A Convergence Analysis of Algorithm 1

First, we present the detailed statements Theorem 2.7.

Theorem A.1. *Let $F(\boldsymbol{x}_0) - F(\boldsymbol{x}^*) \leq \Delta_F$. Under Assumption 2.2,2.3 and consider Algorithm 1 with momentum method for Hessian inverse approximation, with $\eta_1 \leq \min\left\{\frac{\mu_f}{L_f^2}, \frac{1}{\mu_f}, \frac{4m}{\mu_f|I_t|}, \frac{B\epsilon^2}{96C_1\mu_f\sigma^2}\right\}$, $\eta_2 \leq \min\left\{\frac{\mu_g}{L_g^2}, \frac{2m}{|I_t|\mu_g}, \frac{\mu_g B\epsilon^2}{48C_2\sigma^2}\right\}$, $\beta_1 \leq \min\left\{1, \frac{B\epsilon^2}{96C_3\sigma^2}\right\}$, $\beta_0 \leq \frac{\min\{|I_t|,B\}\epsilon^2}{12C_4}$, $\eta_0 \leq \min\left\{\frac{1}{2L/f}, \frac{\beta_0}{\sqrt{80}L_F}, \frac{\eta_1\mu_f|I_t|}{\sqrt{640}m\sqrt{C_1}C_\alpha}, \frac{|I_t|\beta_1}{\sqrt{640}m\sqrt{C_3}L_{gyy}\sqrt{(1+C_y^2)}}, \frac{|I_t|\eta_2\mu_g}{\sqrt{160}m\sqrt{C_2}C_y}\right\}$, $T \geq \max\left\{\frac{30\Delta_F}{\eta_0\epsilon^2}, \frac{15\mathbb{E}[\|\nabla F(\boldsymbol{x}_0)-\boldsymbol{z}_1\|^2]}{\beta_0\epsilon^2}, \frac{60C_1\delta_{\alpha,0}}{\mu_f|I_t|\eta_1\epsilon^2}, \frac{30C_2\delta_{y,0}}{|I_t|\eta_2\mu_g\epsilon^2}, \frac{60C_3\delta_{gyy,0}}{|I_t|\beta_1\epsilon^2}\right\}$ we have*

$$\mathbb{E}[\|\nabla F(\boldsymbol{x}_\tau)\|^2] \leq \epsilon^2, \quad \mathbb{E}[\|\nabla F(\boldsymbol{x}_\tau) - \boldsymbol{z}_{\tau+1})\|^2] < 2\epsilon^2,$$

where τ is randomly sampled from $\{0,\ldots,T\}$, C_1, C_2, C_3, C_4 are constants defined in the proof, and L_F is the Lipschitz continuity constant of $\nabla F(\boldsymbol{x})$.

To prove Theorem A.1, we need the following Lemmas.

Lemma A.2. *Under Assumption 2.2 and 2.3, $F(x)$ is L_F-smooth for some constant $L_F \in \mathbb{R}$.*

Lemma A.3. *Consider the update $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_0 \boldsymbol{z}_{t+1}$. Then under Assumption 2.2, with $\eta_0 L_F \leq \frac{1}{2}$, we have*

$$F(\boldsymbol{x}_{t+1}) \leq F(\boldsymbol{x}_t) + \frac{\eta_0}{2}\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{z}_{t+1}\|^2 - \frac{\eta_0}{2}\|\nabla F(\boldsymbol{x}_t)\|^2 - \frac{\eta_0}{4}\|\boldsymbol{z}_{t+1}\|^2.$$

Lemma A.4. *[Lemma 4.3 [25]] Under Assumption 2.2, $\boldsymbol{y}_i(\boldsymbol{x})$ is $C_y = L_g/\mu_g$-Lipschitz-continuous for all i. Define $\alpha_i(\boldsymbol{x}, \boldsymbol{y}_i) := \arg\max_{\alpha_i \in \mathcal{A}} f_i(\boldsymbol{x}, \alpha_i, \boldsymbol{y}_i)$. Then $\alpha_i(\boldsymbol{x}, \boldsymbol{y})$ is $C_\alpha = L_f/\mu_f$-Lipschitz continuous.*

Proof of Theorem A.1. First, recall and define the following notations

$$\nabla F(\mathbf{x}_t) = \frac{1}{m}\sum_{i\in\mathcal{S}} \nabla_x f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t)) - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))[\nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1} \nabla_y f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t))$$

$$\nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t) := \frac{1}{m}\sum_{i\in\mathcal{S}} \nabla F_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) := \nabla_x f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i^t)\mathbb{E}_t[H_i^t]\nabla_y f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t)$$

$$\Delta^{t+1} = \frac{1}{|I_t|}\sum_{i\in I_t} \Delta_i^{t+1} := \nabla_x f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t; \mathcal{B}_i^t) - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i^t; \tilde{\mathcal{B}}_i^t) H_i^t \nabla_y f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t; \mathcal{B}_i^t)$$

Consider the update $\mathbf{z}_{t+1} = (1-\beta_0)\mathbf{z}_t + \beta_0\Delta^{t+1}$ in Algorithm 1, we have

$$\begin{aligned}
&\mathbb{E}_t[\|\nabla F(\mathbf{x}_t) - \mathbf{z}_{t+1}\|^2] \\
&= \mathbb{E}_t[\|\nabla F(\mathbf{x}_t) - (1-\beta_0)\mathbf{z}_t - \beta_0\Delta^{t+1}\|^2] \\
&= \mathbb{E}_t[\|(1-\beta_0)(\nabla F(\mathbf{x}_{t-1}) - \mathbf{z}_t) + (1-\beta_0)(\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})) + \beta_0(\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t)) \\
&\quad + \beta_0(\nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t) - \Delta^{t+1})\|^2] \\
&\overset{(a)}{=} \|(1-\beta_0)(\nabla F(\mathbf{x}_{t-1}) - \mathbf{z}_t) + (1-\beta_0)(\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})) + \beta_0(\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t))\|^2 \\
&\quad + \beta_0^2\mathbb{E}_t[\|\nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t) - \Delta^{t+1}\|^2] \\
&\overset{(b)}{\leq} (1+\beta_0)(1-\beta_0)^2\|\nabla F(\mathbf{x}_{t-1}) - \mathbf{z}_t\|^2 + 2(1+\frac{1}{\beta_0})\left[\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})\|^2 + \beta_0^2\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t)\|^2\right] \\
&\quad + \beta_0^2\mathbb{E}_t[\|\nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t) - \Delta^{t+1}\|^2] \\
&\overset{(c)}{\leq} (1-\beta_0)\|\nabla F(\mathbf{x}_{t-1}) - \mathbf{z}_t\|^2 + \frac{4L_F^2}{\beta_0}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + 4\beta_0 \underbrace{\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t)\|^2}_{ⓐ} \\
&\quad + \beta_0^2 \underbrace{\mathbb{E}_t[\|\nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t) - \Delta^{t+1}\|^2]}_{ⓑ}
\end{aligned} \tag{4}$$

where (a) follows from $\mathbb{E}_t[\nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t)] = \Delta^{t+1}$, (b) is due to $\|a+b\|^2 \leq (1+\beta)\|a\|^2 + (1+\frac{1}{\beta})\|b\|^2$, and (c) uses the assumption $\beta_0 \leq 1$ and Lemma A.2.
Furthermore, one may bound the last two terms in 4 as following

$$
\begin{aligned}
\text{ⓐ} &= \|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t)\|^2 \\
&= \|\frac{1}{m}\sum_{i\in\mathcal{S}} \nabla_x f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t)) - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))[\nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1} \nabla_y f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t)) \\
&\quad - \frac{1}{m}\sum_{i\in\mathcal{S}} \nabla_x f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i^t)\mathbb{E}_t[H_i^t]\nabla_y f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t)\|^2 \\
&\leq \frac{1}{m}\sum_{i\in\mathcal{S}} 2\|\nabla_x f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t)) - \nabla_x f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t)\|^2 \\
&\quad + 6\|\nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i^t)\mathbb{E}_t[H_i^t]\nabla_y f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i^t)[\nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1}\nabla_y f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t)\|^2 \\
&\quad + 6\|\nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i^t)[\nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1}\nabla_y f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))[\nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1}\nabla_y f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t)\|^2 \\
&\quad + 6\|\nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))[\nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1}\nabla_y f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) \\
&\quad - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))[\nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1}\nabla_y f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t))\|^2 \\
&\leq \frac{1}{m}\sum_{i\in\mathcal{S}} 2L_f^2[\|\alpha_i(\mathbf{x}_t) - \alpha_i^t\|^2 + \|\mathbf{y}_i(\mathbf{x}_t) - \mathbf{y}_i^t\|^2] + 6C^2_{gxy}C_f^2\|\mathbb{E}_t[H_i^t] - [\nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1}\|^2 \\
&\quad + \frac{6L^2_{gxy}C_f^2}{\mu_g^2}\|\mathbf{y}_i^t - \mathbf{y}_i(\mathbf{x}_t)\|^2 + \frac{6C^2_{gxy}L_f^2}{\mu_g^2}[\|\alpha_i^t - \alpha_i(\mathbf{x}_t)\|^2 + \|\mathbf{y}_i^t - \mathbf{y}_i(\mathbf{x}_t)\|^2] \\
&= \frac{1}{m}(2L_f^2 + \frac{6C^2_{gxy}L_f^2}{\mu_g^2})\|\boldsymbol{\alpha}^t - \boldsymbol{\alpha}(\mathbf{x}_t)\|^2 + \frac{1}{m}(2L_f^2 + \frac{6L^2_{gxy}C_f^2}{\mu_g^2} + \frac{6C^2_{gxy}L_f^2}{\mu_g^2})\|\mathbf{y}^t - \mathbf{y}(\mathbf{x}_t)\|^2 \\
&\quad + \frac{6C^2_{gxy}C_f^2}{m}\sum_{i\in\mathcal{S}}\|\mathbb{E}_t[H_i^t] - [\nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1}\|^2
\end{aligned}
$$

$$
\begin{aligned}
\text{ⓑ} &= \mathbb{E}_t[\|\nabla F(\mathbf{x}_t, \alpha^t, \mathbf{y}^t) - \Delta^{t+1}\|^2] \\
&\leq \mathbb{E}_t\left[2\left\|\frac{1}{m}\sum_{i\in\mathcal{S}}\nabla F_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \frac{1}{|I_t|}\sum_{i\in I_t}\nabla F_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t)\right\|^2 + 2\left\|\frac{1}{|I_t|}\sum_{i\in I_t}\nabla F_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \frac{1}{|I_t|}\sum_{i\in I_t}\Delta_i^{t+1}\right\|^2\right] \\
&\leq \frac{8B_F}{B} + \frac{2}{m}\sum_{i\in\mathcal{S}}\mathbb{E}_{\mathcal{B}_i^t}\Big[2\|\nabla_x f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \nabla_x f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t; \mathcal{B}_i^t)\|^2 \\
&\quad + 6\|\nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i^t)\mathbb{E}_t[H_i^t]\nabla_y f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i^t; \tilde{\mathcal{B}}_i^t)\mathbb{E}_t[H_i^t]\nabla_y f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t)\|^2 \\
&\quad + 6\|\nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i^t; \tilde{\mathcal{B}}_i^t)\mathbb{E}_t[H_i^t]\nabla_y f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i^t; \tilde{\mathcal{B}}_i^t)H_i^t\nabla_y f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t)\|^2 \\
&\quad + 6\|\nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i^t; \tilde{\mathcal{B}}_i^t)H_i^t\nabla_y f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i^t; \tilde{\mathcal{B}}_i^t)H_i^t\nabla_y f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t; \mathcal{B}_i^t)\|^2\Big] \\
&\leq \frac{8B_F}{|I_t|} + \frac{4\sigma^2}{B} + \frac{12\sigma^2C_f^2}{B}\|\mathbb{E}_t[H_i^t]\|^2 + 48(C^2_{gxy} + \sigma^2)\|\mathbb{E}_t[H_i^t] - H_i^t\|^2C_f^2 + \frac{12(C^2_{gxy} + \sigma^2)\sigma^2}{B}\|H_i^t\|^2
\end{aligned}
$$

where $B_F = 2C_f^2 + \frac{2C^2_{gxy}C_f^2}{\mu_g^2}$ is the upper bound of $\|\nabla F_i(\mathbf{x}, \alpha_i, \mathbf{y}_i)\|^2$.

