A. Proof of Lemma 2

We recall the standard Descent Lemma (Nesterov, 2018), i.e., $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$,

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

since the global function F is L-smooth. Setting $\mathbf{y} = \overline{\mathbf{x}}_{t+1}$ and $\mathbf{x} = \overline{\mathbf{x}}_t$ in (15) and using (5) , we have: $\forall t \geq 0$,

$$\begin{aligned} F(\overline{\mathbf{x}}_{t+1}) &\leq F(\overline{\mathbf{x}}_t) - \big\langle \nabla F(\overline{\mathbf{x}}_t), \overline{\mathbf{x}}_{t+1} - \overline{\mathbf{x}}_t \big\rangle + \frac{L}{2} \|\overline{\mathbf{x}}_{t+1} - \overline{\mathbf{x}}_t\|^2 \\ &\leq F(\overline{\mathbf{x}}_t) - \alpha \big\langle \nabla F(\overline{\mathbf{x}}_t), \overline{\mathbf{v}}_t \big\rangle + \frac{L\alpha^2}{2} \|\overline{\mathbf{v}}_t\|^2. \end{aligned} \tag{16}$$

Using $\langle \mathbf{a}, \mathbf{b} \rangle = \frac{1}{2}\left(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2\right), \forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^p$, in (16) gives: for $0 < \alpha \leq \frac{1}{2L}$ and $\forall t \geq 0$,

$$\begin{aligned} F(\overline{\mathbf{x}}_{t+1}) &\leq F(\overline{\mathbf{x}}_t) - \frac{\alpha}{2} \|\nabla F(\overline{\mathbf{x}}_t)\|^2 - \left(\frac{\alpha}{2} - \frac{L\alpha^2}{2}\right) \|\overline{\mathbf{v}}_t\|^2 + \frac{\alpha}{2} \|\overline{\mathbf{v}}_t - \nabla F(\overline{\mathbf{x}}_t)\|^2, \\ &\leq F(\overline{\mathbf{x}}_t) - \frac{\alpha}{2} \|\nabla F(\overline{\mathbf{x}}_t)\|^2 - \left(\frac{\alpha}{2} - \frac{L\alpha^2}{2}\right) \|\overline{\mathbf{v}}_t\|^2 + \alpha \left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2 + \alpha \left\|\overline{\nabla \mathbf{f}}(\mathbf{x}_t) - \nabla F(\overline{\mathbf{x}}_t)\right\|^2, \\ &\overset{(i)}{\leq} F(\overline{\mathbf{x}}_t) - \frac{\alpha}{2} \|\nabla F(\overline{\mathbf{x}}_t)\|^2 - \frac{\alpha}{4} \|\overline{\mathbf{v}}_t\|^2 + \alpha \left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2 + \frac{\alpha L^2}{n} \|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2, \end{aligned} \tag{17}$$

where (i) is due to Lemma 1(c) and that $\frac{L\alpha^2}{2} \leq \frac{\alpha}{4}$ since $0 < \alpha \leq \frac{1}{2L}$. Rearranging (17), we have: for $0 < \alpha \leq \frac{1}{2L}$ and $\forall t \geq 0$,

$$\|\nabla F(\overline{\mathbf{x}}_t)\|^2 \leq \frac{2(F(\overline{\mathbf{x}}_t) - F(\overline{\mathbf{x}}_{t+1}))}{\alpha} - \frac{1}{2} \|\overline{\mathbf{v}}_t\|^2 + 2 \left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2 + \frac{2L^2}{n} \|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2. \tag{18}$$

Taking the telescoping sum of (18) over t from 0 to T, $\forall T \geq 0$ and using the fact that F bounded below by F^* in the resulting inequality finishes the proof.

B. Proof of Lemma 3

B.1. Proof of Eq. (7)

We recall that the update of each local stochastic gradient estimator $\mathbf{v}_t^i, \forall t \geq 1$, in (2) may be written equivalently as follows:

$$\mathbf{v}_t^i = \beta \mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) + (1 - \beta)\Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) + \mathbf{v}_{t-1}^i\Big),$$

where $\beta \in (0, 1)$. We have: $\forall t \geq 1$ and $\forall i \in \mathcal{V}$,

$$\begin{aligned} \mathbf{v}_t^i - \nabla f_i(\mathbf{x}_t^i) &= \beta \mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) + (1 - \beta)\Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) + \mathbf{v}_{t-1}^i\Big) - \beta \nabla f_i(\mathbf{x}_t^i) - (1 - \beta) \nabla f_i(\mathbf{x}_t^i) \\ &= \beta\Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i)\Big) + (1 - \beta)\Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) + \mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_t^i)\Big) \\ &= \beta\Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i)\Big) + (1 - \beta)\Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) + \nabla f_i(\mathbf{x}_{t-1}^i) - \nabla f_i(\mathbf{x}_t^i)\Big) \\ &\quad + (1 - \beta)\Big(\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\Big). \end{aligned} \tag{19}$$

In (19), we observe that $\forall t \geq 1$ and $\forall i \in \mathcal{V}$,

$$\mathbb{E}\Big[\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i) | \mathcal{F}_t\Big] = \mathbf{0}_p, \tag{20}$$

$$\mathbb{E}\Big[\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) + \nabla f_i(\mathbf{x}_{t-1}^i) - \nabla f_i(\mathbf{x}_t^i) | \mathcal{F}_t\Big] = \mathbf{0}_p, \tag{21}$$

by the definition of the filtration $\mathcal{F}_t$ in (1). Averaging (19) over i from 1 to n gives: $\forall t \geq 0$,

$$\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t) = (1-\beta)\Big(\overline{\mathbf{v}}_{t-1} - \overline{\nabla \mathbf{f}}(\mathbf{x}_{t-1})\Big) + \beta \cdot \underbrace{\frac{1}{n}\sum_{i=1}^{n}\Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i)\Big)}_{=:\mathbf{s}_t}$$
$$+ (1-\beta) \cdot \underbrace{\frac{1}{n}\sum_{i=1}^{n}\Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) + \nabla f_i(\mathbf{x}_{t-1}^i) - \nabla f_i(\mathbf{x}_t^i)\Big)}_{=:\mathbf{z}_t}. \tag{22}$$

