Appendix

Here we present the omitted proofs of convergence rates. In Section A we give the proof of convergence in strongly-convex-strongly-concave setting. Section B includes the proof for nonconvex-strongly-concave functions, and in Section C we present proof of local SGDA+ for nonconvex-PL objectives. Finally, in Section D we provide the proof of local SGDA+ on nonconvex-one-point-concave setting.

A Strongly-Convex-Strongly-Concave Setting

A.1 Overview of proof techniques

Before we dive into the proof we first sketch the proof of convergence of local SGDA under strongly-convex-strongly-concave setting. We define the following notions to denote the (virtual) average primal and dual solution at tth iteration:

$$\boldsymbol{x}^{(t)} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^{(t)}, \quad \boldsymbol{y}^{(t)} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i^{(t)},$$

and the deviation between local primal and dual solutions and their corresponding averages:

$$\delta_{\boldsymbol{x}}^{(t)} = \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}^{(t)}\right\|^2, \quad \delta_{\boldsymbol{y}}^{(t)} = \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2.$$

Homogeneous setting In homogeneous setting, we first study the behavior of local SGDA for one iteration. With the help of strong convexity, concavity and smoothness we can show that:

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(t+1)} - \boldsymbol{y}^*\right\|^2\right] &\leq \left(1 - \frac{1}{2}\mu\eta\right)\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*\right\|^2\right]\right) \\ &\quad - 2\eta\mathbb{E}\left(F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) - F(\boldsymbol{x}^*, \boldsymbol{y}^{(t)})\right) \\ &\quad + \frac{2\eta^2\sigma^2}{n} + \frac{16\eta_t L^2}{\mu}\mathbb{E}\left(\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right) + 8\eta^2 L^2\mathbb{E}\left(\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right).\end{aligned}$$

Then, to bound $\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}$, with the help of strong convexity and smoothness, we can indeed show that it decreases in the order of $O(\tau(1 + (L-\mu)\eta)^{2\tau}\eta^2\sigma^2)$. By properly choosing τ and η, we recover the rate $O(\tau\eta^2\sigma^2)$ as desired.

Heterogeneous setting Similarly to homogeneous setting, we first do the one iteration analysis

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(t+1)} - \boldsymbol{y}^*\right\|^2\right] &\leq \left(1 - \frac{1}{2}\mu\eta_t\right)\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*\right\|^2\right]\right) \\ &\quad - 2\eta_t\mathbb{E}\left(F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) - F(\boldsymbol{x}^*, \boldsymbol{y}^{(t)})\right) \\ &\quad + \frac{2\eta_t^2\sigma^2}{n} + \frac{16\eta_t L^2}{\mu}\mathbb{E}\left(\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right) + 8\eta_t^2 L^2\mathbb{E}\left(\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right).\end{aligned}$$

Next we need to bound deviation $\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}$, which is also our main technical contribution in this section. We consider the interval of τ steps, if we choose step size to be small enough and properly choose quadratic weights $w_t = (t+a)^2$, to make sure the deviation changes slowly, we can finally prove the following statement:

$$\begin{aligned}\sum_{t=s\tau}^{(s+1)\tau} w_t\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right] &\leq \frac{\mu}{128L^2}\sum_{j=s\tau}^{(s+1)\tau}\mu\eta_j\frac{w_j}{\eta_j}\mathbb{E}\left[\left\|\boldsymbol{x}^{(j)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(j)} - \boldsymbol{y}^*\right\|^2\right] \\ &\quad + 64\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^2\left(\Delta_x + \Delta_y\right) + 32\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^2\sigma^2,\end{aligned}$$

where we related the deviation to the gap between current iterates and saddle points, and heterogeneity at global optimum.

A.2 Proof in homogeneous setting

In this section we are going to present the proof in homogeneous case. Let us introduce some technical lemmas first which will help our proof.

A.2.1 Proof of technical lemmas

The following lemma performs one iteration analysis of local SGDA, on strongly convex function.

Lemma A.1. *For local-SGDA, under Theorem 4.1's assumptions, the following relation holds true:*

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^*\right\|^2\right]+\mathbb{E}\left[\left\|\boldsymbol{y}^{(t+1)}-\boldsymbol{y}^*\right\|^2\right]\leq&\left(1-\frac{1}{2}\mu\eta\right)\left[\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\|^2\right]+\mathbb{E}\left[\left\|\boldsymbol{y}^{(t)}-\boldsymbol{y}^*\right\|^2\right]\right]\\&-2\eta\left(\mathbb{E}\left[F(\boldsymbol{x}^{(t)},\boldsymbol{y}^*)\right]-\mathbb{E}\left[F(\boldsymbol{x}^*,\boldsymbol{y}^{(t)})\right]\right)\\&+\frac{2\eta^2\sigma^2}{n}+\frac{16\eta_t L^2}{\mu}\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)}\right]+8\eta^2L^2\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)}\right],\end{aligned}$$

where $\delta_{\boldsymbol{x}}^{(t)}=\frac{1}{n}\sum_{i=1}^n\left\|\boldsymbol{x}_i^{(t)}-\boldsymbol{x}^{(t)}\right\|^2,\quad \delta_{\boldsymbol{y}}^{(t)}=\frac{1}{n}\sum_{i=1}^n\left\|\boldsymbol{y}_i^{(t)}-\boldsymbol{y}^{(t)}\right\|^2.$

Proof. According to updating rule and strong convexity we have:

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^*\right\|^2\right]&=\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)}-\eta\frac{1}{n}\sum_{i=1}^n\nabla_xF(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)};\xi_i^{(t)})-\boldsymbol{x}^*\right\|^2\right]\\&\leq\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\|^2\right]-2\eta\mathbb{E}\left\langle\frac{1}{n}\sum_{i=1}^n\nabla_xF(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)}),\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\rangle\\&\quad+\frac{\eta^2\sigma^2}{n}+\eta^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\nabla_xF(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right]\\&\leq\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\|^2\right]-2\eta\left\langle\nabla_xF(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)}),\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\rangle\\&\quad-2\eta\mathbb{E}\left\langle\frac{1}{n}\sum_{i=1}^n\nabla_xF(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_xF(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)}),\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\rangle\\&\quad+\frac{\eta^2\sigma^2}{n}+\eta^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\nabla_xF(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right]\\&\leq(1-\mu\eta)\,\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\|^2\right]-2\eta\mathbb{E}\left(F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})-F(\boldsymbol{x}^*,\boldsymbol{y}^{(t)})\right)\\&\quad+\eta\left(\frac{4}{\mu}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n\nabla_xF(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_xF(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2+\frac{\mu}{4}\mathbb{E}\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\|^2\right)\\&\quad+\frac{\eta^2\sigma^2}{n}+\eta^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\nabla_xF(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right].\end{aligned}$$

We now proceed to bound terms $\left\|\frac{1}{n}\sum_{i=1}^n\nabla_xF(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_xF(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2$ and $\left\|\frac{1}{n}\sum_{i=1}^n\nabla_xF(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2$.

By applying Jensen's inequality on $\left\|\frac{1}{n}\sum_{i=1}^n\nabla_xF(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_xF(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2$ we have:

$$\left\|\frac{1}{n}\sum_{i=1}^n\nabla_xF(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_xF(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla_x F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)}) - \nabla_x F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left(2\left\|\nabla_x F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)}) - \nabla_x F(\boldsymbol{x}^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2 + 2\left\|\nabla_x F(\boldsymbol{x}^{(t)},\boldsymbol{y}_i^{(t)}) - \nabla_x F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2\right) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left(2L^2\left\|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}^{(t)}\right\|^2 + 2L^2\left\|\boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2\right) \\
&\leq 2L^2(\delta_x^{(t)} + \delta_y^{(t)}),
\end{aligned}
$$

where we use the smoothness in the second last inequality.

Then we switch to bound $\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2$ as follows:

$$
\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2 &= \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla_x F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla_x F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)}) - \nabla_x F(\boldsymbol{x}^*,\boldsymbol{y}^*)\right\|^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n}2\left(\left\|\nabla_x F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)}) - \nabla_x F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2 + \left\|\nabla_x F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)}) - \nabla_x F(\boldsymbol{x}^*,\boldsymbol{y}^*)\right\|^2\right) \\
&\leq L^2\frac{1}{n}\sum_{i=1}^{n}4\left(\left\|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}^{(t)}\right\|^2 + \left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2 + \left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*\right\|^2\right).
\end{aligned}
$$

where in the second equality we used the fact that $\nabla_x F(\boldsymbol{x}^*,\boldsymbol{y}^*) = \mathbf{0}$.

Putting these pieces together yields:

$$
\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\right\|^2\right] \leq &\left(1 - \frac{3}{4}\mu\eta\right)\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|^2\right] - 2\eta\mathbb{E}\left(F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)}) - F(\boldsymbol{x}^*,\boldsymbol{y}^{(t)})\right) \\
&+ \frac{8}{\mu}\eta_t L^2(\delta_x^{(t)} + \delta_y^{(t)}) + \frac{\eta^2\sigma^2}{n} \\
&+ 4\eta^2 L^2\mathbb{E}\left(\delta_x^{(t)} + \left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^* - \boldsymbol{y}^{(t)}\right\|^2 + \delta_y^{(t)}\right).
\end{aligned}
$$

Similarly, we can get:

$$
\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{y}^{(t+1)} - \boldsymbol{y}^*\right\|^2\right] \leq &\left(1 - \frac{3}{4}\mu\eta\right)\mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*\right\|^2\right] - 2\eta\mathbb{E}\left(F(\boldsymbol{x}^{(t)},\boldsymbol{y}^*) - F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right) \\
&+ \frac{8}{\mu}\eta L^2\mathbb{E}(\delta_x^{(t)} + \delta_y^{(t)}) + \frac{\eta^2\sigma^2}{n} \\
&+ 4\eta^2 L^2\mathbb{E}\left(\delta_y^{(t)} + \left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*\right\|^2 + \left\|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right\|^2 + \delta_x^{(t)}\right).
\end{aligned}
$$

Adding above two inequalities up yields:

$$
\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\right\|^2\right] + \mathbb{E}\left[\left\|\boldsymbol{y}^{(t+1)} - \boldsymbol{y}^*\right\|^2\right] \leq &\left(1 - \frac{3}{4}\mu\eta\right)\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|^2\right] + \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*\right\|^2\right]\right) \\
&- 2\eta\mathbb{E}\left(F(\boldsymbol{x}^{(t)},\boldsymbol{y}^*) - F(\boldsymbol{x}^*,\boldsymbol{y}^{(t)})\right) + \frac{16}{\mu}\eta L^2\mathbb{E}(\delta_x^{(t)} + \delta_y^{(t)}) + \frac{2\eta^2\sigma^2}{n} \\
&+ 8\eta^2 L^2\left(\mathbb{E}\left[\delta_x^{(t)} + \delta_y^{(t)}\right] + \left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|^2\right] + \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*\right\|^2\right]\right)\right).
\end{aligned}
$$

Since $\eta \leq \frac{\sqrt{\mu}}{4\sqrt{2}L}$, we have $8\eta^2 L^2 \leq \frac{\mu\eta}{4}$, then we can conclude:

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^*\right\|^2\right]+\mathbb{E}\left[\left\|\boldsymbol{y}^{(t+1)}-\boldsymbol{y}^*\right\|^2\right] \leq \left(1-\frac{1}{2}\mu\eta\right)\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\|^2\right]+\mathbb{E}\left[\left\|\boldsymbol{y}^{(t)}-\boldsymbol{y}^*\right\|^2\right]\right)&\\ -2\eta\mathbb{E}\left(F(\boldsymbol{x}^{(t)},\boldsymbol{y}^*)-F(\boldsymbol{x}^*,\boldsymbol{y}^{(t)})\right)+\frac{16}{\mu}\eta L^2\mathbb{E}(\delta_x^{(t)}+\delta_y^{(t)})+\frac{2\eta^2\sigma^2}{n}&\\ +8\eta^2L^2\mathbb{E}\left(\delta_x^{(t)}+\delta_y^{(t)}\right).&\end{aligned}$$

□

The next lemma characterizes the local model deviation during the dynamics of local SGDA.

Lemma A.2. *For local-SGDA, under Theorem 4.1's assumptions, the following relation holds true for any $i, j \in [n]$:*

$$\mathbb{E}\left[\|\boldsymbol{x}_i^{(t)}-\boldsymbol{x}_j^{(t)}\|^2\right]+\mathbb{E}\left[\|\boldsymbol{y}_i^{(t)}-\boldsymbol{y}_j^{(t)}\|^2\right] \leq \tau(1+(L-\mu)\eta)^{2\tau}8\eta^2\sigma^2.$$

Proof. Let $i, j \in [n]$, and define $\varepsilon^i_{\sigma,x} = \nabla_x F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)}) - \nabla_x F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)};\xi_i^{(t)})$, $\varepsilon^i_{\sigma,y} = \nabla_y F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)}) - \nabla_y F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)};\xi_i^{(t)})$. Then according to the updating rule, we have:

$$\begin{aligned}\boldsymbol{x}_i^{(t+1)}-\boldsymbol{x}_j^{(t+1)} &= \boldsymbol{x}_i^{(t)}-\eta\nabla_x F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)};\xi_i^{(t)})-\boldsymbol{x}_j^{(t)}+\eta\nabla_x F(\boldsymbol{x}_j^{(t)},\boldsymbol{y}_j^{(t)};\xi_j^{(t)})\\ &= \boldsymbol{x}_i^{(t)}-\boldsymbol{x}_j^{(t)}-\eta\left(\nabla_x F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_x F(\boldsymbol{x}_j^{(t)},\boldsymbol{y}_j^{(t)})\right)\\ &\quad+\eta\left(\nabla_x F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_x F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)};\xi_i^{(t)})\right)+\eta\left(\nabla_x F(\boldsymbol{x}_j^{(t)},\boldsymbol{y}_j^{(t)};\xi_j^{(t)})-\nabla_x F(\boldsymbol{x}_j^{(t)},\boldsymbol{y}_j^{(t)})\right)\\ &= \boldsymbol{x}_i^{(t)}-\boldsymbol{x}_j^{(t)}-\eta\left(\nabla_x F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_x F(\boldsymbol{x}_j^{(t)},\boldsymbol{y}_i^{(t)})\right)-\eta\left(\nabla_x F(\boldsymbol{x}_j^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_x F(\boldsymbol{x}_j^{(t)},\boldsymbol{y}_j^{(t)})\right)\\ &\quad+\eta\varepsilon^i_{\sigma,x}-\eta\varepsilon^j_{\sigma,x}\\ &= (1-\eta_t\mathbf{H}_1)\left(\boldsymbol{x}_i^{(t)}-\boldsymbol{x}_j^{(t)}\right)-\eta\mathbf{H}_2\left(\boldsymbol{y}_i^{(t)}-\boldsymbol{y}_j^{(t)}\right)+\eta\varepsilon^i_{\sigma,x}-\eta\varepsilon^j_{\sigma,x},\end{aligned}$$

where we used the μ-strong-convexity and L-smoothness assumptions, that imply $\mu\mathbf{I} \preccurlyeq \mathbf{H}_1 \preccurlyeq L\mathbf{I}$ and $\mu\mathbf{I} \preccurlyeq \mathbf{H}_2 \preccurlyeq L\mathbf{I}$. We similarly continue to bound $\boldsymbol{y}_i^{(t+1)}-\boldsymbol{y}_j^{(t+1)}$:

$$\begin{aligned}\boldsymbol{y}_i^{(t+1)}-\boldsymbol{y}_j^{(t+1)} &= \boldsymbol{y}_i^{(t)}+\eta\nabla_y F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)};\xi_i^{(t)})-\boldsymbol{y}_j^{(t)}-\eta\nabla_y F(\boldsymbol{x}_j^{(t)},\boldsymbol{y}_j^{(t)};\xi_j^{(t)})\\ &= \boldsymbol{y}_i^{(t)}-\boldsymbol{y}_j^{(t)}+\eta\left(\nabla_y F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_y F(\boldsymbol{x}_j^{(t)},\boldsymbol{y}_j^{(t)})\right)\\ &\quad-\eta\left(\nabla_y F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_y F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)};\xi_i^{(t)})\right)-\eta\left(\nabla_y F(\boldsymbol{x}_j^{(t)},\boldsymbol{y}_j^{(t)};\xi_j^{(t)})-\nabla_y F(\boldsymbol{x}_j^{(t)},\boldsymbol{y}_j^{(t)})\right)\\ &= \boldsymbol{y}_i^{(t)}-\boldsymbol{y}_j^{(t)}+\eta\left(\nabla_y F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_y F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_j^{(t)})\right)+\eta\left(\nabla_y F(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_j^{(t)})-\nabla_y F(\boldsymbol{x}_j^{(t)},\boldsymbol{y}_j^{(t)})\right)\\ &\quad-\eta\varepsilon^i_{\sigma,y}+\eta\varepsilon^j_{\sigma,y}\\ &= (1-\eta\mathbf{H}_3)\left(\boldsymbol{y}_i^{(t)}-\boldsymbol{y}_j^{(t)}\right)-\eta\mathbf{H}_4\left(\boldsymbol{x}_i^{(t)}-\boldsymbol{x}_j^{(t)}\right)-\eta\varepsilon^i_{\sigma,y}+\eta\varepsilon^j_{\sigma,y},\end{aligned}$$

where $\mu\mathbf{I} \preccurlyeq \mathbf{H}_3 \preccurlyeq L\mathbf{I}$ and $\mu\mathbf{I} \preccurlyeq \mathbf{H}_4 \preccurlyeq L\mathbf{I}$.

Let $\varepsilon_x^t = \boldsymbol{x}_i^{(t)}-\boldsymbol{x}_j^{(t)}$, $\varepsilon_y^t = \boldsymbol{y}_i^{(t)}-\boldsymbol{y}_j^{(t)}$. Writing the above inequalities into compact matrix form, we have:

$$\begin{bmatrix}\varepsilon_x^{t+1}\\ \varepsilon_y^{t+1}\end{bmatrix} = \mathcal{A}^t\begin{bmatrix}\varepsilon_x^{t}\\ \varepsilon_y^{t}\end{bmatrix}+\begin{bmatrix}\eta\mathbf{I}, & 0\\ 0, & \eta\mathbf{I}\end{bmatrix}\begin{bmatrix}\varepsilon^i_{\sigma,x}-\varepsilon^j_{\sigma,x}\\ \varepsilon^j_{\sigma,y}-\varepsilon^i_{\sigma,y}\end{bmatrix}, \tag{4}$$

where:

$$\mathcal{A}^t = \begin{bmatrix}(1-\eta\mathbf{H}_1), & -\eta\mathbf{H}_2\\ -\eta\mathbf{H}_4, & (1-\eta\mathbf{H}_3)\end{bmatrix}. \tag{5}$$

Taking squared norm and expectation over (4) yields:

$$\mathbb{E}\left[\left\|\begin{bmatrix}\varepsilon_x^{t+1}\\ \varepsilon_y^{t+1}\end{bmatrix}\right\|^2\right] = \mathbb{E}\left[\left\|\mathcal{A}^t\begin{bmatrix}\varepsilon_x^{t}\\ \varepsilon_y^{t}\end{bmatrix}\right\|^2\right] + \mathbb{E}\left[\left\|\begin{bmatrix}\eta\mathbf{I}, & 0\\ 0, & \eta\mathbf{I}\end{bmatrix}\begin{bmatrix}\varepsilon_{\sigma,x}^{i} - \varepsilon_{\sigma,x}^{j}\\ \varepsilon_{\sigma,y}^{j} - \varepsilon_{\sigma,y}^{i}\end{bmatrix}\right\|^2\right]$$

$$\leq \mathbb{E}\left[\left\|\mathcal{A}^t\right\|^2\right]\mathbb{E}\left[\left\|\begin{bmatrix}\varepsilon_x^{t}\\ \varepsilon_y^{t}\end{bmatrix}\right\|^2\right] + 8\eta^2\sigma^2. \tag{6}$$

Now let us examine the upper bound of $\|\mathcal{A}^t\|^2$. According to [54] (Lemma G.1), we have:

$$\|\mathcal{A}^t\| = \left\|\begin{bmatrix}(1-\eta\mathbf{H}_1), & -\eta\mathbf{H}_2\\ -\eta\mathbf{H}_4, & (1-\eta\mathbf{H}_3)\end{bmatrix}\right\| \leq \max\{\|1-\eta\mathbf{H}_1\|, \|1-\eta\mathbf{H}_3\|\} + \max\{\|\eta\mathbf{H}_2\|, \|\eta\mathbf{H}_4\|\} = 1 + (L-\mu)\eta.$$

So $\|\mathcal{A}^t\|^2 \leq (1+(L-\mu)\eta)^2$. Letting t_0 denote the latest synchronization stage, and plugging $\|\mathcal{A}^t\|^2 \leq (1+(L-\mu)\eta)^2$ back to (6) we have:

$$\begin{aligned}\mathbb{E}\left[\left\|\begin{bmatrix}\varepsilon_x^{t+1}\\ \varepsilon_y^{t+1}\end{bmatrix}\right\|^2\right] &\leq (1+(L-\mu)\eta)^2\mathbb{E}\left[\left\|\begin{bmatrix}\varepsilon_x^{t}\\ \varepsilon_y^{t}\end{bmatrix}\right\|^2\right] + 8\eta^2\sigma^2\\ &\leq \sum_{t'=0}^{t-t_0}(1+(L-\mu)\eta)^{2t'}8\eta^2\sigma^2\\ &\leq \tau(1+(L-\mu)\eta)^{2\tau}8\eta^2\sigma^2,\end{aligned}$$

where we use the fact $\left\|\begin{bmatrix}\varepsilon_x^{t_0}\\ \varepsilon_y^{t_0}\end{bmatrix}\right\|^2 = 0$ at second inequality.

