A Proofs of Theorem 4.6 and Theorem 4.7

In this section, we provide the detailed proofs for Theorem 4.6 and Theorem 4.7. We first give a basic property for weakly-convex functions.

Proposition A.1 (Proposition 2.1 in [7]). *Suppose function $g : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is lower-semicontinuous. Then g is ρ-weakly-convex if and only if*

$$g(y) \geq g(x) + \langle v, y - x \rangle - \frac{\rho}{2}\|y - x\|^2 \tag{10}$$

holds for all vectors $v \in \partial g(x)$ and $x, y \in \mathbb{R}^d$.

A.1 Proof of Theorem 4.6

Note that the proof of Lemma 4.5 also implies the following squared-norm error bound,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i\in\mathcal{S}}\|u_{i,t+1} - g_i(\mathbf{w}_{t+1})\|^2\right] \leq (1 - \frac{B_1\tau}{2n})^{t+1}\frac{1}{n}\sum_{i\in\mathcal{S}}\|u_{i,0} - g_i(\mathbf{w}_0)\|^2 + \frac{4\tau\sigma^2}{B_2} + \frac{16n^2C_g^2M^2\eta^2}{B_1^2\tau}.$$

Proof of Theorem 4.6. Define $\hat{\mathbf{w}}_t := \text{prox}_{F/\bar{\rho}}(\mathbf{w}_t)$. For a given $i \in \mathcal{S}$, we have

$$\begin{aligned}
&f_i(g_i(\hat{\mathbf{w}}_t)) - f_i(u_{i,t}) \\
&\overset{(a)}{\geq} \partial f_i(u_{i,t})^\top(g_i(\hat{\mathbf{w}}_t) - u_{i,t}) - \frac{\rho_f}{2}\|g_i(\hat{\mathbf{w}}_t) - u_{i,t}\|^2 \\
&\geq \partial f_i(u_{i,t})^\top(g_i(\hat{\mathbf{w}}_t) - u_{i,t}) - \rho_f\|g_i(\hat{\mathbf{w}}_t) - g_i(\mathbf{w}_t)\|^2 - \rho_f\|g_i(\mathbf{w}_t) - u_{i,t}\|^2 \\
&\geq \partial f_i(u_{i,t})^\top(g_i(\hat{\mathbf{w}}_t) - u_{i,t}) - \rho_f C_g^2\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 - \rho_f\|g_i(\mathbf{w}_t) - u_{i,t}\|^2 \\
&\overset{(b)}{\geq} \partial f_i(u_{i,t})^\top\left[g_i(\mathbf{w}_t) - u_{i,t} + \partial g_i(\mathbf{w}_t)^\top(\hat{\mathbf{w}}_t - \mathbf{w}_t) - \frac{\rho_g}{2}\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2\right] \\
&\quad - \rho_f C_g^2\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 - \rho_f\|g_i(\mathbf{w}_t) - u_{i,t}\|^2 \\
&\overset{(c)}{\geq} \partial f_i(u_{i,t})^\top(g_i(\mathbf{w}_t) - u_{i,t}) + \partial f_i(u_{i,t})^\top\partial g_i(\mathbf{w}_t)^\top(\hat{\mathbf{w}}_t - \mathbf{w}_t) - (\frac{\rho_g C_f}{2} + \rho_f C_g^2)\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 \\
&\quad - \rho_f\|g_i(\mathbf{w}_t) - u_{i,t}\|^2
\end{aligned}$$

where (a) follows from the ρ_f-weak-convexity of f_i, (b) follows from that $f_i(\cdot)$ is non-decreasing and the weak convexity of g_i, (c) is due to $0 \leq \partial f_i(u_{i,t}) \leq C_f$. Then it follows

$$\begin{aligned}
&\frac{1}{n}\sum_{i\in\mathcal{S}}\partial f_i(u_{i,t})^\top\partial g_i(\mathbf{w}_t)^\top(\hat{\mathbf{w}}_t - \mathbf{w}_t) \\
&\leq \frac{1}{n}\sum_{i\in\mathcal{S}}\left[f_i(g_i(\hat{\mathbf{w}}_t)) - f_i(u_{i,t}) - \partial f_i(u_{i,t})^\top(g_i(\mathbf{w}_t) - u_{i,t}) + (\frac{\rho_g C_f}{2} + \rho_f C_g^2)\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2\right. \\
&\quad \left. + \rho_f\|g_i(\mathbf{w}_t) - u_{i,t}\|^2\right]
\end{aligned} \tag{11}$$