Note that $\mathbb{E}[\mathbf{s}_t|\mathcal{F}_t] = \mathbb{E}[\mathbf{z}_t|\mathcal{F}_t] = \mathbf{0}_p$ by (20) and (21). In light of (22), we have: $\forall t \geq 1$,

$$\begin{aligned}
\mathbb{E}\Big[\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\|^2|\mathcal{F}_t\Big] &= (1-\beta)^2\|\overline{\mathbf{v}}_{t-1} - \overline{\nabla \mathbf{f}}(\mathbf{x}_{t-1})\|^2 + \mathbb{E}\left[\|\beta\mathbf{s}_t + (1-\beta)\mathbf{z}_t\|^2|\mathcal{F}_t\right] \\
&\quad + 2\mathbb{E}\Big[\Big\langle(1-\beta)\left(\overline{\mathbf{v}}_{t-1} - \overline{\nabla \mathbf{f}}(\mathbf{x}_{t-1})\right), \beta\mathbf{s}_t + (1-\beta)\mathbf{z}_t\Big\rangle|\mathcal{F}_t\Big] \\
&\overset{(i)}{=} (1-\beta)^2\|\overline{\mathbf{v}}_{t-1} - \overline{\nabla \mathbf{f}}(\mathbf{x}_{t-1})\|^2 + \mathbb{E}\left[\|\beta\mathbf{s}_t + (1-\beta)\mathbf{z}_t\|^2|\mathcal{F}_t\right] \\
&\leq (1-\beta)^2\|\overline{\mathbf{v}}_{t-1} - \overline{\nabla \mathbf{f}}(\mathbf{x}_{t-1})\|^2 + 2\beta^2\mathbb{E}\left[\|\mathbf{s}_t\|^2|\mathcal{F}_t\right] + 2(1-\beta)^2\mathbb{E}\left[\|\mathbf{z}_t\|^2|\mathcal{F}_t\right],
\end{aligned} \tag{23}$$

where (i) is due to

$$\mathbb{E}\Big[\Big\langle(1-\beta)\left(\overline{\mathbf{v}}_{t-1} - \overline{\nabla \mathbf{f}}(\mathbf{x}_{t-1})\right), \beta\mathbf{s}_t + (1-\beta)\mathbf{z}_t\Big\rangle|\mathcal{F}_t\Big] = 0,$$

since $\mathbb{E}[\mathbf{s}_t|\mathcal{F}_t] = \mathbb{E}[\mathbf{z}_t|\mathcal{F}_t] = \mathbf{0}_p$ and $(\overline{\mathbf{v}}_{t-1} - \overline{\nabla \mathbf{f}}(\mathbf{x}_{t-1}))$ is $\mathcal{F}_t$-measurable. We next bound the second and the third term in (23) respectively. For the second term in (23), we observe that $\forall t \geq 1$,

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{s}_t\|^2\right] &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i)\right\|^2\right] + \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}\left[\Big\langle\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i), \mathbf{g}_j(\mathbf{x}_t^j, \boldsymbol{\xi}_t^j) - \nabla f_j(\mathbf{x}_t^j)\Big\rangle\right] \\
&\overset{(i)}{=} \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i)\right\|^2\right] \\
&\leq \frac{\overline{\nu}^2}{n}.
\end{aligned} \tag{24}$$

We note that (i) in (24) uses that whenever $i \neq j$,

$$\begin{aligned}
&\mathbb{E}\left[\Big\langle\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i), \mathbf{g}_j(\mathbf{x}_t^j, \boldsymbol{\xi}_t^j) - \nabla f_j(\mathbf{x}_t^j)\Big\rangle|\mathcal{F}_t\right] \\
&\overset{(ii)}{=} \mathbb{E}\left[\Big\langle\mathbb{E}\left[\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i)|\sigma(\boldsymbol{\xi}_t^j, \mathcal{F}_t)\right] - \nabla f_i(\mathbf{x}_t^i), \mathbf{g}_j(\mathbf{x}_t^j, \boldsymbol{\xi}_t^j) - \nabla f_j(\mathbf{x}_t^j)\Big\rangle|\mathcal{F}_t\right] \\
&\overset{(iii)}{=} \mathbb{E}\left[\Big\langle\mathbb{E}\left[\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i)|\mathcal{F}_t\right] - \nabla f_i(\mathbf{x}_t^i), \mathbf{g}_j(\mathbf{x}_t^j, \boldsymbol{\xi}_t^j) - \nabla f_j(\mathbf{x}_t^j)\Big\rangle\Big|\mathcal{F}_t\right] \\
&= 0,
\end{aligned} \tag{25}$$

where (ii) is due to the tower property of the conditional expectation and (iii) uses that $\boldsymbol{\xi}_t^j$ is independent of $\{\boldsymbol{\xi}_t^i, \mathcal{F}_t\}$ and $\mathbf{x}_t^i$ is $\mathcal{F}_t$-measurable. Towards the third term (23), we define for the ease of exposition, $\forall t \geq 1$,

$$\widehat{\nabla}_t^i := \nabla f_i(\mathbf{x}_t^i) - \nabla f_i(\mathbf{x}_{t-1}^i)$$

and recall that $\mathbb{E}\left[\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) | \mathcal{F}_t\right] = \widehat{\nabla}_t^i$. Observe that $\forall t \geq 1$,

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{z}_t\|^2 | \mathcal{F}_t\right] &= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) - \widehat{\nabla}_t^i\Big)\right\|^2 \Big| \mathcal{F}_t\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) - \widehat{\nabla}_t^i\right\|^2 | \mathcal{F}_t\right] \\
&\quad + \frac{1}{n^2}\sum_{i \neq j} \underbrace{\mathbb{E}\left[\left\langle \mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) - \widehat{\nabla}_t^i, \mathbf{g}_j(\mathbf{x}_t^j, \boldsymbol{\xi}_t^j) - \mathbf{g}_j(\mathbf{x}_{t-1}^j, \boldsymbol{\xi}_t^j) - \widehat{\nabla}_t^j \right\rangle | \mathcal{F}_t\right]}_{=0} \\
&\overset{(i)}{=} \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) - \widehat{\nabla}_t\right\|^2 | \mathcal{F}_t\right], \\
&\overset{(ii)}{\leq} \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i)\right\|^2 \Big| \mathcal{F}_t\right],
\end{aligned} \tag{26}
$$

where (i) follows from a similar line of arguments as (25) and (ii) uses the conditional variance decomposition, i.e., for any random vector $\mathbf{a} \in \mathbb{R}^p$ consisted of square-integrable random variables,

$$
\mathbb{E}\left[\left\|\mathbf{a} - \mathbb{E}\left[\mathbf{a}|\mathcal{F}_t\right]\right\|^2 | \mathcal{F}_t\right] = \mathbb{E}\left[\|\mathbf{a}\|^2 | \mathcal{F}_t\right] - \|\mathbb{E}\left[\mathbf{a}|\mathcal{F}_t\right]\|^2. \tag{27}
$$

To proceed from (26), we take its expectation and observe that $\forall t \geq 1$,