□

A.2.2 Proof of Theorem 4.1

Now we can proceed to the proof of Theorem 4.1.

Proof. According to Lemma A.1 we have:

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\right\|^2\right] + \mathbb{E}\left[\left\|\boldsymbol{y}^{(t+1)} - \boldsymbol{y}^*\right\|^2\right] &\leq \left(1 - \frac{1}{2}\mu\eta\right)\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|^2\right] + \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*\right\|^2\right]\right)\\ &\quad - 2\eta\left(\mathbb{E}\left[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*)\right] - \mathbb{E}\left[F(\boldsymbol{x}^*, \boldsymbol{y}^{(t)})\right]\right)\\ &\quad + \frac{2\eta^2\sigma^2}{n} + \frac{16\eta_t L^2}{\mu}\left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right]\right) + 8\eta^2L^2\left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right]\right).\end{aligned} \tag{7}$$

Notice that $F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) - F(\boldsymbol{x}^*, \boldsymbol{y}^{(t)}) = F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) - F(\boldsymbol{x}^*, \boldsymbol{y}^*) + F(\boldsymbol{x}^*, \boldsymbol{y}^*) - F(\boldsymbol{x}^*, \boldsymbol{y}^{(t)}) \geq 0$, we can omit this term. We plug Lemma A.2 into (7) to get:

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\right\|^2\right] + \mathbb{E}\left[\left\|\boldsymbol{y}^{(t+1)} - \boldsymbol{y}^*\right\|^2\right] &\leq \left(1 - \frac{1}{2}\mu\eta\right)\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|^2\right] + \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*\right\|^2\right]\right)\\ &\quad + \frac{2\eta^2\sigma^2}{n} + \left(\frac{16\eta L^2}{\mu} + 8\eta^2L^2\right)\left(\tau(1+(L-\mu)\eta)^{2\tau}8\eta^2\sigma^2\right).\end{aligned}$$

Unrolling the recursion yields:

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{x}^{(T)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(T)} - \boldsymbol{y}^*\right\|^2\right] &\leq \left(1 - \frac{1}{2}\mu\eta\right)^T\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(0)} - \boldsymbol{y}^*\right\|^2\right]\right)\\ &\quad + \frac{2\eta\sigma^2}{\mu n} + \left(\frac{32L^2}{\mu^2} + \frac{16\eta L^2}{\mu}\right)\left(\tau(1+(L-\mu)\eta)^{2\tau}8\eta^2\sigma^2\right).\end{aligned}$$

Plugging in $\tau = \frac{T}{n \log T}$ and $\eta = \frac{4 \log T}{\mu T}$, we have:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{x}^{(T)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(T)} - \boldsymbol{y}^*\right\|^2\right] \\
&\leq \left(1 - \frac{2\log T}{T}\right)^T \left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(0)} - \boldsymbol{y}^*\right\|^2\right]\right) \\
&\quad + \frac{8\log T \sigma^2}{\mu^2 n T} + \left(\frac{32L^2}{\mu^2} + 16\frac{4\log T}{\mu^2 T}L^2)\right)\left(\frac{T}{n\log T}\left(1 + (L-\mu)\frac{4\log T}{\mu T}\right)^{2\frac{T}{n\log T}} \frac{128\log^2 T}{\mu^2 T^2}\sigma^2\right) \\
&\leq \exp(-\log T^2)\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(0)} - \boldsymbol{y}^*\right\|^2\right]\right) \\
&\quad + \frac{8\log T \sigma^2}{\mu^2 n T} + \left(\frac{32L^2}{\mu^2} + 16\frac{4\log T}{\mu^2 T}L^2\right)\left(\frac{T}{n\log T}\left(1 + (L-\mu)\frac{4\log T}{\mu T}\right)^{2\frac{T}{n\log T}} \frac{128\log^2 T}{\mu^2 T^2}\sigma^2\right).
\end{aligned}
$$

Notice that:

$$
\left(1 + (L-\mu)\frac{4\log T}{\mu T}\right)^{\frac{2T}{n\log T}} = \left(1 + (L-\mu)\frac{4\log T}{\mu T}\right)^{\frac{\mu T}{4(L-\mu)\log T}\frac{2T}{n\log T}\frac{4(L-\mu)\log T}{\mu T}} \leq \exp\left(\frac{8(L-\mu)}{\mu n}\right).
$$

So we can conclude the proof:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{x}^{(T)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(T)} - \boldsymbol{y}^*\right\|^2\right] \\
&\leq \frac{\mathbb{E}\left[\left\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(0)} - \boldsymbol{y}^*\right\|^2\right]}{T^2} \\
&\quad + \frac{8\log T\sigma^2}{\mu^2 n T} + \left(\frac{32L^2}{\mu^2} + 16\frac{4\log T}{\mu^2 T}L^2\right)\left(\frac{T}{n\log T}\exp\left(\frac{8(L-\mu)}{\mu n}\right)\frac{128\log^2 T}{\mu^2 T^2}\sigma^2\right) \\
&\leq \tilde{O}\left(\frac{1}{T^2} + \frac{\sigma^2}{\mu^2 n T} + \frac{\kappa^2\sigma^2}{\mu^2 n T} + \frac{\kappa^2\sigma^2}{\mu^2 n T^2}\right).
\end{aligned}
$$

as stated where we used $\tilde{O}(\cdot)$ in last inequality to keep key parameters. □

A.3 Proof in heterogeneous setting

In this section we are going to present the proof in heterogeneous case. Let us introduce some technical lemmas first which will help our proof.

A.3.1 Proof of technical lemmas

The following lemma performs one iteration analysis:

Lemma A.3. *For local-SGDA, under Theorem 4.2's assumptions, the following relation holds true:*

$$
\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\right\|^2\right] + \mathbb{E}\left[\left\|\boldsymbol{y}^{(t+1)} - \boldsymbol{y}^*\right\|^2\right] &\leq \left(1 - \frac{1}{2}\mu\eta_t\right)\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|^2\right] + \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*\right\|^2\right]\right) \\
&\quad - 2\eta_t\left(F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) - F(\boldsymbol{x}^*, \boldsymbol{y}^{(t)})\right) + \frac{16}{\mu}\eta_t L^2(\delta_x^{(t)} + \delta_y^{(t)}) + \frac{2\eta_t^2\sigma^2}{n} \\
&\quad + 8\eta_t^2 L^2\left(\delta_x^{(t)} + \delta_y^{(t)}\right).
\end{aligned}
$$

Proof. According to updating rule and strong convexity:

$$
\mathbb{E}\left[\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\right\|^2\right] = \mathbb{E}\left[\left\|\boldsymbol{x}^{(t)} - \eta_t \frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)}; \xi_i^{(t)}) - \boldsymbol{x}^*\right\|^2\right]
$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\|^2\right]-2\eta_t\left\langle\frac{1}{n}\sum_{i=1}^n\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)}),\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\rangle\\
&+\frac{\eta_t^2\sigma^2}{n}+\eta_t^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right]\\
&\leq \mathbb{E}\left[\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\|^2\right]-2\eta_t\left\langle\nabla_x F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)}),\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\rangle\\
&-2\eta_t\left\langle\frac{1}{n}\sum_{i=1}^n\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_x F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)}),\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\rangle\\
&+\frac{\eta_t^2\sigma^2}{n}+\eta_t^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right]\\
&\leq (1-\mu\eta_t)\,\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\|^2\right]-2\eta_t\left(F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})-F(\boldsymbol{x}^*,\boldsymbol{y}^{(t)})\right)\\
&+\eta_t\mathbb{E}\left(\frac{4}{\mu}\left\|\frac{1}{n}\sum_{i=1}^n\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_x F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2+\frac{\mu}{4}\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\|^2\right)\\
&+\frac{\eta_t^2\sigma^2}{n}+\eta_t^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right].
\end{aligned}
$$

Now we are going to bound terms $\left\|\frac{1}{n}\sum_{i=1}^n\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_x F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2$ and $\left\|\frac{1}{n}\sum_{i=1}^n\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2$.

By applying Jensen's inequality on $\left\|\frac{1}{n}\sum_{i=1}^n\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_x F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2$ we have:

$$
\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^n\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_x F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2=\frac{1}{n}\sum_{i=1}^n\left\|\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_x f_i(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2\\
&\leq\frac{1}{n}\sum_{i=1}^n\left(2\left\|\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_x f_i(\boldsymbol{x}^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2+2\left\|\nabla_x f_i(\boldsymbol{x}^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_x f_i(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2\right)\\
&\leq\frac{1}{n}\sum_{i=1}^n\left(2L^2\left\|\boldsymbol{x}_i^{(t)}-\boldsymbol{x}^{(t)}\right\|^2+2L^2\left\|\boldsymbol{y}_i^{(t)}-\boldsymbol{y}^{(t)}\right\|^2\right)\\
&\leq 2L^2(\delta_x^{(t)}+\delta_y^{(t)}),
\end{aligned}
$$

where we use the smoothness in the second last inequality.

Then we switch to bound $\left\|\frac{1}{n}\sum_{i=1}^n\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2$:

$$
\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^n\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2&=\frac{1}{n}\sum_{i=1}^n\left\|\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\\
&=2\left\|\frac{1}{n}\sum_{i=1}^n\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})-\nabla_x F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2+2\left\|\nabla_x F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})-\nabla_x F(\boldsymbol{x}^*,\boldsymbol{y}^*)\right\|^2\\
&\leq L^2\frac{1}{n}\sum_{i=1}^n 4\left(\left\|\boldsymbol{x}_i^{(t)}-\boldsymbol{x}^{(t)}\right\|^2+\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\|^2+\left\|\boldsymbol{y}_i^{(t)}-\boldsymbol{y}^{(t)}\right\|^2+\left\|\boldsymbol{y}^{(t)}-\boldsymbol{y}^*\right\|^2\right).
\end{aligned}
$$

Putting these pieces together yields:

$$
\mathbb{E}\left[\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^*\right\|^2\right]\leq\left(1-\frac{3}{4}\mu\eta_t\right)\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\|^2\right]-2\eta_t\left(\mathbb{E}\left[F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})-F(\boldsymbol{x}^*,\boldsymbol{y}^{(t)})\right]\right)
$$

$$
\begin{aligned}
&+ \frac{8}{\mu}\eta_t L^2 \mathbb{E}(\delta_x^{(t)} + \delta_y^{(t)}) + \frac{\eta_t^2\sigma^2}{n} \\
&+ 4\eta_t^2 L^2 \mathbb{E}\left(\delta_x^{(t)} + \left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^* - \boldsymbol{y}^{(t)}\right\|^2 + \delta_y^{(t)}\right).
\end{aligned}
$$

Similarly, we can get:

$$
\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{y}^{(t+1)} - \boldsymbol{y}^*\right\|^2\right] \le \left(1 - \frac{3}{4}\mu\eta_t\right)\mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*\right\|^2\right] &- 2\eta_t \mathbb{E}\left(F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) - F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\right) \\
&+ \frac{8}{\mu}\eta_t L^2 \mathbb{E}(\delta_x^{(t)} + \delta_y^{(t)}) + \frac{\eta_t^2\sigma^2}{n} \\
&+ 4\eta_t^2 L^2 \mathbb{E}\left(\delta_y^{(t)} + \left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*\right\|^2 + \left\|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right\|^2 + \delta_x^{(t)}\right).
\end{aligned}
$$

Combining the above two inequalities yields:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\right\|^2\right] + \mathbb{E}\left[\left\|\boldsymbol{y}^{(t+1)} - \boldsymbol{y}^*\right\|^2\right] \\
&\le \left(1 - \frac{3}{4}\mu\eta_t\right)\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|^2\right] + \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*\right\|^2\right]\right) \\
&\quad - 2\eta_t \mathbb{E}\left(F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) - F(\boldsymbol{x}^*, \boldsymbol{y}^{(t)})\right) + \frac{16}{\mu}\eta_t L^2(\delta_x^{(t)} + \delta_y^{(t)}) + \frac{2\eta_t^2\sigma^2}{n} \\
&\quad + 8\eta_t^2 L^2\left(\mathbb{E}\left[\delta_x^{(t)} + \delta_y^{(t)}\right] + \left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|^2\right] + \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*\right\|^2\right]\right)\right).
\end{aligned}
$$

Since $\eta_t = \frac{8}{\mu(t+a)}$ and $a = \max\{2048\kappa^2\tau, 1024\sqrt{2}\tau\kappa^2, 256\kappa^2\}$, so we have $8\eta_t^2 L^2 \le \frac{\mu\eta_t}{4}$, then we can conclude:

$$
\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\right\|^2\right] + \mathbb{E}\left[\left\|\boldsymbol{y}^{(t+1)} - \boldsymbol{y}^*\right\|^2\right] \le \left(1 - \frac{1}{2}\mu\eta_t\right)&\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|^2\right] + \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*\right\|^2\right]\right) \\
&- 2\eta_t\left(\mathbb{E}\left[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) - F(\boldsymbol{x}^*, \boldsymbol{y}^{(t)})\right]\right) + \frac{16}{\mu}\eta_t L^2(\mathbb{E}\left[\delta_x^{(t)} + \delta_y^{(t)}\right]) + \frac{2\eta_t^2\sigma^2}{n} \\
&+ 8\eta_t^2 L^2\left(\mathbb{E}\left[\delta_x^{(t)} + \delta_y^{(t)}\right]\right).
\end{aligned}
$$

□

The next lemma upper bounds the weighted accumulative local model deviations between two communication rounds in strongly convex setting under heterogeneous data assumption.

Lemma A.4. *For local-SGDA, under Theorem 4.2's assumption, by letting $w_t = (t+a)^2$, the following inequality holds:*

$$
\begin{aligned}
\sum_{t=s\tau}^{(s+1)\tau} w_t(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right]) \le \frac{\mu}{64L^2}\sum_{j=s\tau}^{(s+1)\tau}\mu\eta_j\frac{w_j}{\eta_j}&\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(j)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(j)} - \boldsymbol{y}^*\right\|^2\right]\right) \\
&+ 64\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^2\left(\Delta_x + \Delta_y\right) + 32\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^2\sigma^2.
\end{aligned}
$$

where $\delta_{\boldsymbol{x}}^{(t)} = \frac{1}{n}\sum_{i=1}^n \left\|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}^{(t)}\right\|^2, \quad \delta_{\boldsymbol{y}}^{(t)} = \frac{1}{n}\sum_{i=1}^n \left\|\boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2.$

Proof. Assume that $s\tau \leq t \leq (s+1)\tau$. According to the updating rule, we have:

$$\begin{aligned}
\delta_{\boldsymbol{x}}^{(t)} &= \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i^{(t)}-\boldsymbol{x}^{(t)}\right\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}^{(s\tau)}-\sum_{j=s\tau}^{t}\eta_j\nabla_x f_i\left(\boldsymbol{x}_i^{(j)},\boldsymbol{y}_i^{(j)};\xi_i^{(j)}\right)-\left(\boldsymbol{x}^{(s\tau)}-\frac{1}{n}\sum_{k=1}^{n}\sum_{j=s\tau}^{t}\eta_j\nabla_x f_k\left(\boldsymbol{x}_i^{(k)},\boldsymbol{y}_i^{(k)};\xi_i^{(k)}\right)\right)\right\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{j=s\tau}^{t-1}\eta_j\nabla_x f_i\left(\boldsymbol{x}_i^{(j)},\boldsymbol{y}_i^{(j)};\xi_i^{(j)}\right)-\frac{1}{n}\sum_{k=1}^{n}\sum_{j=s\tau}^{t-1}\eta_j\nabla_x f_k\left(\boldsymbol{x}_i^{(k)},\boldsymbol{y}_i^{(k)};\xi_i^{(k)}\right)\right\|^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{j=s\tau}^{t-1}\eta_j\nabla_x f_i\left(\boldsymbol{x}_i^{(j)},\boldsymbol{y}_i^{(j)};\xi_i^{(j)}\right)\right\|^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\tau\sum_{j=s\tau}^{(s+1)\tau}\eta_j^2\left(2\left\|\nabla_x f_i\left(\boldsymbol{x}_i^{(j)},\boldsymbol{y}_i^{(j)}\right)\right\|^2+2\sigma^2\right).
\end{aligned}$$

By applying Jensen's inequality to $\left\|\nabla_x f_i\left(\boldsymbol{x}_i^{(j)},\boldsymbol{y}_i^{(j)}\right)\right\|^2$:

$$\begin{aligned}
&\left\|\nabla_x f_i\left(\boldsymbol{x}_i^{(j)},\boldsymbol{y}_i^{(j)}\right)\right\|^2 \leq 4\left\|\nabla_x f_i\left(\boldsymbol{x}_i^{(j)},\boldsymbol{y}_i^{(j)}\right)-\nabla_x f_i\left(\boldsymbol{x}^{(j)},\boldsymbol{y}^{(j)}\right)\right\|^2+4\left\|\nabla_x f_i\left(\boldsymbol{x}^{(j)},\boldsymbol{y}^{(j)}\right)-\nabla_x f_i\left(\boldsymbol{x}^{*},\boldsymbol{y}^{(j)}\right)\right\|^2 \\
&+4\left\|\nabla_x f_i\left(\boldsymbol{x}^{*},\boldsymbol{y}^{(j)}\right)-\nabla_x f_i\left(\boldsymbol{x}^{*},\boldsymbol{y}^{*}\right)\right\|^2+4\left\|\nabla_x f_i\left(\boldsymbol{x}^{*},\boldsymbol{y}^{*}\right)\right\|^2 \\
&\leq 8L^2\left(\left\|\boldsymbol{x}_i^{(j)}-\boldsymbol{x}^{(j)}\right\|^2+\left\|\boldsymbol{y}_i^{(j)}-\boldsymbol{y}^{(j)}\right\|^2\right)+4L^2\left\|\boldsymbol{x}^{(j)}-\boldsymbol{x}^{*}\right\|^2 \\
&+4L^2\left\|\boldsymbol{y}^{(j)}-\boldsymbol{y}^{*}\right\|^2+4\left\|\nabla_x f_i\left(\boldsymbol{x}^{*},\boldsymbol{y}^{*}\right)\right\|^2.
\end{aligned}$$

Plugging back and taking expectation yields:

$$\begin{aligned}
\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)}\right] &\leq \frac{1}{n}\sum_{i=1}^{n}\tau\sum_{j=s\tau}^{(s+1)\tau} \\
&\quad\times\eta_j^2\left(16L^2\left(\mathbb{E}\left[\left\|\boldsymbol{x}_i^{(j)}-\boldsymbol{x}^{(j)}\right\|^2+\left\|\boldsymbol{y}_i^{(j)}-\boldsymbol{y}^{(j)}\right\|^2\right]\right)+\mathbb{E}\left[8L^2\left\|\boldsymbol{x}^{(j)}-\boldsymbol{x}^{*}\right\|^2+8L^2\left\|\boldsymbol{y}^{(j)}-\boldsymbol{y}^{*}\right\|^2\right]\right) \\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\tau\sum_{j=s\tau}^{(s+1)\tau}\left(\eta_j^2 8\left\|\nabla_x f_i\left(\boldsymbol{x}^{*},\boldsymbol{y}^{*}\right)\right\|^2+2\sigma^2\right) \\
&\leq \tau\sum_{j=s\tau}^{(s+1)\tau}\eta_j^2\left(16L^2\left(\delta_{\boldsymbol{x}}^{(j)}+\delta_{\boldsymbol{y}}^{(j)}\right)+8L^2\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(j)}-\boldsymbol{x}^{*}\right\|^2+\left\|\boldsymbol{y}^{(j)}-\boldsymbol{y}^{*}\right\|^2\right]\right)\right) \\
&\quad+8\tau\sum_{j=s\tau}^{(s+1)\tau}\eta_j^2\Delta_x+2\tau\sum_{j=s\tau}^{(s+1)\tau}\eta_j^2\sigma^2.
\end{aligned}$$

Then multiplying w_t on both sides and summing from $t=s\tau$ to $(s+1)\tau$ yields:

$$\begin{aligned}
\sum_{t=s\tau}^{(s+1)\tau}w_t\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)}\right] &\leq \sum_{j=s\tau}^{(s+1)\tau}w_t\tau\sum_{j=s\tau}^{(s+1)\tau}\eta_j^2\left(16L^2\left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(j)}+\delta_{\boldsymbol{y}}^{(j)}\right]\right)+8L^2\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(j)}-\boldsymbol{x}^{*}\right\|^2+\left\|\boldsymbol{y}^{(j)}-\boldsymbol{y}^{*}\right\|^2\right]\right)\right) \\
&\quad+8\sum_{t=s\tau}^{(s+1)\tau}w_t\tau\sum_{j=s\tau}^{(s+1)\tau}\eta_j^2\Delta_x+2\sum_{t=s\tau}^{(s+1)\tau}w_t\tau\sum_{j=s\tau}^{(s+1)\tau}\eta_j^2\sigma^2.
\end{aligned}$$

Notice that $w_t = (t+a)^2$ and $a \geq \tau$, so $w_t < w_{(s+1)\tau} \leq 4w_j$, $\forall t, j$ such that $s\tau \leq t, j \leq (s+1)\tau$. So we have:

$$\sum_{t=s\tau}^{(s+1)\tau} w_t \mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)}\right] \leq \tau^2 \sum_{j=s\tau}^{(s+1)\tau} 4w_j\eta_j^2 \left(16L^2\left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(j)} + \delta_{\boldsymbol{y}}^{(j)}\right]\right) + 8L^2\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(j)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(j)} - \boldsymbol{y}^*\right\|^2\right]\right)\right)$$
$$+ 32\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^2\Delta_x + 8\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^2\sigma^2.$$