Now we consider the change in the Moreau envelope:

$$\begin{aligned}
\mathbb{E}_t[F_{1/\bar{\rho}}(\mathbf{w}_{t+1})] &= \mathbb{E}_t\left[\min_{\tilde{\mathbf{w}}} F(\tilde{\mathbf{w}}) + \frac{\bar{\rho}}{2}\|\tilde{\mathbf{w}} - \mathbf{w}_{t+1}\|^2\right] \\
&\leq \mathbb{E}_t\left[F(\hat{\mathbf{w}}_t) + \frac{\bar{\rho}}{2}\|\hat{\mathbf{w}}_t - \mathbf{w}_{t+1}\|^2\right] \\
&= F(\hat{\mathbf{w}}_t) + \mathbb{E}_t\left[\frac{\bar{\rho}}{2}\|\hat{\mathbf{w}}_t - (\mathbf{w}_t - \eta G_t)\|^2\right] \\
&\leq F(\hat{\mathbf{w}}_t) + \frac{\bar{\rho}}{2}\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 + \bar{\rho}\mathbb{E}_t[\eta\langle\hat{\mathbf{w}}_t - \mathbf{w}_t, G_t\rangle] + \frac{\eta^2\bar{\rho}M^2}{2} \\
&= F_{1/\bar{\rho}}(\mathbf{w}_t) + \bar{\rho}\eta\langle\hat{\mathbf{w}}_t - \mathbf{w}_t, \mathbb{E}_t[G_t]\rangle + \frac{\eta^2\bar{\rho}M^2}{2}
\end{aligned} \tag{12}$$

where

$$\mathbb{E}_t[G_t] = \frac{1}{n}\sum_{i\in\mathcal{S}} \partial g_i(\mathbf{w}_t)\partial f_i(u_{i,t}),$$

and the second inequality uses the bound of $\mathbb{E}[\|G_t\|^2]$, which follows from the Lipschitz continuity and bounded variance assumptions and is denoted by M.

Combining inequality 29 and 30 yields

$$\begin{aligned}
&\mathbb{E}_t[F_{1/\bar\rho}(\mathbf{w}_{t+1})] \\
&\leq F_{1/\bar\rho}(\mathbf{w}_t) + \frac{\eta^2\bar\rho M^2}{2} + \frac{\bar\rho\eta}{n}\sum_{i\in\mathcal{S}}\bigg[f_i(g_i(\hat{\mathbf{w}}_t)) - f_i(u_{i,t}) \\
&\quad - \partial f_i(u_{i,t})^\top(g_i(\mathbf{w}_t) - u_{i,t}) + (\frac{\rho_g C_f}{2} + \rho_f C_g^2)\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 + \rho_f\|g_i(\mathbf{w}_t) - u_{i,t}\|^2\bigg] \qquad (13)\\
&= F_{1/\bar\rho}(\mathbf{w}_t) + \frac{\eta^2\bar\rho M^2}{2} + \frac{\bar\rho\eta}{n}\sum_{i\in\mathcal{S}}\bigg[F_i(\hat{\mathbf{w}}_t) - F_i(\mathbf{w}_t) + f_i(g_i(\mathbf{w}_t)) - f_i(u_{i,t}) \\
&\quad - \partial f_i(u_{i,t})^\top(g_i(\mathbf{w}_t) - u_{i,t}) + (\frac{\rho_g C_f}{2} + \rho_f C_g^2)\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 + \rho_f\|g_i(\mathbf{w}_t) - u_{i,t}\|^2\bigg]
\end{aligned}$$