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{z}_t\|^2\right] &\leq \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i)\right\|^2\right] \\
&\overset{(i)}{\leq} \frac{L^2}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{x}_t^i - \mathbf{x}_{t-1}^i\right\|^2\right] \\
&= \frac{L^2}{n^2}\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\right] \\
&= \frac{L^2}{n^2}\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t + \mathbf{J}\mathbf{x}_t - \mathbf{J}\mathbf{x}_{t-1} + \mathbf{J}\mathbf{x}_{t-1} - \mathbf{x}_{t-1}\|^2\right] \\
&\leq \frac{3L^2}{n^2}\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2 + n\|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_{t-1}\|^2 + \|\mathbf{x}_{t-1} - \mathbf{J}\mathbf{x}_{t-1}\|^2\right] \\
&\overset{(ii)}{=} \frac{3L^2\alpha^2}{n}\mathbb{E}\left[\|\overline{\mathbf{v}}_{t-1}\|^2\right] + \frac{3L^2}{n^2}\Big(\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2 + \|\mathbf{x}_{t-1} - \mathbf{J}\mathbf{x}_{t-1}\|^2\right]\Big),
\end{aligned} \tag{28}
$$

where (i) uses the mean-squared smoothness of each $\mathbf{g}_i$ and (ii) uses the update of $\overline{\mathbf{x}}_t$ in (5). The proof follows by taking the expectation (23) and then using (24) and (28) in the resulting inequality.

B.2. Proof of Eq. (8)

We recall from (19) the following relationship: $\forall t \geq 1$,

$$
\begin{aligned}
\mathbf{v}_t^i - \nabla f_i(\mathbf{x}_t^i) &= \beta\Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i)\Big) + (1-\beta)\Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) + \nabla f_i(\mathbf{x}_{t-1}^i) - \nabla f_i(\mathbf{x}_t^i)\Big) \\
&\quad + (1-\beta)\Big(\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\Big).
\end{aligned} \tag{29}
$$

We recall that the conditional expectation of the first and the second term in (29) with respect to $\mathcal{F}_t$ is zero and that the third term in (29) is $\mathcal{F}_t$-measurable. Following a similar procedure in the proof of (23), we have: $\forall t \geq 1$,

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{v}_t^i - \nabla f_i(\mathbf{x}_t^i)\|^2|\mathcal{F}_t\right] &\leq (1-\beta)^2\left\|\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\right\|^2 + 2\beta^2\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i,\boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i)\right\|^2|\mathcal{F}_t\right] \\
&\quad + 2(1-\beta)^2\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i,\boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i,\boldsymbol{\xi}_t^i) - \left(\nabla f_i(\mathbf{x}_t^i) - \nabla f_i(\mathbf{x}_{t-1}^i)\right)\right\|^2|\mathcal{F}_t\right] \\
&\overset{(i)}{\leq} (1-\beta)^2\left\|\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\right\|^2 + 2\beta^2\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i,\boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i)\right\|^2|\mathcal{F}_t\right] \\
&\quad + 2(1-\beta)^2\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i,\boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i,\boldsymbol{\xi}_t^i)\right\|^2|\mathcal{F}_t\right]
\end{aligned} \tag{30}$$

where (i) uses the conditional variance decomposition (27). We then take the expectation of (30) with the help of the mean-squared smoothness and the bounded variance of each $\mathbf{g}_i$ to proceed: $\forall t \geq 1$,

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{v}_t^i - \nabla f_i(\mathbf{x}_t^i)\right\|^2\right] &\leq (1-\beta)^2\mathbb{E}\left[\left\|\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\right\|^2\right] + 2\beta^2\nu_i^2 + 2(1-\beta)^2L^2\mathbb{E}\left[\left\|\mathbf{x}_t^i - \mathbf{x}_{t-1}^i\right\|^2\right] \\
&\leq (1-\beta)^2\mathbb{E}\left[\left\|\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\right\|^2\right] + 2\beta^2\nu_i^2 \\
&\quad + 6(1-\beta)^2L^2\left(\mathbb{E}\left[\left\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\right\|^2 + \|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_{t-1}\|^2 + \left\|\overline{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}^i\right\|^2\right]\right), \\
&= (1-\beta)^2\mathbb{E}\left[\left\|\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\right\|^2\right] + 2\beta^2\nu_i^2 + 6(1-\beta)^2L^2\alpha^2\mathbb{E}\left[\|\overline{\mathbf{v}}_{t-1}\|^2\right] \\
&\quad + 6(1-\beta)^2L^2\mathbb{E}\left[\left\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\right\|^2 + \left\|\mathbf{x}_{t-1}^i - \overline{\mathbf{x}}_{t-1}\right\|^2\right],
\end{aligned} \tag{31}$$

where the last line uses the $\overline{\mathbf{x}}_t$-update in (5). Summing up (31) over i from 1 to n completes the proof.

C. Proof of Lemma 5

C.1. Proof of Lemma 5(a)

Recall the initialization of `GT-HSGD` that $\mathbf{v}_{-1} = \mathbf{0}_{np}$, $\mathbf{y}_0 = \mathbf{0}_{np}$, and $\mathbf{v}_0^i = \frac{1}{b_0}\sum_{r=1}^{b_0}\mathbf{g}_i(\mathbf{x}_0^i,\boldsymbol{\xi}_{0,r}^i)$. Using the gradient tracking update (4a) at iteration $t = 0$, we have:

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{y}_1 - \mathbf{J}\mathbf{y}_1\|^2\right] &= \mathbb{E}\left[\|\mathbf{W}\left(\mathbf{y}_0 + \mathbf{v}_0 - \mathbf{v}_{-1}\right) - \mathbf{J}\mathbf{W}\left(\mathbf{y}_0 + \mathbf{v}_0 - \mathbf{v}_{-1}\right)\|^2\right] \\
&\overset{(i)}{=} \mathbb{E}\left[\|\left(\mathbf{W} - \mathbf{J}\right)\mathbf{v}_0\|^2\right] \\
&\overset{(ii)}{\leq} \lambda^2\mathbb{E}\left[\|\mathbf{v}_0 - \nabla\mathbf{f}(\mathbf{x}_0) + \nabla\mathbf{f}(\mathbf{x}_0)\|^2\right] \\
&= \lambda^2\sum_{i=1}^n\mathbb{E}\left[\left\|\mathbf{v}_0^i - \nabla f_i(\mathbf{x}_0^i)\right\|^2\right] + \lambda^2\|\nabla\mathbf{f}(\mathbf{x}_0)\|^2 \\
&\overset{(iii)}{=} \lambda^2\sum_{i=1}^n\mathbb{E}\left[\left\|\frac{1}{b_0}\sum_{r=1}^{b_0}\left(\mathbf{g}_i(\mathbf{x}_0^i,\boldsymbol{\xi}_{0,r}^i) - \nabla f_i(\mathbf{x}_0^i)\right)\right\|^2\right] + \lambda^2\|\nabla\mathbf{f}(\mathbf{x}_0)\|^2 \\
&\overset{(iv)}{=} \frac{\lambda^2}{b_0^2}\sum_{i=1}^n\sum_{r=1}^{b_0}\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_0^i,\boldsymbol{\xi}_{0,r}^i) - \nabla f_i(\mathbf{x}_0^i)\right\|^2\right] + \lambda^2\|\nabla\mathbf{f}(\mathbf{x}_0)\|^2,
\end{aligned} \tag{32}$$