Since $\eta_t = \frac{8}{\mu(t+a)}$ and $a = \max\{2048\kappa^2\tau, 1024\sqrt{2}\tau\kappa^2, 256\kappa^2\}$, we have the following facts:

$$\eta_t < \eta_{s\tau} \leq 2\eta_j, \quad \forall t, j \text{ such that } s\tau \leq t, j \leq (s+1)\tau,$$
$$256\eta_t^2\tau^2L^2 \leq \frac{1}{4},$$
$$128\eta_t^2\tau^2L^2 \leq \frac{\mu^2}{256L^2}.$$

Hence:

$$\sum_{t=s\tau}^{(s+1)\tau} w_t \mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)}\right] \leq 4\eta_t^2\tau^2 \sum_{j=s\tau}^{(s+1)\tau} 4w_j \left(16L^2\left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(j)} + \delta_{\boldsymbol{y}}^{(j)}\right]\right) + 8L^2\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(j)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(j)} - \boldsymbol{y}^*\right\|^2\right]\right)\right)$$
$$+ 32\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^2\Delta_x + 8\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^2\sigma^2.$$
$$\leq \frac{1}{4}\sum_{j=s\tau}^{(s+1)\tau} w_j \left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(j)} + \delta_{\boldsymbol{y}}^{(j)}\right]\right) + 128\eta_t^2\tau^2L^2 \sum_{j=s\tau}^{(s+1)\tau} w_j \left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(j)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(j)} - \boldsymbol{y}^*\right\|^2\right]\right)$$
$$+ 32\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^2\Delta_x + 8\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^2\sigma^2$$
$$\leq \frac{1}{4}\sum_{j=s\tau}^{(s+1)\tau} w_j \left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(j)} + \delta_{\boldsymbol{y}}^{(j)}\right]\right) + \frac{\mu^2}{256L^2} \sum_{j=s\tau}^{(s+1)\tau} w_j \left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(j)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(j)} - \boldsymbol{y}^*\right\|^2\right]\right)$$
$$+ 32\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^2\Delta_x + 8\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^2\sigma^2$$
$$\leq \frac{1}{4}\sum_{j=s\tau}^{(s+1)\tau} w_j \left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(j)} + \delta_{\boldsymbol{y}}^{(j)}\right]\right) + \frac{\mu}{256L^2} \sum_{j=s\tau}^{(s+1)\tau} \mu\eta_j \frac{w_j}{\eta_j} \left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(j)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(j)} - \boldsymbol{y}^*\right\|^2\right]\right)$$
$$+ 32\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^2\Delta_x + 8\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^2\sigma^2.$$

Similarly, we get:

$$\sum_{t=s\tau}^{(s+1)\tau} w_t \mathbb{E}[\delta_{\boldsymbol{y}}^{(t)}] \leq \frac{1}{4}\sum_{j=s\tau}^{(s+1)\tau} w_j \left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(j)} + \delta_{\boldsymbol{y}}^{(j)}\right]\right) + \frac{\mu}{256L^2} \sum_{j=s\tau}^{(s+1)\tau} \mu\eta_j \frac{w_j}{\eta_j} \left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(j)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(j)} - \boldsymbol{y}^*\right\|^2\right]\right)$$
$$+ 32\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^2\Delta_y + 8\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^2\sigma^2.$$

Adding the two inequalities up gives:

$$\sum_{t=s\tau}^{(s+1)\tau} w_t \left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right]\right) \leq \frac{1}{2}\sum_{j=s\tau}^{(s+1)\tau} w_j \left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(j)} + \delta_{\boldsymbol{y}}^{(j)}\right]\right) + \frac{\mu}{128L^2} \sum_{j=s\tau}^{(s+1)\tau} \mu\eta_j \frac{w_j}{\eta_j} \left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(j)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(j)} - \boldsymbol{y}^*\right\|^2\right]\right)$$

$$+ 32\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j \eta_j^2 (\Delta_x + \Delta_y) + 16\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j \eta_j^2 \sigma^2$$

$$\iff \frac{1}{2} \sum_{t=s\tau}^{(s+1)\tau} w_t \left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)} \right] \right) \leq \frac{\mu}{128L^2} \sum_{j=s\tau}^{(s+1)\tau} \mu\eta_j \frac{w_j}{\eta_j} \left(\mathbb{E}\left[\left\| \boldsymbol{x}^{(j)} - \boldsymbol{x}^* \right\|^2 + \left\| \boldsymbol{y}^{(j)} - \boldsymbol{y}^* \right\|^2 \right] \right)$$

$$+ 32\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j \eta_j^2 (\Delta_x + \Delta_y) + 16\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j \eta_j^2 \sigma^2$$

$$\iff \sum_{t=s\tau}^{(s+1)\tau} w_t \left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)} \right] \right) \leq \frac{\mu}{64L^2} \sum_{j=s\tau}^{(s+1)\tau} \mu\eta_j \frac{w_j}{\eta_j} \left(\mathbb{E}\left[\left\| \boldsymbol{x}^{(j)} - \boldsymbol{x}^* \right\|^2 + \left\| \boldsymbol{y}^{(j)} - \boldsymbol{y}^* \right\|^2 \right] \right)$$

$$+ 64\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j \eta_j^2 (\Delta_x + \Delta_y) + 32\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j \eta_j^2 \sigma^2.$$

□

The following lemma also gives the upper bound for weighted local model deviations, but the weights multiplied in front of $\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right]$ is different from Lemma A.4.

Lemma A.5. *For local-SGDA, under Theorem 4.2's assumption, by letting $w_t = (t+a)^2$, the following holds:*

$$\sum_{t=s\tau}^{(s+1)\tau} w_t \eta_t \left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)} \right] \right) \leq \frac{1}{64L^2} \sum_{j=s\tau}^{(s+1)\tau} \mu\eta_j \frac{w_j}{\eta_j} \left(\mathbb{E}\left[\left\| \boldsymbol{x}^{(j)} - \boldsymbol{x}^* \right\|^2 + \left\| \boldsymbol{y}^{(j)} - \boldsymbol{y}^* \right\|^2 \right] \right)$$

$$+ 128\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j \eta_j^3 (\Delta_x + \Delta_y) + 64\tau^2 \sum_{j=s\tau}^{(s+1)\tau} w_j \eta_j^3 \sigma^2.$$

Proof. According to Lemma A.4, we have:

$$\sum_{t=s\tau}^{(s+1)\tau} w_t \eta_t \mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} \right] \leq \sum_{t=s\tau}^{(s+1)\tau} w_t \tau \eta_t \sum_{j=s\tau}^{(s+1)\tau} \eta_j^2 \left(16L^2 \left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(j)} + \delta_{\boldsymbol{y}}^{(j)} \right] \right) + 8L^2 \left(\mathbb{E}\left[\left\| \boldsymbol{x}^{(j)} - \boldsymbol{x}^* \right\|^2 + \left\| \boldsymbol{y}^{(j)} - \boldsymbol{y}^* \right\|^2 \right] \right) \right)$$

$$+ 8 \sum_{t=s\tau}^{(s+1)\tau} w_t \tau \sum_{j=s\tau}^{(s+1)\tau} \eta_j^2 \Delta_x + 2 \sum_{t=s\tau}^{(s+1)\tau} w_t \tau \sum_{j=s\tau}^{(s+1)\tau} \eta_j^2 \sigma^2. \tag{8}$$

Notice that $w_t = (t+a)^2$ and $a \geq \tau$, so $w_t < w_{(s+1)\tau} \leq 4w_j$, $\forall t, j$ such that $s\tau \leq t, j \leq (s+1)\tau$. So we have:

$$\sum_{t=s\tau}^{(s+1)\tau} w_t \eta_t \mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} \right] \leq \tau^2 \eta_t \sum_{j=s\tau}^{(s+1)\tau} 4 w_j \eta_j^2 \left(16L^2 \left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(j)} + \delta_{\boldsymbol{y}}^{(j)} \right] \right) + 8L^2 \left(\mathbb{E}\left[\left\| \boldsymbol{x}^{(j)} - \boldsymbol{x}^* \right\|^2 + \left\| \boldsymbol{y}^{(j)} - \boldsymbol{y}^* \right\|^2 \right] \right) \right)$$

$$+ 32\tau^2 \eta_t \sum_{j=s\tau}^{(s+1)\tau} w_j \eta_j^2 \Delta_x + 8\tau^2 \eta_t \sum_{j=s\tau}^{(s+1)\tau} w_j \eta_j^2 \sigma^2.$$

$$\leq \tau^2 \sum_{j=s\tau}^{(s+1)\tau} 4 w_j \eta_j^2 \left(16L^2 \left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(j)} + \delta_{\boldsymbol{y}}^{(j)} \right] \right) + 8L^2 \left(\mathbb{E}\left[\left\| \boldsymbol{x}^{(j)} - \boldsymbol{x}^* \right\|^2 + \left\| \boldsymbol{y}^{(j)} - \boldsymbol{y}^* \right\|^2 \right] \right) \right) \tag{9}$$

$$+ 32\tau^2 \eta_t \sum_{j=s\tau}^{(s+1)\tau} w_j \eta_j^2 \Delta_x + 8\tau^2 \eta_t \sum_{j=s\tau}^{(s+1)\tau} w_j \eta_j^2 \sigma^2,$$

where we omit a η_t in (9) since $\eta_t \leq 1$.

Since $\eta_t = \frac{8}{\mu(t+a)}$ and $a = \max\{2048\kappa^2\tau, 1024\sqrt{2}\tau\kappa^2, 256\kappa^2\}$, we have the following facts:

$$\eta_t < \eta_{s\tau} \leq 2\eta_j, \quad \forall t, j \text{ such that } s\tau \leq t, j \leq (s+1)\tau,$$
$$256\eta_t^2\tau^2 \leq \frac{1}{4},$$
$$128\eta_t^2\tau^2L^2 \leq \frac{\mu}{256L^2}.$$

Hence:

$$\begin{aligned}
\sum_{t=s\tau}^{(s+1)\tau} w_t\eta_t\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)}\right] &\leq 4\eta_t^2\tau^2\sum_{j=s\tau}^{(s+1)\tau} 4w_j\left(16L^2\left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(j)}+\delta_{\boldsymbol{y}}^{(j)}\right]\right)+8L^2\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(j)}-\boldsymbol{x}^*\right\|^2+\left\|\boldsymbol{y}^{(j)}-\boldsymbol{y}^*\right\|^2\right]\right)\right) \\
&\quad + 64\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^3\Delta_x + 16\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^3\sigma^2. \\
&\leq \frac{1}{4}\sum_{j=s\tau}^{(s+1)\tau} w_j\left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(j)}+\delta_{\boldsymbol{y}}^{(j)}\right]\right)+128\eta_t^2\tau^2L^2\sum_{j=s\tau}^{(s+1)\tau} w_j\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(j)}-\boldsymbol{x}^*\right\|^2+\left\|\boldsymbol{y}^{(j)}-\boldsymbol{y}^*\right\|^2\right]\right) \\
&\quad + 64\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^3\Delta_x + 16\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^2\sigma^2 \\
&\leq \frac{1}{4}\sum_{j=s\tau}^{(s+1)\tau} w_j\left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(j)}+\delta_{\boldsymbol{y}}^{(j)}\right]\right)+\frac{\mu}{256L^2}\sum_{j=s\tau}^{(s+1)\tau} w_j\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(j)}-\boldsymbol{x}^*\right\|^2+\left\|\boldsymbol{y}^{(j)}-\boldsymbol{y}^*\right\|^2\right]\right) \\
&\quad + 64\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^3\Delta_x + 16\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^3\sigma^2 \\
&\leq \frac{1}{4}\sum_{j=s\tau}^{(s+1)\tau} w_j\left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(j)}+\delta_{\boldsymbol{y}}^{(j)}\right]\right)+\frac{1}{256L^2}\sum_{j=s\tau}^{(s+1)\tau} \mu\eta_j\frac{w_j}{\eta_j}\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(j)}-\boldsymbol{x}^*\right\|^2+\left\|\boldsymbol{y}^{(j)}-\boldsymbol{y}^*\right\|^2\right]\right) \\
&\quad + 64\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^3\Delta_x + 16\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^3\sigma^2.
\end{aligned}$$

Similarly, we get:

$$\begin{aligned}
\sum_{t=s\tau}^{(s+1)\tau} w_t\eta_t\mathbb{E}\left[\delta_{\boldsymbol{y}}^{(t)}\right] &\leq \frac{1}{4}\sum_{j=s\tau}^{(s+1)\tau} w_j\left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(j)}+\delta_{\boldsymbol{y}}^{(j)}\right]\right)+\frac{1}{256L^2}\sum_{j=s\tau}^{(s+1)\tau} \mu\eta_j\frac{w_j}{\eta_j}\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(j)}-\boldsymbol{x}^*\right\|^2+\left\|\boldsymbol{y}^{(j)}-\boldsymbol{y}^*\right\|^2\right]\right) \\
&\quad + 64\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^3\Delta_y + 16\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^3\sigma^2.
\end{aligned}$$

Combining the two inequalities yields:

$$\begin{aligned}
\sum_{t=s\tau}^{(s+1)\tau} w_t\eta_t(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)}\right]) &\leq \frac{1}{2}\sum_{j=s\tau}^{(s+1)\tau} w_j\left(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(j)}+\delta_{\boldsymbol{y}}^{(j)}\right]\right)+\frac{1}{128L^2}\sum_{j=s\tau}^{(s+1)\tau} \mu\eta_j\frac{w_j}{\eta_j}\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(j)}-\boldsymbol{x}^*\right\|^2+\left\|\boldsymbol{y}^{(j)}-\boldsymbol{y}^*\right\|^2\right]\right) \\
&\quad + 64\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^3\left(\Delta_x+\Delta_y\right) + 32\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^3\sigma^2 \\
\Longleftrightarrow \frac{1}{2}\sum_{t=s\tau}^{(s+1)\tau} w_t\eta_t(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)}\right]) &\leq \frac{1}{128L^2}\sum_{j=s\tau}^{(s+1)\tau} \mu\eta_j\frac{w_j}{\eta_j}\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(j)}-\boldsymbol{x}^*\right\|^2+\left\|\boldsymbol{y}^{(j)}-\boldsymbol{y}^*\right\|^2\right]\right) \\
&\quad + 64\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^3\left(\Delta_x+\Delta_y\right) + 32\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^3\sigma^2
\end{aligned}$$

$$\iff \sum_{t=s\tau}^{(s+1)\tau} w_t\eta_t(\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)}\right]) \leq \frac{1}{64L^2}\sum_{j=s\tau}^{(s+1)\tau}\mu\eta_j\frac{w_j}{\eta_j}\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(j)}-\boldsymbol{x}^*\right\|^2+\left\|\boldsymbol{y}^{(j)}-\boldsymbol{y}^*\right\|^2\right]\right)$$
$$+128\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^3\left(\Delta_x+\Delta_y\right)+64\tau^2\sum_{j=s\tau}^{(s+1)\tau} w_j\eta_j^3\sigma^2.$$

□

A.3.2 Proof of Theorem 4.2

Now we are going to proof Theorem 4.2.

Proof. According to Lemma A.3 we have:

$$\mathbb{E}\left[\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^*\right\|^2\right]+\mathbb{E}\left[\left\|\boldsymbol{y}^{(t+1)}-\boldsymbol{y}^*\right\|^2\right] \leq \left(1-\frac{1}{2}\mu\eta_t\right)\left(\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\|^2\right]+\mathbb{E}\left[\left\|\boldsymbol{y}^{(t)}-\boldsymbol{y}^*\right\|^2\right]\right)$$
$$-2\eta_t\mathbb{E}\left(F(\boldsymbol{x}^{(t)},\boldsymbol{y}^*)-F(\boldsymbol{x}^*,\boldsymbol{y}^{(t)})\right)+\frac{16}{\mu}\eta_t L^2\mathbb{E}(\delta_x^{(t)}+\delta_y^{(t)})+\frac{2\eta_t^2\sigma^2}{n}$$
$$+8\eta_t^2L^2\mathbb{E}\left(\delta_x^{(t)}+\delta_y^{(t)}\right).$$

Then, letting $w_t=(t+a)^2$ and multiplying $\frac{w_t}{\eta_t}$ on both sides, and summing up from $t=1$ to T:

$$\begin{aligned}
&\sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau}\frac{w_t}{\eta_t}\mathbb{E}\left(\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^*\right\|^2+\left\|\boldsymbol{y}^{(t+1)}-\boldsymbol{y}^*\right\|^2\right)\\
&\leq \sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau}\left(1-\frac{1}{2}\mu\eta_t\right)\frac{w_t}{\eta_t}\mathbb{E}\left(\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\|^2+\left\|\boldsymbol{y}^{(t)}-\boldsymbol{y}^*\right\|^2\right)\\
&\quad -2\sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau} w_t\mathbb{E}\left(F(\boldsymbol{x}^{(t)},\boldsymbol{y}^*)-F(\boldsymbol{x}^*,\boldsymbol{y}^{(t)})\right)+\sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau}\frac{2w_t\eta_t\sigma^2}{n}\\
&\quad +\underbrace{\frac{16L^2}{\mu}\sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau} w_t\mathbb{E}\left(\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)}\right)}_{T_1}+\underbrace{8L^2\sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau} w_t\eta_t\mathbb{E}\left(\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)}\right)}_{T_2}.
\end{aligned} \tag{10}$$

Then we use Lemmas A.4 and A.5 in T_1 and T_2 to get:

$$T_1=\sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau}\frac{1}{8}\mu\eta_t\frac{w_t}{\eta_t}\mathbb{E}\left(\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\|^2+\left\|\boldsymbol{y}^{(t)}-\boldsymbol{y}^*\right\|^2\right)$$
$$+\frac{1024\tau^2L^2}{\mu}\sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau} w_t\eta_t^2\left(\Delta_x+\Delta_y\right)+\frac{512\tau^2L^2}{\mu}\sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau} w_t\eta_t^2\sigma^2$$

$$T_2=\sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau}\frac{1}{8}\mu\eta_t\frac{w_t}{\eta_t}\mathbb{E}\left(\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\|^2+\left\|\boldsymbol{y}^{(t)}-\boldsymbol{y}^*\right\|^2\right)$$
$$+1024L^2\tau^2\sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau} w_t\eta_t^3\left(\Delta_x+\Delta_y\right)+512L^2\tau^2\sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau} w_t\eta_t^3\sigma^2.$$

Plugging T_1 and T_2 back into (10) yields:

$$\sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau}\frac{w_t}{\eta_t}\mathbb{E}\left(\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^*\right\|^2+\left\|\boldsymbol{y}^{(t+1)}-\boldsymbol{y}^*\right\|^2\right)\leq\sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau}\left(1-\frac{1}{4}\mu\eta_t\right)\frac{w_t}{\eta_t}\mathbb{E}\left(\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\right\|^2+\left\|\boldsymbol{y}^{(t)}-\boldsymbol{y}^*\right\|^2\right)$$

$$- 2\sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau} w_t \mathbb{E}\left(F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) - F(\boldsymbol{x}^*, \boldsymbol{y}^{(t)})\right) + \sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau} \frac{2w_t\eta_t\sigma^2}{n}$$
$$+ \left(\frac{1024\tau^2 L^2}{\mu} + 1024L^2\tau^2\right)(\Delta_x + \Delta_y)\sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau} w_t\left(\eta_t^2 + \eta_t^3\right)$$
$$+ \left(\frac{512\tau^2 L^2}{\mu} + 512L^2\tau^2\right)\sigma^2\sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau} w_t\left(\eta_t^2 + \eta_t^3\right).$$

Using the fact that $\left(1 - \frac{1}{4}\mu\eta_t\right)\frac{w_t}{\eta_t} \leq \frac{w_{t-1}}{\eta_{t-1}}$, we can cancel up the terms:

$$\frac{w_T}{\eta_T}\mathbb{E}\left(\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(T+1)} - \boldsymbol{y}^*\right\|^2\right)$$
$$\leq \frac{w_0}{\eta_0}\left(\left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(1)} - \boldsymbol{y}^*\right\|^2\right)$$
$$+ \left(\frac{1024\tau^2 L^2}{\mu} + 1023L^2\tau^2\right)(\Delta_x + \Delta_y)\sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau} w_t\left(\eta_t^2 + \eta_t^3\right)$$
$$+ \left(\frac{512\tau^2 L^2}{\mu} + 512L^2\tau^2\right)\sigma^2\sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau} w_t\left(\eta_t^2 + \eta_t^3\right) + \sum_{s=0}^{S-1}\sum_{t=s\tau}^{(s+1)\tau}\frac{2w_t\eta_t\sigma^2}{n}.$$

Dividing both side by $\frac{w_T}{\eta_T}$ yields:

$$\mathbb{E}\left[\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(T+1)} - \boldsymbol{y}^*\right\|^2\right]$$
$$\leq \frac{8}{\mu(T+a)^3}\frac{w_0}{\eta_0}\left(\left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}^{(1)} - \boldsymbol{y}^*\right\|^2\right)$$
$$+ \frac{8}{\mu(T+a)^3}\left(\frac{1024\tau^2 L^2}{\mu} + 1024L^2\tau^2\right)(\Delta_x + \Delta_y)\left(\frac{64T}{\mu^2} + \frac{\Theta(\ln T)}{\mu^3}\right)$$
$$+ \frac{8}{\mu(T+a)^3}\left(\frac{512\tau^2 L^2}{\mu} + 512L^2\tau^2\right)\sigma^2\left(\frac{64T}{\mu^2} + \frac{\Theta(\ln T)}{\mu^3}\right) + \frac{8}{\mu(T+a)^2}\frac{16T\sigma^2}{\mu n}$$
$$\leq O\left(\frac{a^3}{T^3}\right) + O\left(\frac{\kappa^2\tau^2(\Delta_x + \Delta_y)}{\mu T^2}\right) + O\left(\frac{\kappa^2\tau^2\sigma^2}{\mu T^2}\right) + O\left(\frac{\sigma^2}{\mu^2 n T}\right).$$

Plugging in $\tau = \sqrt{T/n}$ concludes the proof. □

B Proof of Nonconvex-Strongly-Concave Case

B.1 Overview of proofs

Now we proceed to the proof of convergence rate in nonconvex-strongly-concave setting. Recall that in this case we study the envelope function $\Phi(\cdot)$ and $\boldsymbol{y}^*(\cdot)$. The following proposition establishes the smoothness property of these auxiliary functions.

