbf{x}}^{(\ell)}\|_F^2+6nH\gamma^2\beta^2\sum_{\ell=\tau(k)}^k\beta^{k-\ell}\|\nabla f(\bar{\boldsymbol{x}}^{(\ell)})\|^2+2n\gamma^2\beta^2C_\beta(3H\hat{b}^2+\sigma^2)\end{aligned}\tag{76}$$

Bound 2. Following (43) and first two lines in (44), it holds for any $k=\tau(k),\cdots,\tau(k)+H-1$ that

$$\mathbb{E}[\|\mathbf{x}^{(k+1)}-\bar{\mathbf{x}}^{(k+1)}\|_F^2|\mathcal{F}^{(k)}]\leq\beta\|\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}\|_F^2+\frac{\beta^2\gamma^2}{1-\beta}\|\nabla f(\mathbf{x}^{(k)})\|_F^2+n\gamma^2\beta^2\sigma^2. \tag{77}$$

Substituting (75) into (77), we get

$$\mathbb{E}[\|\mathbf{x}^{(k+1)}-\bar{\mathbf{x}}^{(k+1)}\|_F^2|\mathcal{F}^{(k)}]\leq\big(\beta+\frac{3\beta^2\gamma^2L^2}{1-\beta}\big)\|\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}\|_F^2+\frac{3n\beta^2\gamma^2\|\nabla f(\bar{\boldsymbol{x}}^k)\|^2}{1-\beta}+n\gamma^2\beta^2\sigma^2+\frac{3n\beta^2\gamma^2\hat{b}^2}{1-\beta}. \tag{78}$$

We next follow (43)–(46) and take expectation on $\mathcal{F}^{(k)}$ to get

$$\begin{aligned}
&\mathbb{E}\|\mathbf{x}^{(k+1)}-\bar{\mathbf{x}}^{(k+1)}\|_F^2\\
&\leq\frac{3\beta^2\gamma^2L^2}{1-\beta}\sum_{\ell=\tau(k)}^{k}\beta^{k-\ell}\,\mathbb{E}\|\mathbf{x}^{(\ell)}-\bar{\mathbf{x}}^{(\ell)}\|_F^2+\frac{3n\beta^2\gamma^2}{1-\beta}\sum_{\ell=\tau(k)}^{k}\beta^{k-\ell}\,\mathbb{E}\|\nabla f(\bar{\boldsymbol{x}}^{(\ell)})\|^2+n\gamma^2\beta^2C_\beta(\frac{3\hat{b}^2}{1-\beta}+\sigma^2)\\
&\leq\frac{6\beta^2\gamma^2L^2}{1-\beta}\sum_{\ell=\tau(k)}^{k}\beta^{k-\ell}\,\mathbb{E}\|\mathbf{x}^{(\ell)}-\bar{\mathbf{x}}^{(\ell)}\|_F^2+\frac{6n\beta^2\gamma^2}{1-\beta}\sum_{\ell=\tau(k)}^{k}\beta^{k-\ell}\,\mathbb{E}\|\nabla f(\bar{\boldsymbol{x}}^{(\ell)})\|^2+2n\gamma^2\beta^2C_\beta(\frac{3\hat{b}^2}{1-\beta}+\sigma^2)
\end{aligned} \tag{79}$$

With bounds (76) and (79), we reach the result (65).

C.4. Proof of Lemma 8.

We first simplify the notation as follows:

$$\begin{aligned}
&A^{(k)}=\mathbb{E}\|\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}\|^2,\quad B^{(k)}=\mathbb{E}\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\|^2,\\
&c_1=6D_\beta\beta^2\gamma^2L^2,\qquad c_2=6nD_\beta\beta^2\gamma^2,\qquad c_3=2n\beta^2\gamma^2C_\beta(3D_\beta\hat{b}^2+\sigma^2).
\end{aligned} \tag{80}$$

With these notations, we can follow the proof of Lemma 5 to get the final result.

C.5. Proof of Theorem 2.