where (i) uses $\mathbf{J}\mathbf{W} = \mathbf{J}$ and the initial condition of $\mathbf{v}_{-1}$ and $\mathbf{y}_0$, (ii) uses $\|\mathbf{W} - \mathbf{J}\| = \lambda$, (iii) is due to the initialization of $\mathbf{v}_0^i$, and (iv) follows from the fact that $\{\boldsymbol{\xi}_{0,1}^i, \boldsymbol{\xi}_{0,2}^i, \cdots, \boldsymbol{\xi}_{0,b_0}^i\}$, $\forall i \in \mathcal{V}$, is an independent family of random vectors, by a similar line of arguments in (24) and (25). The proof then follows by using the bounded variance of each $\mathbf{g}_i$ in (32).

C.2. Proof of Lemma 5(b)

Following the gradient tracking update (4a), we have: $\forall t \geq 1$,

$$\begin{aligned}
\left\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\right\|^2 &= \left\|\mathbf{W}\left(\mathbf{y}_t + \mathbf{v}_t - \mathbf{v}_{t-1}\right) - \mathbf{J}\mathbf{W}\left(\mathbf{y}_t + \mathbf{v}_t - \mathbf{v}_{t-1}\right)\right\|^2 \\
&\overset{(i)}{=} \left\|\mathbf{W}\mathbf{y}_t - \mathbf{J}\mathbf{y}_t + \left(\mathbf{W} - \mathbf{J}\right)\left(\mathbf{v}_t - \mathbf{v}_{t-1}\right)\right\|^2 \\
&= \left\|\mathbf{W}\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\right\|^2 + 2\Big\langle \mathbf{W}\mathbf{y}_t - \mathbf{J}\mathbf{y}_t, \left(\mathbf{W} - \mathbf{J}\right)\left(\mathbf{v}_t - \mathbf{v}_{t-1}\right) \Big\rangle + \left\|\left(\mathbf{W} - \mathbf{J}\right)\left(\mathbf{v}_t - \mathbf{v}_{t-1}\right)\right\|^2 \\
&\overset{(ii)}{\leq} \lambda^2 \left\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\right\|^2 + \underbrace{2\Big\langle \mathbf{W}\mathbf{y}_t - \mathbf{J}\mathbf{y}_t, \left(\mathbf{W} - \mathbf{J}\right)\left(\mathbf{v}_t - \mathbf{v}_{t-1}\right) \Big\rangle}_{=:A_t} + \lambda^2 \left\|\mathbf{v}_t - \mathbf{v}_{t-1}\right\|^2,
\end{aligned} \tag{33}$$

where (i) uses $\mathbf{J}\mathbf{W} = \mathbf{J}$ and (ii) is due to $\|\mathbf{W} - \mathbf{J}\| = \lambda$. In the following, we bound A_t and the last term in (33) respectively. We recall the update of each local stochastic gradient estimator $\mathbf{v}_t^i$ in (2): $\forall t \geq 1$,

$$\mathbf{v}_t^i = \mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) + (1-\beta)\mathbf{v}_{t-1}^i - (1-\beta)\mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i).$$

We observe that $\forall t \geq 1$ and $\forall i \in \mathcal{V}$,

$$\begin{aligned}
\mathbf{v}_t^i - \mathbf{v}_{t-1}^i &= \mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \beta\mathbf{v}_{t-1}^i - (1-\beta)\mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) \\
&= \mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) - \beta\mathbf{v}_{t-1}^i + \beta\mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) \\
&= \mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) - \beta\Big(\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\Big) + \beta\Big(\mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_{t-1}^i)\Big).
\end{aligned} \tag{34}$$

Moreover, we observe from (34) that $\forall t \geq 1$,

$$\mathbb{E}\left[\mathbf{v}_t - \mathbf{v}_{t-1} | \mathcal{F}_t\right] = \nabla\mathbf{f}(\mathbf{x}_t) - \nabla\mathbf{f}(\mathbf{x}_{t-1}) - \beta\Big(\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\Big). \tag{35}$$

Towards A_t, we have: $\forall t \geq 1$,

$$\begin{aligned}
\mathbb{E}\left[A_t | \mathcal{F}_t\right] &\overset{(i)}{=} 2\Big\langle \mathbf{W}\mathbf{y}_t - \mathbf{J}\mathbf{y}_t, \left(\mathbf{W} - \mathbf{J}\right)\mathbb{E}\left[\mathbf{v}_t - \mathbf{v}_{t-1} | \mathcal{F}_t\right] \Big\rangle \\
&\overset{(ii)}{=} 2\Big\langle \mathbf{W}\mathbf{y}_t - \mathbf{J}\mathbf{y}_t, \left(\mathbf{W} - \mathbf{J}\right)\Big(\nabla\mathbf{f}(\mathbf{x}_t) - \nabla\mathbf{f}(\mathbf{x}_{t-1}) - \beta\big(\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\big)\Big) \Big\rangle \\
&\overset{(iii)}{\leq} 2\lambda \left\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\right\| \cdot \lambda \left\|\nabla\mathbf{f}(\mathbf{x}_t) - \nabla\mathbf{f}(\mathbf{x}_{t-1}) - \beta\Big(\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\Big)\right\| \\
&\overset{(iv)}{\leq} \frac{1-\lambda^2}{2}\left\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\right\|^2 + \frac{2\lambda^4}{1-\lambda^2}\left\|\nabla\mathbf{f}(\mathbf{x}_t) - \nabla\mathbf{f}(\mathbf{x}_{t-1}) - \beta\Big(\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\Big)\right\|^2, \\
&\overset{(v)}{\leq} \frac{1-\lambda^2}{2}\left\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\right\|^2 + \frac{4\lambda^4 L^2}{1-\lambda^2}\left\|\mathbf{x}_t - \mathbf{x}_{t-1}\right\|^2 + \frac{4\lambda^4\beta^2}{1-\lambda^2}\left\|\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\right\|^2,
\end{aligned} \tag{36}$$

where (i) is due to the $\mathcal{F}_t$-measurability of $\mathbf{y}_t$, (ii) uses (35), (iii) is due to the Cauchy-Schwarz inequality and $\|\mathbf{W} - \mathbf{J}\| = \lambda$, (iv) uses the elementary inequality that $2ab \leq \eta a^2 + b^2/\eta$, with $\eta = \frac{1-\lambda^2}{2\lambda^2}$ for any $a, b \in \mathbb{R}$, and (v) holds since each f_i is L-smooth. Next, towards the last term in (33), we take the expectation of (34) to obtain: $\forall t \geq 1$ and $\forall i \in \mathcal{V}$,