Proposition 1 (Lin et al [29]). *If a function $f(\boldsymbol{x}, \cdot)$ is μ-strongly concave and L smooth, then $\Phi(\boldsymbol{x})$ is $\beta = \kappa L + L$ smooth and $\boldsymbol{y}^*(\boldsymbol{x})$ is κ-Lipschitz where $\kappa = L/\mu$.*

Since Φ is β-smooth, then the starting point is to conduct the standard analysis scheme for nonconvex smooth function on one iteration as follows:

$$\mathbb{E}\left[\Phi(\boldsymbol{x}^{(t+1)})\right] - \mathbb{E}\left[\Phi(\boldsymbol{x}^{(t)})\right] \leq -\frac{\eta}{2}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right] - \left(\eta_x - 3\beta\eta_x^2\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\right\|^2\right]$$

$$+ \left(2\eta + 3\beta\eta_x^2\right) L^2 \mathbb{E}\left[\left(\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right)\right] + \frac{\eta_x L^2}{2}\mathbb{E}\left[\left\|\boldsymbol{y}^*(\boldsymbol{x}^{(t)}) - \boldsymbol{y}^{(t)}\right\|^2\right] + \frac{3\beta\eta_x^2\sigma^2}{2n}.$$

We can see the convergence depends on $\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}$, and a new term: $\left\|\boldsymbol{y}^*(\boldsymbol{x}^{(t)}) - \boldsymbol{y}^{(t)}\right\|^2$. The bound we derived for $\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}$ is no longer suitable here since in nonconvex objective, convergence to global saddle point is NP-hard. Instead, we derive the following deviation bound with the help of *gradient dissimilarity*:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right) \leq 10\tau^2(\eta_x^2 + \eta_y^2)\left(\sigma^2 + \frac{\sigma^2}{n}\right) + 10\tau^2\eta_x^2\zeta_x + 10\tau^2\eta_y^2\zeta_y.$$

Another thing is to bound the gap of current dual iterate and optimal dual variable: $\left\|\boldsymbol{y}^*(\boldsymbol{x}^{(t)}) - \boldsymbol{y}^{(t)}\right\|^2$. [29] has established the convergence of it, but they use a fairly large dual step size $O(1/L)$. However, in the local descent method, due to the issue of local model drifting, we are forced to stick with a small step size. Thus, as our main contribution in this part, we established the convergence of $\left\|\boldsymbol{y}^*(\boldsymbol{x}^{(t)}) - \boldsymbol{y}^{(t)}\right\|^2$ using a smaller dual step size:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*(\boldsymbol{x}^{(t)})\right\|^2\right] &\leq \frac{2C\kappa}{T}\mathbb{E}\left[\left\|\boldsymbol{y}^{(0)} - \boldsymbol{y}^*(\boldsymbol{x}^{(0)})\right\|^2\right] + O\left(\frac{C\eta_y^2\sigma^2}{n}\right)\\ &\quad + \frac{1}{T}\sum_{t=1}^{T} O\left(C\left(\eta_y + \eta_y^2\right) + C^2\eta_x^2\right)\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right]\\ &\quad + \frac{1}{T}\sum_{t=1}^{T} O\left(C^2\eta_x^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\right\|^2\right]\right),\end{aligned}$$

where $C = \frac{2}{\eta_y L}$. C could be large if we choose η_y to be small, and will thus negatively affect convergence rate, which means we trade some rate for communication efficiency.

Putting these piece together, and letting η_x and η_y to be sufficiently small, we can cancel up the term $\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\right\|^2\right]$ and establish the convergence rate.

B.2 Proof of technical lemmas

Before proceeding to the main proof of theorem, let us introduce a few useful intermediate results. The following lemma shows the analysis for one iteration of local SGDA, on nonconvex-strongly-concave function.

Lemma B.1. *For local-SGDA, under the assumptions in Theorem 5.1, the following statement holds:*

$$\begin{aligned}\mathbb{E}\left[\Phi(\boldsymbol{x}^{(t+1)})\right] - \mathbb{E}\left[\Phi(\boldsymbol{x}^{(t)})\right] &\leq -\frac{\eta}{2}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right] - \left(\eta_x - 3\beta\eta_x^2\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\right\|^2\right]\\ &\quad + \left(2\eta + 3\beta\eta_x^2\right)L^2\mathbb{E}\left[\left(\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right)\right] + \frac{\eta_x L^2}{2}\mathbb{E}\left[\left\|\boldsymbol{y}^*(\boldsymbol{x}^{(t)}) - \boldsymbol{y}^{(t)}\right\|^2\right] + \frac{3}{2n}\beta\eta_x^2\sigma^2,\end{aligned}$$

where $\beta = L + \kappa L$, and $\delta_{\boldsymbol{x}}^{(t)} = \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}^{(t)}\right\|^2$, $\quad \delta_{\boldsymbol{y}}^{(t)} = \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2$.

Proof. According to [29], $\Phi(\cdot)$ is $\beta = L + \kappa L$-smooth, together with updating rule, so we have:

$$\begin{aligned}\Phi(\boldsymbol{x}^{(t+1)}) &\leq \Phi(\boldsymbol{x}^{(t)}) + \left\langle\nabla\Phi(\boldsymbol{x}^{(t)}), \boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right\rangle + \frac{\beta}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right\|^2\\ &\leq \Phi(\boldsymbol{x}^{(t)}) - \eta_x\left\langle\nabla\Phi(\boldsymbol{x}^{(t)}), \frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)}; \xi_i^t)\right\rangle + \frac{\beta}{2}\eta^2\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)}; \xi_i^t)\right\|^2.\end{aligned}$$

Taking expectation on both sides yields:

$$\mathbb{E}\left[\Phi(\boldsymbol{x}^{(t+1)})\right] \leq \mathbb{E}\left[\Phi(\boldsymbol{x}^{(t)})\right] - \eta_x \left\langle \nabla\Phi(\boldsymbol{x}^{(t)}), \frac{1}{n}\sum_{i=1}^{n} \nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)}) \right\rangle + \frac{\beta}{2}\eta_x^2 \mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^{n} \nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)}; \xi_i^t)\right\|^2\right]$$

$$\leq \mathbb{E}\left[\Phi(\boldsymbol{x}^{(t)})\right] - \eta_x \left\langle \nabla\Phi(\boldsymbol{x}^{(t)}), \frac{1}{n}\sum_{i=1}^{n} \nabla_x f_i(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \right\rangle + \frac{\beta}{2}\eta_x^2 \mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^{n} \nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)}; \xi_i^t)\right\|^2\right]$$

$$- \eta_x \left\langle \nabla\Phi(\boldsymbol{x}^{(t)}), \frac{1}{n}\sum_{i=1}^{n} \nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)}) - \frac{1}{n}\sum_{i=1}^{n} \nabla_x f_i(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \right\rangle.$$

Using the identity $\langle \boldsymbol{a}, \boldsymbol{b}\rangle = -\frac{1}{2}\|\boldsymbol{a}-\boldsymbol{b}\|^2 + \frac{1}{2}\|\boldsymbol{a}\|^2 + \frac{1}{2}\|\boldsymbol{b}\|^2$, we have:

$$\mathbb{E}\left[\Phi(\boldsymbol{x}^{(t+1)})\right] - \mathbb{E}\left[\Phi(\boldsymbol{x}^{(t)})\right]$$

$$\leq -\frac{\eta_x}{2}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right] - \frac{\eta_x}{2}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\right\|^2\right] + \frac{\eta_x}{2}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)}) - \frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\right\|^2\right]$$

$$+ \frac{\eta_x}{2}\left(\frac{1}{2}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right] + 2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)}) - \frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\right\|^2\right]\right)$$

$$+ \frac{\beta}{2}\eta_x^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)}; \xi_i^t)\right\|^2\right]$$

$$\leq -\frac{\eta_x}{4}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right] - \frac{\eta_x}{2}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\right\|^2\right] + \frac{\eta_x L^2}{2}\mathbb{E}\left[\left\|\boldsymbol{y}^*(\boldsymbol{x}^{(t)}) - \boldsymbol{y}^{(t)}\right\|^2\right]$$

$$+ \eta_x L^2 \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[2\left\|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}^{(t)}\right\|^2 + 2\left\|\boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2\right]$$

$$+ \frac{\beta}{2}\eta_x^2\mathbb{E}\left[3\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\right\|^2 + 3\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)}) - \frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\right\|^2 + 3\sigma^2\right]$$

$$\leq -\frac{\eta_x}{4}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right] - \left(\frac{\eta}{2} - \frac{3\beta}{2}\eta^2\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\right\|^2\right] + \frac{\eta_x L^2}{2}\mathbb{E}\left[\left\|\boldsymbol{y}^*(\boldsymbol{x}^{(t)}) - \boldsymbol{y}^{(t)}\right\|^2\right]$$

$$+ (2\eta_x + 3\beta\eta_x^2)L^2\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right] + \frac{3\beta}{2n}\eta_x^2\sigma^2.$$

□

The following lemma characterizes the local model deviation bound for nonconvex-strongly-concave function.

Lemma B.2. *For local-SGDA, under assumptions of Theorem 5.1, the following statement holds true:*

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}_i^{(t)}\right\|^2\right] + \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}_i^{(t)}\right\|^2\right] \leq 10\tau^2(\eta_x^2 + \eta_y^2)\left(\sigma^2 + \frac{\sigma^2}{n}\right) + 10\tau^2\eta_x^2\zeta_x + 10\tau^2\eta_y^2\zeta_y.$$

Proof. We start to prove the first statement here. For the simplicity of notations, we define $\delta^t = \mathbb{E}[\delta_x^t + \delta_y^t] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}_i^{(t)}\right\|^2\right] + \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}_i^{(t)}\right\|^2\right]$. Assume $s\tau + 1 \leq t \leq (s+1)\tau$. Notice that:

$$\delta^t = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\boldsymbol{x}^{s\tau} - \sum_{j=s\tau}^{(s+1)\tau}\frac{\eta_x}{n}\sum_{k=1}^{n}\nabla_x f_k(\boldsymbol{x}_k^{(j)}, \boldsymbol{y}_k^{(j)}; \xi_k^j) - \left(\boldsymbol{x}^{s\tau} - \sum_{j=s\tau}^{(s+1)\tau}\eta_x\nabla_x f_i(\boldsymbol{x}_i^{(j)}, \boldsymbol{y}_i^{(j)}; \xi_i^j)\right)\right\|^2\right]$$

$$
\begin{aligned}
&+ \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\boldsymbol{y}^{s\tau} - \sum_{j=s\tau}^{(s+1)\tau}\frac{\eta_y}{n}\sum_{k=1}^{n}\nabla_y f_k(\boldsymbol{x}_k^{(j)},\boldsymbol{y}_k^{(j)};\xi_k^j) - \left(\boldsymbol{y}^{s\tau} - \sum_{j=s\tau}^{(s+1)\tau}\eta_y\nabla_y f_i(\boldsymbol{x}_i^{(j)},\boldsymbol{y}_i^{(j)};\xi_i^j)\right)\right\|^2\right] \\
&= \tau\sum_{j=s\tau}^{(s+1)\tau}\frac{\eta_x^2}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{k=1}^{n}\nabla_x f_k(\boldsymbol{x}_k^{(j)},\boldsymbol{y}_k^{(j)};\xi_k^j) - \nabla_x f_i(\boldsymbol{x}_i^{(j)},\boldsymbol{y}_i^{(j)};\xi_i^j)\right\|^2\right] \\
&+ \tau\sum_{j=s\tau}^{(s+1)\tau}\frac{\eta_y}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{k=1}^{n}\nabla_y f_k(\boldsymbol{x}_k^{(j)},\boldsymbol{y}_k^{(j)};\xi_k^j) - \nabla_y f_i(\boldsymbol{x}_i^{(j)},\boldsymbol{y}_i^{(j)};\xi_i^j)\right\|^2\right] \\
&= \tau\sum_{j=s\tau}^{(s+1)\tau}\frac{\eta_x^2}{n}\sum_{i=1}^{n}\mathbb{E}\Bigg[\Bigg\|\frac{1}{n}\sum_{k=1}^{n}\nabla_x f_k(\boldsymbol{x}_k^{(j)},\boldsymbol{y}_k^{(j)};\xi_k^j) - \nabla_x f_k(\boldsymbol{x}_k^{(j)},\boldsymbol{y}_k^{(j)}) + \nabla_x f_k(\boldsymbol{x}_k^{(j)},\boldsymbol{y}_k^{(j)}) - \nabla_x f_k(\boldsymbol{x}^{(j)},\boldsymbol{y}^{(j)}) \\
&+\nabla_x f_k(\boldsymbol{x}^{(j)},\boldsymbol{y}^{(j)}) - \nabla_x f_i(\boldsymbol{x}^{(j)},\boldsymbol{y}^{(j)}) + \nabla_x f_i(\boldsymbol{x}^{(j)},\boldsymbol{y}^{(j)}) - \nabla_x f_i(\boldsymbol{x}_i^{(j)},\boldsymbol{y}_i^{(j)}) + \nabla_x f_i(\boldsymbol{x}_i^{(j)},\boldsymbol{y}_i^{(j)}) - \nabla_x f_i(\boldsymbol{x}_i^{(j)},\boldsymbol{y}_i^{(j)};\xi_i^t)\Bigg\|^2\Bigg] \\
&+ \tau\sum_{j=s\tau}^{(s+1)\tau}\frac{\eta_y^2}{n}\sum_{i=1}^{n}\mathbb{E}\Bigg[\Bigg\|\frac{1}{n}\sum_{k=1}^{n}\nabla_y f_k(\boldsymbol{x}_k^{(j)},\boldsymbol{y}_k^{(j)};\xi_k^j) - \nabla_y f_k(\boldsymbol{x}_k^{(j)},\boldsymbol{y}_k^{(j)}) + \nabla_y f_k(\boldsymbol{x}_k^{(j)},\boldsymbol{y}_k^{(j)}) - \nabla_y f_k(\boldsymbol{x}^{(j)},\boldsymbol{y}^{(j)}) \\
&+\nabla_y f_k(\boldsymbol{x}^{(j)},\boldsymbol{y}^{(j)}) - \nabla_y f_i(\boldsymbol{x}^{(j)},\boldsymbol{y}^{(j)}) + \nabla_y f_i(\boldsymbol{x}^{(j)},\boldsymbol{y}^{(j)}) - \nabla_y f_i(\boldsymbol{x}_i^{(j)},\boldsymbol{y}_i^{(j)}) + \nabla_y f_i(\boldsymbol{x}_i^{(j)},\boldsymbol{y}_i^{(j)}) - \nabla_y f_i(\boldsymbol{x}_i^{(j)},\boldsymbol{y}_i^{(j)};\xi_i^t)\Bigg\|^2\Bigg] \\
&\le \sum_{j=s\tau}^{(s+1)\tau} 5\eta_x^2\left(\sigma^2 + \frac{\sigma^2}{n} + 2L^2\delta^j + \zeta_x\right) + 5\eta_y^2\left(\sigma^2 + \frac{\sigma^2}{n} + 2L^2\delta^j + \zeta_y\right).
\end{aligned}
$$

Summing over t from $s\tau$ to $(s+1)\tau$ yields:

$$
\begin{aligned}
\sum_{t=s\tau}^{(s+1)\tau}\delta^t &\le \sum_{t=s\tau}^{(s+1)\tau}\sum_{j=s\tau}^{(s+1)\tau} 5\tau\eta_x^2\left(\sigma^2 + \frac{\sigma^2}{n} + 2L^2\delta^j + \zeta_x\right) + 5\tau\eta_y^2\left(\sigma^2 + \frac{\sigma^2}{n} + 2L^2\delta^j + \zeta_y\right) \\
&\le 10L^2\tau^2(\eta_x^2+\eta_y^2)\sum_{j=s\tau}^{(s+1)\tau}\delta^j + 5\tau^2(\eta_x^2+\eta_y^2)\left(\sigma^2+\frac{\sigma^2}{n}\right) + 5\tau^2\eta_x^2\zeta_x + 5\tau^2\eta_y^2\zeta_y. \qquad (11)
\end{aligned}
$$

Since $\tau = \frac{T^{1/3}}{n^{1/3}}, \eta_x = \frac{n^{1/3}}{LT^{2/3}}, \eta_y = \frac{2}{LT^{1/2}}$ and $T \ge \max\left\{\frac{160^3}{n^2}, 40^{3/2}\right\}$, then $10L^2\tau^2(\eta_x^2+\eta_y^2) \le \frac{1}{2}$, by re-arranging the terms we have:

$$
\sum_{t=s\tau+1}^{(s+1)\tau}\delta^t \le 10\tau^3(\eta_x^2+\eta_y^2)\left(\sigma^2+\frac{\sigma^2}{n}\right) + 10\tau^3\eta_x^2\zeta_x + 10\tau^3\eta_y^2\zeta_y.
$$

Summing over s from 0 to $T/\tau - 1$, and dividing both sides by T can conclude the proof of the first statement:

$$
\frac{1}{T}\sum_{t=1}^{T}\delta^t \le 10\tau^2(\eta_x^2+\eta_y^2)\left(\sigma^2+\frac{\sigma^2}{n}\right) + 10\tau^2\eta_x^2\zeta_x + 10\tau^2\eta_y^2\zeta_y.
$$

□

The next lemma establishes an upper bound on the dual optimality gap.

Lemma B.3. *For local-SGDA, if we choose $\eta_y = \frac{2}{CL}$, then under assumptions of Theorem 5.1, the gap between $\boldsymbol{y}^t$ and $\boldsymbol{y}^*(\boldsymbol{x}^{(t)})$ can be bounded as follows:*

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*(\boldsymbol{x}^{(t)})\right\|^2\right] &\le \frac{2C\kappa}{T}\mathbb{E}\left[\left\|\boldsymbol{y}^{(0)} - \boldsymbol{y}^*(\boldsymbol{x}^{(0)})\right\|^2\right] + 2C\kappa\left(1+\frac{1}{2(C\kappa-1)}\right)\frac{4\eta_y^2\sigma^2}{n} \\
&+ \frac{1}{T}\sum_{t=1}^{T}2C\kappa\left(1+\frac{1}{2(C\kappa-1)}\right)\left(\frac{4\eta_y L^2}{\mu} + 8\eta_y^2L^2\right)\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right]
\end{aligned}
$$

$$+\frac{1}{T}\sum_{t=1}^{T}4C^2\kappa^4\eta_x^2\mathbb{E}\left[3\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2+6L^2(\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)})+\frac{3\sigma^2}{n}\right]. \tag{12}$$

where $\delta_{\boldsymbol{x}}^{(t)}=\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i^{(t)}-\boldsymbol{x}^{(t)}\right\|^2$ *and* $\delta_{\boldsymbol{y}}^{(t)}=\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^{(t)}-\boldsymbol{y}^{(t)}\right\|^2$.

Proof. According to arithmetic and geometric inequality and Cauchy's inequality: $\|\boldsymbol{a}+\boldsymbol{b}\|^2\le\|\boldsymbol{a}\|^2+2\|\boldsymbol{a}\|\|\boldsymbol{b}\|+\|\boldsymbol{b}\|^2\le\left(1+\frac{1}{q}\right)\|\boldsymbol{a}\|^2+(1+q)\|\boldsymbol{b}\|^2$, we have:

$$\mathbb{E}\left[\left\|\boldsymbol{y}^*(\boldsymbol{x}^{(t)})-\boldsymbol{y}^{(t)}\right\|^2\right]\le\left(1+\frac{1}{2(C\kappa-1)}\right)\mathbb{E}\left[\left\|\boldsymbol{y}^*(\boldsymbol{x}^{(t-1)})-\boldsymbol{y}^{(t)}\right\|^2\right]+(1+2(C\kappa-1))\,\mathbb{E}\left[\left\|\boldsymbol{y}^*(\boldsymbol{x}^{(t)})-\boldsymbol{y}^*(\boldsymbol{x}^{(t-1)})\right\|^2\right].$$

Then we are going to bound $\left\|\boldsymbol{y}^*(\boldsymbol{x}^{(t-1)})-\boldsymbol{y}^{(t)}\right\|^2$ and $\left\|\boldsymbol{y}^*(\boldsymbol{x}^{(t)})-\boldsymbol{y}^*(\boldsymbol{x}^{(t-1)})\right\|^2$ separately.