Following the notation in (80), we further define $F^{(k)}:=\mathbb{E}f(\bar{\boldsymbol{x}}^{(k)})$. With these notations, the inequality (64) becomes

$$B^{(k)}\leq\frac{4F^{(k)}}{\gamma}-\frac{4F^{(k+1)}}{\gamma}+\frac{2\gamma L\sigma^2}{n}+\frac{3L^2}{n}A^{(k)} \tag{81}$$

Taking the weighted running average over the above inequality and divide $T+1$ to get

$$\begin{aligned}
\frac{1}{T+1}\sum_{k=0}^{T}B^{(k)}&\leq\frac{4F^{(0)}}{(T+1)\gamma}+\frac{3L^2}{n(T+1)}\sum_{k=0}^{T}A^{(k)}+\frac{2\gamma L\sigma^2}{n}\\
&\overset{(66)}{\leq}\frac{4F^{(0)}}{(T+1)\gamma}+\frac{6\beta^2L^2Hc_2}{n(T+1)}\sum_{k=0}^{T}B^{(k)}+\frac{6L^2c_3}{n}+\frac{2\gamma L\sigma^2}{n}\\
&\leq\frac{4F^{(0)}}{(T+1)\gamma}+\frac{1}{2(T+1)}\sum_{k=0}^{T}B^{(k)}+\frac{6L^2c_3}{n}+\frac{2\gamma L\sigma^2}{n}
\end{aligned} \tag{82}$$

where the last inequality holds when $\gamma\leq\frac{1}{9LH\beta}$. Substituting c_3 into the above inequality, we have

$$\frac{1}{T+1}\sum_{k=0}^{T}B^{(k)}\leq\frac{8F^{(0)}}{(T+1)\gamma}+\frac{4\gamma L\sigma^2}{n}+24L^2\gamma^2\beta^2C_\beta\sigma^2+72L^2\gamma^2\beta^2C_\beta D_\beta\hat{b}^2. \tag{83}$$

By following the arguments (57) – (63), we reach the result in (8).

D. Transient stage and transient time

D.1. Transient stage derivation.

(i) Gossip SGD. We first consider the iid scenario where $b^2 = 0$. To make the first term dominate the other terms (see the first line in Table 4), T has to be sufficiently large such that (ignoring the affects of σ)

$$\max\Big\{\frac{\beta^{\frac{2}{3}}}{T^{\frac{2}{3}}(1-\beta)^{\frac{1}{3}}}, \frac{\beta}{(1-\beta)T}\Big\} \le \frac{1}{\sqrt{nT}} \quad \Longrightarrow \quad T \ge \max\Big\{\frac{n^3\beta^4}{(1-\beta)^2}, \frac{n\beta^2}{(1-\beta)^2}\Big\}. \tag{84}$$

We next consider the non-iid scenario where $b^2 \neq 0$. To make the first term dominate the other terms, T has to be sufficiently large such that (ignoring the affects of σ and b)

$$\max\Big\{\frac{\beta^{\frac{2}{3}}}{T^{\frac{2}{3}}(1-\beta)^{\frac{1}{3}}}, \frac{\beta^{\frac{2}{3}}}{T^{\frac{2}{3}}(1-\beta)^{\frac{2}{3}}}, \frac{\beta}{(1-\beta)T}\Big\} \le \frac{1}{\sqrt{nT}} \quad \Longrightarrow \quad T \ge \max\Big\{\frac{n^3\beta^4}{(1-\beta)^2}, \frac{n^3\beta^4}{(1-\beta)^4}, \frac{n\beta^2}{(1-\beta)^2}\Big\}. \tag{85}$$

When $n\beta > 1$ which usually holds for most commonly-used network topologies, inequalities (84) and (85) will result in the transient stage $T = \Omega(\frac{n^3\beta^4}{(1-\beta)^2})$ and $T = \Omega(\frac{n^3\beta^4}{(1-\beta)^4})$ for iid and non-iid scenarios, respectively.