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{v}_t^i - \mathbf{v}_{t-1}^i\right\|^2\right] &\leq 3\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i)\right\|^2\right] + 3\beta^2\mathbb{E}\left[\left\|\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\right\|^2\right] \\
&\quad + 3\beta^2\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_{t-1}^i)\right\|^2\right] \\
&\leq 3L^2\mathbb{E}\left[\left\|\mathbf{x}_t^i - \mathbf{x}_{t-1}^i\right\|^2\right] + 3\beta^2\mathbb{E}\left[\left\|\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\right\|^2\right] + 3\beta^2\nu_i^2,
\end{aligned} \tag{37}$$

where (37) is due to the mean-squared smoothness and the bounded variance of each $\mathbf{g}_i$. Summing up (37) over i from 1 to n gives an upper bound on the last term in (33): $\forall t \geq 1$,

$$\lambda^2\mathbb{E}\left[\left\|\mathbf{v}_t - \mathbf{v}_{t-1}\right\|^2\right] \leq 3\lambda^2L^2\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{x}_{t-1}\right\|^2\right] + 3\lambda^2\beta^2\mathbb{E}\left[\left\|\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\right\|^2\right] + 3\lambda^2 n\beta^2\overline{\nu}^2. \tag{38}$$

We now use (36) and (38) in (33) to obtain: $\forall t \geq 1$,

$$\mathbb{E}\left[\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2\right] \leq \frac{1+\lambda^2}{2}\mathbb{E}\left[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\right] + \frac{7\lambda^2 L^2}{1-\lambda^2}\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\right] + \frac{7\lambda^2\beta^2}{1-\lambda^2}\mathbb{E}\left[\|\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\|^2\right] + 3\lambda^2 n\beta^2\overline{\nu}^2. \quad (39)$$

Towards the second term in (39), we use (10) to obtain: $\forall t \geq 1$,

$$\begin{aligned}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 &= \|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t + \mathbf{J}\mathbf{x}_t - \mathbf{J}\mathbf{x}_{t-1} + \mathbf{J}\mathbf{x}_{t-1} - \mathbf{x}_{t-1}\|^2 \\ &\overset{(i)}{\leq} 3\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2 + 3n\alpha^2\|\overline{\mathbf{v}}_{t-1}\|^2 + 3\|\mathbf{x}_{t-1} - \mathbf{J}\mathbf{x}_{t-1}\|^2 \\ &\leq 6\lambda^2\alpha^2\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2 + 3n\alpha^2\|\overline{\mathbf{v}}_{t-1}\|^2 + 9\|\mathbf{x}_{t-1} - \mathbf{J}\mathbf{x}_{t-1}\|^2, \end{aligned} \quad (40)$$

where (i) uses the $\overline{\mathbf{x}}_t$-update in (5). Finally, we use (40) in (39) to obtain: $\forall t \geq 1$,

$$\mathbb{E}\left[\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2\right] \leq \left(\frac{1+\lambda^2}{2} + \frac{42\lambda^4 L^2\alpha^2}{1-\lambda^2}\right)\mathbb{E}\left[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\right] + \frac{21\lambda^2 nL^2\alpha^2}{1-\lambda^2}\mathbb{E}\left[\|\overline{\mathbf{v}}_{t-1}\|^2\right] + \frac{63\lambda^2 L^2}{1-\lambda^2}\mathbb{E}\left[\|\mathbf{x}_{t-1} - \mathbf{J}\mathbf{x}_{t-1}\|^2\right] + \frac{7\lambda^2\beta^2}{1-\lambda^2}\mathbb{E}\left[\|\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\|^2\right] + 3\lambda^2 n\beta^2\overline{\nu}^2.$$

The proof is completed by the fact that $\frac{1+\lambda^2}{2} + \frac{42\lambda^4 L^2\alpha^2}{1-\lambda^2} \leq \frac{3+\lambda^2}{4}$ if $0 < \alpha \leq \frac{1-\lambda^2}{2\sqrt{42}\lambda^2 L}$.

D. Proof of Lemma 6

D.1. Proof of Eq. (11)

We recursively apply the inequality on V_t from t to 0 to obtain: $\forall t \geq 1$,

$$\begin{aligned} V_t &\leq qV_{t-1} + qR_{t-1} + Q_t + C \\ &\leq q^2 V_{t-2} + (q^2 R_{t-2} + qR_{t-1}) + (qQ_{t-1} + Q_t) + (qC + C) \\ &\cdots \\ &\leq q^t V_0 + \sum_{i=0}^{t-1} q^{t-i}R_i + \sum_{i=1}^{t} q^{t-i}Q_i + C\sum_{i=0}^{t-1} q^i. \end{aligned} \quad (41)$$

Summing up (41) over t from 1 to T gives: $\forall T \geq 1$,

$$\begin{aligned}\sum_{t=0}^{T} V_t &\leq V_0\sum_{t=0}^{T} q^t + \sum_{t=1}^{T}\sum_{i=0}^{t-1} q^{t-i}R_i + \sum_{t=1}^{T}\sum_{i=1}^{t} q^{t-i}Q_i + C\sum_{t=1}^{T}\sum_{i=0}^{t-1} q^i \\ &\leq V_0\sum_{t=0}^{\infty} q^t + \sum_{t=0}^{T-1}\left(\sum_{i=0}^{\infty} q^i\right)R_t + \sum_{t=1}^{T}\left(\sum_{i=0}^{\infty} q^i\right)Q_t + C\sum_{t=1}^{T}\sum_{i=0}^{\infty} q^i, \end{aligned}$$

and the proof follows by $\sum_{i=0}^{\infty} q^i = (1-q)^{-1}$.

D.2. Proof of Eq. (12)

We recursively apply the inequality on V_t from $t+1$ to 1 to obtain: $\forall t \geq 1$,

$$\begin{aligned} V_{t+1} &\leq qV_t + R_{t-1} + C \\ &\leq q^2 V_{t-1} + (qR_{t-2} + R_{t-1}) + (qC + C) \\ &\cdots \\ &\leq q^t V_1 + \sum_{i=0}^{t-1} q^{t-1-i}R_i + C\sum_{i=0}^{t-1} q^i. \end{aligned} \quad (42)$$

We sum up (42) over t from 1 to $T-1$ to obtain: $\forall T \geq 2$,

$$\begin{aligned}\sum_{t=0}^{T-1} V_{t+1} &\leq V_1 \sum_{t=0}^{T-1} q^t + \sum_{t=1}^{T-1}\sum_{i=0}^{t-1} q^{t-1-i} R_i + C\sum_{t=1}^{T-1}\sum_{i=0}^{t-1} q^i \\ &\leq V_1 \sum_{t=0}^{\infty} q^t + \sum_{t=0}^{T-2}\left(\sum_{i=0}^{\infty} q^i\right) R_t + C\sum_{t=1}^{T-1}\sum_{i=0}^{\infty} q^i,\end{aligned}$$

and the proof follows by $\sum_{i=0}^{\infty} q^i = (1-q)^{-1}$.