Due to the ρ_F-weak convexity of $F_i(\mathbf{w})$, we have $(\bar\rho - \rho_F)$-strong convexity of $\mathbf{w} \mapsto F_i(\mathbf{w}) + \frac{\bar\rho}{2}\|\mathbf{w}_t - \mathbf{w}\|^2$. Then it follows

$$\begin{aligned}
F_i(\hat{\mathbf{w}}_t) - F_i(\mathbf{w}_t) &= \bigg[F_i(\hat{\mathbf{w}}_t) + \frac{\bar\rho}{2}\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2\bigg] - \bigg[F_i(\mathbf{w}_t) + \frac{\bar\rho}{2}\|\mathbf{w}_t - \mathbf{w}_t\|^2\bigg] - \frac{\bar\rho}{2}\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2 \\
&\leq (\frac{\rho_F}{2} - \bar\rho)\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2
\end{aligned} \qquad (14)$$

Plugging inequality 32 into inequality 31 yields

$$\begin{aligned}
\mathbb{E}_t[F_{1/\bar\rho}(\mathbf{w}_{t+1})] \leq \mathbb{E}[F_{1/\bar\rho}(\mathbf{w}_t)] + \frac{\eta^2\bar\rho M^2}{2} + \frac{\bar\rho\eta}{n}\sum_{i\in\mathcal{S}}\bigg[(\frac{\rho_F}{2} - \bar\rho)\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2 \\
+ f_i(g_i(\mathbf{w}_t)) - f_i(u_{i,t}) - \partial f_i(u_{i,t})^\top(g_i(\mathbf{w}_t) - u_{i,t}) \\
+ (\frac{\rho_g C_f}{2} + \rho_f C_g^2)\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 + \rho_f\|g_i(\mathbf{w}_t) - u_{i,t}\|^2\bigg]
\end{aligned} \qquad (15)$$

Set $\bar\rho = \rho_F + \rho_g C_f + 2\rho_f C_g^2$. We have

$$\begin{aligned}
\mathbb{E}_t[F_{1/\bar\rho}(\mathbf{w}_{t+1})] &\leq F_{1/\bar\rho}(\mathbf{w}_t) + \frac{\eta^2\bar\rho M^2}{2} + \frac{\bar\rho\eta}{n_+}\sum_{i\in\mathcal{S}}\bigg[-\frac{\bar\rho}{2}\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2 + f_i(g_i(\mathbf{w}_t)) - f_i(u_{i,t}) \\
&\quad - \partial f_i(u_{i,t})^\top(g_i(\mathbf{w}_t) - u_{i,t}) + \rho_f\|g_i(\mathbf{w}_t) - u_{i,t}\|^2\bigg] \\
&\overset{(a)}{\leq} F_{1/\bar\rho}(\mathbf{w}_t) + \frac{\eta^2\bar\rho M^2}{2} - \frac{\eta}{2}\|\nabla F_{1/\bar\rho}(\mathbf{w}_t)\|^2 + \frac{\bar\rho\eta}{n}\sum_{i\in\mathcal{S}}\bigg[f_i(g_i(\mathbf{w}_t)) - f_i(u_{i,t}) \\
&\quad - \partial f_i(u_{i,t})^\top(g_i(\mathbf{w}_t) - u_{i,t}) + \rho_f\|g_i(\mathbf{w}_t) - u_{i,t}\|^2\bigg]
\end{aligned}$$

where inequality (a) follows from Lemma 3.2.

Using the Lipschitz continuity of f_i, we have

$$\begin{aligned}
\mathbb{E}_t[F_{1/\bar\rho}(\mathbf{w}_{t+1})] \leq F_{1/\bar\rho}(\mathbf{w}_t) + \frac{\eta^2\bar\rho M^2}{2} - \frac{\eta}{2}\|\nabla F_{1/\bar\rho}(\mathbf{w}_t)\|^2 + \frac{\bar\rho\eta}{n}\sum_{i\in\mathcal{S}} 2C_f\|g_i(\mathbf{w}_t) - u_{i,t}\| \\
+ \frac{\bar\rho\eta}{n}\sum_{i\in\mathcal{S}}\rho_f\|g_i(\mathbf{w}_t) - u_{i,t}\|^2
\end{aligned}$$