Since H_i^t is irrelevant to the randomness at the t-th iteration, we have $\mathbb{E}_t[H_i^t] = H_i^t$. Thus

$$\begin{aligned}
\text{ⓐ} &\leq \frac{1}{m}(2L_f^2 + \frac{6C_{gxy}^2 L_f^2}{\mu_g^2})\|\boldsymbol{\alpha}^t - \boldsymbol{\alpha}(\mathbf{x}_t)\|^2 + \frac{1}{m}(2L_f^2 + \frac{6L_{gxy}^2 C_f^2}{\mu_g^2} + \frac{6C_{gxy}^2 L_f^2}{\mu_g^2})\|\mathbf{y}^t - \mathbf{y}(\mathbf{x}_t)\|^2 \\
&\quad + \frac{6C_{gxy}^2 C_f^2}{m}\sum_{i\in\mathcal{S}}\|H_i^t - [\nabla_{yy}^2 g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1}\|^2 \\
&\leq \frac{1}{m}(2L_f^2 + \frac{6C_{gxy}^2 L_f^2}{\mu_g^2})\|\boldsymbol{\alpha}^t - \boldsymbol{\alpha}(\mathbf{x}_t)\|^2 + \frac{1}{m}(2L_f^2 + \frac{6L_{gxy}^2 C_f^2}{\mu_g^2} + \frac{6C_{gxy}^2 L_f^2}{\mu_g^2})\|\mathbf{y}^t - \mathbf{y}(\mathbf{x}_t)\|^2 \\
&\quad + \frac{6C_{gxy}^2 C_f^2}{\mu_g^4 m}\sum_{i\in\mathcal{S}}\|s_i^t - \nabla_{yy}^2 g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))\|^2 \\
&=: \frac{C_1}{4m}\|\boldsymbol{\alpha}^t - \boldsymbol{\alpha}(\mathbf{x}_t)\|^2 + \frac{\widetilde{C}_2}{4m}\|\mathbf{y}^t - \mathbf{y}(\mathbf{x}_t)\|^2 + \frac{C_3}{4m}\|s_i^t - \nabla_{yy}^2 g(\mathbf{x}_t, \mathbf{y}(\mathbf{x}_t))\|^2
\end{aligned} \tag{5}$$

and

$$\text{ⓑ} \leq \frac{8B_F}{|I_t|} + \frac{4\sigma^2}{B} + \frac{12\sigma^2 C_f^2}{B\mu_g^2} + \frac{12(C_{gxy}^2 + \sigma^2)\sigma^2}{B\mu_g^2} =: \frac{C_4}{\min\{|I_t|, B\}} \tag{6}$$

Thus, combining inequalities 4, 5 and 6, we have

$$\begin{aligned}
\mathbb{E}_t[\|\nabla F(\mathbf{x}_t) - \mathbf{z}_{t+1}\|^2] &\leq (1-\beta_0)\|\nabla F(\mathbf{x}_{t-1}) - \mathbf{z}_t\|^2 + \frac{4L_F^2\eta_0^2}{\beta_0}\|\mathbf{z}_t\|^2 + 4\beta_0\bigg[\frac{C_1}{4m}\|\boldsymbol{\alpha}^t - \boldsymbol{\alpha}(\mathbf{x}_t)\|^2 \\
&\quad + \frac{\widetilde{C}_2}{4m}\|\mathbf{y}^t - \mathbf{y}(\mathbf{x}_t)\|^2 + \frac{C_3}{4m}\|s^t - \nabla_{yy}^2 g(\mathbf{x}_t, \mathbf{y}(\mathbf{x}_t))\|^2\bigg] + \frac{C_4\beta_0^2}{\min\{|I_t|, B\}}
\end{aligned} \tag{7}$$

For simplicity, denote $\delta_{\alpha,t} := \|\boldsymbol{\alpha}^t - \boldsymbol{\alpha}(\mathbf{x}_t)\|^2$, $\delta_{y,t} := \|\mathbf{y}^t - \mathbf{y}(\mathbf{x}_t)\|^2$ and $\delta_{gyy,t} := \|s^t - \nabla_{yy}^2 g(\mathbf{x}_t, \mathbf{y}(\mathbf{x}_t))\|^2$. Take expectation over all randomness and summation over $t = 1, \ldots, T$ to get

$$\begin{aligned}
\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{z}_{t+1}\|^2] &\leq \frac{1}{\beta_0}\mathbb{E}[\|\nabla F(\mathbf{x}_0) - \mathbf{z}_1\|^2] + \frac{4L_F^2\eta_0^2}{\beta_0^2}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{z}_t\|^2] + \frac{C_1}{m}\sum_{t=1}^{T}\mathbb{E}[\delta_{\alpha,t}] \\
&\quad + \frac{\widetilde{C}_2}{m}\sum_{t=1}^{T}\mathbb{E}[\delta_{y,t}] + \frac{C_3}{m}\sum_{t=1}^{T}\mathbb{E}[\delta_{gyy,t}] + \frac{C_4\beta_0 T}{\min\{|I_t|, B\}}
\end{aligned} \tag{8}$$

Recall that from Lemma 2.5, Lemma 2.4 and Lemma 2, we have

$$\sum_{t=0}^{T}\mathbb{E}[\delta_{\alpha,t}] \leq \frac{4m}{\mu_f|I_t|\eta_1}\delta_{\alpha,0} + \frac{24L_f^2}{\mu_f^2}\sum_{t=0}^{T-1}\mathbb{E}[\delta_{y,t}] + \frac{8m\mu_f\eta_1\sigma^2 T}{B} + \frac{32m^3C_\alpha^2\eta_0^2}{\eta_1^2\mu_f^2|I_t|^2}\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2] \tag{9}$$

$$\sum_{t=0}^{T}\mathbb{E}[\delta_{y,t}] \leq \frac{2m}{|I_t|\eta_2\mu_g}\delta_{y,0} + \frac{4m\eta_2 T\sigma^2}{\mu_g B} + \frac{8m^3C_y^2\eta_0^2}{|I_t|^2\eta_2^2\mu_g^2}\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2] \tag{10}$$

$$\sum_{t=0}^{T}\mathbb{E}[\delta_{gyy,t}] \leq \frac{4m}{|I_t|\beta_1}\delta_{gyy,0} + 32L_{gyy}^2\sum_{t=0}^{T-1}\mathbb{E}[\delta_{y,t}] + 8m\beta_1 T\frac{\sigma^2}{B} + \frac{32m^3L_{gyy}^2(1+C_y^2)\eta_0^2}{|I_t|^2\beta_1^2}\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2] \tag{11}$$

Combining inequalities 8, 9, 10 and 11, we obtain

$$\begin{aligned}
&\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)-\mathbf{z}_{t+1}\|^2]\\
&\leq \frac{1}{\beta_0}\mathbb{E}[\|\nabla F(\mathbf{x}_0)-\mathbf{z}_1\|^2]+\frac{4L_F^2\eta_0^2}{\beta_0^2}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{z}_t\|^2]+\frac{C_1}{m}\left[\frac{4m}{\mu_f|I_t|\eta_1}\delta_{\alpha,0}+\frac{8m\mu_f\eta_1\sigma^2T}{B}+\frac{32m^3C_\alpha^2\eta_0^2}{\eta_1^2\mu_f^2|I_t|^2}\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2]\right]\\
&\quad+\frac{C_2}{m}\sum_{t=1}^{T}\mathbb{E}[\delta_{y,t}]+\frac{C_3}{m}\left[\frac{4m}{|I_t|\beta_1}\delta_{gyy,0}+8m\beta_1T\frac{\sigma^2}{B}+\frac{32m^3L_{gyy}^2(1+C_y^2)\eta_0^2}{|I_t|^2\beta_1^2}\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2]\right]+\frac{C_4\beta_0T}{\min\{|I_t|,B\}}\\
&\leq \frac{\mathbb{E}[\|\nabla F(\mathbf{x}_0)-\mathbf{z}_1\|^2]}{\beta_0}+\frac{4C_1\delta_{\alpha,0}}{\mu_f|I_t|\eta_1}+\frac{8C_1\mu_f\eta_1\sigma^2T}{B}+\frac{2C_2\delta_{y,0}}{|I_t|\eta_2\mu_g}+\frac{4C_2\eta_2T\sigma^2}{\mu_gB}+\frac{4C_3\delta_{gyy,0}}{|I_t|\beta_1}+8C_3\beta_1T\frac{\sigma^2}{B}\\
&\quad+\frac{C_4\beta_0T}{\min\{|I_t|,B\}}+\left(\frac{4L_F^2\eta_0^2}{\beta_0^2}+\frac{32m^2C_1C_\alpha^2\eta_0^2}{\eta_1^2\mu_f^2|I_t|^2}+\frac{32m^2C_3L_{gyy}^2(1+C_y^2)\eta_0^2}{|I_t|^2\beta_1^2}+\frac{8m^2C_2C_y^2\eta_0^2}{|I_t|^2\eta_2^2\mu_g^2}\right)\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{z}_t\|^2]
\end{aligned}\tag{12}$$

where $C_2:=\frac{24L_f^2C_1}{\mu_f^2}+\widetilde{C}_2+32L_{gyy}^2C_3$.
Recall Lemma A.3, we have

$$F(\mathbf{x}_{t+1})\leq F(\mathbf{x}_t)+\frac{\eta_x}{2}\|\nabla F(\mathbf{x}_t)-\mathbf{z}_{t+1}\|^2-\frac{\eta_0}{2}\|\nabla F(\mathbf{x}_t)\|^2-\frac{\eta_0}{4}\|\mathbf{z}_{t+1}\|^2$$