E. Proof of Lemma 7

E.1. Proof of Eq. (13)

We first observe that $\frac{1}{1-(1-\beta)^2} \leq \frac{1}{\beta}$ for $\beta \in (0,1)$. Applying (11) to (7) gives: $\forall T \geq 1$,

$$\begin{aligned}&\sum_{t=0}^{T} \mathbb{E}\left[\left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2\right] \\ &\leq \frac{\mathbb{E}\left[\|\overline{\mathbf{v}}_0 - \overline{\nabla \mathbf{f}}(\mathbf{x}_0)\|^2\right]}{\beta} + \frac{6L^2\alpha^2}{n\beta}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{6L^2}{n^2\beta}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\mathbf{x}_{t+1} - \mathbf{J}\mathbf{x}_{t+1}\right\|^2 + \left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right] + \frac{2\beta\overline{\nu}^2 T}{n} \\ &\leq \frac{\mathbb{E}\left[\|\overline{\mathbf{v}}_0 - \overline{\nabla \mathbf{f}}(\mathbf{x}_0)\|^2\right]}{\beta} + \frac{6L^2\alpha^2}{n\beta}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{12L^2}{n^2\beta}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right] + \frac{2\beta\overline{\nu}^2 T}{n}.\end{aligned} \tag{43}$$

Towards the first term in (43), we observe that

$$\begin{aligned}\mathbb{E}\left[\left\|\overline{\mathbf{v}}_0 - \overline{\nabla \mathbf{f}}(\mathbf{x}_0)\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{b_0}\sum_{r=1}^{b_0}\left(\mathbf{g}_i(\mathbf{x}_0^i, \boldsymbol{\xi}_{0,r}^i) - \nabla f_i(\mathbf{x}_0^i)\right)\right\|^2\right] \\ &\overset{(i)}{=} \frac{1}{n^2 b_0^2}\sum_{i=1}^{n}\sum_{r=1}^{b_0}\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_0^i, \boldsymbol{\xi}_{0,r}^i) - \nabla f_i(\mathbf{x}_0^i)\right\|^2\right] \leq \frac{\overline{\nu}^2}{n b_0},\end{aligned} \tag{44}$$

where (i) follows from a similar line of arguments in (25). Then (13) follows from using (44) in (43).

E.2. Proof of Eq. (14)

We apply (11) to (8) to obtain: $\forall T \geq 1$,

$$\begin{aligned}&\sum_{t=0}^{T} \mathbb{E}\left[\left\|\mathbf{v}_t - \nabla \mathbf{f}(\mathbf{x}_t)\right\|^2\right] \\ &\leq \frac{\mathbb{E}\left[\|\mathbf{v}_0 - \nabla \mathbf{f}(\mathbf{x}_0)\|^2\right]}{\beta} + \frac{6nL^2\alpha^2}{\beta}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{6L^2}{\beta}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\mathbf{x}_{t+1} - \mathbf{J}\mathbf{x}_{t+1}\right\|^2 + \left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right] + 2n\beta\overline{\nu}^2 T \\ &\leq \frac{\mathbb{E}\left[\|\mathbf{v}_0 - \nabla \mathbf{f}(\mathbf{x}_0)\|^2\right]}{\beta} + \frac{6nL^2\alpha^2}{\beta}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{12L^2}{\beta}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right] + 2n\beta\overline{\nu}^2 T.\end{aligned} \tag{45}$$

In (45), we observe that

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{v}_0 - \nabla \mathbf{f}(\mathbf{x}_0)\|^2\right] &= \sum_{i=1}^{n}\mathbb{E}\left[\left\|\frac{1}{b_0}\sum_{r=1}^{b_0}\left(\mathbf{g}_i(\mathbf{x}_0^i, \boldsymbol{\xi}_{0,r}^i) - \nabla f_i(\mathbf{x}_0^i)\right)\right\|^2\right] \\ &\overset{(i)}{=} \frac{1}{b_0^2}\sum_{i=1}^{n}\sum_{r=1}^{b_0}\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_0^i, \boldsymbol{\xi}_{0,r}^i) - \nabla f_i(\mathbf{x}_0^i)\right\|^2\right] \leq \frac{n\overline{\nu}^2}{b_0},\end{aligned} \tag{46}$$

where (i) follows from a similar line of arguments in (25). Then (14) follows from using (46) in (45).

F. Proof of Lemma 8

We recall that $\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\| = 0$, since it is assumed without generality that $\mathbf{x}_0^i = \mathbf{x}_0^j$ for any $i, j \in \mathcal{V}$. Applying (11) to (9) yields: $\forall T \geq 1$,

$$\sum_{t=0}^{T} \|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2 \leq \frac{4\lambda^2\alpha^2}{(1-\lambda^2)^2} \sum_{t=1}^{T} \|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2. \tag{47}$$

To further bound $\sum_{t=1}^{T} \|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2$, we apply (12) in Lemma 5(b) to obtain: if $0 < \alpha \leq \frac{1-\lambda^2}{2\sqrt{42}\lambda^2 L}$, then $\forall T \geq 2$,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\right] \\
&\leq \frac{4\mathbb{E}\left[\|\mathbf{y}_1 - \mathbf{J}\mathbf{y}_1\|^2\right]}{1-\lambda^2} + \frac{84\lambda^2 n L^2 \alpha^2}{(1-\lambda^2)^2} \sum_{t=0}^{T-2} \mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{252\lambda^2 L^2}{(1-\lambda^2)^2} \sum_{t=0}^{T-2} \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right] \\
&\quad + \frac{28\lambda^2\beta^2}{(1-\lambda^2)^2} \sum_{t=0}^{T-2} \mathbb{E}\left[\|\mathbf{v}_t - \nabla \mathbf{f}(\mathbf{x}_t)\|^2\right] + \frac{12\lambda^2 n \beta^2 \overline{\nu}^2 T}{1-\lambda^2} \\
&\leq \frac{84\lambda^2 n L^2 \alpha^2}{(1-\lambda^2)^2} \sum_{t=0}^{T-2} \mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{252\lambda^2 L^2}{(1-\lambda^2)^2} \sum_{t=0}^{T-2} \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right] \\
&\quad + \frac{28\lambda^2\beta^2}{(1-\lambda^2)^2} \sum_{t=0}^{T-2} \mathbb{E}\left[\|\mathbf{v}_t - \nabla \mathbf{f}(\mathbf{x}_t)\|^2\right] + \frac{12\lambda^2 n \beta^2 \overline{\nu}^2 T}{1-\lambda^2} + \frac{4\lambda^2 \left\|\nabla \mathbf{f}(\mathbf{x}_0)\right\|^2}{1-\lambda^2} + \frac{4\lambda^2 n \overline{\nu}^2}{(1-\lambda^2) b_0},
\end{aligned} \tag{48}$$