By Lemma 4.5, the error bound of the MSVR update gives

$$\mathbb{E}\left[\frac{1}{n}\sum_{i\in\mathcal{S}}\|u_{i,t}-g_i(\mathbf{w}_t)\|\right] \leq (1-\mu)^t\frac{1}{n}\sum_{i\in\mathcal{S}}\|u_{i,0}-g_i(\mathbf{w}_0)\|+R_1,$$

$$\mathbb{E}\left[\frac{1}{n}\sum_{i\in\mathcal{S}}\|u_{i,t}-g_i(\mathbf{w}_t)\|^2\right] \leq (1-\mu)^t\frac{1}{n}\sum_{i\in\mathcal{S}}\|u_{i,0}-g_i(\mathbf{w}_0)\|^2+R_2,$$

where

$$\mu=\frac{B_1\tau}{2n},\quad R_1=\frac{2\tau^{1/2}\sigma}{B_2^{1/2}}+\frac{4nC_gM\eta}{B_1\tau^{1/2}},\quad R_2=\frac{4\tau\sigma^2}{B_2}+\frac{16n^2C_g^2M^2\eta^2}{B_1^2\tau}$$

Then

$$\begin{aligned}\mathbb{E}[F_{1/\bar\rho}(\mathbf{w}_{t+1})] &\leq F_{1/\bar\rho}(\mathbf{w}_t)+\frac{\eta^2\bar\rho M^2}{2}-\frac{\eta}{2}\mathbb{E}[\|\nabla F_{1/\bar\rho}(\mathbf{w}_t)\|^2]\\ &\quad+2C_f\bar\rho\eta\left((1-\mu)^t\frac{1}{n}\sum_{i\in\mathcal{S}}\|g_i(\mathbf{w}_0)-u_{i,0}\|+R_1\right)\\ &\quad+C\rho_f\bar\rho\eta\left((1-\mu)^t\frac{1}{n}\sum_{i\in\mathcal{S}}\|g_i(\mathbf{w}_0)-u_{i,0}\|^2+R_2\right)\end{aligned} \tag{16}$$

Taking summation from $t=0$ to $T-1$ yields

$$\begin{aligned}&\mathbb{E}[F_{1/\bar\rho}(\mathbf{w}_T)]\\ &\leq F_{1/\bar\rho}(\mathbf{w}_0)+\frac{\eta^2\bar\rho M^2T}{2}-\frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F_{1/\bar\rho}(\mathbf{w}_t)\|^2]\\ &\quad+2C_f\bar\rho\eta\left(\sum_{t=0}^{T-1}(1-\mu)^t\frac{1}{n}\sum_{i\in\mathcal{S}}\|g_i(\mathbf{w}_0)-u_{i,0}\|+R_1T\right)\\ &\quad+C\rho_f\bar\rho\eta\left((1-\mu)^t\frac{1}{n}\sum_{i\in\mathcal{S}}\|g_i(\mathbf{w}_0)-u_{i,0}\|^2+R_2T\right)\\ &\overset{(a)}{\leq} F_{1/\bar\rho}(\mathbf{w}_0)+\frac{\eta^2\bar\rho M^2T}{2}-\frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F_{1/\bar\rho}(\mathbf{w}_t)\|^2]\\ &\quad+\frac{2C_f\bar\rho\eta}{n\mu}\sum_{i\in\mathcal{S}}\|g_i(\mathbf{w}_0)-u_{i,0}\|+2C_f\bar\rho\eta R_1T+\frac{\rho_f\bar\rho\eta}{n\mu}\sum_{i\in\mathcal{S}}\|g_i(\mathbf{w}_0)-u_{i,0}\|^2+2\rho_f\bar\rho\eta R_2T,\end{aligned} \tag{17}$$

where (a) uses $\sum_{t=0}^{T-1}(1-\mu)^t\leq\frac{1}{\mu}$.