(ii) Gossip-PGA. We first consider the iid scenario where $b^2 = 0$. To make the first term dominate the other terms (see the first line in Table 4), T has to be sufficiently large such that (ignoring the affects of σ)

$$\max\Big\{\frac{C_\beta^{\frac{1}{3}}\beta^{\frac{2}{3}}}{T^{\frac{2}{3}}}, \frac{\beta D_\beta}{T}\Big\} \le \frac{1}{\sqrt{nT}} \quad \Longrightarrow \quad T \ge \max\big\{n^3\beta^4C_\beta^2, n\beta^2D_\beta^2\big\} = \Omega(n^3\beta^4C_\beta^2). \tag{86}$$

We next consider the non-iid scenario where $b^2 \neq 0$. To make the first term dominate the other terms, T has to be sufficiently large such that (ignoring the affects of σ and b)

$$\max\Big\{\frac{C_\beta^{\frac{1}{3}}\beta^{\frac{2}{3}}}{T^{\frac{2}{3}}}, \frac{C_\beta^{\frac{1}{3}}D_\beta^{\frac{1}{3}}\beta^{\frac{2}{3}}}{T^{\frac{2}{3}}}, \frac{\beta D_\beta}{T}\Big\} \le \frac{1}{\sqrt{nT}} \quad \Longrightarrow \quad T \ge \max\big\{n^3\beta^4C_\beta^2, n^3\beta^4C_\beta^2D_\beta^2, n\beta^2D_\beta^2\big\} = \Omega(n^3\beta^4C_\beta^2D_\beta^2) \tag{87}$$

when $n\beta > 1$.

(iii) Local SGD. We first consider the iid scenario where $b^2 = 0$. To make the first term dominate the other terms (see the first line in Table 4), T has to be sufficiently large such that (ignoring the affects of σ)

$$\max\Big\{\frac{H^{\frac{1}{3}}}{T^{\frac{2}{3}}}, \frac{H}{T}\Big\} \le \frac{1}{\sqrt{nT}} \quad \Longrightarrow \quad T \ge \max\big\{n^3H^2, nH^2\big\} = \Omega(n^3H^2). \tag{88}$$

We next consider the non-iid scenario where $b^2 \neq 0$. To make the first term dominate the other terms, T has to be sufficiently large such that (ignoring the affects of σ and b)

$$\max\Big\{\frac{H^{\frac{1}{3}}}{T^{\frac{2}{3}}}, \frac{H^{\frac{2}{3}}}{T^{\frac{2}{3}}}, \frac{H}{T}\Big\} \le \frac{1}{\sqrt{nT}} \quad \Longrightarrow \quad T \ge \max\big\{n^3H^2, n^3H^4, nH^2\big\} = \Omega(n^3H^4) \tag{89}$$

D.2. Transient time comparison

The transient time comparisons between Gossip SGD and Gossip-PGA for the iid or non-iid scenario over the grid or ring topology are listed in Tables 12, 13 and 14.

E. Proof of Corollary 1

The proof of Corollary 1 closely follows Theorem 1. First, the descent lemma 7 still holds for time-varying period. Second, with the facts that $k+1-\tau(k) \le H_{\max}$, $\sum_{\ell=\tau(k)}^{k}\beta^{k+1-\ell} \le \sum_{k=0}^{H_{\max}}\beta^k := C_\beta$, and $\sum_{k=\ell+1}^{\tau(\ell)+H^{(\ell)}-1}\beta^{k-1-\ell} \le$

$\sum_{k=0}^{H_{\max}} \beta^k = C_\beta$, we follow Appendix C.3 and C.4 to reach the consensus distance inequality:

$$\frac{1}{T+1}\sum_{k=0}^{T}\mathbb{E}\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|^2 \leq \frac{2c_2 D_\beta}{T+1}\sum_{k=0}^{T}\mathbb{E}\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\|^2 + 2c_3 \tag{90}$$

where c_2 and c_3 are constants defined as

$$c_2 = 6nD_\beta\beta^2\gamma^2, \tag{91}$$

$$c_3 = 2n\beta^2\gamma^2 C_\beta(3D_\beta\hat{b}^2 + \sigma^2) \tag{92}$$

and $D_\beta = \min\{1/(1-\beta), H_{\max}\}$, $C_\beta = \sum_{k=0}^{H_{\max}} \beta^k$. With Lemma 6 and inequality (90), we can follow Appendix B.5 to reach the result in Corollary 1.