where the last inequality is due to Lemma 5(a). To proceed, we use (14), an upper bound on $\sum_t \mathbb{E}\left[\|\mathbf{v}_t - \nabla \mathbf{f}(\mathbf{x}_t)\|^2\right]$, in (48) to obtain: if $0 < \alpha \leq \frac{1-\lambda^2}{2\sqrt{42}\lambda^2 L}$ and $\beta \in (0,1)$, then $\forall T \geq 2$,

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\left[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\right] &\leq \frac{252\lambda^2 n L^2 \alpha^2}{(1-\lambda^2)^2} \sum_{t=0}^{T-2} \mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{588\lambda^2 L^2}{(1-\lambda^2)^2} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right] \\
&\quad + \frac{28\lambda^2 n \beta \overline{\nu}^2}{(1-\lambda^2)^2 b_0} + \frac{56\lambda^2 n \beta^3 \overline{\nu}^2 T}{(1-\lambda^2)^2} + \frac{12\lambda^2 n \beta^2 \overline{\nu}^2 T}{1-\lambda^2} + \frac{4\lambda^2 \left\|\nabla \mathbf{f}(\mathbf{x}_0)\right\|^2}{1-\lambda^2} + \frac{4\lambda^2 n \overline{\nu}^2}{(1-\lambda^2) b_0} \\
&= \frac{252\lambda^2 n L^2 \alpha^2}{(1-\lambda^2)^2} \sum_{t=0}^{T-2} \mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{588\lambda^2 L^2}{(1-\lambda^2)^2} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right] \\
&\quad + \left(\frac{7\beta}{1-\lambda^2} + 1\right) \frac{4\lambda^2 n \overline{\nu}^2}{(1-\lambda^2) b_0} + \left(\frac{14\beta}{1-\lambda^2} + 3\right) \frac{4\lambda^2 n \beta^2 \overline{\nu}^2 T}{1-\lambda^2} + \frac{4\lambda^2 \left\|\nabla \mathbf{f}(\mathbf{x}_0)\right\|^2}{1-\lambda^2}.
\end{aligned} \tag{49}$$

Finally, we use (49) in (47) to obtain: $\forall T \geq 2$,

$$\begin{aligned}
\sum_{t=0}^{T} \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right] &\leq \frac{1008\lambda^4 n L^2 \alpha^4}{(1-\lambda^2)^4} \sum_{t=0}^{T-2} \mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{2352\lambda^4 L^2 \alpha^2}{(1-\lambda^2)^4} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right] \\
&\quad + \left(\frac{7\beta}{1-\lambda^2} + 1\right) \frac{16\lambda^4 n \overline{\nu}^2 \alpha^2}{(1-\lambda^2)^3 b_0} + \left(\frac{14\beta}{1-\lambda^2} + 3\right) \frac{16\lambda^4 n \beta^2 \overline{\nu}^2 \alpha^2 T}{(1-\lambda^2)^3} + \frac{16\lambda^4 \left\|\nabla \mathbf{f}(\mathbf{x}_0)\right\|^2 \alpha^2}{(1-\lambda^2)^3},
\end{aligned}$$

which may be written equivalently as

$$\begin{aligned}
\left(1 - \frac{2352\lambda^4 L^2 \alpha^2}{(1-\lambda^2)^4}\right) \sum_{t=0}^{T} \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right] &\leq \frac{1008\lambda^4 n L^2 \alpha^4}{(1-\lambda^2)^4} \sum_{t=0}^{T-2} \mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \left(\frac{7\beta}{1-\lambda^2} + 1\right) \frac{16\lambda^4 n \overline{\nu}^2 \alpha^2}{(1-\lambda^2)^3 b_0} \\
&\quad + \left(\frac{14\beta}{1-\lambda^2} + 3\right) \frac{16\lambda^4 n \beta^2 \overline{\nu}^2 \alpha^2 T}{(1-\lambda^2)^3} + \frac{16\lambda^4 \left\|\nabla \mathbf{f}(\mathbf{x}_0)\right\|^2 \alpha^2}{(1-\lambda^2)^3}.
\end{aligned} \tag{50}$$

We observe in (50) that $\frac{2352\lambda^4 L^2 \alpha^2}{(1-\lambda^2)^4} \leq \frac{1}{2}$ if $0 < \alpha \leq \frac{(1-\lambda^2)^2}{70\lambda^2 L}$, and the proof follows.

G. Proof of Theorem 1

For the ease of presentation, we denote $\Delta_0 := F(\overline{\mathbf{x}}_0) - F^*$ in the following. We apply (13) to Lemma 2 to obtain: if $0 < \alpha \leq \frac{1}{2L}$, then $\forall T \geq 1$,

$$\begin{aligned}
\sum_{t=0}^{T} \mathbb{E}\left[\|\nabla F(\overline{\mathbf{x}}_t)\|^2\right] &\leq \frac{2\Delta_0}{\alpha} - \frac{1}{2}\sum_{t=0}^{T}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{2L^2}{n}\sum_{t=0}^{T}\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right] \\
&\quad + \frac{2\overline{\nu}^2}{\beta b_0 n} + \frac{12L^2\alpha^2}{n\beta}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{24L^2}{n^2\beta}\sum_{t=0}^{T}\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right] + \frac{4\beta\overline{\nu}^2 T}{n} \\
&\leq \frac{2\Delta_0}{\alpha} - \frac{1}{4}\sum_{t=0}^{T}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{2L^2}{n}\left(1 + \frac{12}{n\beta}\right)\sum_{t=0}^{T}\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right] \\
&\quad + \frac{2\overline{\nu}^2}{\beta b_0 n} + \frac{4\beta\overline{\nu}^2 T}{n} - \left(\frac{1}{4} - \frac{12L^2\alpha^2}{n\beta}\right)\sum_{t=0}^{T}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right]. \qquad (51)
\end{aligned}$$