Lower bounding the left-hand-side by $\min_{\mathbf{w}}F(\mathbf{w})$, we obtain

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F_{1/\bar\rho}(\mathbf{w}_t)\|^2]\\ &\leq\frac{2}{\eta T}\left[F_{1/\bar\rho}(\mathbf{w}_0)-\min_{\mathbf{w}}F(\mathbf{w})+\frac{\eta^2\bar\rho M^2T}{2}+\frac{2C_f\bar\rho\eta}{n}\sum_{i\in\mathcal{S}}\|g_i(\mathbf{w}_0)-u_{i,0}\|+2C_f\bar\rho\eta R_1T\right.\\ &\quad\left.+\frac{\rho_f\bar\rho\eta}{n}\sum_{i\in\mathcal{S}}\|g_i(\mathbf{w}_0)-u_{i,0}\|^2+\rho_f\bar\rho\eta R_2T\right]\\ &\leq\frac{2\Delta}{\eta T}+\eta\bar\rho M^2+\frac{4C_f\bar\rho}{\mu Tn}\sum_{i\in\mathcal{S}}\|g_i(\mathbf{w}_0)-u_{i,0}\|+4C_f\bar\rho R_1+\frac{2\rho_f\bar\rho}{\mu Tn}\sum_{i\in\mathcal{S}}\|g_i(\mathbf{w}_0)-u_{i,0}\|^2+2\rho_f\bar\rho R_2\\ &\leq\frac{C}{T}(\frac{1}{\eta}+\frac{1}{\mu})+C(\eta+R_1+R_2)\end{aligned}$$

where we assume $F_{1/\bar\rho}(\mathbf{w}_0,\mathbf{s}_0,s_0')-\min_{\mathbf{w},\mathbf{s},s'}F(\mathbf{w},\mathbf{s},s')\leq\Delta$ and

$$C=\max\{8\Delta,12\bar\rho M^2,\frac{16C_f\bar\rho}{n}\sum_{i\in\mathcal{S}}\|g_i(\mathbf{w}_0)-u_{i,0}\|,\frac{8\rho_f\bar\rho}{n}\sum_{i\in\mathcal{S}}\|g_i(\mathbf{w}_0)-u_{i,0}\|^2,16C_f\bar\rho,8\rho_f\bar\rho\}.$$

Thus

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t)\|^2] \\
&\leq \frac{C}{T}(\frac{1}{\eta}+\frac{2n}{B_1\tau})+C(\eta+\frac{2\tau^{1/2}\sigma}{B_2^{1/2}}+\frac{4nC_gM\eta}{B_1\tau^{1/2}}+\frac{4\tau\sigma^2}{B_2}+\frac{16n^2C_g^2M^2\eta^2}{B_1^2\tau}) \\
&= \mathcal{O}\left(\frac{1}{T}(\frac{1}{\eta}+\frac{n}{B_1\tau})+(\eta+\frac{\tau^{1/2}\sigma}{B_2^{1/2}}+\frac{n\eta}{B_1\tau^{1/2}}+\frac{\tau\sigma^2}{B_2}+\frac{n^2\eta^2}{B_1^2\tau})\right)
\end{aligned}$$

Setting

$$\tau=\mathcal{O}(B_2\epsilon^4),\quad \eta=\mathcal{O}\left(\frac{B_1B_2^{1/2}\epsilon^4}{n}\right)$$

To reach an ϵ-stationary point, we need

$$T=\mathcal{O}\left(\frac{n}{B_1B_2^{1/2}\epsilon^6}\right)$$

□

A.2 Proof of Theorem 4.7

A formal statement in given below.

Theorem A.2. *Under Assumption 4.3, with* $\gamma_1=\frac{n_1n_2-B_1B_2}{B_1B_2(1-\tau_1)}+(1-\tau_1)$*,* $\gamma_2=\frac{n_1-B_1}{B_1(1-\tau_2)}+(1-\tau_2)$*,* $\tau_1=\mathcal{O}\left(\min\{B_3,\frac{B_1^{1/2}n_2^{1/2}}{n_1^{1/2}}\}\epsilon^4\right)\leq\frac{1}{2}$*,* $\tau_2=\mathcal{O}(B_2\epsilon^4)\leq\frac{1}{2}$*,* $\eta=\mathcal{O}\left(\min\left\{B_3^{1/2},\frac{B_1^{1/4}n_2^{1/4}}{n_1^{1/4}},\frac{B_1^{1/2}n_2^{1/2}}{n_1^{1/2}}\right\}\frac{B_1B_2}{n_1n_2}\epsilon^4\right)$*, and* $\bar{\rho}=\rho_F+4\rho_fC_g^2+2\rho_gC_fC_h^2+C_fC_gL_h$*, Algorithm 2 converges to an* ϵ*-stationary point of the Moreau envelope* $F_{1/\bar{\rho}}$ *in* $T=\mathcal{O}\left(\max\left\{\frac{1}{B_3^{1/2}},\frac{n_1^{1/4}}{B_1^{1/4}n_2^{1/4}},\frac{n_1^{1/2}}{B_1^{1/2}n_2^{1/2}}\right\}\frac{n_1n_2}{B_1B_2}\epsilon^{-6}\right)$ *iterations.*