Combining with 12, we obtain

$$\begin{aligned}
&\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)\|^2]\\
&\leq \frac{2\mathbb{E}[F(\mathbf{x}_0)-F(\mathbf{x}_{T+1})]}{\eta_0T}+\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)-\mathbf{z}_{t+1}\|^2]-\frac{1}{2T}\sum_{t=0}^{T}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2]\\
&\leq \frac{1}{T}\left[\frac{2\mathbb{E}[F(\mathbf{x}_0)-F(\mathbf{x}^*)]}{\eta_0}+\frac{\mathbb{E}[\|\nabla F(\mathbf{x}_0)-\mathbf{z}_1\|^2]}{\beta_0}+\frac{4C_1\delta_{\alpha,0}}{\mu_f|I_t|\eta_1}+\frac{2C_2\delta_{y,0}}{|I_t|\eta_2\mu_g}+\frac{4C_3\delta_{gyy,0}}{|I_t|\beta_1}\right]\\
&\quad+\frac{8C_1\mu_f\eta_1\sigma^2}{B}+\frac{4C_2\eta_2\sigma^2}{\mu_gB}+8C_3\beta_1\frac{\sigma^2}{B}+\frac{C_4\beta_0}{\min\{|I_t|,B\}}\\
&\quad+\frac{1}{T}\left(\frac{4L_F^2\eta_0^2}{\beta_0^2}+\frac{32m^2C_1C_\alpha^2\eta_0^2}{\eta_1^2\mu_f^2|I_t|^2}+\frac{32m^2C_3L_{gyy}^2(1+C_y^2)\eta_0^2}{|I_t|^2\beta_1^2}+\frac{8m^2C_2C_y^2\eta_0^2}{|I_t|^2\eta_2^2\mu_g^2}-\frac{1}{2}\right)\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{z}_t\|^2]
\end{aligned}\tag{13}$$

By setting

$$\eta_0^2\leq\min\left\{\frac{\beta_0^2}{80L_F^2},\frac{\eta_1^2\mu_f^2|I_t|^2}{640m^2C_1C_\alpha^2},\frac{|I_t|^2\beta_1^2}{640m^2C_3L_{gyy}^2(1+C_y^2)},\frac{|I_t|^2\eta_2^2\mu_g^2}{160m^2C_2C_y^2}\right\}$$

we have

$$\frac{4L_F^2\eta_0^2}{\beta_0^2}+\frac{32m^2C_1C_\alpha^2\eta_0^2}{\eta_1^2\mu_f^2|I_t|^2}+\frac{32m^2C_3L_{gyy}^2(1+C_y^2)\eta_0^2}{|I_t|^2\beta_1^2}+\frac{8m^2C_2C_y^2\eta_0^2}{|I_t|^2\eta_2^2\mu_g^2}-\frac{1}{4}\leq 0$$

which implies that the last term of the right hand side of inequality 13 is less or equal to zero. Hence

$$\begin{aligned}
&\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)\|^2]\\
&\leq \frac{1}{T}\left[\frac{2\mathbb{E}[F(\mathbf{x}_0)-F(\mathbf{x}^*)]}{\eta_0}+\frac{\mathbb{E}[\|\nabla F(\mathbf{x}_0)-\mathbf{z}_1\|^2]}{\beta_0}+\frac{4C_1\delta_{\alpha,0}}{\mu_f|I_t|\eta_1}+\frac{2C_2\delta_{y,0}}{|I_t|\eta_2\mu_g}+\frac{4C_3\delta_{gyy,0}}{|I_t|\beta_1}\right]\\
&\quad+\frac{8C_1\mu_f\eta_1\sigma^2}{B}+\frac{4C_2\eta_2\sigma^2}{\mu_gB}+8C_3\beta_1\frac{\sigma^2}{B}+\frac{C_4\beta_0}{\min\{|I_t|,B\}}
\end{aligned}\tag{14}$$

With

$$\eta_1 \leq \frac{B\epsilon^2}{96C_1\mu_f\sigma^2}, \eta_2 \leq \frac{\mu_g B\epsilon^2}{48C_2\sigma^2}, \beta_1 \leq \frac{B\epsilon^2}{96C_3\sigma^2}, \beta_0 \leq \frac{\min\{|I_t|, B\}\epsilon^2}{12C_4}$$

$$T \geq \max\left\{\frac{30\mathbb{E}[F(\mathbf{x}_0) - F(\mathbf{x}^*)]}{\eta_0\epsilon^2}, \frac{15\mathbb{E}[\|\nabla F(\mathbf{x}_0) - \mathbf{z}_1\|^2]}{\beta_0\epsilon^2}, \frac{60C_1\delta_{\alpha,0}}{\mu_f|I_t|\eta_1\epsilon^2}, \frac{30C_2\delta_{y,0}}{|I_t|\eta_2\mu_g\epsilon^2}, \frac{60C_3\delta_{gyy,0}}{|I_t|\beta_1\epsilon^2}\right\}$$

we have

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)\|^2] \leq \frac{\epsilon^2}{3} + \frac{\epsilon^2}{3} < \epsilon^2$$

Furthermore, to show the second part of the theorem, following from inequality 12, we have

$$\begin{aligned}
&\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{z}_{t+1}\|^2] \\
&\leq \frac{\mathbb{E}[\|\nabla F(\mathbf{x}_0) - \mathbf{z}_1\|^2]}{\beta_0} + \frac{4C_1\delta_{\alpha,0}}{\mu_f|I_t|\eta_1} + \frac{8C_1\mu_f\eta_1\sigma^2 T}{B} + \frac{2C_2\delta_{y,0}}{|I_t|\eta_2\mu_g} + \frac{4C_2\eta_2 T\sigma^2}{\mu_g B} + \frac{4C_3\delta_{gyy,0}}{|I_t|\beta_1} \\
&\quad + 8C_3\beta_1 T\frac{\sigma^2}{B} + \frac{C_4\beta_0 T}{\min\{|I_t|, B\}} + \frac{1}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\mathbf{x}_t)\|^2] + \mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{z}_{t+1}\|^2]
\end{aligned} \tag{15}$$

With parameter set above, we have

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{z}_{t+1}\|^2] < 2\epsilon^2$$

□

A.1 Proof of Lemma A.2

Proof. Take arbitrary $\mathbf{x}_1, \mathbf{x}_2$. Then with Assumption 2.2 and Lemma A.4 we have

$$\begin{aligned}
&\|\nabla F(\mathbf{x}_1) - \nabla F(\mathbf{x}_2)\|^2 \\
&\leq \left\|\frac{1}{m}\sum_{i\in\mathcal{S}}\nabla_x f_i(\mathbf{x}_1, \alpha_i(\mathbf{x}_1), \mathbf{y}_i(\mathbf{x}_1)) - \nabla^2_{xy} g_i(\mathbf{x}_1, \mathbf{y}_i(\mathbf{x}_1))[\nabla^2_{yy} g_i(\mathbf{x}_1, \mathbf{y}_i(\mathbf{x}_1))]^{-1}\nabla_y f_i(\mathbf{x}_1, \alpha_i(\mathbf{x}_1), \mathbf{y}_i(\mathbf{x}_1))\right. \\
&\quad \left. - \frac{1}{m}\sum_{i\in\mathcal{S}}\nabla_x f_i(\mathbf{x}_2, \alpha_i(\mathbf{x}_2), \mathbf{y}_i(\mathbf{x}_2)) - \nabla^2_{xy} g_i(\mathbf{x}_2, \mathbf{y}_i(\mathbf{x}_2))[\nabla^2_{yy} g_i(\mathbf{x}_2, \mathbf{y}_i(\mathbf{x}_2))]^{-1}\nabla_y f_i(\mathbf{x}_2, \alpha_i(\mathbf{x}_2), \mathbf{y}_i(\mathbf{x}_2))\right\|^2 \\
&\leq \left(2L_f^2(1 + C_\alpha^2 + C_y^2) + \frac{6L_{gxy}^2(1 + C_y^2)C_f^2}{\mu_g^2} + \frac{6C_{gxy}^2 L_{gyy}^2(1 + C_y^2)C_f^2}{\mu_g^4} + \frac{6C_{gxy}^2 L_f^2(1 + C_\alpha^2 + C_y^2)}{\mu_g^2}\right)\|\mathbf{x}_1 - \mathbf{x}_2\|^2 \\
&=: L_F^2\|\mathbf{x}_1 - \mathbf{x}_2\|^2
\end{aligned}$$

□

A.2 Proof of Lemma A.3

Proof. By L_F-smoothness of $F(\mathbf{x})$, with $\eta_0 \leq \frac{1}{2L_F}$, we have

$$\begin{aligned}
F(\mathbf{x}_{t+1}) &\leq F(\mathbf{x}_t) + \nabla F(\mathbf{x}_t)^T(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L_F}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&= F(\mathbf{x}_t) - \eta_0\nabla F(\mathbf{x}_t)^T\mathbf{z}_{t+1} + \frac{L_F}{2}\eta_0^2\|\mathbf{z}_{t+1}\|^2 \\
&= F(\mathbf{x}_t) + \frac{\eta_0}{2}\|\nabla F(\mathbf{x}_t) - \mathbf{z}_{t+1}\|^2 - \frac{\eta_0}{2}\|\nabla F(\mathbf{z}_t)\|^2 + \left(\frac{L_F}{2}\eta_0^2 - \frac{\eta_0}{2}\right)\|\mathbf{z}_{t+1}\|^2.
\end{aligned}$$

□

A.3 Proof of Lemma 2.4

This proof follows from the proof of Lemma 8 in [29].