Therefore, if $0 < \alpha < \frac{1}{4\sqrt{3}L}$ and $\frac{48L^2\alpha^2}{n} \leq \beta < 1$, i.e., $\frac{1}{4} - \frac{12L^2\alpha^2}{n\beta} \geq 0$, we may drop the last term in (51) to obtain: $\forall T \geq 1$,

$$\sum_{t=0}^{T}\mathbb{E}\left[\|\nabla F(\overline{\mathbf{x}}_t)\|^2\right] \leq \frac{2\Delta_0}{\alpha} - \frac{1}{4}\sum_{t=0}^{T}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{2L^2}{n}\left(1 + \frac{12}{n\beta}\right)\sum_{t=0}^{T}\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right] + \frac{2\overline{\nu}^2}{\beta b_0 n} + \frac{4\beta\overline{\nu}^2 T}{n}. \qquad (52)$$

Moreover, we observe: $\forall T \geq 1$,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t^i)\right\|^2\right] &\leq \frac{2}{n}\sum_{i=1}^{n}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t^i) - \nabla F(\overline{\mathbf{x}}_t)\right\|^2 + \|\nabla F(\overline{\mathbf{x}}_t)\|^2\right] \\
&= \frac{2L^2}{n}\sum_{t=0}^{T}\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right] + 2\sum_{t=0}^{T}\mathbb{E}\left[\|\nabla F(\overline{\mathbf{x}}_t)\|^2\right], \qquad (53)
\end{aligned}$$

where the last line uses the L-smoothness of F. Using (52) in (53) yields: if $0 < \alpha < \frac{1}{4\sqrt{3}L}$ and $48L^2\alpha^2/n \leq \beta < 1$, then

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t^i)\right\|^2\right] \leq \frac{4\Delta_0}{\alpha} - \frac{1}{2}\sum_{t=0}^{T}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{6L^2}{n}\left(1 + \frac{8}{n\beta}\right)\sum_{t=0}^{T}\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right] + \frac{4\overline{\nu}^2}{\beta b_0 n} + \frac{8\beta\overline{\nu}^2 T}{n}. \qquad (54)$$

According to (54), if $0 < \alpha < \frac{1}{4\sqrt{3}L}$ and $\beta = 48L^2\alpha^2/n$, we have: $\forall T \geq 1$,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t^i)\right\|^2\right] &\leq \frac{4\Delta_0}{\alpha} - \frac{1}{2}\sum_{t=0}^{T}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{6L^2}{n}\left(1 + \frac{1}{6L^2\alpha^2}\right)\sum_{t=0}^{T}\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right] + \frac{4\overline{\nu}^2}{\beta b_0 n} + \frac{8\beta\overline{\nu}^2 T}{n} \\
&\leq \frac{4\Delta_0}{\alpha} \underbrace{- \frac{1}{2}\sum_{t=0}^{T}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{2}{n\alpha^2}\sum_{t=0}^{T}\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right]}_{=:\Phi_T} + \frac{4\overline{\nu}^2}{\beta b_0 n} + \frac{8\beta\overline{\nu}^2 T}{n}, \qquad (55)
\end{aligned}$$

where the last line is due to $6L^2\alpha^2 < 1/8$. To simplify Φ_T, we use Lemma 8 to obtain: if $0 < \alpha \leq \frac{(1-\lambda^2)^2}{70\lambda^2 L}$ then $\forall T \geq 2$,

$$\begin{aligned}
\Phi_T \leq &-\frac{1}{2}\left(1 - \frac{8064\lambda^4L^2\alpha^2}{(1-\lambda^2)^4}\right)\sum_{t=0}^{T}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{64\lambda^4}{(1-\lambda^2)^3}\frac{\left\|\nabla \mathbf{f}(\mathbf{x}_0)\right\|^2}{n} \\
&+ \left(\frac{7\beta}{1-\lambda^2} + 1\right)\frac{64\lambda^4\overline{\nu}^2}{(1-\lambda^2)^3 b_0} + \left(\frac{14\beta}{1-\lambda^2} + 3\right)\frac{64\lambda^4\beta^2\overline{\nu}^2 T}{(1-\lambda^2)^3}. \qquad (56)
\end{aligned}$$

In (56), we observe that if $0 < \alpha \leq \frac{(1-\lambda^2)^2}{90\lambda^2 L}$, then $1 - \frac{8064\lambda^4 L^2\alpha^2}{(1-\lambda^2)^4} \geq 0$ and thus the first term in (56) may be dropped; moreover, if $0 < \alpha \leq \frac{\sqrt{n(1-\lambda^2)}}{26\lambda L}$, then $\beta = \frac{48L^2\alpha^2}{n} \leq \frac{1-\lambda^2}{14\lambda^2}$. Hence, if $0 < \alpha \leq \min\left\{\frac{(1-\lambda^2)^2}{90\lambda^2}, \frac{\sqrt{n(1-\lambda^2)}}{26\lambda}\right\}\frac{1}{L}$, then (56) reduces to: $\forall T \geq 2$,

$$\Phi_T \leq \frac{64\lambda^4}{(1-\lambda^2)^3}\frac{\left\|\nabla \mathbf{f}\big(\mathbf{x}_0\big)\right\|^2}{n} + \frac{96\lambda^2\overline{\nu}^2}{(1-\lambda^2)^3 b_0} + \frac{256\lambda^2\beta^2\overline{\nu}^2 T}{(1-\lambda^2)^3}. \tag{57}$$

Finally, we use (57) in (55) to obtain: if $0 < \alpha < \min\left\{\frac{1}{4\sqrt{3}}, \frac{(1-\lambda^2)^2}{90\lambda^2}, \frac{\sqrt{n(1-\lambda^2)}}{26\lambda}\right\}\frac{1}{L}$, we have: $\forall T \geq 2$,

$$\begin{aligned}\frac{1}{n(T+1)}\sum_{i=1}^{n}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t^i)\right\|^2\right] \leq \frac{4\Delta_0}{\alpha T} + \frac{4\overline{\nu}^2}{\beta b_0 n T} + \frac{8\beta\overline{\nu}^2}{n} \\ + \frac{64\lambda^4}{(1-\lambda^2)^3 T}\frac{\left\|\nabla \mathbf{f}\big(\mathbf{x}_0\big)\right\|^2}{n} + \frac{96\lambda^2\overline{\nu}^2}{(1-\lambda^2)^3 b_0 T} + \frac{256\lambda^2\beta^2\overline{\nu}^2}{(1-\lambda^2)^3}.\end{aligned} \tag{58}$$

The proof follows by (58) and that $\mathbb{E}[\|\nabla F(\widetilde{\mathbf{x}}_T)\|^2] = \frac{1}{n(T+1)}\sum_{i=1}^{n}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t^i)\|^2]$ since $\widetilde{\mathbf{x}}_T$ is chosen uniformly at random from $\{\mathbf{x}_t^i : \forall i \in \mathcal{V}, 0 \leq t \leq T\}$.