We first define constant $M^2\geq\max\{\frac{3C_f^2C_g^2\sigma^2}{B_3}+\frac{3C_f^2C_g^2C_h^2}{B_2}+\frac{3C_f^2C_g^2C_h^2}{B_1},\tilde{C}_h^2+\sigma^2\}$ so that $\mathbb{E}_t[\|G_t\|^2]\leq M^2$ and $\|v_{i,j,t}\|^2\leq M^2$ for all $i\in\mathcal{S}_1,j\in\mathcal{S}_2$ and t. Then to prove Theorem A.2, we need the following Lemmas.

Lemma A.3. *Consider MSVR update for* v*. Assume* $h_{i,j}(\mathbf{w};\xi)$ *is* C_h*-Lipshitz for all* $(i,j)\in S_1\times S_2$ *, and* $\mathbb{E}[\|G_t\|^2]\leq M^2$*. With* $\gamma_1=\frac{n_1n_2-B_1B_2}{B_1B_2(1-\tau_1)}+(1-\tau_1)$*, and* $\tau_1\leq\frac{1}{2}$*, we have*

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}\|v_{i,j,t+1}-h_{i,j}(\mathbf{w}_{t+1})\|\right] \\
&\leq(1-\frac{B_1B_2\tau_1}{2n_1n_2})^{t+1}\frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}\|v_{i,j,0}-h_{i,j}(\mathbf{w}_0)\|+\frac{2\tau_1^{1/2}\sigma}{B_3^{1/2}}+\frac{4n_1n_2C_hM\eta}{B_1B_2\tau_1^{1/2}} \\
&\mathbb{E}\left[\frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}\|v_{i,j,t+1}-h_{i,j}(\mathbf{w}_{t+1})\|^2\right] \\
&\leq(1-\frac{B_1B_2\tau_1}{2n_1n_2})^{2(t+1)}\frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}\|v_{i,j,0}-h_{i,j}(\mathbf{w}_0)\|^2+\frac{4\tau_1\sigma^2}{B_3}+\frac{16n_1^2n_2^2C_h^2M^2\eta^2}{B_1^2B_2^2\tau_1}
\end{aligned}$$

Lemma A.4. *Consider MSVR update for* u*. Assume* $g_i(\cdot)$ *is* C_g*-Lipshitz for all* $i \in S_1$*. With* $\gamma_2 = \frac{n_+ - B_1}{B_1(1-\tau_2)} + (1-\tau_2)$ *and* $\tau_2 \le \frac{1}{2}$*, we have*

$$\mathbb{E}\left[\frac{1}{n_1}\sum_{i\in S_1}\|u_{i,t+1} - \frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t+1})\|\right]$$

$$\le (1 - \frac{B_1\tau_2}{2n_1})^{t+1}\frac{1}{n_1}\sum_{i\in S_1}\|u_{i,0} - \frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,0})\| + \frac{2\tau_2^{1/2}\sigma}{B_2^{1/2}} + \frac{C_2 n_1^{1/2} B_2^{1/2}\tau_1}{B_1^{1/2} n_2^{1/2}\tau_2^{1/2}} + \frac{C_2 n_1^{3/2} n_2^{1/2}\eta}{B_1^{3/2} B_2^{1/2}\tau_2^{1/2}}$$