First, according to updating rule for $\boldsymbol{y}$ and strong concavity, we have:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{y}^{(t)}-\boldsymbol{y}^*(\boldsymbol{x}^{(t-1)})\right\|^2\right]\\
&=\mathbb{E}\left[\left\|\boldsymbol{y}^{(t-1)}+\eta_y\frac{1}{n}\sum_{i=1}^{n}\nabla_y f_i(\boldsymbol{x}_i^{(t-1)},\boldsymbol{y}_i^{(t-1)};\xi_i^t)-\boldsymbol{y}^*(\boldsymbol{x}^{(t-1)})\right\|^2\right]\\
&\le\mathbb{E}\left[\left\|\boldsymbol{y}^{(t-1)}-\boldsymbol{y}^*(\boldsymbol{x}^{(t-1)})\right\|^2\right]+\eta_y^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_y f_i(\boldsymbol{x}_i^{(t-1)},\boldsymbol{y}_i^{(t-1)};\xi_i^t)\right\|^2\right]\\
&\quad+2\eta_y\mathbb{E}\left[\left\langle\frac{1}{n}\sum_{i=1}^{n}\nabla_y f_i(\boldsymbol{x}_i^{(t-1)},\boldsymbol{y}_i^{(t-1)}),\boldsymbol{y}^{(t-1)}-\boldsymbol{y}^*(\boldsymbol{x}^{(t-1)})\right\rangle\right]\\
&\le\mathbb{E}\left[\left\|\boldsymbol{y}^{(t-1)}-\boldsymbol{y}^*(\boldsymbol{x}^{(t-1)})\right\|^2\right]\\
&\quad+\eta_y^2\left(4\underbrace{\mathbb{E}\left[\left\|\nabla_y F(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^*(\boldsymbol{x}^{(t-1)}))\right\|^2\right]}_{=0}+4\mathbb{E}\left[\left\|\nabla_y F(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)})-\nabla_y F(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^*(\boldsymbol{x}^{(t-1)}))\right\|^2\right]\right)\\
&\quad+\eta_y^2\frac{1}{n}\sum_{i=1}^{n}\left(4\mathbb{E}\left[\left\|\nabla_y f_i(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)})-\nabla_y f_i(\boldsymbol{x}_i^{(t-1)},\boldsymbol{y}_i^{(t-1)})\right\|^2\right]+4\frac{\sigma^2}{n}\right)\\
&\quad+2\eta_y\mathbb{E}\left[\left\langle\frac{1}{n}\sum_{i=1}^{n}\nabla_y f_i(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)}),\boldsymbol{y}^{(t-1)}-\boldsymbol{y}^*(\boldsymbol{x}^{(t-1)})\right\rangle\right]\\
&\quad+2\eta_y\mathbb{E}\left[\left\langle\frac{1}{n}\sum_{i=1}^{n}\nabla_y f_i(\boldsymbol{x}_i^{(t-1)},\boldsymbol{y}_i^{(t-1)})-\nabla_x f_i(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)}),\boldsymbol{y}^{(t-1)}-\boldsymbol{y}^*(\boldsymbol{x}^{(t-1)})\right\rangle\right]\\
&\le(1-\mu\eta_y)\,\mathbb{E}\left[\left\|\boldsymbol{y}^{(t-1)}-\boldsymbol{y}^*(\boldsymbol{x}^{(t-1)})\right\|^2\right]+4\eta_y^2\frac{\sigma^2}{n}+\frac{\mu\eta_y}{2}\mathbb{E}\left[\left\|\boldsymbol{y}^{(t-1)}-\boldsymbol{y}^*(\boldsymbol{x}^{(t-1)})\right\|^2\right]\\
&\quad+2\underbrace{(\eta_y-4\eta_y^2L)}_{\ge0}\underbrace{\mathbb{E}\left[F(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)})-F(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^*(\boldsymbol{x}^{(t-1)}))\right]}_{\le0}\\
&\quad+\left(\frac{2\eta_y}{\mu}+4\eta_y^2\right)\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla_y f_i(\boldsymbol{x}_i^{(t-1)},\boldsymbol{y}_i^{(t-1)})-\nabla_y f_i(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)})\right\|^2\right]\\
&\le\left(1-\frac{\mu\eta_y}{2}\right)\mathbb{E}\left[\left\|\boldsymbol{y}^{(t-1)}-\boldsymbol{y}^*(\boldsymbol{x}^{(t-1)})\right\|^2\right]+\frac{4\eta_y^2\sigma^2}{n}+\left(\frac{4\eta_yL^2}{\mu}+8\eta_y^2L^2\right)\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t-1)}+\delta_{\boldsymbol{y}}^{(t-1)}\right].
\end{aligned} \tag{13}$$

Then, for the term $\left\|\boldsymbol{y}^*(\boldsymbol{x}^{(t)}) - \boldsymbol{y}^*(\boldsymbol{x}^{(t-1)})\right\|^2$, since $\boldsymbol{y}^*(\cdot)$ is κ-Lipschitz, we have:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{y}^*(\boldsymbol{x}^{(t)}) - \boldsymbol{y}^*(\boldsymbol{x}^{(t-1)})\right\|^2\right] \leq \kappa^2 \mathbb{E}\left[\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|^2\right] \\
&= \kappa^2\eta_x^2 \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t-1)}, \boldsymbol{y}_i^{(t-1)}; \xi_i^t)\right\|^2\right] \\
&\leq \kappa^2\eta_x^2 \mathbb{E}\left[3\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)})\right\|^2 + 3\frac{1}{n}\sum_{i=1}^n \left\|\nabla_x f_i(\boldsymbol{x}_i^{(t-1)}, \boldsymbol{y}_i^{(t-1)}) - \nabla_x f_i(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)})\right\|^2 + \frac{3\sigma^2}{n}\right] \\
&\leq \kappa^2\eta_x^2 \mathbb{E}\left[3\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)})\right\|^2 + 6L^2(\delta_{\boldsymbol{x}}^{(t-1)} + \delta_{\boldsymbol{y}}^{(t-1)}) + \frac{3\sigma^2}{n}\right].
\end{aligned} \tag{14}$$

Recall that we choose $\eta_y = \frac{2}{CL}$, $C > 0$. Combining (13) and (14) yields:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{y}^*(\boldsymbol{x}^{(t)}) - \boldsymbol{y}^{(t)}\right\|^2\right] \\
&\leq \left(1 + \frac{1}{2(C\kappa-1)}\right)\left(\left(1 - \frac{\mu\eta_y}{2}\right)\mathbb{E}\left[\left\|\boldsymbol{y}^{(t-1)} - \boldsymbol{y}^*(\boldsymbol{x}^{(t-1)})\right\|^2\right] + \frac{4\eta_y^2\sigma^2}{n} + \left(\frac{4\eta_y L^2}{\mu} + 8\eta_y^2 L^2\right)\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t-1)} + \delta_{\boldsymbol{y}}^{(t-1)}\right]\right) \\
&\quad + (1 + 2(C\kappa - 1))\,\kappa^2\eta_x^2 \mathbb{E}\left[3\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)})\right\|^2 + 6L^2(\delta_{\boldsymbol{x}}^{(t-1)} + \delta_{\boldsymbol{y}}^{(t-1)}) + \frac{3\sigma^2}{n}\right] \\
&\leq \left(1 + \frac{1}{2(C\kappa-1)}\right)\left(1 - \frac{1}{C\kappa}\right)\mathbb{E}\left[\left\|\boldsymbol{y}^{(t-1)} - \boldsymbol{y}^*(\boldsymbol{x}^{(t-1)})\right\|^2\right] \\
&\quad + \left(1 + \frac{1}{2(C\kappa-1)}\right)\left(\frac{4\eta_y^2\sigma^2}{n} + \left(\frac{4\eta_y L^2}{\mu} + 8\eta_y^2 L^2\right)\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t-1)} + \delta_{\boldsymbol{y}}^{(t-1)}\right]\right) \\
&\quad + (1 + 2(C\kappa - 1))\,\kappa^2\eta_x^2 \mathbb{E}\left[3\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)})\right\|^2 + 6L^2(\delta_{\boldsymbol{x}}^{(t-1)} + \delta_{\boldsymbol{y}}^{(t-1)}) + \frac{3\sigma^2}{n}\right].
\end{aligned}$$

Using the fact $\left(1 + \frac{1}{2(C\kappa-1)}\right)\left(1 - \frac{1}{C\kappa}\right) = \left(1 - \frac{1}{2C\kappa}\right)$, and unrolling the recursion yields:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{y}^*(\boldsymbol{x}^{(t)}) - \boldsymbol{y}^{(t)}\right\|^2\right] \\
&\leq \left(1 - \frac{1}{2C\kappa}\right)\mathbb{E}\left[\left\|\boldsymbol{y}^{(t-1)} - \boldsymbol{y}^*(\boldsymbol{x}^{(t-1)})\right\|^2\right] + \left(1 + \frac{1}{2(C\kappa-1)}\right)\left(\frac{4\eta_y^2\sigma^2}{n} + \left(\frac{4\eta_y L^2}{\mu} + 8\eta_y^2 L^2\right)\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t-1)} + \delta_{\boldsymbol{y}}^{(t-1)}\right]\right) \\
&\quad + (1 + 2(C\kappa - 1))\,\kappa^2\eta_x^2 \mathbb{E}\left[3\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)})\right\|^2 + 6L^2(\delta_{\boldsymbol{x}}^{(t-1)} + \delta_{\boldsymbol{y}}^{(t-1)}) + \frac{3\sigma^2}{n}\right] \\
&\leq \left(1 - \frac{1}{2C\kappa}\right)^t \mathbb{E}\left[\left\|\boldsymbol{y}^{(0)} - \boldsymbol{y}^*(\boldsymbol{x}^{(0)})\right\|^2\right] \\
&\quad + \sum_{j=1}^t \left(1 - \frac{1}{2C\kappa}\right)^{t-j}\left(1 + \frac{1}{2(C\kappa-1)}\right)\left(\frac{4\eta_y^2\sigma^2}{n} + \left(\frac{4\eta_y L^2}{\mu} + 8\eta_y^2 L^2\right)\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t-1)} + \delta_{\boldsymbol{y}}^{(t-1)}\right]\right) \\
&\quad + \sum_{j=1}^t \left(1 - \frac{1}{2C\kappa}\right)^{t-j}(1 + 2(C\kappa - 1))\,\kappa^2\eta_x^2 \mathbb{E}\left[3\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}^{(j-1)}, \boldsymbol{y}^{(t-1)})\right\|^2 + 6L^2(\delta_x^{j-1} + \delta_y^{j-1}) + \frac{3\sigma^2}{n}\right] \\
&\leq \left(1 - \frac{1}{2C\kappa}\right)^t \mathbb{E}\left[\left\|\boldsymbol{y}^{(0)} - \boldsymbol{y}^*(\boldsymbol{x}^{(0)})\right\|^2\right] + 2C\kappa\left(1 + \frac{1}{2(C\kappa-1)}\right)\left(\frac{4\eta_y^2\sigma^2}{n}\right)
\end{aligned}$$

$$+\sum_{j=1}^{t}\left(1-\frac{1}{2C\kappa}\right)^{t-j}\left(1+\frac{1}{2(C\kappa-1)}\right)\left(\left(\frac{4\eta_y L^2}{\mu}+8\eta_y^2 L^2\right)\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t-1)}+\delta_{\boldsymbol{y}}^{(t-1)}\right]\right)$$
$$+\sum_{j=1}^{t}\left(1-\frac{1}{2C\kappa}\right)^{t-j}(1+2(C\kappa-1))\,\kappa^2\eta_x^2\mathbb{E}\left[3\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}^{(j-1)},\boldsymbol{y}^{(j-1)})\right\|^2+6L^2(\delta_x^{j-1}+\delta_y^{j-1})+\frac{3\sigma^2}{n}\right].$$

Summing from $t=1$ to T, and dividing by T yields:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{y}^*(\boldsymbol{x}^{(t)})-\boldsymbol{y}^{(t)}\right\|^2\right]$$
$$\leq\frac{1}{T}\sum_{t=1}^{T}\left(1-\frac{1}{2C\kappa}\right)^{t}\mathbb{E}\left[\left\|\boldsymbol{y}^{(0)}-\boldsymbol{y}^*(\boldsymbol{x}^{(0)})\right\|^2\right]+2C\kappa\left(1+\frac{1}{2(C\kappa-1)}\right)\frac{4\eta_y^2\sigma^2}{n}$$
$$+\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{t}\left(1-\frac{1}{2C\kappa}\right)^{t-j}\left(1+\frac{1}{2(C\kappa-1)}\right)\left(\frac{4\eta_y L^2}{\mu}+8\eta_y^2 L^2\right)\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t-1)}+\delta_{\boldsymbol{y}}^{(t-1)}\right]$$
$$+\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{t}\left(1-\frac{1}{2C\kappa}\right)^{t-j}(1+2(C\kappa-1))\,\kappa^2\eta_x^2\mathbb{E}\left[3\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}^{(j-1)},\boldsymbol{y}^{(t-1)})\right\|^2+6L^2(\delta_x^{j-1}+\delta_y^{j-1})+\frac{3\sigma^2}{n}\right].$$
$$\leq\frac{2C\kappa}{T}\mathbb{E}\left[\left\|\boldsymbol{y}^{(0)}-\boldsymbol{y}^*(\boldsymbol{x}^{(0)})\right\|^2\right]+2C\kappa\left(1+\frac{1}{2(C\kappa-1)}\right)\frac{4\eta_y^2\sigma^2}{n}$$
$$+\frac{1}{T}\sum_{t=0}^{T}2C\kappa\left(1+\frac{1}{2(C\kappa-1)}\right)\left(\frac{4\eta_y L^2}{\mu}+8\eta_y^2 L^2\right)\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)}\right]$$
$$+\frac{1}{T}\sum_{t=0}^{T}4C^2\kappa^4\eta_x^2\mathbb{E}\left[3\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2+6L^2(\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)})+\frac{3\sigma^2}{n}\right].$$

□

B.3 Proof of Theorem 5.1

Now we provide the proof of Theorem 5.1. In Lemma B.1, summing over $t=1$ to T and divding both sides by T yields:

$$\frac{1}{T}\left(\mathbb{E}\left[\Phi(\boldsymbol{x}^{(T+1)})\right]-\mathbb{E}\left[\Phi(\boldsymbol{x}^{(0)})\right]\right)$$
$$\leq-\frac{\eta_x}{2}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right]-\left(\eta_x-3\beta\eta_x^2\right)\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2\right]$$
$$+\left(2\eta_x+3\beta\eta_x^2\right)L^2\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)}\right)\right]+\frac{\eta_x L^2}{2}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{y}^*(\boldsymbol{x}^{(t)})-\boldsymbol{y}^{(t)}\right\|^2\right]+\frac{3}{2}\beta\eta_x^2\frac{\sigma^2}{n}.$$

For the simplicity of the notation, we let $\Re=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2\right]$. Re-arranging the terms and plugging in Lemma B.2 and Lemma B.3 gives:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right]$$
$$\leq\frac{2}{\eta_x T}\mathbb{E}\left[\Phi(\boldsymbol{x}^{(0)})\right]-2\left(1-3\beta\eta_x\right)\Re$$
$$+2\left(2+3\beta\eta_x\right)L^2\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)}\right)\right]+L^2\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{y}^*(\boldsymbol{x}^{(t)})-\boldsymbol{y}^{(t)}\right\|^2\right]+3\beta\eta_x\frac{\sigma^2}{n}$$

$$\leq \frac{2}{\eta_x T}\mathbb{E}\left[\Phi(\boldsymbol{x}^{(0)})\right] - 2\left(1-3\beta\eta_x\right)\Re + 3\beta\eta_x\frac{\sigma^2}{n}$$
$$+\left(4+6\beta\eta_x\right)L^2\left[10\tau^2(\eta_x^2+\eta_y^2)\left(\sigma^2+\frac{\sigma^2}{n}\right)+10\tau^2\eta_x^2\zeta_x+10\tau^2\eta_y^2\zeta_y\right]$$
$$+\frac{2L^2C\kappa}{T}\mathbb{E}\left[\left\|\boldsymbol{y}^{(0)}-\boldsymbol{y}^*(\boldsymbol{x}^{(0)})\right\|^2\right]+\left(\frac{2C^2\kappa^2L^2}{C\kappa-1}\right)\frac{4\eta_y^2\sigma^2}{n}$$
$$+4C^2\kappa^4\eta_x^2L^2\left(\Re+6L^2\left[10\tau^2(\eta_x^2+\eta_y^2)\left(\sigma^2+\frac{\sigma^2}{n}\right)+10\tau^2\eta_x^2\zeta_x+10\tau^2\eta_y^2\zeta_y\right]+\frac{3\sigma^2}{n}\right)$$
$$+\left(\frac{2C^2\kappa^2L^2}{C\kappa-1}\right)\left(\frac{4\eta_yL^2}{\mu}+8\eta_y^2L^2\right)\left[10\tau^2(\eta_x^2+\eta_y^2)\left(\sigma^2+\frac{\sigma^2}{n}\right)+10\tau^2\eta_x^2\zeta_x+10\tau^2\eta_y^2\zeta_y\right].$$
$$\leq \frac{2}{\eta_x T}\mathbb{E}\left[\Phi(\boldsymbol{x}^{(0)})\right]+\frac{2L^2C\kappa}{T}\mathbb{E}\left[\left\|\boldsymbol{y}^{(0)}-\boldsymbol{y}^*(\boldsymbol{x}^{(0)})\right\|^2\right]-2\left(1-3\beta\eta_x-4C^2\kappa^4\eta_x^2L^2\right)\Re$$
$$+10\left(4+6\beta\eta_x+24C^2\kappa^4\eta_x^2L^2+\left(\frac{2C^2\kappa^2}{C\kappa-1}\right)\left(\frac{4\eta_yL^2}{\mu}+8\eta_y^2L^2\right)\right)L^2\left[\tau^2(\eta_x^2+\eta_y^2)\left(\sigma^2+\frac{\sigma^2}{n}\right)+\tau^2\eta_x^2\zeta_x+\tau^2\eta_y^2\zeta_y\right]$$
$$+\frac{12C^2\kappa^4\eta_x^2L^2\sigma^2}{n}+3\beta\eta_x\frac{\sigma^2}{n}+\left(\frac{2C^2\kappa^2L^2}{C\kappa-1}\right)\frac{4\eta_y^2\sigma^2}{n}.$$

By choosing $\eta_x=\frac{n^{1/3}}{LT^{2/3}}, C=T^{1/2}$ and $T\geq\max\left\{\left(\frac{16n^{4/3}\kappa^4+\sqrt{16n^{4/3}\kappa^8-12\beta n^{1/3}/L}}{2}\right)^3, 40^{3/2}, \frac{160^3}{n^2}\right\}$ in Theorem 5.1 such that

$$1-3\beta\eta_x-4C^2\kappa^4\eta_x^2L^2\geq 0,$$

holds, then we have:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right]\leq\frac{2}{\eta_xT}\mathbb{E}\left[\Phi(\boldsymbol{x}^{(0)})\right]+\frac{2L^2C\kappa}{T}\mathbb{E}\left[\left\|\boldsymbol{y}^{(0)}-\boldsymbol{y}^*(\boldsymbol{x}^{(0)})\right\|^2\right]$$
$$+10\left(4+6\beta\eta_x+24C^2\kappa^4\eta_x^2L^2+\left(\frac{2C^2\kappa^2}{C\kappa-1}\right)\left(\frac{4\eta_yL^2}{\mu}+8\eta_y^2L^2\right)\right)L^2\left[\tau^2(\eta_x^2+\eta_y^2)\left(\sigma^2+\frac{\sigma^2}{n}\right)+\tau^2\eta_x^2\zeta_x+\tau^2\eta_y^2\zeta_y\right]$$
$$+\frac{12C^2\kappa^4\eta_x^2L^2\sigma^2}{n}+\frac{3\beta\eta_x\sigma^2}{n}+\left(\frac{2C^2\kappa^2L^2}{C\kappa-1}\right)\frac{4\eta_y^2\sigma^2}{n}.$$

Plugging in $\tau=\frac{T^{1/3}}{n^{1/3}}$ and $\eta_x=\frac{n^{1/3}}{LT^{2/3}}$,$\eta_y=\frac{2}{LT^{\frac{1}{2}}}$, will conclude the proof:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right]\leq O\left(\frac{L}{(nT)^{1/3}}+\frac{\kappa^4L^2\sigma^2}{(nT)^{1/3}}+\frac{L^2\zeta_x}{T^{2/3}}+\frac{L^2\zeta_y}{n^{2/3}T^{1/3}}+\frac{L^2\kappa}{T^{1/2}}\right).$$

□

C Proof of Local SGDA+ under Nonconvex-PL Setting

C.1 Overview of proofs

Now we proceed to the proof of convergence rate in nonconvex-PL setting. In this case we still study the envelope function $\Phi(\cdot)$. The following proposition establishes the smoothness property of these auxiliary functions.

Proposition 2 (Nouiehed et al [39]). *If a function $F(\boldsymbol{x},\cdot)$ satisfies μ-PL condition and L smooth, then $\Phi(\boldsymbol{x})$ is $\beta=\kappa L/2+L$ smooth where $\kappa=L/\mu$.*

Since Φ is β-smooth, then the starting point is similar to what we did in nonconvex-strongly-concave case, to conduct the one iteration analysis scheme for nonconvex smooth function on one iteration as follows:

$$\mathbb{E}\left[\Phi(\boldsymbol{x}^{(t+1)})\right]-\mathbb{E}\left[\Phi(\boldsymbol{x}^{(t)})\right]\leq-\frac{\eta_x}{2}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right]-\left(\frac{\eta_x}{2}-\frac{\beta\eta_x^2}{2}\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_xf_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right]$$

$$+\frac{2\eta_x L^2}{\mu}\mathbb{E}\left[(\Phi(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}))\right] + 2\eta_x L^2 \mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right] + \frac{\beta\eta_x^2\sigma^2}{2n}.$$

We can see the convergence depends on $\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}$, and $\mathbb{E}\left[(\Phi(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}))\right]$. For $\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}$, we bound it in an analogous way to nonconvex-strongly-concave case.