A.4 Proof of Lemma 2.5

Proof. Define $\widetilde{\alpha}_i^t := \Pi_{\mathcal{A}}[\alpha_i^t + \eta_1 \nabla_\alpha f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t; \mathcal{B}_i^t)]$. Note that $\alpha_i(\mathbf{x}_t) = \arg\max_{\alpha \in \mathcal{A}} f_i(\mathbf{x}_t, \alpha, \mathbf{y}_i(\mathbf{x}_t))$. Since $\alpha_i(\mathbf{x}_t) = \Pi_{\mathcal{A}}[\alpha_i(\mathbf{x}_t) + \eta_1 \nabla_\alpha f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t))]$, take $i \in \mathcal{S}$, then

$$
\begin{aligned}
&\mathbb{E}_t[\|\widetilde{\alpha}_i^t - \alpha_i(\mathbf{x}_t)\|^2] \\
&= \mathbb{E}_t[\|\Pi_{\mathcal{A}}[\alpha_i^t + \eta_1 \nabla_\alpha f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t; \mathcal{B}_i^t)] - \Pi_{\mathcal{A}}[\alpha_i(\mathbf{x}_t) + \eta_1 \nabla_\alpha f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t))]\|^2] \\
&\leq \mathbb{E}_t[\|[\alpha_i^t + \eta_1 \nabla_\alpha f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t; \mathcal{B}_i^t)] - [\alpha_i(\mathbf{x}_t) + \eta_1 \nabla_\alpha f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t))]\|^2] \\
&= \mathbb{E}_t[\|\alpha_i^t - \alpha_i(\mathbf{x}_t) + \eta_1 \nabla_\alpha f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \eta_1 \nabla_\alpha f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t)) \\
&\quad + \eta_1 \nabla_\alpha f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t; \mathcal{B}_i^t) - \eta_1 \nabla_\alpha f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t)\|^2] \\
&\leq \|\alpha_i^t - \alpha_i(\mathbf{x}_t) + \eta_1 \nabla_\alpha f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \eta_1 \nabla_\alpha f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t))\|^2 + \frac{\eta_1^2 \sigma^2}{B} \\
&\leq \|\alpha_i^t - \alpha_i(\mathbf{x}_t)\|^2 + \eta_1^2 \|\nabla_\alpha f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \nabla_\alpha f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t))\|^2 \\
&\quad + 2\eta_1 \langle \alpha_i^t - \alpha_i(\mathbf{x}_t), \nabla_\alpha f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i(\mathbf{x}_t)) - \nabla_\alpha f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t)) \rangle \\
&\quad + 2\eta_1 \langle \alpha_i^t - \alpha_i(\mathbf{x}_t), \nabla_\alpha f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \nabla_\alpha f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i(\mathbf{x}_t)) \rangle + \frac{\eta_1^2 \sigma^2}{B} \\
&\overset{(a)}{\leq} \|\alpha_i^t - \alpha_i(\mathbf{x}_t)\|^2 + \eta_1^2 \|\nabla_\alpha f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \nabla_\alpha f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t))\|^2 - 2\eta_1 \mu_f \|\alpha_i^t - \alpha_i(\mathbf{x}_t)\|^2 \\
&\quad + 2\eta_1 \left[\frac{\mu_f}{4} \|\alpha_i^t - \alpha_i(\mathbf{x}_t)\|^2 + \frac{1}{\mu_f} \|\nabla_\alpha f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \nabla_\alpha f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i(\mathbf{x}_t))\|^2 \right] + \frac{\eta_1^2 \sigma^2}{B} \\
&\overset{(b)}{\leq} (1 - \frac{\eta_1 \mu_f}{2}) \|\alpha_i^t - \alpha_i(\mathbf{x}_t)\|^2 + \frac{3\eta_1 L_f^2}{\mu_f} \|\mathbf{y}_i^t - \mathbf{y}_i(\mathbf{x}_t)\|^2 + \frac{\eta_1^2 \sigma^2}{B}
\end{aligned}
\tag{16}
$$

where inequality (a) uses the standard inequality $\langle a, b \rangle \leq \frac{\beta}{2}\|a\|^2 + \frac{1}{2\beta}\|b\|^2$ and the strong monotonicity of $-f_i(\mathbf{x}_t, \cdot, \mathbf{y}_{i,t})$ as it is assumed to be μ_f-strongly convex, and (b) uses the assumption $\eta_1 \leq \min\{\mu_f / L_f^2, 1/\mu_f\}$. Note that

$$
\mathbb{E}_t[\|\alpha_i^{t+1} - \alpha_i(\mathbf{x}_t)\|^2] = \frac{|I_t|}{m} \mathbb{E}_t[\|\widetilde{\alpha}_i^t - \alpha_i(\mathbf{x}_t)\|^2] + \frac{m - |I_t|}{m} \|\alpha_i^t - \alpha_i(\mathbf{x}_t)\|^2
$$

which implies

$$
\mathbb{E}_t[\|\widetilde{\alpha} i^t - \alpha_i(\mathbf{x}_t)\|^2] = \frac{m}{|I_t|} \mathbb{E}_t[\|\alpha_i^{t+1} - \alpha_i(\mathbf{x}_t)\|^2] - \frac{m - |I_t|}{|I_t|} \|\alpha_i^t - \alpha_i(\mathbf{x}_t)\|^2 \tag{17}
$$

Thus combining inequalities 16 and 17 gives

$$
\begin{aligned}
&\frac{m}{|I_t|} \mathbb{E}_t[\|\alpha_i^{t+1} - \alpha_i(\mathbf{x}_t)\|^2] - \frac{m - |I_t|}{|I_t|} \|\alpha_i^t - \alpha_i(\mathbf{x}_t)\|^2 \\
&\leq (1 - \frac{\eta_1 \mu_f}{2}) \|\alpha_i^t - \alpha_i(\mathbf{x}_t)\|^2 + \frac{3\eta_1 L_f^2}{\mu_f} \|\mathbf{y}_i^t - \mathbf{y}_i(\mathbf{x}_t)\|^2 + \frac{\eta_1^2 \sigma^2}{B}
\end{aligned}
\tag{18}
$$

Rearrange it to get

$$
\mathbb{E}_t[\|\alpha_i^{t+1} - \alpha_i(\mathbf{x}_t)\|^2] \leq \left(1 - \frac{\eta_1 \mu_f |I_t|}{2m}\right) \|\alpha_i^t - \alpha_i(\mathbf{x}_t)\|^2 + \frac{3\eta_1 L_f^2 |I_t|}{\mu_f m} \|\mathbf{y}_i^t - \mathbf{y}_i(\mathbf{x}_t)\|^2 + \frac{\eta_1^2 \sigma^2 |I_t|}{mB} \tag{19}
$$

Thus

$$\begin{aligned}
&\mathbb{E}_t[\|\alpha_i^{t+1} - \alpha_i(\mathbf{x}_{t+1})\|^2] \\
&\leq \left(1 + \frac{\eta_1\mu_f|I_t|}{4m}\right)\mathbb{E}_t[\|\alpha_i^{t+1} - \alpha_i(\mathbf{x}_t)\|^2] + \left(1 + \frac{4m}{\eta_1\mu_f|I_t|}\right)\mathbb{E}_t[\|\alpha_i(\mathbf{x}_{t+1}) - \alpha_i(\mathbf{x}_t)\|^2] \\
&\leq \left(1 - \frac{\eta_1\mu_f|I_t|}{4m}\right)\|\alpha_i^t - \alpha_i(\mathbf{x}_t)\|^2 + \frac{6\eta_1 L_f^2|I_t|}{\mu_f m}\|\mathbf{y}_i^t - \mathbf{y}_i(\mathbf{x}_t)\|^2 + \frac{2\eta_1^2\sigma^2|I_t|}{mB} + \frac{8mC_\alpha^2}{\eta_1\mu_f|I_t|}\mathbb{E}_t[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2]
\end{aligned} \tag{20}$$

where we use the assumption $\eta_1 \leq \frac{4m}{\mu_f|I_t|}$ i.e. $\frac{\eta_1\mu_f|I_t|}{4m} \leq 1$. Taking summation over all tasks $i \in \mathcal{S}$, we obtain

$$\begin{aligned}
&\mathbb{E}_t[\|\boldsymbol{\alpha}^{t+1} - \boldsymbol{\alpha}(\mathbf{x}_{t+1})\|^2] \\
&\leq \left(1 - \frac{\eta_1\mu_f|I_t|}{4m}\right)\|\boldsymbol{\alpha}^t - \boldsymbol{\alpha}(\mathbf{x}_t)\|^2 + \frac{6\eta_1 L_f^2|I_t|}{\mu_f m}\|\mathbf{y}^t - \mathbf{y}(\mathbf{x}_t)\|^2 + \frac{2\eta_1^2\sigma^2|I_t|}{B} + \frac{8m^2C_\alpha^2}{\eta_1\mu_f|I_t|}\mathbb{E}_t[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2]
\end{aligned} \tag{21}$$

Taking summation over $t = 0, \ldots, T-1$ and taking expectation over all randomness, we obtain

$$\begin{aligned}
&\sum_{t=0}^{T}\mathbb{E}[\|\boldsymbol{\alpha}^t - \boldsymbol{\alpha}(\mathbf{x}_t)\|^2] \\
&\leq \frac{4m}{\eta_1\mu_f|I_t|}\|\boldsymbol{\alpha}^0 - \boldsymbol{\alpha}(\mathbf{x}_0)\|^2 + \frac{24L_f^2}{\mu_f^2}\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{y}^t - \mathbf{y}(\mathbf{x}_t)\|^2] + \frac{8m\mu_f\eta_1\sigma^2 T}{B} + \frac{32m^3C_\alpha^2}{\eta_1^2\mu_f^2|I_t|^2}\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2]
\end{aligned} \tag{22}$$

□

A.5 Proof of Lemma 2.6

This proof follows from the proof of Lemma 10 in [29].

B Convergence Analysis of Algorithm 2

First, we note that the bounded variance of $\nabla_v\gamma_i(v, \mathbf{x}, \alpha_i, \mathbf{y}_i; \mathcal{B}_i)$ can be derived as

$$\begin{aligned}
&\mathbb{E}_{\mathcal{B}_i^t}[\|\nabla_v\gamma_i(\mathbf{v}_i^t, \mathbf{x}_t, \mathbf{y}_i^t; \mathcal{B}_i^t) - \nabla_v\gamma_i(\mathbf{v}_i^t, \mathbf{x}_t, \mathbf{y}_i^t)\|^2] \\
&= \mathbb{E}_{\mathcal{B}_i^t}[\|\nabla_{yy}^2 g_i(\mathbf{x}_t, \mathbf{y}_i^t; \mathcal{B}_i^t)\mathbf{v}_i^t - \nabla_y f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t; \mathcal{B}_i^t) - \nabla_{yy}^2 g_i(\mathbf{x}_t, \mathbf{y}_i^t)\mathbf{v}_i^t + \nabla_y f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t)\|^2] \\
&\leq \mathbb{E}_{\mathcal{B}_i^t}[2\|\nabla_{yy}^2 g_i(\mathbf{x}_t, \mathbf{y}_i^t; \mathcal{B}_i^t)\mathbf{v}_i^t - \nabla_{yy}^2 g_i(\mathbf{x}_t, \mathbf{y}_i^t)\mathbf{v}_i^t\|^2 + 2\|\nabla_y f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \nabla_y f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t; \mathcal{B}_i^t)\|^2] \\
&\leq \frac{2\sigma^2}{B}\|\mathbf{v}_i^t\|^2 + \frac{2\sigma^2}{B} \leq (1 + \Gamma^2)\frac{2\sigma^2}{B}.
\end{aligned}$$