F. Additional Experiments

F.1. Implementation Details.

We implement all the aforementioned algorithms with PyTorch (Paszke et al., 2019) 1.5.1 using NCCL 2.5.7 (CUDA 10.1) as the communication backend. Each server contains 8 V100 GPUs in our cluster and is treated as one node. The inter-node network fabrics are chosen from 25 Gbps TCP (which is a common distributed training platform setting) and 4×100 Gbps RoCE (which is a high-performance distributed training platform setting).

All deep learning experiments are trained in the mixed precision using Pytorch extension package NVIDIA apex (https://github.com/NVIDIA/apex). For Gossip SGD related training, we use the time-varying one-peer exponential graph following (Assran et al., 2019). Workers send and receive a copy of the model's parameters to and from its peer, thus keeping the load balancing among workers. All data are stored in the cloud storage service and downloaded to workers using HTTP during training.

Image Classfication The Nesterov momentum SGD optimizer is used with a linear scaling learning rate strategy. 32 nodes (each node is with 8 V100 GPUs) are used in all the experiments and the batch-size is set as 256 per node (8,192 in total). The learning rate is warmed up in the first 5 epochs and is decayed by a factor of 10 at 30, 60 and 90 epochs. We train 120 epochs by default (unless specified otherwise) in every experiment and record the epoch and runtime when a 76% top-1 accuracy in the validation set has reached. 25 Gbps TCP network is used for inter-node communication in ResNet-50 training. In 4×100 Gbps RoCE network, the communication overhead is negligible given the high computation-to-communication ratio nature of ResNet models and Parallel SGD with computation and communication pipeline is recommended. We use a period 6 for both Local SGD and Gossip-PGA. In Gossip-AGA, the averaging period is set to 4 in the warm-up stage and changed adaptively afterwards, roughly 9% iterations conduct global averaging.

Language Modeling All experiments are based on NVIDIA BERT implementation with mixed precision support and LAMB optimizer (You et al., 2019). 8 nodes are used in all the experiments with a batch-size 64 per GPU (4096 in total). We do not use gradient accumulation as it is not vertical with Local SGD. We only do phase 1 training and indicate the decreasing of training loss as convergence speed empirically. The learning rate is scaled to $3.75e^{-4}$ initially and decayed in a polynomial policy with warm-up. The phase 1 training consists of 112,608 steps in all experiments. We use a period 6 for both Local SGD and Gossip-PGA. In Gossip-AGA, the averaging period is set to 4 in the warm-up phase and changed adaptively afterwards, roughly 9.6% iterations conduct global averaging.

	GOSSIP SGD	GOSSIP-PGA
TRANSIENT ITER.	$O(n^5)$	$O(n^4)$
SINGLE COMM.	$O(\theta d + \alpha)$	$O(\theta d + \sqrt{n}\alpha)$
TRANSIENT TIME	$O(n^5\theta d + n^5\alpha)$	$O(n^4\theta d + n^{4.5}\alpha)$

Table 12. Transient time comparison between Gossip SGD and Gossip-PGA for iid scenario over the specific grid ($1-\beta = O(1/n)$) topologiy. We choose $H = \sqrt{n}$ as the period in Gossip-PGA.

	Gossip SGD	Gossip-PGA
Transient iter.	$O(n^{11})$	$O(n^5)$
Single comm.	$O(\theta d+\alpha)$	$O(\theta d+\sqrt{n}\alpha)$
Transient time	$O(n^{11}\theta d+n^{11}\alpha)$	$O(n^5\theta d+n^{5.5}\alpha)$

Table 13. Transient time comparison between Gossip SGD and Gossip-PGA for non-iid scenario over the specific ring ($1-\beta=O(1/n^2)$) topologiy. We choose $H=\sqrt{n}$ as the period in Gossip-PGA.