Another thing is to bound the gap $\mathbb{E}\left[(\Phi(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}))\right]$. Here we borrow the proof idea from [43]:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\Phi(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\right] \\
&\leq \frac{2\mathbb{E}\left[\Phi(\boldsymbol{x}^{(0)}) - F(\boldsymbol{x}^{(0)}, \boldsymbol{y}^{(0)})\right]}{\mu\eta_y T} + \frac{2}{\mu T}\sum_{t=1}^{T}\left(L^2\eta_x^2\frac{\sigma^2}{n} + 2L^2 S\eta_x^2(G_x^2 + \sigma^2) + 2L^2\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right]\right) \\
&+ \left[\frac{2(1-\mu\eta_y)}{\mu\eta_y}\left(\frac{\eta_x^2 L}{2} + \frac{\beta\eta_x^2}{2}\right) + L^2\eta_x^2\right]\frac{1}{T}\left(\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(0)}, \boldsymbol{y}_i^{(0)})\right\|^2\right] + \sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)})\right\|^2\right]\right) \\
&+ \frac{2(1-\mu\eta_y)}{\mu\eta_y}\frac{1}{T}\left(\sum_{t=1}^{T}\left(\frac{1}{2}\eta_x\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right] + \frac{\eta_x^2 L\sigma^2}{2n}\right) + \mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(0)})\right\|^2\right]\right) \\
&+ \frac{2(1-\mu\eta_y)}{\mu\eta_y}\frac{1}{T}\sum_{t=1}^{T}\left(2\eta_x L^2\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right] + \frac{\beta\eta_x^2\sigma^2}{2n}\right) + \frac{\eta_y L\sigma^2}{n}.
\end{aligned}$$

Putting these piece together, concludes the proof.

C.2 Proof of technical lemmas

We first introduce some useful lemmas. The following lemma performs one iteration analysis of local SGDA+, on nonconvex-PL objective.

Lemma C.1. *For local-SGDA+, under the assumptions in Theorem 6.1, the following statement holds:*

$$\begin{aligned}
\mathbb{E}\left[\Phi(\boldsymbol{x}^{(t+1)})\right] - \mathbb{E}\left[\Phi(\boldsymbol{x}^{(t)})\right] \leq -\frac{\eta_x}{2}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right] - \left(\frac{\eta_x}{2} - \frac{\beta\eta_x^2}{2}\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)})\right\|^2\right] \\
+\frac{2\eta_x L^2}{\mu}\mathbb{E}\left[(\Phi(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}))\right] + 2\eta_x L^2\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right] + \frac{\beta\eta_x^2\sigma^2}{2n}.
\end{aligned}$$

where $\beta = L + \kappa L/2$.

Proof. Since $\Phi(\cdot)$ is $\beta = L + \kappa L$-smooth, we have:

$$\begin{aligned}
\Phi(\boldsymbol{x}^{(t+1)}) &\leq \Phi(\boldsymbol{x}^{(t)}) + \left\langle\nabla\Phi(\boldsymbol{x}^{(t)}), \boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right\rangle + \frac{\beta}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right\|^2 \\
&\leq \Phi(\boldsymbol{x}^{(t)}) - \eta_x\left\langle\nabla\Phi(\boldsymbol{x}^{(t)}), \frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)}; \xi_i^t)\right\rangle + \frac{\beta}{2}\eta^2\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)}; \xi_i^t)\right\|^2.
\end{aligned}$$

Taking expectation on both sides yields:

$$\mathbb{E}\left[\Phi(\boldsymbol{x}^{(t+1)})\right] \leq \mathbb{E}\left[\Phi(\boldsymbol{x}^{(t)})\right] - \eta_x\left\langle\nabla\Phi(\boldsymbol{x}^{(t)}), \frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)})\right\rangle + \frac{\beta}{2}\eta_x^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)}; \xi_i^t)\right\|^2\right].$$

Using the identity $\langle\boldsymbol{a}, \boldsymbol{b}\rangle = -\frac{1}{2}\|\boldsymbol{a} - \boldsymbol{b}\|^2 + \frac{1}{2}\|\boldsymbol{a}\|^2 + \frac{1}{2}\|\boldsymbol{b}\|^2$, we have:

$$\mathbb{E}\left[\Phi(\boldsymbol{x}^{(t+1)})\right] - \mathbb{E}\left[\Phi(\boldsymbol{x}^{(t)})\right]$$

$$\leq -\frac{\eta_x}{2}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right] - \frac{\eta_x}{2}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right] + \eta_x\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)}) - \frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2\right]$$
$$+ \eta_x\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)}) - \frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2\right]$$
$$+ \frac{\beta}{2}\eta_x^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right] + \frac{\beta\eta_x^2\sigma^2}{2n}$$
$$\leq -\frac{\eta_x}{2}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right] - \left(\frac{\eta_x}{2} - \frac{\beta\eta_x^2}{2}\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right] + \eta_x L^2\mathbb{E}\left[\left\|\phi(\boldsymbol{x}^{(t)}) - \boldsymbol{y}^{(t)}\right\|^2\right]$$
$$+ \eta_x L^2\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[2\left\|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}^{(t)}\right\|^2 + 2\left\|\boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2\right] + \frac{\beta\eta_x^2\sigma^2}{2n}$$
$$\leq -\frac{\eta_x}{2}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right] - \left(\frac{\eta_x}{2} - \frac{\beta\eta_x^2}{2}\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right] + \eta_x L^2\mathbb{E}\left[\left\|\phi(\boldsymbol{x}^{(t)}) - \boldsymbol{y}^{(t)}\right\|^2\right]$$
$$+ 2\eta_x L^2\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right] + \frac{\beta\eta_x^2\sigma^2}{2n}.$$

According to [13], PL condition implies quadratic growth, we have:

$$\left\|\phi(\boldsymbol{x}^{(t)}) - \boldsymbol{y}^{(t)}\right\|^2 \leq \frac{2}{\mu}(F(\boldsymbol{x}^{(t)},\phi(\boldsymbol{x}^{(t)})) - F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})) = \frac{2}{\mu}(\Phi(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})), \tag{15}$$

which concludes the proof.

□

The following lemma characterizes the sub-linear convergence of gap $\mathbb{E}\left[\Phi(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right]$.

Lemma C.2. *For local-SGDA+, under the assumptions in Theorem 6.1, the following statement holds:*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\Phi(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right]$$
$$\leq \frac{2\mathbb{E}\left[\Phi(\boldsymbol{x}^{(0)}) - F(\boldsymbol{x}^{(0)},\boldsymbol{y}^{(0)})\right]}{\mu\eta_y T} + \frac{2}{\mu T}\sum_{t=1}^T\left(L^2\eta_x^2\frac{\sigma^2}{n} + 2L^2S\eta_x^2(G_x^2+\sigma^2) + 2L^2\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right]\right)$$
$$+ \left[\frac{2(1-\mu\eta_y)}{\mu\eta_y}\left(\frac{\eta_x^2 L}{2} + \frac{\beta\eta_x^2}{2}\right) + L^2\eta_x^2\right]\frac{1}{T}\left(\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(0)},\boldsymbol{y}_i^{(0)})\right\|^2\right] + \sum_{t=1}^T\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right]\right)$$
$$+ \frac{2(1-\mu\eta_y)}{\mu\eta_y}\frac{1}{T}\left(\sum_{t=1}^T\left(\frac{1}{2}\eta_x\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right] + \frac{\eta_x^2 L\sigma^2}{2n}\right) + \mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(0)})\right\|^2\right]\right)$$
$$+ \frac{2(1-\mu\eta_y)}{\mu\eta_y}\frac{1}{T}\sum_{t=1}^T\left(2\eta_x L^2\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right] + \frac{\beta\eta_x^2\sigma^2}{2n}\right) + \frac{\eta_y L\sigma^2}{n}.$$

Proof. According to smoothness of $F(\boldsymbol{x},\cdot)$, we have

$$F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)}) \leq F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t+1)}) - \left\langle\nabla_y F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)}),\boldsymbol{y}^{(t+1)} - \boldsymbol{y}^{(t)}\right\rangle + \frac{L}{2}\left\|\frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\tilde{\boldsymbol{x}},\boldsymbol{y}_i^{(t)};\xi_i^t)\right\|^2$$
$$\leq F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t+1)}) - \eta_y\left\langle\nabla_y F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)}),\frac{1}{n}\nabla_y f_i(\tilde{\boldsymbol{x}},\boldsymbol{y}_i^{(t)};\xi_i^t)\right\rangle + \frac{\eta_y^2 L}{2}\left\|\frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\tilde{\boldsymbol{x}},\boldsymbol{y}_i^{(t)};\xi_i^t)\right\|^2$$

Taking expectation on both sides yields:

$$\begin{aligned}
\mathbb{E}[F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)})] &\leq \mathbb{E}[F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t+1)})] - \eta_y \mathbb{E}\left[\left\langle \nabla_y F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)}), \frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\tilde{\boldsymbol{x}},\boldsymbol{y}_i^{(t)})\right\rangle\right] \\
&+ \frac{\eta_y^2 L}{2}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\tilde{\boldsymbol{x}},\boldsymbol{y}_i^{(t)};\xi_i^t)\right\|^2\right] \\
&\leq \mathbb{E}[F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t+1)})] - \frac{\eta_y}{2}\mathbb{E}\left[\left\|\nabla_y F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}_i^{(t)})\right\|^2\right] + \frac{1}{2}\eta_y \mathbb{E}\left[\left\|\nabla_y F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)}) - \frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\tilde{\boldsymbol{x}},\boldsymbol{y}_i^{(t)})\right\|^2\right] \\
&- \left(\frac{\eta_y}{2} - \frac{\eta_y^2 L}{2}\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\tilde{\boldsymbol{x}},\boldsymbol{y}_i^{(t)})\right\|^2\right] + \frac{\eta_y^2 L\sigma^2}{2n} \qquad (16) \\
&\leq \mathbb{E}[F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t+1)})] - \frac{\eta_y}{2}\mathbb{E}\left[\left\|\nabla_y F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}_i^{(t)})\right\|^2\right] - \left(\frac{\eta_y}{2} - \frac{\eta_y^2 L}{2}\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\tilde{\boldsymbol{x}},\boldsymbol{y}_i^{(t)})\right\|^2\right] + \frac{\eta_y^2 L\sigma^2}{2n} \\
&+ \frac{1}{2}\eta_y \mathbb{E}\left[\left\|\nabla_y F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)}) - \nabla_y F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)}) + \nabla_y F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)}) - \frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\tilde{\boldsymbol{x}},\boldsymbol{y}_i^{(t)})\right\|^2\right] \\
&\leq \mathbb{E}[F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t+1)})] - \frac{\eta_y}{2}\mathbb{E}\left[\left\|\nabla_y F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}_i^{(t)})\right\|^2\right] - \left(\frac{\eta_y}{2} - \frac{\eta_y^2 L}{2}\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\tilde{\boldsymbol{x}},\boldsymbol{y}_i^{(t)})\right\|^2\right] + \frac{\eta_y^2 L\sigma^2}{2n} \\
&+ \eta_y \underbrace{\mathbb{E}\left[\left\|\nabla_y F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)}) - \nabla_y F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})\right\|^2\right]}_{T_1} + \eta_y \underbrace{\mathbb{E}\left[\left\|\nabla_y F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)}) - \frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\tilde{\boldsymbol{x}},\boldsymbol{y}_i^{(t)})\right\|^2\right]}_{T_2},
\end{aligned}$$

where we use the identity $\langle \boldsymbol{a},\boldsymbol{b}\rangle = -\frac{1}{2}\|\boldsymbol{a}-\boldsymbol{b}\|^2 + \frac{1}{2}\|\boldsymbol{a}\|^2 + \frac{1}{2}\|\boldsymbol{b}\|^2$.

To bound T_1, we notice that:

$$T_1 \leq L^2\mathbb{E}\left[\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right\|^2\right] \leq L^2\eta_x^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right] + L^2\eta_x^2\frac{\sigma^2}{n}.$$

For T_2, we bound it as follows:

$$\begin{aligned}
T_2 &\leq 2\mathbb{E}\left[\left\|\nabla_y F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)}) - \nabla_y F(\tilde{\boldsymbol{x}},\boldsymbol{y}^{(t)})\right\|^2\right] + 2\mathbb{E}\left[\left\|\nabla_y F(\tilde{\boldsymbol{x}},\boldsymbol{y}^{(t)}) - \frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\tilde{\boldsymbol{x}},\boldsymbol{y}_i^{(t)})\right\|^2\right] \\
&\leq 2L^2\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)} - \tilde{\boldsymbol{x}}\right\|^2\right] + 2L^2\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}_i^{(t)}\right\|^2\right] \\
&\leq 2L^2 S\eta_x^2(G_x^2+\sigma^2) + 2L^2\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}_i^{(t)}\right\|^2\right]
\end{aligned}$$

Putting these pieces together yields:

$$\begin{aligned}
\mathbb{E}[F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)})] &\leq \mathbb{E}[F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t+1)})] - \frac{\eta_y}{2}\mathbb{E}\left[\left\|\nabla_y F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)})\right\|^2\right] - \left(\frac{\eta_y}{2} - \frac{\eta_y^2 L}{2}\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\tilde{\boldsymbol{x}},\boldsymbol{y}_i^{(t)})\right\|^2\right] + \frac{\eta_y^2 L\sigma^2}{2n} \\
&+ \eta_y\left(L^2\eta_x^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right] + L^2\eta_x^2\frac{\sigma^2}{n}\right)
\end{aligned}$$

$$
\begin{aligned}
&+\eta_y\left(2L^2S\eta_x^2(G_x^2+\sigma^2)+2L^2\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\boldsymbol{y}^{(t)}-\boldsymbol{y}_i^{(t)}\right\|^2\right]\right)\\
&\leq \mathbb{E}[F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t+1)})]-\frac{\eta_y}{2}\mathbb{E}\left[\left\|\nabla_y F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)})\right\|^2\right]-\left(\frac{\eta_y}{2}-\frac{\eta_y^2L}{2}\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\nabla_y f_i(\tilde{\boldsymbol{x}},\boldsymbol{y}_i^{(t)})\right\|^2\right]+\frac{\eta_y^2L\sigma^2}{2n}\\
&+\eta_y\left(L^2\eta_x^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right]+L^2\eta_x^2\frac{\sigma^2}{n}\right)\\
&+\eta_y\left(2L^2S\eta_x^2(G_x^2+\sigma^2)+2L^2\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\boldsymbol{y}^{(t)}-\boldsymbol{y}_i^{(t)}\right\|^2\right]\right).
\end{aligned}
$$

Now, applying the PL condition to substitute $\left\|\nabla_y F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)})\right\|^2$:

$$
\left\|\nabla_y F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)})\right\|^2\geq 2\mu(\Phi(\boldsymbol{x}^{(t+1)})-F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)})). \tag{17}
$$

Thus we have:

$$
\begin{aligned}
\eta_y\mu\mathbb{E}\left[(\Phi(\boldsymbol{x}^{(t+1)})-F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)}))\right]&\leq \mathbb{E}[F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t+1)})]-\mathbb{E}[F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)})]+\frac{\eta_y^2L\sigma^2}{2n}\\
&+\eta_y\left(L^2\eta_x^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right]+L^2\eta_x^2\frac{\sigma^2}{n}\right)\\
&+\eta_y\left(2L^2S\eta_x^2(G_x^2+\sigma^2)+2L^2\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\boldsymbol{y}^{(t)}-\boldsymbol{y}_i^{(t)}\right\|^2\right]\right).
\end{aligned}
$$

Re-arranging the terms yields:

$$
\begin{aligned}
\mathbb{E}\left[(\Phi(\boldsymbol{x}^{(t+1)})-F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t+1)}))\right]&\leq(1-\mu\eta_y)\mathbb{E}\left[(\Phi(\boldsymbol{x}^{(t+1)})-F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)}))\right]+\frac{\eta_y^2L\sigma^2}{2n}\\
&+\eta_y\left(L^2\eta_x^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right]+L^2\eta_x^2\frac{\sigma^2}{n}\right)\\
&+\eta_y\left(2L^2S\eta_x^2(G_x^2+\sigma^2)+2L^2\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\boldsymbol{y}^{(t)}-\boldsymbol{y}_i^{(t)}\right\|^2\right]\right).
\end{aligned}
$$

Notice that in RHS:

$$
\mathbb{E}[\Phi(\boldsymbol{x}^{(t+1)})-F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)})]=\mathbb{E}[\Phi(\boldsymbol{x}^{(t)})-F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})]+\underbrace{\mathbb{E}[\Phi(\boldsymbol{x}^{(t+1)})-\Phi(\boldsymbol{x}^{(t)})]}_{T_3}+\underbrace{\mathbb{E}[F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})-F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)})]}_{T_4} \tag{18}
$$

According to Lemma C.1 we can bound T_3 as:

$$
\begin{aligned}
T_3\leq&-\frac{\eta_x}{2}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right]-\left(\frac{\eta_x}{2}-\frac{\beta\eta_x^2}{2}\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right]\\
&+\frac{2\eta_xL^2}{\mu}\mathbb{E}\left[(\Phi(\boldsymbol{x}^{(t)})-F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)}))\right]+2\eta_xL^2\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)}\right]+\frac{\beta\eta_x^2\sigma^2}{2n}.
\end{aligned}
$$

For T_4, applying smoothness of $F(\cdot,\boldsymbol{y}^{(t)})$ gives:

$$
T_4=\mathbb{E}[F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})-F(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^{(t)})]\leq\mathbb{E}[-\left\langle\nabla_x F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)}),\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t)}\right\rangle]+\frac{L}{2}\mathbb{E}\left[\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t)}\right\|^2\right]
$$

$$
\begin{aligned}
&= \eta_x \mathbb{E}\left[\left\langle \nabla_x F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}), \frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)})\right\rangle\right] + \frac{\eta_x^2 L}{2}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)})\right\|^2\right] + \frac{\eta_x^2 L\sigma^2}{2n} \\
&\leq \frac{1}{2}\eta_x \mathbb{E}\left[\left\|\nabla_x F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\right\|^2\right] + \frac{1}{2}\eta_x \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)})\right\|^2\right] + \frac{\eta_x^2 L}{2}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)})\right\|^2\right] + \frac{\eta_x^2 L\sigma^2}{2n} \\
&\leq \eta_x \mathbb{E}\left[\left\|\nabla \Phi(\boldsymbol{x}^{(t)})\right\|^2\right] + \eta_x \mathbb{E}\left[\left\|\nabla_x F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right] + \left(\frac{1}{2}\eta_x + \frac{\eta_x^2 L}{2}\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)})\right\|^2\right] + \frac{\eta_x^2 L\sigma^2}{2n}.
\end{aligned}
$$

For $\mathbb{E}\left[\left\|\nabla_x F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right]$, we apply the smoothness of F and quadratic growth of $F(\boldsymbol{x}, \cdot)$ to get:

$$
\mathbb{E}\left[\left\|\nabla_x F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right] \leq L^2 \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*(\boldsymbol{x}^{(t)})\right\|^2\right] \leq \frac{2L^2}{\mu}\mathbb{E}\left[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*(\boldsymbol{x}^{(t)})) - F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\right].
$$

Using above bound to replace $\mathbb{E}\left[\left\|\nabla_x F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right]$ we can finally bound T_4 as:

$$
\begin{aligned}
T_4 \leq\ & \eta_x \mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right] + \eta_x \frac{2L^2}{\mu}\mathbb{E}\left[\Phi(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\right] \\
&+ \left(\frac{1}{2}\eta_x + \frac{\eta_x^2 L}{2}\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)})\right\|^2\right] + \frac{\eta_x^2 L\sigma^2}{2n}.
\end{aligned}
$$

Plugging T_3 and T_4 back yields:

$$
\begin{aligned}
&\mathbb{E}\left[(\Phi(\boldsymbol{x}^{(t+1)}) - F(\boldsymbol{x}^{(t+1)}, \boldsymbol{y}^{(t+1)}))\right] \\
&\leq (1-\mu\eta_y)\left(1 + \eta_x \frac{4L^2}{\mu}\right)\mathbb{E}\left[(\Phi(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}))\right] + \frac{\eta_y^2 L\sigma^2}{2n} \\
&\quad + (1-\mu\eta_y)\left(\eta_x \mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right] + \left(\frac{1}{2}\eta_x + \frac{\eta_x^2 L}{2}\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)})\right\|^2\right] + \frac{\eta_x^2 L\sigma^2}{2n}\right) \\
&\quad + (1-\mu\eta_y)\left(-\frac{\eta_x}{2}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right] - \left(\frac{\eta_x}{2} - \frac{\beta}{2}\eta_x^2\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)})\right\|^2\right] + 2\eta_x L^2 \mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right] + \frac{\beta\eta_x^2\sigma^2}{2n}\right) \\
&\quad + \eta_y\left(L^2\eta_x^2 \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)})\right\|^2\right] + L^2\eta_x^2\frac{\sigma^2}{n}\right) \\
&\quad + \eta_y\left(2L^2 S\eta_x^2(G_x^2 + \sigma^2) + 2L^2\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}_i^{(t)}\right\|^2\right]\right) \\
&\leq \left(1 - \frac{\mu\eta_y}{2}\right)\mathbb{E}\left[(\Phi(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}))\right] + \frac{\eta_y^2 L\sigma^2}{2n} \\
&\quad + (1-\mu\eta_y)\left(\frac{1}{2}\eta_x \mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right] + \frac{\eta_x^2 L\sigma^2}{2n}\right) + \left[(1-\mu\eta_y)\left(\frac{\eta_x^2 L}{2} + \frac{\beta\eta_x^2}{2}\right) + \eta_y L^2\eta_x^2\right]\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)})\right\|^2\right] \\
&\quad + (1-\mu\eta_y)\left(2\eta_x L^2 \mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)} + \delta_{\boldsymbol{y}}^{(t)}\right] + \frac{\beta\eta_x^2\sigma^2}{2n}\right) \\
&\quad + \eta_y\left(L^2\eta_x^2\frac{\sigma^2}{n} + 2L^2 S\eta_x^2(G_x^2 + \sigma^2) + 2L^2\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}_i^{(t)}\right\|^2\right]\right),
\end{aligned}
$$

where we use the fact $(1-\mu\eta_y)(1+\frac{4L^2\eta_x}{\mu}) \leq (1-\frac{\mu\eta_y}{2})$ due to $\eta_x \leq \frac{\mu\eta_y}{2(4L^2/\mu - 4L^2\eta_y)}$. Denote $A_t = \mathbb{E}\left[(\Phi(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}))\right]$. It is obvious that $A_t \geq 0$ for all t. Then, based on the above inequality and do the summation:

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} A_t \leq & \frac{1}{T}\sum_{t=0}^{T-1}\left(1-\frac{\mu\eta_y}{2}\right) A_t + \frac{1}{T}\sum_{t=0}^{T-1}\left[(1-\mu\eta_y)\left(\frac{\eta_x^2 L}{2}+\frac{\beta\eta_x^2}{2}\right)+\eta_y L^2\eta_x^2\right]\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right] \\
& + \frac{1}{T}\sum_{t=0}^{T-1}(1-\mu\eta_y)\left(\frac{1}{2}\eta_x\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right]+\frac{\eta_x^2 L\sigma^2}{2n}\right)+\frac{\eta_y^2 L\sigma^2}{2n} \\
& + \frac{1}{T}\sum_{t=0}^{T-1}(1-\mu\eta_y)\left(2\eta_x L^2\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)}\right]+\frac{\beta\eta_x^2\sigma^2}{2n}\right) \\
& + \frac{1}{T}\sum_{t=0}^{T-1}\eta_y\left(L^2\eta_x^2\frac{\sigma^2}{n}+2L^2 S\eta_x^2(G_x^2+\sigma^2)+2L^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\boldsymbol{y}^{(t)}-\boldsymbol{y}_i^{(t)}\right\|^2\right]\right) \\
& \leq (1-\frac{\mu\eta_y}{2})\frac{1}{T}\left(A_0+\sum_{t=1}^{T}A_t\right)+\frac{\eta_y^2 L\sigma^2}{2n} \\
& + \left[(1-\mu\eta_y)\left(\frac{\eta_x^2 L}{2}+\frac{\beta\eta_x^2}{2}\right)+\eta_y L^2\eta_x^2\right]\frac{1}{T}\left(\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(0)},\boldsymbol{y}_i^{(0)})\right\|^2\right]+\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right]\right) \\
& + (1-\mu\eta_y)\frac{1}{T}\left(\sum_{t=1}^{T}\left(\frac{1}{2}\eta_x\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right]+\frac{\eta_x^2 L\sigma^2}{2n}\right)+\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(0)})\right\|^2\right]\right) \\
& + (1-\mu\eta_y)\frac{1}{T}\sum_{t=1}^{T}\left(2\eta_x L^2\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)}\right]+\frac{\beta\eta_x^2\sigma^2}{2n}\right) \\
& + \eta_y\frac{1}{T}\sum_{t=1}^{T}\left(L^2\eta_x^2\frac{\sigma^2}{n}+2L^2 S\eta_x^2(G_x^2+\sigma^2)+2L^2\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)}\right]\right)
\end{aligned}
$$

Re-arranging the terms will conclude the proof:

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} A_t \leq & \frac{2A_0}{\mu\eta_y T}+\frac{\eta_y L\sigma^2}{n} \\
& + \left[\frac{2(1-\mu\eta_y)}{\mu\eta_y}\left(\frac{\eta_x^2 L}{2}+\frac{\beta\eta_x^2}{2}\right)+L^2\eta_x^2\right]\frac{1}{T}\left(\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(0)},\boldsymbol{y}_i^{(0)})\right\|^2\right]+\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right]\right) \\
& + \frac{2(1-\mu\eta_y)}{\mu\eta_y}\frac{1}{T}\left(\sum_{t=1}^{T}\left(\frac{1}{2}\eta_x\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right]+\frac{\eta_x^2 L\sigma^2}{2n}\right)+\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(0)})\right\|^2\right]\right) \\
& + \frac{2(1-\mu\eta_y)}{\mu\eta_y}\frac{1}{T}\sum_{t=1}^{T}\left(2\eta_x L^2\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)}\right]+\frac{\beta\eta_x^2\sigma^2}{2n}\right) \\
& + \frac{2}{\mu T}\sum_{t=1}^{T}\left(L^2\eta_x^2\frac{\sigma^2}{n}+2L^2 S\eta_x^2(G_x^2+\sigma^2)+2L^2\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)}\right]\right).
\end{aligned}
$$

□

The next lemma bounds the local model deviations on nonconvex-PL objective.

Lemma C.3. *For local-SGDA+, under assumptions of Theorem 6.1, the following statement holds true:*

$$
\frac{1}{T}\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}_i^{(t)}\right\|^2\right]+\mathbb{E}\left[\left\|\boldsymbol{y}^{(t)}-\boldsymbol{y}_i^{(t)}\right\|^2\right]\leq 10\tau^2(\eta_x^2+\eta_y^2)\left(\sigma^2+\frac{\sigma^2}{n}\right)+10\tau^2\eta_x^2\zeta_x+10\tau^2\eta_y^2\zeta_y.
$$

Proof. Similarly, for the second statement, we define $\gamma^t = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}_i^{(t)}\right\|^2\right] + \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}_i^{(t)}\right\|^2\right]$, then we have:

$$\begin{aligned}
\gamma^t \leq & \frac{1}{n}\sum_{i=1}^n \frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[\left\|\boldsymbol{x}^{r\tau} - \sum_{j=r\tau}^{(r+1)\tau} \eta_x \nabla_x f_k(\boldsymbol{x}_k^{(j)}, \boldsymbol{y}_k^{(j)}; \xi_k^j) - \left(\boldsymbol{x}^{r\tau} - \sum_{j=r\tau}^{(r+1)\tau} \eta_x \nabla_x f_i(\boldsymbol{x}_i^{(j)}, \boldsymbol{y}_i^{(j)}; \xi_i^j)\right)\right\|^2\right] \\
& + \frac{1}{n}\sum_{i=1}^n \frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[\left\|\boldsymbol{y}^{r\tau} - \sum_{j=r\tau}^{(r+1)\tau} \eta_y \nabla_y f_k(\tilde{\boldsymbol{x}}, \boldsymbol{y}_k^{(j)}; \xi_k^j) - \left(\boldsymbol{y}^{r\tau} - \sum_{j=r\tau}^{(r+1)\tau} \eta_y \nabla_y f_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}_i^{(j)}; \xi_i^j)\right)\right\|^2\right] \\
\leq & \tau \sum_{j=r\tau}^{(r+1)\tau} \frac{\eta_x^2}{n}\sum_{i=1}^n \frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[\left\|\nabla_x f_k(\boldsymbol{x}_k^{(j)}, \boldsymbol{y}_k^{(j)}; \xi_k^j) - \nabla_x f_i(\boldsymbol{x}_i^{(j)}, \boldsymbol{y}_i^{(j)}; \xi_i^j)\right\|^2\right] \\
& + \tau \sum_{j=r\tau}^{(r+1)\tau} \frac{\eta_y}{n}\sum_{i=1}^n \frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[\left\|\nabla_y f_k(\tilde{\boldsymbol{x}}, \boldsymbol{y}_k^{(j)}; \xi_k^j) - \nabla_y f_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}_i^{(j)}; \xi_i^j)\right\|^2\right] \\
\leq & \tau \sum_{j=r\tau}^{(r+1)\tau} \frac{\eta_x^2}{n}\sum_{i=1}^n \frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[\left\|\nabla_x f_k(\boldsymbol{x}_k^{(j)}, \boldsymbol{y}_k^{(j)}; \xi_k^j) - \nabla_x f_k(\boldsymbol{x}_k^{(j)}, \boldsymbol{y}_k^{(j)}) + \nabla_x f_k(\boldsymbol{x}_k^{(j)}, \boldsymbol{y}_k^{(j)}) - \nabla_x f_k(\boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)})\right.\right. \\
& \left.\left. + \nabla_x f_k(\boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)}) - \nabla_x f_i(\boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)}) + \nabla_x f_i(\boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)}) - \nabla_x f_i(\boldsymbol{x}_i^{(j)}, \boldsymbol{y}_i^{(j)}) + \nabla_x f_i(\boldsymbol{x}_i^{(j)}, \boldsymbol{y}_i^{(j)}) - \nabla_x f_i(\boldsymbol{x}_i^{(j)}, \boldsymbol{y}_i^{(j)}; \xi_i^t)\right\|^2\right] \\
& + \tau \sum_{j=r\tau}^{(r+1)\tau} \frac{\eta_y^2}{n}\sum_{i=1}^n \frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[\left\|\nabla_y f_k(\tilde{\boldsymbol{x}}, \boldsymbol{y}_k^{(j)}; \xi_k^j) - \nabla_y f_k(\tilde{\boldsymbol{x}}, \boldsymbol{y}_k^{(j)}) + \nabla_y f_k(\tilde{\boldsymbol{x}}, \boldsymbol{y}_k^{(j)}) - \nabla_y f_k(\tilde{\boldsymbol{x}}, \boldsymbol{y}^{(j)})\right.\right. \\
& \left.\left. + \nabla_y f_k(\tilde{\boldsymbol{x}}, \boldsymbol{y}^{(j)}) - \nabla_y f_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}^{(j)}) + \nabla_y f_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}^{(j)}) - \nabla_y f_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}_i^{(j)}) + \nabla_y f_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}_i^{(j)}) - \nabla_y f_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}_i^{(j)}; \xi_i^j)\right\|^2\right] \\
\leq & \sum_{j=r\tau}^{(r+1)\tau} 5\eta_x^2\left(\sigma^2 + \frac{\sigma^2}{n} + 2L^2\gamma^j + \zeta_x\right) + 5\eta_y^2\left(\sigma^2 + \frac{\sigma^2}{n} + 2L^2\gamma^j + \zeta_y\right).
\end{aligned}$$

Summing over t from $r\tau$ to $(r+1)\tau$ yields:

$$\begin{aligned}
\sum_{t=r\tau}^{(r+1)\tau} \gamma^t &\leq \sum_{t=r\tau}^{(r+1)\tau}\sum_{j=r\tau}^{(r+1)\tau} 5\tau\eta_x^2\left(\sigma^2 + \frac{\sigma^2}{n} + 2L^2\gamma^j + \zeta_x\right) + 5\tau\eta_y^2\left(\sigma^2 + \frac{\sigma^2}{n} + 2L^2\gamma^j + \zeta_y\right) \\
&\leq 10L^2\tau^2(\eta_x^2 + \eta_y^2)\sum_{j=r\tau}^{(r+1)\tau}\gamma^j + 5\tau^2(\eta_x^2 + \eta_y^2)\left(\sigma^2 + \frac{\sigma^2}{n}\right) + 5\tau^2\eta_x^2\zeta_x + 5\tau^2\eta_y^2\zeta_y. \qquad (19)
\end{aligned}$$

Since $10L^2\tau^2(\eta_x^2 + \eta_y^2) \leq \frac{1}{2}$, by re-arranging the terms we have:

$$\sum_{t=r\tau+1}^{(r+1)\tau} \gamma^t \leq 10\tau^3(\eta_x^2 + \eta_y^2)\left(\sigma^2 + \frac{\sigma^2}{n}\right) + 10\tau^3\eta_x^2\zeta_x + 10\tau^3\eta_y^2\zeta_y.$$

Summing over r from 0 to $T/\tau - 1$, and dividing both sides by T can conclude the proof of the first statement:

$$\frac{1}{T}\sum_{t=1}^{T}\gamma^t \leq 10\tau^2(\eta_x^2 + \eta_y^2)\left(\sigma^2 + \frac{\sigma^2}{n}\right) + 10\tau^2\eta_x^2\zeta_x + 10\tau^2\eta_y^2\zeta_y.$$

□

C.3 Proof of Theorem 6.1

According to Lemma C.1, we sum over $t=1$ to T, and divide both sides with T:

$$\frac{1}{T}\left(\mathbb{E}\left[\Phi(\boldsymbol{x}^{(T+1)})\right]-\mathbb{E}\left[\Phi(\boldsymbol{x}^{(1)})\right]\right)\leq-\frac{\eta_x}{2}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right]-\left(\frac{\eta_x}{2}-\frac{\beta\eta_x^2}{2}\right)\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right]$$
$$+\frac{2\eta_x L^2}{\mu}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[(\Phi(\boldsymbol{x}^{(t)})-F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)}))\right]+\frac{1}{T}\sum_{t=1}^{T}2\eta_x L^2\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)}\right]+\frac{\beta\eta_x^2\sigma^2}{2n}.$$

Plugging in Lemma C.2 yields:

$$\frac{1}{T}\left(\mathbb{E}\left[\Phi(\boldsymbol{x}^{(T+1)})\right]-\mathbb{E}\left[\Phi(\boldsymbol{x}^{(1)})\right]\right)$$
$$\leq-\underbrace{\left(\frac{\eta_x}{2}-\frac{4(1-\mu\eta_y)L^2}{\mu^2\eta_y}\eta_x^2\right)}_{\spadesuit}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right]+\frac{8\eta_x^3L^4}{\mu^2}S(G_x^2+\sigma^2)+\frac{2\eta_x L^2}{\mu}\frac{\eta_y L\sigma^2}{n}$$
$$-\underbrace{\left(\frac{\eta_x}{2}-\frac{\beta\eta_x^2}{2}-\frac{2\eta_x L^2}{\mu}\left[\frac{2(1-\mu\eta_y)}{\mu\eta_y}\left(\frac{\eta_x^2 L}{2}+\frac{\beta\eta_x^2}{2}\right)+L^2\eta_x^2\right]\right)}_{\clubsuit}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}_i^{(t)},\boldsymbol{y}_i^{(t)})\right\|^2\right]$$
$$+\left(2\eta_x L^2+\frac{8\eta_x L^4}{\mu^2}+\frac{8(1-\mu\eta_y)L^4}{\mu^2\eta_y}\eta_x^2\right)\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\delta_{\boldsymbol{x}}^{(t)}+\delta_{\boldsymbol{y}}^{(t)}\right]+\left(\frac{8(1-\mu\eta_y)\eta_x L^2(L+\beta)}{\mu^2\eta_y}+\beta+\frac{8\eta_x L^4}{\mu^2}\right)\frac{\eta_x^2\sigma^2}{2n}$$
$$+\frac{2\eta_x L^2}{\mu}\left(\frac{\mathbb{E}\left[\Phi(\boldsymbol{x}^{(0)})\right]-\mathbb{E}\left[F(\boldsymbol{x}^{(0)},\boldsymbol{y}^{(0)})\right]}{\mu\eta_y T}+\left[\frac{2(1-\mu\eta_y)}{\mu\eta_y}\left(\frac{\eta_x^2 L}{2}+\frac{\beta\eta_x^2}{2}\right)+L^2\eta_x^2\right]\frac{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\boldsymbol{x}^{(0)},\boldsymbol{y}^{(0)})\right\|^2\right]}{T}\right)$$
$$+\frac{2\eta_x L^2}{\mu}\frac{2(1-\mu\eta_y)}{\mu\eta_y}\left(\frac{\mathbb{E}\left[\left\|\nabla_x\Phi(\boldsymbol{x}^{(0)})\right\|^2\right]}{T}\right).$$

Recall that we choose: $\eta_x=\frac{n^{1/3}}{LT^{2/3}}$, $\eta_y=\frac{n^{1/3}}{LT^{1/2}}$, $\tau=\frac{T^{1/3}}{n^{2/3}}$, $S=\frac{T^{1/3}}{n^{2/3}}$, and

$$T\geq\max\left\{\left(\frac{\beta n^{1/3}}{2L}+\sqrt{\frac{\beta^2n^{2/3}}{4L^2}+\frac{8L(L+\beta)n^{1/3}}{\mu^2}+\frac{4L^2n^{2/3}}{\mu}}\right)^{3/2},(8\kappa^2)^6\right\},$$

so we know that $\spadesuit\geq\frac{\eta_x}{4}$ and $\clubsuit\geq 0$. Plugging in η_x,η_y,τ,S, and plugging in Lemma C.3 will conclude the proof for Theorem 6.1:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}^{(t)})\right\|^2\right]\leq O\left(\frac{\beta\sigma^2}{(nT)^{1/3}}+\frac{\kappa^2L^2\zeta_y}{n^{2/3}T^{1/3}}+\frac{\kappa^2L^2\zeta_x}{n^{2/3}T}+\frac{\kappa^2L^2G_x^2}{T}+\frac{\kappa^2}{n^{1/3}T^{1/2}}\right).\tag{20}$$

□

D Proof of Local SGDA+ under Nonconvex-One-Point-Concave Setting

D.1 Overview of the proof techniques

In this section we are going to present the proof of convergence of local SGDA+, under the setting that F is nonconvex in $\boldsymbol{x}$ but one point concave in $\boldsymbol{y}$. In this setting, $\Phi(\boldsymbol{x})$ is no longer smooth any more, and $\boldsymbol{y}^*(\boldsymbol{x})$ is not Lipschitz. As we mentioned in the main paper, we study the Moreau evenlope function: $\Phi_{1/2L}(\boldsymbol{x})$. The proof mainly contains two parts: **one iteration analysis of Moreau envelope** and **Convergence of SGA under one point concave condition**.

Step I: One iteration analysis of Moreau envelope. By examining one iteration of local SGDA+, we have the following relation:

$$\mathbb{E}[\Phi_{1/2L}(\boldsymbol{x}^{(t)})] \leq \mathbb{E}\left[\Phi_{1/2L}(\boldsymbol{x}^{(t-1)})\right] + L\eta_x^2(G_x^2+\sigma_x^2) + 2\eta_x L^2 \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|\boldsymbol{x}_i^{(t-1)} - \boldsymbol{x}^{(t-1)}\right\|^2 + \left\|\boldsymbol{y}_i^{(t-1)} - \boldsymbol{y}^{(t-1)}\right\|^2\right]$$
$$+ 2L\eta_x\left(\mathbb{E}\left[\Phi(\boldsymbol{x}^{(t-1)})\right] - \mathbb{E}\left[F(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)})\right]\right) - \frac{\eta_x}{8}\mathbb{E}\left[\left\|\nabla\Phi_{1/2L}(\boldsymbol{x}^{(t-1)})\right\|^2\right].$$

It turns out our next job is to bound local model deviation $\mathbb{E}\left[\left\|\boldsymbol{x}_i^{(t-1)} - \boldsymbol{x}^{(t-1)}\right\| + \left\|\boldsymbol{y}_i^{(t-1)} - \boldsymbol{y}^{(t-1)}\right\|\right]$ and the gap $\mathbb{E}[\Phi(\boldsymbol{x}^{(t-1)})] - \mathbb{E}[F(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)})]$. The the analysis of deviation term is similar to what we did in nonconvex-strongly-concave setting. The remaining tricky part is how to bound $\mathbb{E}[\Phi(\boldsymbol{x}^{(t-1)})] - \mathbb{E}[F(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)})]$.

Step II: Convergence of SGA under one point concave condition. To deal with $\mathbb{E}[\Phi(\boldsymbol{x}^{(t)})] - \mathbb{E}[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})]$, we first notice that:

$$\begin{aligned}
\mathbb{E}[\Phi(\boldsymbol{x}^{(t)})] - \mathbb{E}[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})] &= \mathbb{E}[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*(\boldsymbol{x}^t))] - \mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\tilde{\boldsymbol{x}}))] + \mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\tilde{\boldsymbol{x}}))] - \mathbb{E}[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})] \\
&\leq \mathbb{E}[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*(\boldsymbol{x}^t))] - \mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\boldsymbol{x}^t))] + \mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\tilde{\boldsymbol{x}}))] - \mathbb{E}[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})] \\
&\leq \underbrace{\mathbb{E}[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*(\boldsymbol{x}^t))] - \mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\boldsymbol{x}^t))]}_{T_1} + \underbrace{\mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\tilde{\boldsymbol{x}}))] - \mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^{(t)})]}_{T_2} \\
&\quad + \underbrace{\mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^{(t)})] - \mathbb{E}[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})]}_{T_3}.
\end{aligned}$$

According to the Lipschitz continuity of F, and the fact that $\tilde{\boldsymbol{x}}$ will be updated every S iterations, we can bound T_1 and T_3 by $\eta_x S G_x\sqrt{G_x^2+\sigma^2}$.

The tricky part is to handle T_2. Basically fixing $\tilde{\boldsymbol{x}}$, we wish to know how fast $\mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^{(t)})]$ converges to $\mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\tilde{\boldsymbol{x}}))]$. Thanks to one point concave property and the updating rule of local SGDA+ where we fixed $\tilde{\boldsymbol{x}}$ while updating $\boldsymbol{y}$, we can show that:

$$\sum_{t=kS+1}^{(k+1)S}\mathbb{E}\left[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\tilde{\boldsymbol{x}})) - F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^{(t)})\right] \leq \frac{D}{\eta_y} + L\sum_{t=kS+1}^{(k+1)S}\frac{1}{n}\mathbb{E}\left[\left\|\boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2\right] + 2\eta_y L^2\sum_{t=kS+1}^{(k+1)S}\frac{1}{n}\mathbb{E}\left[\left\|\boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2\right] + \frac{\eta_y S\sigma^2}{n}.$$

Putting these pieces together will conclude the proof.