We present the detailed statement of Theorem 2.8

Theorem B.1. *Let* $F(\boldsymbol{x}_0) - F(\boldsymbol{x}^*) \leq \Delta_F$. *Under Assumption 2.2,2.3 and consider Algorithm 2, with* $\eta_1 \leq \min\left\{\frac{\mu_f}{L_f^2}, \frac{1}{\mu_f}, \frac{4m}{\mu_f|I_t|}, \frac{B\epsilon^2}{96C_1\mu_f\sigma^2}\right\}$, $\eta_3 \leq \min\left\{\frac{\mu_\gamma}{L_\gamma^2}, \frac{1}{\mu_\gamma}, \frac{4m}{\mu_\gamma|I_t|}, \frac{B\epsilon^2}{96C_3\mu_g\sigma^2}\right\}$, $\eta_2 \leq \min\{\frac{\mu_g}{L_g^2}, \frac{2m}{|I_t|\mu_g}, \frac{\mu_g B\epsilon^2}{48C_2\sigma^2}\}$, $\beta_0 \leq \frac{\min\{|I_t|, B\}\epsilon^2}{12C_4}$, $\eta_0 \leq \min\left\{\frac{1}{2L_F}, \frac{\beta_0}{8L_F}, \frac{\eta_3\mu_g|I_t|}{32mC_v\sqrt{C_3}}, \frac{\eta_1\mu_f|I_t|}{32mC_\alpha\sqrt{C_1}}, \frac{|I_t|\eta_2\mu_g}{16mC_y\sqrt{C_2}}\right\}$, $T \geq \max\left\{\frac{32[F(\boldsymbol{x}_0) - F(\boldsymbol{x}^*)]}{\eta_0\epsilon^2}, \frac{15\mathbb{E}[\|\nabla F(\boldsymbol{x}_0) - \boldsymbol{z}_1\|^2]}{\beta_0\epsilon^2}, \frac{60C_1\delta_{\alpha,0}}{\mu_f|I_t|\eta_1\epsilon^2}, \frac{60C_3\delta_{v,0}}{\eta_3\mu_g|I_t|\epsilon^2}, \frac{30C_2\delta_{y,0}}{|I_t|\eta_2\mu_g\epsilon^2}\right\}$

we have

$$\mathbb{E}[\|\nabla F(\boldsymbol{x}_\tau)\|^2] \leq \epsilon^2, \quad \mathbb{E}[\|\nabla F(\boldsymbol{x}_\tau) - \boldsymbol{z}_{\tau+1})\|^2] < 2\epsilon^2,$$

where τ *is randomly sampled from* $\{0, \ldots, T\}$, C_1, C_2, C_3, C_4 *are constants defined in the proof, and* L_F *is the Lipschitz continuity constant of* $\nabla F(\boldsymbol{x})$.

To prove Theorem B.1, we need the following Lemmas.

Lemma B.2 (Lemma 4.3 [25]). *Under Assumption 2.2, $\boldsymbol{v}_i(\boldsymbol{x}, \alpha_i, \boldsymbol{y}_i)$ is $C_v = L_g/\mu_g$-Lipschitz-continuous for all i.*

Lemma B.3. *Consider the updates for $\boldsymbol{v}_i^t$ in Algorithm 2, under Assumption 2.2, 2.3, with $\eta_3 \leq \min\left\{\frac{\mu_\gamma}{L_\gamma^2}, \frac{1}{\mu_\gamma}, \frac{4m}{\mu_\gamma |I_t|}\right\}$, we have*

$$\sum_{t=0}^{T} \mathbb{E}_t[\delta_{\boldsymbol{v},t}] \leq \frac{4m}{\eta_3 \mu_g |I_t|}\delta_{\boldsymbol{v},0} + \frac{24L_g^2}{\mu_g^2}\sum_{t=0}^{T-1}\mathbb{E}[\delta_{y,t} + \delta_{\alpha,t}] + \frac{8m\mu_g\eta_3\sigma^2 T}{B} + \frac{32m^3 C_v^2 \eta_0^2}{\eta_3^2 \mu_g^2 |I_t|^2}\sum_{t=0}^{T-1}\mathbb{E}[\|\boldsymbol{z}_{t+1}\|^2].$$

Proof of Theorem B.1. First, recall and define the following notations

$$\nabla F(\mathbf{x}_t) = \frac{1}{m}\sum_{i\in\mathcal{S}} \nabla_x f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t)) - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))[\nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1}\nabla_y f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t))$$
$$\nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t, \mathbf{v}_i^t) := \frac{1}{m}\sum_{i\in\mathcal{S}} \nabla F_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) := \nabla_x f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i^t)\mathbf{v}_i^t$$
$$\Delta^{t+1} = \frac{1}{|I_t|}\sum_{i\in I_t}\Delta_i^{t+1} := \nabla_x f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t; \mathcal{B}_i^t) - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i^t; \tilde{\mathcal{B}}_i^t)\mathbf{v}_i^t$$

Consider the update $\mathbf{z}_{t+1} = (1-\beta_0)\mathbf{z}_t + \beta_0\Delta^{t+1}$ in Algorithm 2, we have

$$\begin{aligned}
&\mathbb{E}_t[\|\nabla F(\mathbf{x}_t) - \mathbf{z}_{t+1}\|^2] \\
&= \mathbb{E}_t[\|\nabla F(\mathbf{x}_t) - (1-\beta_0)\mathbf{z}_t - \beta_0\Delta^{t+1}\|^2] \\
&= \mathbb{E}_t[\|(1-\beta_0)(\nabla F(\mathbf{x}_{t-1}) - \mathbf{z}_t) + (1-\beta_0)(\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})) + \beta_0(\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t, \mathbf{v}^t)) \\
&\quad + \beta_0(\nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t, \mathbf{v}^t) - \Delta^{t+1})\|^2] \\
&\overset{(a)}{=} \|(1-\beta_0)(\nabla F(\mathbf{x}_{t-1}) - \mathbf{z}_t) + (1-\beta_0)(\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})) + \beta_0(\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t, \mathbf{v}_i^t))\|^2 \\
&\quad + \beta_0^2\mathbb{E}_t[\|\nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t, \mathbf{v}^t) - \Delta^{t+1}\|^2] \\
&\overset{(b)}{\leq} (1+\beta_0)(1-\beta_0)^2\|\nabla F(\mathbf{x}_{t-1}) - \mathbf{z}_t\|^2 + 2(1+\frac{1}{\beta_0})\left[\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})\|^2 + \beta_0^2\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t, \mathbf{v}^t)\|^2\right] \\
&\quad + \beta_0^2\mathbb{E}_t[\|\nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t, \mathbf{v}^t) - \Delta^{t+1}\|^2] \\
&\overset{(c)}{\leq} (1-\beta_0)\|\nabla F(\mathbf{x}_{t-1}) - \mathbf{z}_t\|^2 + \frac{4L_F^2}{\beta_0}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + 4\beta_0\underbrace{\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t, \mathbf{v}^t)\|^2}_{ⓐ} \\
&\quad + \beta_0^2\underbrace{\mathbb{E}_t[\|\nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t, \mathbf{v}^t) - \Delta^{t+1}\|^2]}_{ⓑ}
\end{aligned} \tag{23}$$

where (a) follows from $\mathbb{E}_t[\nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t)] = \Delta^{t+1}$, (b) is due to $\|a+b\|^2 \leq (1+\beta)\|a\|^2 + (1+\frac{1}{\beta})\|b\|^2$, and (c) uses the assumption $\beta_0 \leq 1$ and Lemma A.2.

Furthermore, one may bound the last two terms in 23 as following

$$
\begin{aligned}
ⓐ &= \|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_t, \boldsymbol{\alpha}^t, \mathbf{y}^t, \mathbf{v}_i^t)\|^2 \\
&= \Big\| \frac{1}{m}\sum_{i\in\mathcal{S}} \nabla_x f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t)) - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))[\nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1} \nabla_y f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t)) \\
&\quad - \frac{1}{m}\sum_{i\in\mathcal{S}} \nabla_x f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i^t)\mathbf{v}_i^t \Big\|^2 \\
&\le \frac{1}{m}\sum_{i\in\mathcal{S}} 2\|\nabla_x f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t)) - \nabla_x f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t)\|^2 \\
&\quad + 4\Big\| \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i^t)\big[\mathbf{v}_i^t - [\nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1}\nabla_y f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t))\big] \Big\|^2 \\
&\quad + 4\Big\| \big[\nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i^t) - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))\big][\nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1}\nabla_y f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t)) \Big\|^2 \\
&\le \frac{1}{m}\sum_{i\in\mathcal{S}} 2L_f^2[\|\alpha_i(\mathbf{x}_t) - \alpha_i^t\|^2 + \|\mathbf{y}_i(\mathbf{x}_t) - \mathbf{y}_i^t\|^2] \\
&\quad + \frac{1}{m}\sum_{i\in\mathcal{S}} 4C^2_{gxy}\big\|\mathbf{v}_i^t - [\nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1}\nabla_y f_i(\mathbf{x}_t, \alpha_i(\mathbf{x}_t), \mathbf{y}_i(\mathbf{x}_t))\big\|^2 + \frac{4L^2_{gxy}C_f^2}{\mu_g^2 m}\|\mathbf{y}^t - \mathbf{y}(\mathbf{x}_t)\|^2 \\
&= \frac{1}{m}\sum_{i\in\mathcal{S}} 2L_f^2[\|\alpha_i(\mathbf{x}_t) - \alpha_i^t\|^2 + \|\mathbf{y}_i(\mathbf{x}_t) - \mathbf{y}_i^t\|^2] + \frac{4C^2_{gxy}}{m}\big\|\mathbf{v}^t - \mathbf{v}(\mathbf{x}_t)\big\|^2 + \frac{4L^2_{gxy}C_f^2}{\mu_g^2 m}\|\mathbf{y}^t - \mathbf{y}(\mathbf{x}_t)\|^2 \\
&= \frac{\tilde{C}_1}{m}\|\alpha(\mathbf{x}_t) - \alpha^t\|^2 + \frac{\tilde{C}_2}{m}\|\mathbf{y}^t - \mathbf{y}(\mathbf{x}_t)\|^2 + \frac{C_3}{m}\big\|\mathbf{v}^t - \mathbf{v}(\mathbf{x}_t)\big\|^2
\end{aligned} \tag{24}
$$

$$
\begin{aligned}
ⓑ &= \mathbb{E}_t[\|\nabla F(\mathbf{x}_t, \alpha^t, \mathbf{y}^t, \mathbf{v}_i^t) - \Delta^{t+1}\|^2] \\
&\le \mathbb{E}_t\left[2\left\|\frac{1}{m}\sum_{i\in\mathcal{S}} \nabla F_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t, \mathbf{v}_i^t) - \frac{1}{|I_t|}\sum_{i\in I_t} \nabla F_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t, \mathbf{v}_i^t)\right\|^2 + 2\left\|\frac{1}{|I_t|}\sum_{i\in I_t} \nabla F_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t, \mathbf{v}_i^t) - \frac{1}{|I_t|}\sum_{i\in I_t}\Delta_i^{t+1}\right\|^2\right] \\
&\le \frac{8B_F}{B} + \frac{2}{m}\sum_{i\in\mathcal{S}} \mathbb{E}_{\mathcal{B}_i^t}\left[2\|\nabla_x f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t) - \nabla_x f_i(\mathbf{x}_t, \alpha_i^t, \mathbf{y}_i^t; \mathcal{B}_i^t)\|^2 + 2\|\big[\nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i^t) - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_i^t; \tilde{\mathcal{B}}_i^t)\big]\mathbf{v}_i^t\|^2\right] \\
&\le \frac{8B_F}{|I_t|} + \frac{4\sigma^2}{B} + \frac{4\sigma^2}{B}\frac{1}{m}\sum_{i\in\mathcal{S}}\|\mathbf{v}_i^t\|^2 \\
&=: \frac{C_4}{\min\{|I_t|, B\}}
\end{aligned} \tag{25}
$$

where B_F is the upper bound of $\|\nabla F_i(\mathbf{x}, \alpha_i, \mathbf{y}_i, \mathbf{v}_i^t)\|^2$.