	Gossip SGD	Gossip-PGA
Transient iter.	$O(n^7)$	$O(n^4)$
Single comm.	$O(\theta d+\alpha)$	$O(\theta d+\sqrt{n}\alpha)$
Transient time	$O(n^7\theta d+n^7\alpha)$	$O(n^4\theta d+n^{4.5}\alpha)$

Table 14. Transient time comparison between Gossip SGD and Gossip-PGA for iid scenario over the specific ring ($1-\beta=O(1/n^2)$) topologiy. We choose $H=\sqrt{n}$ as the period in Gossip-PGA.

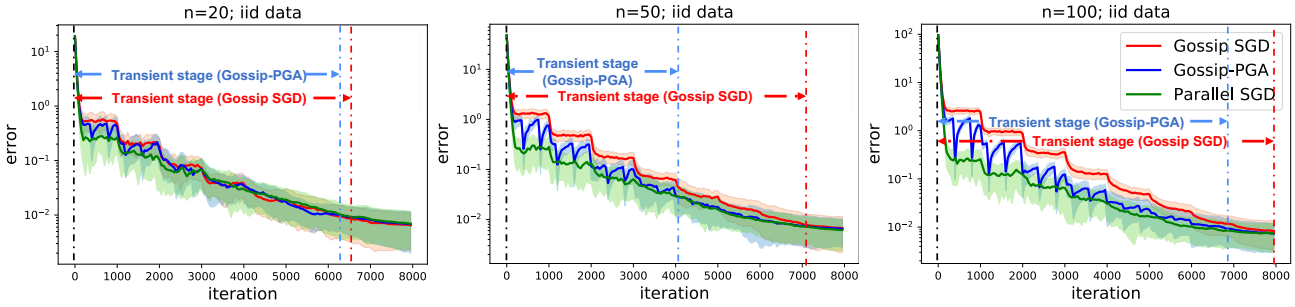

Figure 4. Convergence comparison between Gossip-PGA, Gossip SGD and parallel SGD on the logistic regression problem in iid data distributed setting over the ring topology.

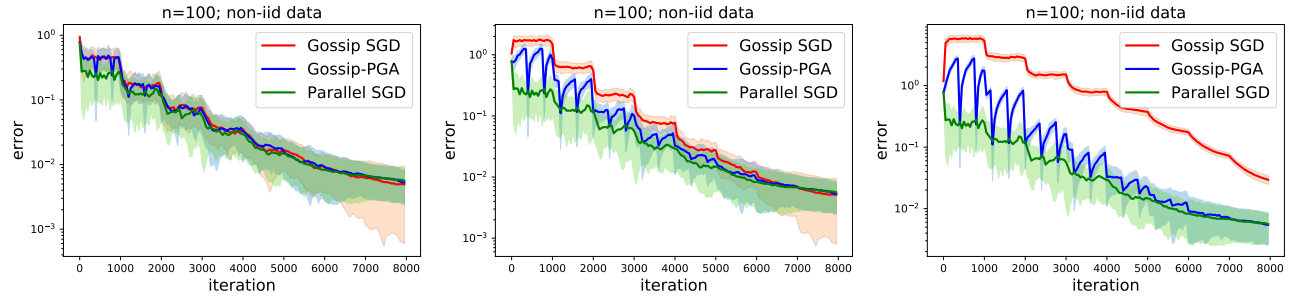

Figure 5. Convergence comparison between Gossip-PGA, Gossip SGD and parallel SGD on the logistic regression problem in non-iid data distributed setting over the exponential graph (left), grid (middle) and ring (right) topology.

F.2. More experiments on convex logistic regression.

In this subsection we will test the performance of Gossip-PGA with iid data distribution and on different topologies. We will also compare it with Local SGD.