D.2 Proof of technical lemmas

Lemma D.1 (One iteration analysis)**.** *For local SGDA+, under Theorem 6.2's assumption, the following statement holds:*

$$\mathbb{E}[\Phi_{1/2L}(\boldsymbol{x}^{(t)})] \leq \mathbb{E}\left[\Phi_{1/2L}(\boldsymbol{x}^{(t-1)})\right] + L\eta_x^2(G_x^2+\sigma_x^2) + 2\eta_x L^2 \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|\boldsymbol{x}_i^{(t-1)} - \boldsymbol{x}^{(t-1)}\right\|^2 + \left\|\boldsymbol{y}_i^{(t-1)} - \boldsymbol{y}^{(t-1)}\right\|^2\right]$$
$$+ 2L\eta_x\left(\mathbb{E}\left[\Phi(\boldsymbol{x}^{(t-1)})\right] - \mathbb{E}\left[F(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)})\right]\right) - \frac{\eta_x}{8}\mathbb{E}\left[\left\|\nabla\Phi_{1/2L}(\boldsymbol{x}^{(t-1)})\right\|^2\right].$$

Proof. Define $\hat{\boldsymbol{x}}^{(t)} = \arg\min_{\boldsymbol{x}\in\mathcal{X}} \Phi(\boldsymbol{x}) + L\|\boldsymbol{x} - \boldsymbol{x}^{(t)}\|^2$, the by the definition of $\Phi_{1/2L}$ we have:

$$\Phi_{1/2L}(\boldsymbol{x}^{(t)}) \leq \Phi(\hat{\boldsymbol{x}}^{(t-1)}) + L\|\hat{\boldsymbol{x}}^{(t-1)} - \boldsymbol{x}^{(t)}\|^2. \tag{21}$$

Meanwhile according to updating rule we have:

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{x}}^{(t-1)} - \boldsymbol{x}^{(t)}\right\|^2\right] = \mathbb{E}\left[\left\|\boldsymbol{x}^{(t-1)} - \eta_x\frac{1}{n}\sum_{i=1}^n\nabla_x f_i(\boldsymbol{x}_i^{(t-1)}, \boldsymbol{y}_i^{(t-1)}; \xi_i^{(t)})\right\|^2\right]$$

$$\leq \mathbb{E}\left[\left\|\hat{\boldsymbol{x}}^{(t-1)}-\boldsymbol{x}^{(t-1)}\right\|^2\right]+\eta_x^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t-1)},\boldsymbol{y}_i^{(t-1)};\xi_i^{(t)})\right\|^2\right]$$
$$+2\eta_x\mathbb{E}\left[\left\langle \hat{\boldsymbol{x}}^{(t-1)}-\boldsymbol{x}^{(t-1)},\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}_i^{(t-1)},\boldsymbol{y}_i^{(t-1)})\right\rangle\right]$$
$$\leq \mathbb{E}\left[\left\|\hat{\boldsymbol{x}}^{(t-1)}-\boldsymbol{x}^{(t-1)}\right\|^2\right]+\eta_x^2(G_w^2+\sigma_w^2)+2\eta_x\left\langle \hat{\boldsymbol{x}}^{(t-1)}-\boldsymbol{x}^{(t-1)},\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)})\right\rangle$$
$$+\eta_x\left(\frac{L}{2}\mathbb{E}\left[\left\|\hat{\boldsymbol{x}}^{(t-1)}-\boldsymbol{x}^{(t-1)}\right\|^2\right]+\frac{2}{L}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|\nabla_x f_i(\boldsymbol{x}_i^{(t-1)},\boldsymbol{y}_i^{(t-1)})-\nabla_x f_i(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)})\right\|^2\right]\right)$$
$$\leq \mathbb{E}\left[\left\|\hat{\boldsymbol{x}}^{(t-1)}-\boldsymbol{x}^{(t-1)}\right\|^2\right]+\eta_x^2(G_w^2+\sigma_w^2)+\eta_x 2L\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\boldsymbol{x}_i^{(t-1)}-\boldsymbol{x}^{(t-1)}\right\|^2+\left\|\boldsymbol{y}_i^{(t-1)}-\boldsymbol{y}^{(t-1)}\right\|^2\right]$$
$$+2\eta_x\mathbb{E}\left[\left\langle \hat{\boldsymbol{x}}^{(t-1)}-\boldsymbol{x}^{(t-1)},\nabla_x F(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)})\right\rangle\right]+\frac{\eta_x L}{2}\mathbb{E}\left[\left\|\hat{\boldsymbol{x}}^{(t-1)}-\boldsymbol{x}^{(t-1)}\right\|^2\right]. \quad (22)$$

According to smoothness of F we obtain:

$$\mathbb{E}\left[\left\langle \hat{\boldsymbol{x}}^{(t-1)}-\boldsymbol{x}^{(t-1)},\nabla_x F(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)})\right\rangle\right]$$
$$\leq \mathbb{E}\left[F(\hat{\boldsymbol{x}}^{(t-1)},\boldsymbol{y}^{(t-1)})\right]-\mathbb{E}\left[F(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)})\right]+\frac{L}{2}\mathbb{E}\left[\left\|\hat{\boldsymbol{x}}^{(t-1)}-\boldsymbol{x}^{(t-1)}\right\|^2\right]$$
$$\leq \mathbb{E}\left[\Phi(\hat{\boldsymbol{x}}^{(t-1)})\right]-\mathbb{E}\left[F(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)})\right]+\frac{L}{2}\mathbb{E}\left[\left\|\hat{\boldsymbol{x}}^{(t-1)}-\boldsymbol{x}^{(t-1)}\right\|^2\right]$$
$$\leq \underbrace{\mathbb{E}\left[\Phi(\hat{\boldsymbol{x}}^{(t-1)})\right]+L\mathbb{E}\left[\left\|\hat{\boldsymbol{x}}^{(t-1)}-\boldsymbol{x}^{(t-1)}\right\|^2\right]}_{\leq \mathbb{E}\left[\Phi(\boldsymbol{x}^{(t-1)})\right]+L\mathbb{E}\left[\left\|\boldsymbol{x}^{(t-1)}-\boldsymbol{x}^{(t-1)}\right\|^2\right]}-\mathbb{E}\left[F(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)})\right]-\frac{L}{2}\mathbb{E}\left[\left\|\hat{\boldsymbol{x}}^{(t-1)}-\boldsymbol{x}^{(t-1)}\right\|^2\right]$$
$$\leq \mathbb{E}\left[\Phi(\boldsymbol{x}^{(t-1)})\right]-\mathbb{E}\left[F(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)})\right]-\frac{L}{2}\mathbb{E}\left[\left\|\hat{\boldsymbol{x}}^{(t-1)}-\boldsymbol{x}^{(t-1)}\right\|^2\right]. \quad (23)$$

Plugging (22) and (23) into (21) yields:

$$\mathbb{E}\left[\Phi_{1/2L}(\boldsymbol{x}^{(t)})\right]\leq \mathbb{E}\left[\Phi(\hat{\boldsymbol{x}}^{(t-1)})\right]+L\mathbb{E}\left[\left\|\hat{\boldsymbol{x}}^{(t-1)}-\boldsymbol{x}^{(t-1)}\right\|^2\right]+2\eta_x L^2\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\boldsymbol{x}_i^{(t-1)}-\boldsymbol{x}^{(t-1)}\right\|^2+\left\|\boldsymbol{y}_i^{(t-1)}-\boldsymbol{y}^{(t-1)}\right\|^2\right]$$
$$+2\eta_x L\left(\mathbb{E}\left[\Phi(\boldsymbol{x}^{(t-1)})\right]-\mathbb{E}\left[F(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)})\right]-\frac{L}{2}\mathbb{E}\left[\left\|\hat{\boldsymbol{x}}^{(t-1)}-\boldsymbol{x}^{(t-1)}\right\|^2\right]\right)$$
$$+L\eta_x^2(G_w^2+\sigma_w^2)+\frac{\eta_x L^2}{2}\mathbb{E}\left[\left\|\hat{\boldsymbol{x}}^{(t-1)}-\boldsymbol{x}^{(t-1)}\right\|^2\right]$$
$$\leq \mathbb{E}\left[\Phi_{1/2L}(\boldsymbol{x}^{(t-1)})\right]+L\eta_x^2(G_w^2+\sigma_w^2)+2\eta_x L^2\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\boldsymbol{x}_i^{(t-1)}-\boldsymbol{x}^{(t-1)}\right\|^2+\left\|\boldsymbol{y}_i^{(t-1)}-\boldsymbol{y}^{(t-1)}\right\|^2\right]$$
$$+2L\eta_x\left(\mathbb{E}\left[\Phi(\boldsymbol{x}^{(t-1)})\right]-\mathbb{E}\left[F(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)})\right]\right)-\frac{\eta_x}{8}\mathbb{E}\left[\left\|\nabla\Phi_{1/2L}(\boldsymbol{x}^{(t-1)})\right\|^2\right],$$

where we use the result from Lemma 2.8 in [29]: $\nabla\Phi_{1/2L}(\boldsymbol{x})=2L(\boldsymbol{x}-\hat{\boldsymbol{x}})$. □

The following lemma derives the convergence rate of the gap $\mathbb{E}[\Phi(\boldsymbol{x}^{(t)})]-\mathbb{E}[F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})]$.

Lemma D.2. *For local SGDA+, under Theorem 6.2's assumption, the following statement holds:*

$$\frac{1}{T}\sum_{t=1}^T\mathbb{E}[\Phi(\boldsymbol{x}^{(t)})]-\mathbb{E}[F(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)})]\leq 2\eta_x SG_x\sqrt{G_x^2+\sigma^2}+\frac{D}{S\eta_y}+(L+4\eta_y L^2)\frac{1}{T}\sum_{t=1}^T\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\boldsymbol{y}_i^{(t)}-\boldsymbol{y}^{(t)}\right\|^2\right]+\frac{\eta_y\sigma^2}{n}.$$

Proof. Consider $t = kS + 1$ to $(k+1)S$. Let $\tilde{\boldsymbol{x}}$ denote the latest snapshot iterate. Observe that:

$$\begin{aligned}
\mathbb{E}[\Phi(\boldsymbol{x}^{(t)})] - \mathbb{E}[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})] &\leq \mathbb{E}[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*(\boldsymbol{x}^t))] - \mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\boldsymbol{x}^t))] + \mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\tilde{\boldsymbol{x}}))] - \mathbb{E}[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})] \\
&\leq G_x \mathbb{E}\|\boldsymbol{x}^{(t)} - \tilde{\boldsymbol{x}}\| + \mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\tilde{\boldsymbol{x}})] - \mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^{(t)})] + \mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^{(t)})] - \mathbb{E}[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}] \\
&\leq 2\eta_x S G_x \sqrt{G_x^2 + \sigma^2} + \mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\tilde{\boldsymbol{x}})] - \mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^{(t)})]. \qquad (24)
\end{aligned}$$

where we use the fact $f(\cdot, \boldsymbol{y})$ is G_x-Lipschitz, so that:

$$\begin{aligned}
\mathbb{E}[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*(\boldsymbol{x}^t))] - \mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\boldsymbol{x}^t))] \leq G_x \mathbb{E}\|\boldsymbol{x}^{(t)} - \tilde{\boldsymbol{x}}\| \leq \eta_x S G_x \sqrt{G_x^2 + \sigma^2}, \\
\mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^{(t)})] - \mathbb{E}[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})] \leq G_x \mathbb{E}\|\boldsymbol{x}^{(t)} - \tilde{\boldsymbol{x}}\| \leq \eta_x S G_x \sqrt{G_x^2 + \sigma^2}.
\end{aligned}$$

Summing over $t = kS + 1$ to $(k+1)S$ in (24), and dividing both sides with T yields:

$$\sum_{t=kS}^{(k+1)S-1} \mathbb{E}[\Phi(\boldsymbol{x}^{(t)})] - \mathbb{E}[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})] \leq 2\eta_x S^2 G_x \sqrt{G_x^2 + \sigma^2} + \sum_{t=kS}^{(k+1)S-1} \mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\tilde{\boldsymbol{x}})] - \mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^{(t)})]. \qquad (25)$$

Now let us study the convergence of $\mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\tilde{\boldsymbol{x}})] - \mathbb{E}[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^{(t)})]$.

By the updating rule of $\boldsymbol{y}$ we have:

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{y}^{(t+1)} - \boldsymbol{y}^*(\tilde{\boldsymbol{x}})\|^2\right] \\
&= \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} + \eta_y \frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}_i^{(t)}; \xi_i^t) - \boldsymbol{y}^*(\tilde{\boldsymbol{x}})\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*(\tilde{\boldsymbol{x}})\right\|^2\right] + 2\eta_y \mathbb{E}\left[\left\langle \frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}_i^{(t)}; \xi_i^t), \boldsymbol{y}^{(t)} - \boldsymbol{y}^*(\tilde{\boldsymbol{x}})\right\rangle\right] + \eta_y^2 \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}_i^{(t)}; \xi_i^t)\right\|^2\right] \\
&\leq \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*(\tilde{\boldsymbol{x}})\right\|^2\right] + 2\eta_y \mathbb{E}\left[\left\langle \frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}_i^{(t)}), \boldsymbol{y}^{(t)} - \boldsymbol{y}_i^{(t)}\right\rangle\right] \\
&+ 2\eta_y \mathbb{E}\left[\left\langle \frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}_i^{(t)}), \boldsymbol{y}_i^{(t)} - \boldsymbol{y}^*(\tilde{\boldsymbol{x}})\right\rangle\right] + \eta_y^2 \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}_i^{(t)})\right\|^2\right] + \frac{\eta_y^2\sigma^2}{n}.
\end{aligned}$$

Applying one point concavity and L-smoothness of $f_i(\tilde{\boldsymbol{x}}, \cdot)$ we have:

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{y}^{(t+1)} - \boldsymbol{y}^*(\tilde{\boldsymbol{x}})\|^2\right] &\leq \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*(\tilde{\boldsymbol{x}})\right\|^2\right] + 2\eta_y \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[f_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}^{(t)}) - f_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\tilde{\boldsymbol{x}}))\right] + \eta_y L \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|\boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2\right] \\
&+ 4\eta_y^2 L \mathbb{E}\left[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\tilde{\boldsymbol{x}})) - F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^{(t)})\right] + 2\eta_y^2 L^2 \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\| \boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2\right] + \frac{\eta_y^2\sigma^2}{n} \\
&\leq \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*(\tilde{\boldsymbol{x}})\right\|^2\right] + \underbrace{(2\eta_y - 4\eta_y^2 L)}_{\geq \eta_y} \mathbb{E}\left[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^{(t)}) - F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\tilde{\boldsymbol{x}}))\right] + \frac{\eta_y^2\sigma^2}{n} \\
&+ \eta_y L \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|\boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2\right] + 2\eta_y^2 L^2 \frac{1}{n} \mathbb{E}\left[\left\| \boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2\right] \\
&\leq \mathbb{E}\left[\left\|\boldsymbol{y}^{(t)} - \boldsymbol{y}^*(\tilde{\boldsymbol{x}})\right\|^2\right] - \eta_y \mathbb{E}\left[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\tilde{\boldsymbol{x}})) - F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^{(t)})\right] + \frac{\eta_y^2\sigma^2}{n} \\
&+ \eta_y L \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|\boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2\right] + 2\eta_y^2 L^2 \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\| \boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2\right].
\end{aligned}$$

Re-arranging the terms, and summing $t = kS+1$ to $(k+1)S$ yields:

$$\sum_{t=kS+1}^{(k+1)S} \mathbb{E}\left[F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^*(\tilde{\boldsymbol{x}})) - F(\tilde{\boldsymbol{x}}, \boldsymbol{y}^{(t)})\right] \leq \frac{1}{\eta_y}\left(\mathbb{E}\left[\left\|\boldsymbol{y}^{(kS+1)} - \boldsymbol{y}^*(\tilde{\boldsymbol{x}})\right\|^2\right] - \mathbb{E}\left[\|\boldsymbol{y}^{((k+1)S)} - \boldsymbol{y}^*(\tilde{\boldsymbol{x}})\|^2\right]\right) + \frac{\eta_y S\sigma^2}{n}$$
$$+ L\sum_{t=kS+1}^{(k+1)S}\frac{1}{n}\mathbb{E}\left[\left\|\boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2\right] + 2\eta_y L^2 \sum_{t=kS+1}^{(k+1)S}\frac{1}{n}\mathbb{E}\left[\left\|\ \boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2\right]$$
$$\leq \frac{D}{\eta_y} + L\sum_{t=kS+1}^{(k+1)S}\frac{1}{n}\mathbb{E}\left[\left\|\boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2\right] + 2\eta_y L^2 \sum_{t=kS+1}^{(k+1)S}\frac{1}{n}\mathbb{E}\left[\left\|\ \boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2\right] + \frac{\eta_y S\sigma^2}{n}.$$

Plugging above bound into (25) yields:

$$\sum_{t=kS}^{(k+1)S-1} \mathbb{E}[\Phi(\boldsymbol{x}^{(t)})] - \mathbb{E}[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})] \leq 2\eta_x S^2 G_x\sqrt{G_x^2+\sigma^2} + \frac{D}{\eta_y} + (L + 4\eta_y L^2)\sum_{t=kS+1}^{(k+1)S}\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2\right] + \frac{S\eta_y\sigma^2}{n}.$$

Finally, summing $k = 0$ to $T/S - 1$, and dividing both sides by T will conclude the proof:

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}[\Phi(\boldsymbol{x}^{(t)})] - \mathbb{E}[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})] \leq 2\eta_x S G_x\sqrt{G_x^2+\sigma^2} + \frac{D}{S\eta_y} + (L + 4\eta_y L^2)\frac{1}{T}\sum_{t=1}^T\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2\right] + \frac{\eta_y\sigma^2}{n}.$$

□

D.3 Proof of Theorem 6.2

In this section we provide the full proof of Theorem 6.2. We first sum over $t = 1$ to T in Lemma D.1, and divide both sides with T:

$$\frac{1}{T}\sum_{t=1}^T\mathbb{E}\left[\left\|\nabla\Phi_{1/2L}(\boldsymbol{x}^{(t)})\right\|^2\right] \leq \frac{8\mathbb{E}[\Phi_{1/2L}(\boldsymbol{x}^{(0)})] - 8\mathbb{E}[\Phi_{1/2L}(\boldsymbol{x}^{(T)})]}{\eta_x T} + 16\frac{1}{T}\sum_{t=1}^T L^2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}^{(t)}\right\| + \left\|\boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t-1)}\right\|\right]$$
$$+ 16L\frac{1}{T}\sum_{t=1}^T\left(\mathbb{E}[\Phi(\boldsymbol{x}^{(t)})] - \mathbb{E}[F(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}]\right) + 8L\eta_x^2(G_x^2+\sigma^2).$$

Plugging in Lemma D.2 and C.3 yields:

$$\frac{1}{T}\sum_{t=1}^T\mathbb{E}\left[\left\|\nabla\Phi_{1/2L}(\boldsymbol{x}^{(t)})\right\|^2\right]$$
$$\leq \frac{8\mathbb{E}[\Phi_{1/2L}(\boldsymbol{x}^{(0)})]}{\eta_x T} + 16L^2\left(10\tau^2(\eta_x^2+\eta_y^2)\left(\sigma^2 + \frac{\sigma^2}{n}\right) + 10\tau^2\eta_x^2\zeta_x + 10\tau^2\eta_y^2\zeta_y\right) + 8L\eta_x(G_x^2+\sigma^2)$$
$$+ 8L\left(2\eta_x S G_x\sqrt{G_x^2+\sigma^2} + \frac{D}{S\eta_y} + (L+4\eta_y L^2)\frac{1}{T}\sum_{t=1}^T\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\boldsymbol{y}_i^{(t)} - \boldsymbol{y}^{(t)}\right\|^2\right] + \frac{\eta_y\sigma^2}{n}\right).$$
$$\leq \frac{8\mathbb{E}[\Phi_{1/2L}(\boldsymbol{x}^{(0)})]}{\eta_x T} + 16L^2\left(10\tau^2(\eta_x^2+\eta_y^2)\left(\sigma^2 + \frac{\sigma^2}{n}\right) + 10\tau^2\eta_x^2\zeta_x + 10\tau^2\eta_y^2\zeta_y\right) + 8L\eta_x(G_x^2+\sigma^2)$$
$$+ 8L\left(2\eta_x S G_x\sqrt{G_x^2+\sigma^2} + \frac{D}{S\eta_y} + (L+4\eta_y L^2)\left(10\tau^2(\eta_x^2+\eta_y^2)\left(\sigma^2 + \frac{\sigma^2}{n}\right) + 10\tau^2\eta_x^2\zeta_x + 10\tau^2\eta_y^2\zeta_y\right) + \frac{\eta_y\sigma^2}{n}\right)$$
$$\leq \frac{8\mathbb{E}[\Phi_{1/2L}(\boldsymbol{x}^{(0)})]}{\eta_x T} + (16L^2 + 8L(L+4\eta_y L^2))\left(10\tau^2(\eta_x^2+\eta_y^2)\left(\sigma^2 + \frac{\sigma^2}{n}\right) + 10\tau^2\eta_x^2\zeta_x + 10\tau^2\eta_y^2\zeta_y\right) + 8L\eta_x(G_x^2+\sigma^2)$$
$$+ 8L\left(2\eta_x S G_x\sqrt{G_x^2+\sigma^2} + \frac{D}{S\eta_y} + \frac{\eta_y\sigma^2}{n}\right)$$

If we choose $\eta_x = \frac{1}{LT^{\frac{5}{6}}}$, $\eta_y = \frac{1}{4LT^{\frac{1}{2}}}$, $\tau = T^{\frac{1}{3}}/n^{\frac{1}{6}}$, $S = T^{\frac{2}{3}}$ we recover the rate:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/2L}(\boldsymbol{x}^{(t)})\right\|^2\right] \leq O\left(\frac{L\sigma^2}{T^{\frac{1}{6}}}\right) + O\left(\frac{D}{T^{\frac{1}{6}}}\right) + O\left(\frac{L^2\sigma^2}{(nT)^{\frac{1}{3}}} + \frac{L^2\zeta_x}{n^{\frac{1}{3}}T} + \frac{L^2\zeta_y}{(nT)^{\frac{1}{3}}}\right) + O\left(\frac{LG_x^2}{T^{\frac{1}{6}}}\right) + O\left(\frac{\sigma^2}{nT^{\frac{1}{6}}}\right),$$

as stated by the theorem. □