Thus combining inequalities 23,24,25, we have

$$
\begin{aligned}
\mathbb{E}_t[\|\nabla F(\mathbf{x}_t) - \mathbf{z}_{t+1}\|^2] \le (1-\beta_0)\|\nabla F(\mathbf{x}_{t-1}) - \mathbf{z}_t\|^2 + \frac{4L_F^2}{\beta_0}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + 4\beta_0\Bigg[\frac{\widetilde{C_1}}{m}\|\alpha(\mathbf{x}_t) - \alpha^t\|^2 \\
+ \frac{\widetilde{C_2}}{m}\|\mathbf{y}^t - \mathbf{y}(\mathbf{x}_t)\|^2 + \frac{C_3}{m}\big\|\mathbf{v}^t - \mathbf{v}(\mathbf{x}_t)\big\|^2\Bigg] + \frac{C_4\beta_0^2}{\min\{|I_t|, B\}}
\end{aligned}
$$

For simplicity, denote $\delta_{\alpha,t} := \|\boldsymbol{\alpha}^t - \boldsymbol{\alpha}(\mathbf{x}_t)\|^2$, $\delta_{y,t} := \|\mathbf{y}^t - \mathbf{y}(\mathbf{x}_t)\|^2$ and $\delta_{v,t} := \sum_{i\in\mathcal{S}} \|\mathbf{v}^t - \mathbf{v}(\mathbf{x}_t)\|^2$. Take expectation over all randomness and summation over $t = 1, \ldots, T$ to get

$$\sum_{t=0}^{T} \mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{z}_{t+1}\|^2] \leq \frac{1}{\beta_0}\mathbb{E}[\|\nabla F(\mathbf{x}_0) - \mathbf{z}_1\|^2] + \frac{4L_F^2\eta_0^2}{\beta_0^2}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{z}_t\|^2] + \frac{\widetilde{C}_1}{m}\sum_{t=1}^{T}\mathbb{E}[\delta_{\alpha,t}] + \frac{\widetilde{C}_2}{m}\sum_{t=1}^{T}\mathbb{E}[\delta_{y,t}] + \frac{C_3}{m}\sum_{t=1}^{T}\mathbb{E}[\delta_{v,t}] + \frac{C_4\beta_0 T}{\min\{|I_t|, B\}} \tag{26}$$

Recall that from Lemma 2.5, Lemma 2.4 and Lemma B.3, we have

$$\sum_{t=0}^{T}\mathbb{E}[\delta_{\alpha,t}] \leq \frac{4m}{\mu_f|I_t|\eta_1}\delta_{\alpha,0} + \frac{24L_f^2}{\mu_f^2}\sum_{t=0}^{T-1}\mathbb{E}[\delta_{y,t}] + \frac{8m\mu_f\eta_1\sigma^2 T}{B} + \frac{32m^3C_\alpha^2\eta_0^2}{\eta_1^2\mu_f^2|I_t|^2}\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2] \tag{27}$$

$$\sum_{t=0}^{T}\mathbb{E}[\delta_{y,t}] \leq \frac{2m}{|I_t|\eta_2\mu_g}\delta_{y,0} + \frac{4m\eta_2 T\sigma^2}{\mu_g B} + \frac{8m^3C_y^2\eta_0^2}{|I_t|^2\eta_2^2\mu_g^2}\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2] \tag{28}$$

$$\sum_{t=0}^{T}\mathbb{E}_t[\delta_{\mathbf{v},t}] \leq \frac{4m}{\eta_3\mu_g|I_t|}\delta_{\mathbf{v},0} + \frac{24L_g^2}{\mu_g^2}\sum_{t=0}^{T-1}\mathbb{E}[\delta_{y,t} + \delta_{\alpha,t}] + \frac{8m\mu_g\eta_3\sigma^2 T}{B} + \frac{32m^3C_v^2\eta_0^2}{\eta_3^2\mu_g^2|I_t|^2}\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2]. \tag{29}$$

Combining inequalities 26,27,28,29, we have

$$
\begin{aligned}
&\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)-\mathbf{z}_{t+1}\|^2]\\
&\le \frac{1}{\beta_0}\mathbb{E}[\|\nabla F(\mathbf{x}_0)-\mathbf{z}_1\|^2]+\frac{4L_F^2\eta_0^2}{\beta_0^2}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{z}_t\|^2]+\frac{\widetilde{C}_1}{m}\sum_{t=1}^{T}\mathbb{E}[\delta_{\alpha,t}]+\frac{\widetilde{C}_2}{m}\sum_{t=1}^{T}\mathbb{E}[\delta_{y,t}]\\
&\quad+\frac{C_3}{m}\left\{\frac{4m}{\eta_3\mu_g|I_t|}\delta_{\mathbf{v},0}+\frac{24L_g^2}{\mu_g^2}\sum_{t=0}^{T-1}\mathbb{E}[\delta_{y,t}+\delta_{\alpha,t}]+\frac{8m\mu_g\eta_3\sigma^2T}{B}+\frac{32m^3C_v^2\eta_0^2}{\eta_3^2\mu_g^2|I_t|^2}\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2]\right\}\\
&\quad+\frac{C_4\beta_0T}{\min\{|I_t|,B\}}\\
&\le \frac{1}{\beta_0}\mathbb{E}[\|\nabla F(\mathbf{x}_0)-\mathbf{z}_1\|^2]+\left(\frac{4L_F^2\eta_0^2}{\beta_0^2}+\frac{32m^2C_v^2\eta_0^2C_3}{\eta_3^2\mu_g^2|I_t|^2}\right)\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{z}_t\|^2]+\left(\frac{\widetilde{C}_1}{m}+\frac{24L_g^2C_3}{\mu_g^2m}\right)\sum_{t=1}^{T}\mathbb{E}[\delta_{\alpha,t}]\\
&\quad+\left(\frac{\widetilde{C}_2}{m}+\frac{24L_g^2C_3}{\mu_g^2m}\right)\sum_{t=1}^{T}\mathbb{E}[\delta_{y,t}]+\frac{4C_3}{\eta_3\mu_g|I_t|}\delta_{\mathbf{v},0}+\frac{8C_3\mu_g\eta_3\sigma^2T}{B}+\frac{C_4\beta_0T}{\min\{|I_t|,B\}}\\
&\le \frac{1}{\beta_0}\mathbb{E}[\|\nabla F(\mathbf{x}_0)-\mathbf{z}_1\|^2]+\left(\frac{4L_F^2\eta_0^2}{\beta_0^2}+\frac{32m^2C_v^2\eta_0^2C_3}{\eta_3^2\mu_g^2|I_t|^2}+\frac{32m^2C_\alpha^2\eta_0^2C_1}{\eta_1^2\mu_f^2|I_t|^2}\right)\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{z}_t\|^2]\\
&\quad+\frac{4C_1}{\mu_f|I_t|\eta_1}\delta_{\alpha,0}+\frac{8C_1\mu_f\eta_1\sigma^2T}{B}+\frac{4C_3}{\eta_3\mu_g|I_t|}\delta_{\mathbf{v},0}+\frac{8C_3\mu_g\eta_3\sigma^2T}{B}+\frac{C_4\beta_0T}{\min\{|I_t|,B\}}\\
&\quad+\left(\frac{\widetilde{C}_2}{m}+\frac{24L_g^2C_3}{\mu_g^2m}+\frac{24L_f^2C_1}{\mu_f^2m}\right)\sum_{t=1}^{T}\mathbb{E}[\delta_{y,t}]\\
&\le \frac{1}{\beta_0}\mathbb{E}[\|\nabla F(\mathbf{x}_0)-\mathbf{z}_1\|^2]+\left(\frac{4L_F^2\eta_0^2}{\beta_0^2}+\frac{32m^2C_v^2\eta_0^2C_3}{\eta_3^2\mu_g^2|I_t|^2}+\frac{32m^2C_\alpha^2\eta_0^2C_1}{\eta_1^2\mu_f^2|I_t|^2}+\frac{8m^2C_y^2\eta_0^2C_2}{|I_t|^2\eta_2^2\mu_g^2}\right)\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{z}_t\|^2]\\
&\quad+\frac{4C_1}{\mu_f|I_t|\eta_1}\delta_{\alpha,0}+\frac{8C_1\mu_f\eta_1\sigma^2T}{B}+\frac{4C_3}{\eta_3\mu_g|I_t|}\delta_{\mathbf{v},0}+\frac{8C_3\mu_g\eta_3\sigma^2T}{B}+\frac{C_4\beta_0T}{\min\{|I_t|,B\}}\\
&\quad+\frac{2C_2}{|I_t|\eta_2\mu_g}\delta_{y,0}+\frac{4C_2\eta_2T\sigma^2}{\mu_gB}
\end{aligned}
\tag{30}
$$

where $C_1 := \widetilde{C}_1 + \frac{24L_g^2C_3}{\mu_g^2}$ and $C_2 := \widetilde{C}_2 + \frac{24L_g^2C_3}{\mu_g^2} + \frac{24L_f^2C_1}{\mu_f^2}$

Recall Lemma A.3, we have

$$
F(\mathbf{x}_{t+1}) \le F(\mathbf{x}_t) + \frac{\eta_x}{2}\|\nabla F(\mathbf{x}_t)-\mathbf{z}_{t+1}\|^2 - \frac{\eta_0}{2}\|\nabla F(\mathbf{x}_t)\|^2 - \frac{\eta_0}{4}\|\mathbf{z}_{t+1}\|^2
$$

Combining with 30, we obtain

$$\begin{aligned}
&\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)\|^2] \\
&\leq \frac{2\mathbb{E}[F(\mathbf{x}_0)-F(\mathbf{x}_{T+1})]}{\eta_0 T}+\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)-\mathbf{z}_{t+1}\|^2]-\frac{1}{2T}\sum_{t=0}^{T}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2] \\
&\leq \frac{1}{T}\left\{\frac{2[F(\mathbf{x}_0)-F(\mathbf{x}^*)]}{\eta_0}+\frac{1}{\beta_0}\mathbb{E}[\|\nabla F(\mathbf{x}_0)-\mathbf{z}_1\|^2]+\frac{4C_1}{\mu_f|I_t|\eta_1}\delta_{\alpha,0}+\frac{4C_3}{\eta_3\mu_g|I_t|}\delta_{\mathbf{v},0}+\frac{2C_2}{|I_t|\eta_2\mu_g}\delta_{y,0}\right\} \\
&\quad+\frac{8C_1\mu_f\eta_1\sigma^2}{B}+\frac{8C_3\mu_g\eta_3\sigma^2}{B}+\frac{4C_2\eta_2\sigma^2}{\mu_g B}+\frac{C_4\beta_0}{\min\{|I_t|,B\}} \\
&\quad+\frac{1}{T}\left(\frac{4L_F^2\eta_0^2}{\beta_0^2}+\frac{32m^2C_v^2\eta_0^2C_3}{\eta_3^2\mu_g^2|I_t|^2}+\frac{32m^2C_\alpha^2\eta_0^2C_1}{\eta_1^2\mu_f^2|I_t|^2}+\frac{8m^2C_y^2\eta_0^2C_2}{|I_t|^2\eta_2^2\mu_g^2}-\frac{1}{2}\right)\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{z}_t\|^2]
\end{aligned} \tag{31}$$