Experiments on iid dataset. Figure 4 illustrates how Gossip SGD and Gossip-PGA converges under the iid data distributed setting over the ring topology. Similar to the non-iid scenario shown in Figure 1, it is observed that Gossip-PGA always converges faster (or has shorter transient stages) than Gossip SGD. When network size gets larger and hence $\beta\to 1$, the superiority of Gossip-PGA gets more evident. Moreover, it is also noticed that the transient stage gap between Gossip SGD

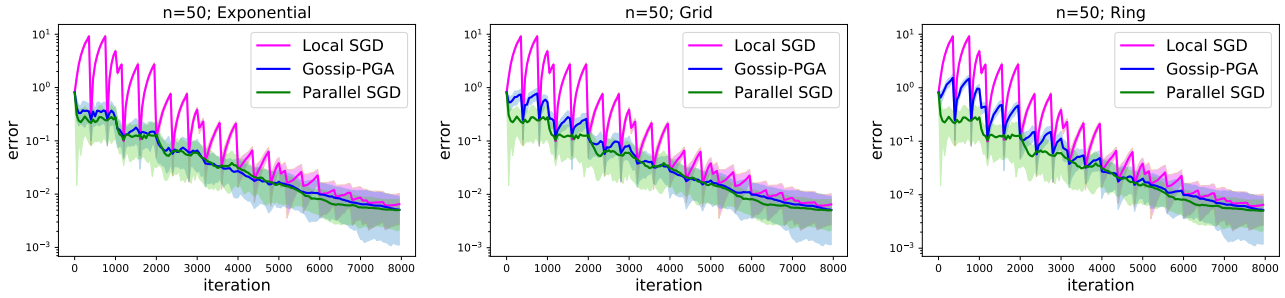

Figure 6. Convergence comparison between Gossip-PGA, Local SGD and parallel SGD on the logistic regression problem in non-iid data distributed setting over the exponential graph (left), grid (middle) and ring (right) topology.

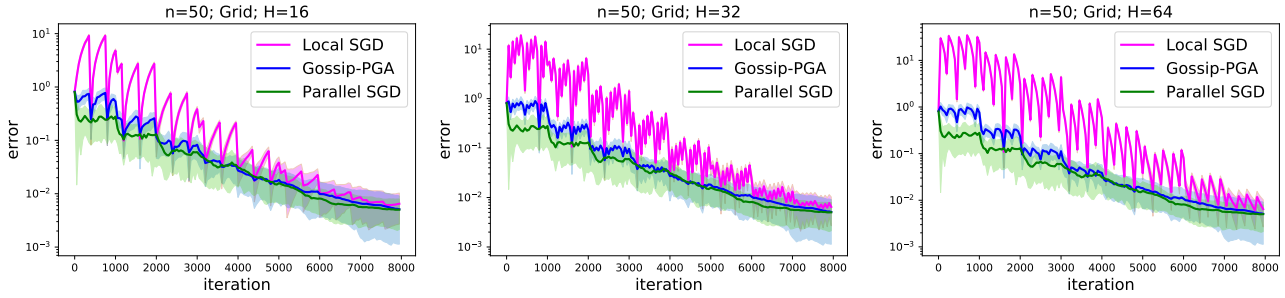

Figure 7. Convergence comparison between Gossip-PGA, Local SGD and parallel SGD on the logistic regression problem in non-iid data distributed setting over the grid topology with period $H = 16$ (left), $H = 32$ (middle), $H = 64$ (right).

and Gossip-PGA is smaller than the non-iid scenario in all three plots in Figure 4. All these observations are consistent with the transient stage comparisons in Table 2.

Experiments on different topologies. Figure 5 illustrates how Gossip SGD and Gossip-PGA converges over the exponential graph, grid and ring topology. For all plots, it is observed that Gossip-PGA is no worse than Gossip SGD. Moreover, as the network gets sparser and hence $\beta \to 1$ from the left plot to right, it is observed that the superiority of Gossip-PGA gets more evident, which is consistent with the transient stage comparisons between Gossip SGD and Gossip-PGA in Table 2.