By setting

$$\eta_0^2\leq\min\left\{\frac{\beta_0^2}{64L_F^2},\frac{\eta_3^2\mu_g^2|I_t|^2}{512m^2C_v^2C_3},\frac{\eta_1^2\mu_f^2|I_t|^2}{512m^2C_\alpha^2C_1},\frac{|I_t|^2\eta_2^2\mu_g^2}{128m^2C_y^2C_2}\right\}$$

we have

$$\frac{4L_F^2\eta_0^2}{\beta_0^2}+\frac{32m^2C_v^2\eta_0^2C_3}{\eta_3^2\mu_g^2|I_t|^2}+\frac{32m^2C_\alpha^2\eta_0^2C_1}{\eta_1^2\mu_f^2|I_t|^2}+\frac{8m^2C_y^2\eta_0^2C_2}{|I_t|^2\eta_2^2\mu_g^2}-\frac{1}{4}\leq 0$$

which implies that the last term of the right hand side of inequality 31 is less than or equal to zero. Hence

$$\begin{aligned}
&\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)\|^2] \\
&\leq \frac{1}{T}\left\{\frac{2[F(\mathbf{x}_0)-F(\mathbf{x}^*)]}{\eta_0}+\frac{1}{\beta_0}\mathbb{E}[\|\nabla F(\mathbf{x}_0)-\mathbf{z}_1\|^2]+\frac{4C_1}{\mu_f|I_t|\eta_1}\delta_{\alpha,0}+\frac{4C_3}{\eta_3\mu_g|I_t|}\delta_{\mathbf{v},0}+\frac{2C_2}{|I_t|\eta_2\mu_g}\delta_{y,0}\right\} \\
&\quad+\frac{8C_1\mu_f\eta_1\sigma^2}{B}+\frac{8C_3\mu_g\eta_3\sigma^2}{B}+\frac{4C_2\eta_2\sigma^2}{\mu_g B}+\frac{C_4\beta_0}{\min\{|I_t|,B\}}
\end{aligned} \tag{32}$$

With

$$\eta_1\leq\frac{B\epsilon^2}{96C_1\mu_f\sigma^2},\eta_3\leq\frac{B\epsilon^2}{96C_3\mu_g\sigma^2},\eta_2\leq\frac{\mu_g B\epsilon^2}{48C_2\sigma^2},\beta_0\leq\frac{\min\{|I_t|,B\}\epsilon^2}{12C_4}$$

$$T\geq\max\left\{\frac{32[F(\mathbf{x}_0)-F(\mathbf{x}^*)]}{\eta_0\epsilon^2},\frac{15\mathbb{E}[\|\nabla F(\mathbf{x}_0)-\mathbf{z}_1\|^2]}{\beta_0\epsilon^2},\frac{60C_1\delta_{\alpha,0}}{\mu_f|I_t|\eta_1\epsilon^2},\frac{60C_3\delta_{\mathbf{v},0}}{\eta_3\mu_g|I_t|\epsilon^2},\frac{30C_2\delta_{y,0}}{|I_t|\eta_2\mu_g\epsilon^2}\right\}$$

we have

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)\|^2]\leq\frac{\epsilon^2}{3}+\frac{\epsilon^2}{3}<\epsilon^2$$

Furthermore, to show the second part of the theorem, following from inequality 30, we have

$$\begin{aligned}
&\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)-\mathbf{z}_{t+1}\|^2] \\
&\leq \frac{\mathbb{E}[\|\nabla F(\mathbf{x}_0)-\mathbf{z}_1\|^2]}{\beta_0}+\frac{4C_1\delta_{\alpha,0}}{\mu_f|I_t|\eta_1}+\frac{8C_1\mu_f\eta_1\sigma^2T}{B}+\frac{4C_3\delta_{\mathbf{v},0}}{\eta_3\mu_g|I_t|}+\frac{8C_3\mu_g\eta_3\sigma^2T}{B}+\frac{C_4\beta_0T}{\min\{|I_t|,B\}} \\
&\quad+\frac{2C_2\delta_{y,0}}{|I_t|\eta_2\mu_g}+\frac{4C_2\eta_2T\sigma^2}{\mu_g B}+\frac{1}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\mathbf{x}_t)\|^2]+\mathbb{E}[\|\nabla F(\mathbf{x}_t)-\mathbf{z}_{t+1}\|^2]
\end{aligned} \tag{33}$$

With parameter set above, we have

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)-\mathbf{z}_{t+1}\|^2]<2\epsilon^2$$

□

B.0.1 Proof of Lemma B.3

This proof is the same as the proof of Lemma 2.5.

C Algorithm and Derivation for Multi-task Deep Partial AUC Maximization

Algorithm 3 Min-Max Bilevel Optimization for pAUC Maximization (MMB-pAUC)

Require: $\boldsymbol{\alpha}^0, \boldsymbol{\lambda}^0, H^0, \mathbf{z}^0, \mathbf{w}^0, \mathbf{a}^0, \mathbf{b}^0$

for $t = 0, 1, \ldots, T$ **do**
  Draw task batch $I_t \subset \mathcal{S}$.
  Draw data sample batch $\mathcal{B}_k^t$ for each $k \in I_t$
  For sampled tasks $k \in I_t$, update
    $\alpha_k^{t+1} \leftarrow \alpha_k^t + \eta_1 G_\alpha(\mathbf{w}^t, \alpha_k^t, \lambda_k^t; \mathcal{B}_k^t)$ ⋄$G_\alpha(\cdot)$ denotes a stochastic gradient w.r.t α_k
    $\lambda_k^{t+1} \leftarrow \lambda_k^t - \eta_2 \nabla_\lambda L_k(\lambda_k^t, \mathbf{w}^t; \mathcal{B}_k^t)$
    $H_k^{t+1} \leftarrow (1-\beta_1)H_k^t + \beta_1 \nabla^2_{\lambda\lambda} L_k(\lambda_k^t, \mathbf{w}^t; \mathcal{B}_k^t)$
  Compute loss G^{t+1} according to 38 ⋄G^{t+1} denotes an appropriate loss
  Update gradient estimator $\Delta^{t+1} \leftarrow$ autograd(G^{t+1})
  $\mathbf{z}^{t+1} \leftarrow (1-\beta_0)\mathbf{z}^t + \beta_0 \Delta^{t+1}$
  $(\mathbf{w}^{t+1}, \mathbf{a}^{t+1}, \mathbf{b}^{t+1}) \leftarrow (\mathbf{w}^t, \mathbf{a}^t, \mathbf{b}^t) - \eta_0 \mathbf{z}^{t+1}$
end for

Let $\mathcal{D}_-[K]$ denote the top-K negative examples according to their prediction scores. Let n_+, n_- denote the number of positive and negative label samples respectively. Then we have partial AUC loss formulated as following

$$\min_{\mathbf{w}} \frac{1}{n_+}\sum_{\mathbf{x}_i\in\mathcal{D}_+}\frac{1}{n_-\rho}\sum_{\mathbf{x}_j\in\mathcal{D}_-[K]}(h_{\mathbf{w}}(\mathbf{x}_j)-h_{\mathbf{w}}(\mathbf{x}_i)+c)^2$$

where $K = n_-\rho$. Let $a(\mathbf{w}) = \frac{1}{n_+}\sum_{\mathbf{x}_i\in\mathcal{D}_+} h_{\mathbf{w}}(\mathbf{x}_i)$ and $b(\mathbf{w}) = \frac{1}{n_-\rho}\sum_{\mathbf{x}_j\in\mathcal{D}_-[K]} h_{\mathbf{w}}(\mathbf{x}_j)$. Then we can transform the objective as

$$\begin{aligned}
&\frac{1}{n_+}\sum_{\mathbf{x}_i\in\mathcal{D}_+}\frac{1}{n_-\rho}\sum_{\mathbf{x}_j\in\mathcal{D}_-[K]}(h_{\mathbf{w}}(\mathbf{x}_j)-b(\mathbf{w})+a(\mathbf{w})-h_{\mathbf{w}}(\mathbf{x}_i)+b(\mathbf{w})-a(\mathbf{w})+c)^2\\
&=\frac{1}{n_+}\sum_{\mathbf{x}_i\in\mathcal{D}_+}(h_{\mathbf{w}}(\mathbf{x}_i)-a(\mathbf{w}))^2+\frac{1}{n_-\rho}\sum_{\mathbf{x}_j\in\mathcal{D}_-[K]}(h_{\mathbf{w}}(\mathbf{x}_j)-b(\mathbf{w}))^2+(b(\mathbf{w})-a(\mathbf{w})+c)^2
\end{aligned}$$

Then we can write the problem as

$$\min_{\mathbf{w},a,b}\frac{1}{n_+}\sum_{\mathbf{x}_i\in\mathcal{D}_+}(h_{\mathbf{w}}(\mathbf{x}_i)-a)^2+\frac{1}{n_-\rho}\sum_{\mathbf{x}_j\in\mathcal{D}_-}\mathbb{I}(\mathbf{x}_j\in\mathcal{D}_-[K])(h_{\mathbf{w}}(\mathbf{x}_j)-b)^2+(b(\mathbf{w})-a(\mathbf{w})+c)^2 \quad (34)$$

Replacing the indicator function by $\mathbb{I}(\mathbf{x}_j\in\mathcal{D}_-[K]) = \psi(h_{\mathbf{w}}(\mathbf{x}_j)-\lambda(\mathbf{w}))$, where $\lambda(\mathbf{w})$ represents the $K+1$-th largest scores among all negative examples, which can be represented as

$$\lambda(\mathbf{w}) = \arg\min_{\lambda}\frac{K+\varepsilon}{n_-}\lambda+\frac{1}{n_-}\sum_{\mathbf{x}\in\mathcal{D}_-}(h_{\mathbf{w}}(\mathbf{x})-\lambda)_+.$$

We smooth the problem as

$$\hat{\lambda}(\mathbf{w}) = \arg\min_{\lambda} L(\lambda,\mathbf{w}) := \frac{K+\varepsilon}{n_-}\lambda+\frac{\tau_2}{2}\lambda^2+\frac{1}{n_-}\sum_{\mathbf{x}\in\mathcal{D}_-}\tau_1\ln(1+\exp((h_{\mathbf{w}}(\mathbf{x})-\lambda)/\tau_1)).$$