Comparison with Local SGD. Figure 6 illustrates how Local SGD and Gossip-PGA converges over the exponential graph, grid and ring topology. The period is set as $H = 16$. In all three plots, Gossip-PGA always converges faster than Local SGD because of the additional gossip communications. Moreover, since the exponential graph has the smallest β, it is observed Gossip-PGA has almost the same convergence performance as parallel SGD. Figure 7 illustrates how Local SGD and Gossip-PGA converges over the grid topology with different periods. It is observed that Gossip-PGA can be significantly faster when H is large. All these observations are consistent with the transient stage comparisons in Table 3.

F.3. More experiments on image classification.

Training accuracy. Figure F.3 shows the iteration-wise and time-wise training accuracy curves of aforementioned algorithms separately. In the left figure, it is observed Gossip-PGA/AGA converges faster (in iteration) and more accurate than local and Gossip SGD, which is consistent with our theory. In the right figure, it is observed that Gossip-PGA/AGA is the fastest method (in time) that can reach the same training accuracy as parallel SGD.

The effect of averaging period. Table 15 compares the top-1 accuracy in the validation set with a different averaging period setting in Gossip-PGA SGD. Compared to Gossip SGD, a relatively large global averaging period (48), roughly 2.1% iterations with global averaging can still result in 0.32% gain in validation accuracy. With a moderate global averaging period (6/12), the validation accuracy is comparable with the parallel SGD baseline. The communication overhead of global averaging can be amortized since it happens every H iterations.

	PARALLEL SGD	GOSSIP SGD	GOSSIP-PGA				
PERIOD H	-	-	3	6	12	24	48
VAL ACC.(%)	76.22	75.34	76.19	76.28	76.04	75.68	75.66

Table 15. Comparison of Top-1 validation accuracy with different averaging period setting in Gossip-PGA.

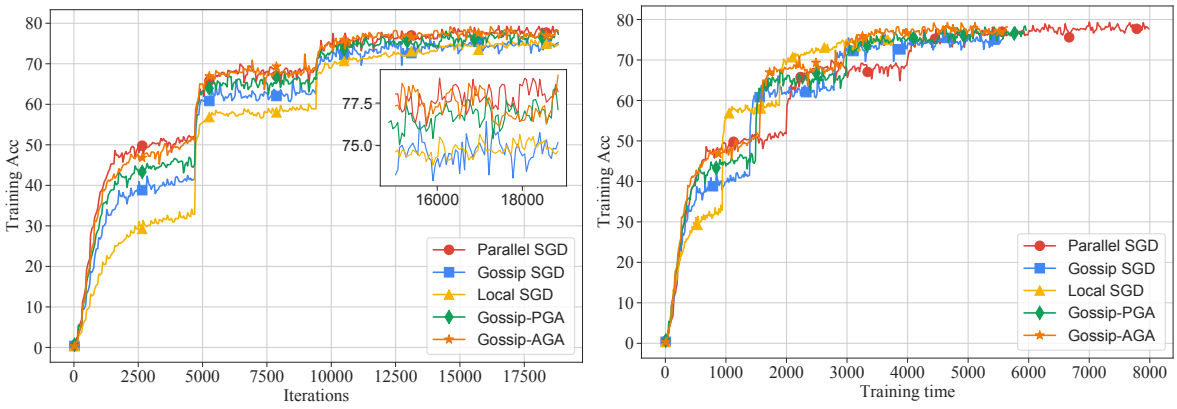

Figure 8. Convergence results on the ImageNet classification task.(a) Iteration-wise convergence in terms of training accuracy. (b) runtime-wise convergence speed in terms of training accuracy.