Due to the fact that the last term $(b(\mathbf{w}) - a(\mathbf{w}) + c)^2$ in Problem 34 cannot be directly obtained since $a(\mathbf{w})$ and $b(\mathbf{w})$ are expectations, one may use $p^2 = \max_\alpha 2p\alpha - \alpha^2$ to get

$$(b(\mathbf{w}) - a(\mathbf{w}) + c)^2 = \max_{\alpha} \mathbb{E}_{\mathbf{x}_j \in \mathcal{D}_-[K], \mathbf{x}_i \in \mathcal{D}_+} \left[2\alpha(h_{\mathbf{w}}(\mathbf{x}_j) - h_{\mathbf{w}}(\mathbf{x}_i) + c) - \alpha^2\right] \tag{35}$$

Then by replacing the top-K selector with $\phi(h_{\mathbf{w}}(\mathbf{x}_j) - \hat{\lambda}(\mathbf{w}))$, the partial AUC minimization problem can be formulated as a min-max bilevel optimization problem.

$$\begin{aligned}
&\min_{\mathbf{w},a,b} \frac{1}{n_+} \sum_{\mathbf{x}_i \in \mathcal{D}_+} (h_{\mathbf{w}}(\mathbf{x}_i) - a)^2 + \frac{1}{n_-\rho} \sum_{\mathbf{x}_j \in \mathcal{D}_-} \phi(h_{\mathbf{w}}(\mathbf{x}_j) - \hat{\lambda}(\mathbf{w}))(h_{\mathbf{w}}(\mathbf{x}_j) - b)^2 \\
&+ \max_{\alpha} 2\alpha \left(\frac{1}{n_-\rho} \sum_{\mathbf{x}_j \in \mathcal{D}_-} \phi(h_{\mathbf{w}}(\mathbf{x}_j) - \hat{\lambda}(\mathbf{w})) h_{\mathbf{w}}(\mathbf{x}_j) - \frac{1}{n_+} \sum_{\mathbf{x}_i \in \mathcal{D}_+} h_{\mathbf{w}}(\mathbf{x}_i) + c \right) - \alpha^2 \\
&s.t., \hat{\lambda}(\mathbf{w}) = \arg\min_{\lambda} L(\lambda, \mathbf{w}) := \frac{K+\varepsilon}{n_-}\lambda + \frac{\tau_2}{2}\lambda^2 + \frac{1}{n_-} \sum_{\mathbf{x}_j \in \mathcal{D}_-} \tau_1 \ln(1 + \exp((h_{\mathbf{w}}(\mathbf{x}_j) - \lambda)/\tau_1))
\end{aligned}$$

We consider multi-task partial AUC maximization, which is then given by

$$\begin{aligned}
&\min_{\mathbf{w},\mathbf{a}\in\mathbb{R}^m,\mathbf{b}\in\mathbb{R}^m} \max_{\boldsymbol{\alpha}\in\mathbb{R}^m} \sum_{k=1}^{m} \left\{ \frac{1}{n_+^k} \sum_{\mathbf{x}_i \in \mathcal{D}_+^k} (h_{\mathbf{w}}(\mathbf{x}_i; k) - a_k)^2 + \frac{1}{n_-^k\rho} \sum_{\mathbf{x}_j \in \mathcal{D}_-^k} \phi(h_{\mathbf{w}}(\mathbf{x}_j; k) - \hat{\lambda}_k(\mathbf{w}))(h_{\mathbf{w}}(\mathbf{x}_j; k) - b_k)^2 \right. \\
&\left. + 2\alpha_k \left(\frac{1}{n_-^k\rho} \sum_{\mathbf{x}_j \in \mathcal{D}_-^k} \phi(h_{\mathbf{w}}(\mathbf{x}_j; k) - \hat{\lambda}_k(\mathbf{w})) h_{\mathbf{w}}(\mathbf{x}_j; k) - \frac{1}{n_+^k} \sum_{\mathbf{x}_i \in \mathcal{D}_+^k} h_{\mathbf{w}}(\mathbf{x}_i; k) + c \right) - \alpha_k^2 \right\} \\
&s.t., \hat{\lambda}_k(\mathbf{w}) = \arg\min_{\lambda} L_k(\lambda, \mathbf{w}) := \frac{K+\varepsilon}{n_-}\lambda + \frac{\tau_2}{2}\lambda^2 + \frac{1}{n_-} \sum_{\mathbf{x}_j \in \mathcal{D}_-^k} \tau_1 \ln(1 + \exp((h_{\mathbf{w}}(\mathbf{x}_j; k) - \lambda)/\tau_1))
\end{aligned}$$

For function ϕ, we use sigmoid function $\phi(s) = \frac{1}{1+\exp(-s)} = \sigma(s)$. $\nabla\phi(h_{\mathbf{w}}(\mathbf{x}_j; k) - \lambda_k(\mathbf{w})) = \sigma(h_{\mathbf{w}}(\mathbf{x}_j; k) - \lambda_k(\mathbf{w}))(1 - \sigma(h_{\mathbf{w}}(\mathbf{x}_j; k) - \lambda_k(\mathbf{w}))(\nabla h_{\mathbf{w}}(\mathbf{x}_j; k) - \nabla\lambda_k(\mathbf{w}))$, where $\nabla\lambda_k(\mathbf{w}) = -(\nabla^2_{\lambda\lambda} L_k(\lambda, \mathbf{w}))^{-1} \nabla^2_{\mathbf{w}\lambda} L_k(\lambda, \mathbf{w})$. Let $H_k = (\nabla^2_{\lambda\lambda} L_k(\lambda, \mathbf{w}))$, and

$$\begin{aligned}
\nabla^2_{\mathbf{w}\lambda} L_k(\lambda, \mathbf{w}) &= \nabla_{\mathbf{w}} \left[\nabla_\lambda L_k(\lambda, \mathbf{w})\right] \\
&= \nabla_{\mathbf{w}} \left(\frac{K+\varepsilon}{n_-} + \tau_2\lambda - \frac{1}{n_-} \sum_{\mathbf{x}_j \in \mathcal{D}_-^k} \frac{\exp((h_{\mathbf{w}}(\mathbf{x}_j; k) - \lambda)/\tau_1)}{1 + \exp((h_{\mathbf{w}}(\mathbf{x}_j; k) - \lambda)/\tau_1)} \right) \\
&= -\frac{1}{n_-} \sum_{\mathbf{x}_j \in \mathcal{D}_-^k} \nabla_{\mathbf{w}} \left(\frac{\exp((h_{\mathbf{w}}(\mathbf{x}_j; k) - \lambda)/\tau_1)}{1 + \exp((h_{\mathbf{w}}(\mathbf{x}_j; k) - \lambda)/\tau_1)} \right)
\end{aligned}$$

As a result in order to compute in Pytorch $\nabla\phi(h_{\mathbf{w}}(\mathbf{x}_j; k) - \lambda(\mathbf{w}))$ we can define the following loss

$$L_\phi^k(\mathbf{w}) = st(\sigma(h_{\mathbf{w}}(\mathbf{x}_j; k) - \lambda_k)(1 - \sigma(h_{\mathbf{w}}(\mathbf{x}_j; k) - \lambda_k)) \cdot \tag{36}$$

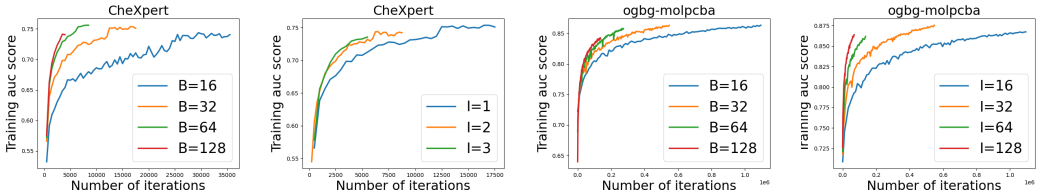
$$\left(h_{\mathbf{w}}(\mathbf{x}_j; k) - st([H_k]^{-1}) \frac{1}{B_-} \sum_{\mathbf{x}_j \in \mathcal{B}_-^k} \frac{\exp((h_{\mathbf{w}}(\mathbf{x}_j; k) - \lambda_k)/\tau_1)}{1 + \exp((h_{\mathbf{w}}(\mathbf{x}_j; k) - \lambda_k)/\tau_1)} \right) \tag{37}$$

so that $\nabla L_\phi^k(\mathbf{w}) = \nabla\phi(h_\mathbf{w}(\mathbf{x}_j;k) - \lambda(\mathbf{w}))$. Hence for updating $\mathbf{w}, \mathbf{a}, \mathbf{b}$ we can define the following loss and the Pytorch will compute the gradient automatically:

$$
\begin{aligned}
G^{t+1} := \frac{1}{|I_t|}\sum_{k\in I_t}\Bigg\{ & \frac{1}{B_+^k}\sum_{\mathbf{x}_i\in\mathcal{B}_+^k}(h_\mathbf{w}(\mathbf{x}_i;k)-a_k)^2 + \frac{1}{B_-^k\rho}\sum_{\mathbf{x}_j\in\mathcal{B}_-^k} st(\phi(h_\mathbf{w}(\mathbf{x}_j;k)-\lambda_k))(h_\mathbf{w}(\mathbf{x}_j;k)-b_k)^2 \\
& + \frac{1}{B_-^k\rho}\sum_{\mathbf{x}_j\in\mathcal{B}_-^k} L_\phi^k(\mathbf{w})st((h_\mathbf{w}(\mathbf{x}_j;k)-b_k)^2) + 2\alpha_k\Bigg(\frac{1}{B_-^k\rho}\sum_{\mathbf{x}_j\in\mathcal{B}_-^k} st(\phi(h_\mathbf{w}(\mathbf{x}_j;k)-\lambda_k))h_\mathbf{w}(\mathbf{x}_j;k) \\
& + \frac{1}{B_-^k\rho}\sum_{\mathbf{x}_j\in\mathcal{B}_-^k} L_\phi^k(\mathbf{w})st(h_\mathbf{w}(\mathbf{x}_j;k)) - \frac{1}{B_+^k}\sum_{\mathbf{x}_i\in\mathcal{B}_+^k} h_\mathbf{w}(\mathbf{x}_i;k) + c\Bigg)\Bigg\}
\end{aligned} \tag{38}
$$

The loss for α_k, λ_k and H_k in the mini-batch for computing the gradient can be easily defined. For practical version the terms that involve $L_\phi^k(\mathbf{w})$ can be ignored.

D Additional Experiments



(a) Varying B (b) Varying I (c) Varying B (d) Varying I

Figure 3: Convergence of our method vs data sample batch sizes B and vs task sample batch size $I := |I_t|$ for multi-task deep AUC maximization.

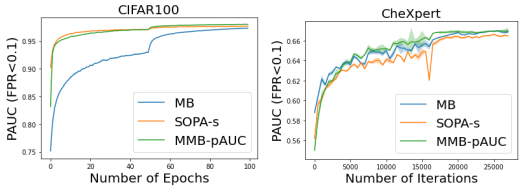


Figure 4: Comparison of Convergence on training data for multi-task deep pAUC maximization on the CIFAR100 and CheXpert datasets.