Experiments on SGD optimizer (without momentum). In previous Imagenet training, Nesterov momentum SGD optimizer is used. Following common practice (Lian et al., 2017; Assran et al., 2019; Tang et al., 2018), we establish the convergence rate of the non-accelerated method while running experiments with momentum. For the sake of clarity, we further add a new series of experiments on Gossip-SGD without momentum, see Table 16. Gossip-PGA still outperform Gossip-SGD utilizing the SGD optimizer.

METHOD	ACC. %
PARALLEL SGD	69.5
GOSSIP SGD	68.47
GOSSIP-PGA	69.21

Table 16. Comparison of validation accuracy of Imagenet training on different algorithms with SGD optimizer.

G. Implementation of Gossip AGA

Practical consideration. The Gossip-AGA algorithm is listed in Algorithm 2. We use a counter C to record the number of gossip iterations since last global averaging. The global averaging period H is initialized to a small value H_{init} (e.g. 2$\sim$4). Once C equals to current H, global averaging happens. In practice, we sample loss scores for the first fewer iterations and get a F_{init} estimation in a running-average fashion. We remove the exponential term in the loss score ratio for flexible period adjustment.

H. Comparison of communication overhead between gossip and All-Reduce

Table 17 compares the overhead of different communication styles in two deep training tasks. The implementation details follow Appendix F. For each profiling, we run a 500 iterations and take their average as the iteration time. As typically All-Reduce implementation containing overlapping between computation and communication, we run a series of separate experiments which do not perform communication (Column 2) to get communication overhead fairly (the figures in the brackets). For ResNet-50 training, gossip introduces 150ms communication overhead while All-Reduce needs 278ms. For

Algorithm 2 Gossip-AGA

Require: Initialize $x_{0,i} = x_0$, learning rate $\gamma > 0$, topology matrix W for all nodes $i \in \{1, 2, ..., n\}$, global averaging period $H = H_{init}$, $C \leftarrow 0$, $F_{init} \leftarrow 0$, warmup iterations K_w

for $k = 0, 1, 2, ..., T-1$, *every node* i **do**

- $C \leftarrow C + 1$
- Sample mini-batch data $\boldsymbol{\xi}_i^{(k+1)}$ from local dataset Compute stochastic gradient $\nabla F_i(\boldsymbol{x}_i^{(k)}; \boldsymbol{\xi}_i^{(k+1)})$ and loss $F_i(\boldsymbol{x}_i^{(k)}; \boldsymbol{\xi}_i^{(k+1)})$
- Conduct local udpate $\boldsymbol{x}_i^{(k+\frac{1}{2})} = \boldsymbol{x}_i^{(k)} - \gamma \nabla F_i(\boldsymbol{x}_i^{(k)}; \boldsymbol{\xi}_i^{(k+1)})$
- **if** $C == H$ **then**
  - $C \leftarrow 0$
  - $x_i^{(k+1)} \leftarrow \frac{1}{n} \sum_{j=1}^{n} x_{k+\frac{1}{2},j}$
  - $F(x_k; \xi_k) = \frac{1}{n} \sum_{i=1}^{n} F_i(x_{k,i}; \xi_{k,i})$
  - **if** $k < K_w$ **then**
    - $F_{init} \leftarrow \frac{1}{2}(F_{init} + F(x_k; \xi_k))$
  - **else**
    - $H \leftarrow \left\lceil \frac{F_{init}}{F(x_k; \xi_k)} H_{init} \right\rceil$
- **else**
  - $\boldsymbol{x}_i^{(k+1)} = \sum_{j \in \mathcal{N}_i} w_{ij} \boldsymbol{x}_j^{(k+\frac{1}{2})}$

MODEL	ITERATION TIME (MS)		
	NO COMMUNICATION	ALL-REDUCE	GOSSIP
RESNET-50	146	424 (278)	296 (150)
BERT	445	1913.8 (1468.8)	1011.5 (566.5)

Table 17. Comparison of communication overhead between gossip and All-Reduce in terms of runtime.

BERT training, gossip introduces 566.5ms communication overhead while All-Reduce needs 1468.8ms with the tremendous model size of BERT-Large.