where C_2 *is a constant defined in the proof.*

Proof of Theorem A.2. Consider the change in the Moreau envelope:

$$\begin{aligned}
\mathbb{E}_t[F_{1/\bar{\rho}}(\mathbf{w}_{t+1})] &= \mathbb{E}_t\left[\min_{\tilde{\mathbf{w}}} F(\tilde{\mathbf{w}}) + \frac{\bar{\rho}}{2}\|\tilde{\mathbf{w}} - \mathbf{w}_{t+1}\|^2\right] \\
&\le \mathbb{E}_t\left[F(\hat{\mathbf{w}}_t) + \frac{\bar{\rho}}{2}\|\hat{\mathbf{w}}_t - \mathbf{w}_{t+1}\|^2\right] \\
&= F(\hat{\mathbf{w}}_t) + \mathbb{E}_t\left[\frac{\bar{\rho}}{2}\|\hat{\mathbf{w}}_t - (\mathbf{w}_t - \eta G_t)\|^2\right] \\
&\le F(\hat{\mathbf{w}}_t) + \frac{\bar{\rho}}{2}\left(\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2\right) + \bar{\rho}\mathbb{E}_t[\eta\langle\hat{\mathbf{w}}_t - \mathbf{w}_t, G_t\rangle] + \frac{\eta^2\bar{\rho}M^2}{2} \\
&= F_{1/\bar{\rho}}(\mathbf{w}_t) + \bar{\rho}\mathbb{E}_t[\eta\langle\hat{\mathbf{w}}_t - \mathbf{w}_t, G_t\rangle] + \frac{\eta^2\bar{\rho}M^2}{2}
\end{aligned} \tag{18}$$

Note that

$$\mathbb{E}_t[G_t] = \frac{1}{n_1}\sum_{i=1}^{n_1}\left[\frac{1}{n_2}\sum_{j=1}^{n_2}\nabla h_{i,j}(\mathbf{w}_t)\partial g_i(v_{i,j,t})\right]\partial f_i\left(u_{i,t}\right),$$

and the second inequality uses the bound of $\mathbb{E}[\|G_t\|^2]$, which follows from the Lipschitz continuity and bounded variance assumptions and is denoted by M.

Define $\hat{\mathbf{w}}_t := \mathrm{prox}_{F/\bar{\rho}}(\mathbf{w}_t)$. For a given $i \in \{1, \dots, m\}$, we have

$$
\begin{aligned}
&\frac{1}{n_1}\sum_{i\in S_1} f_i(\frac{1}{n_2}\sum_{j\in S_2} g_i(h_{i,j}(\hat{\mathbf{w}}_t))) - \frac{1}{n_1}\sum_{i\in S_1} f_i(u_{i,t}) \\
&\overset{(a)}{\geq} \frac{1}{n_1}\sum_{i\in S_1} \partial f_i(u_{i,t})^\top(\frac{1}{n_2}\sum_{j\in S_2} g_i(h_{i,j}(\hat{\mathbf{w}}_t)) - u_{i,t}) - \frac{1}{n_1}\sum_{i\in S_1}\frac{\rho_f}{2}\|\frac{1}{n_2}\sum_{j\in S_2} g_i(h_{i,j}(\hat{\mathbf{w}}_t)) - u_{i,t}\|^2 \\
&\geq \frac{1}{n_1}\sum_{i\in S_1} \partial f_i(u_{i,t})^\top(\frac{1}{n_2}\sum_{j\in S_2} g_i(h_{i,j}(\hat{\mathbf{w}}_t)) - u_{i,t}) \\
&\quad - \frac{1}{n_1}\sum_{i\in S_1}\rho_f\|\frac{1}{n_2}\sum_{j\in S_2} g_i(h_{i,j}(\hat{\mathbf{w}}_t)) - \frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t})\|^2 - \frac{1}{n_1}\sum_{i\in S_1}\rho_f\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\|^2 \\
&\geq \frac{1}{n_1}\sum_{i\in S_1} \partial f_i(u_{i,t})^\top(\frac{1}{n_2}\sum_{j\in S_2} g_i(h_{i,j}(\hat{\mathbf{w}}_t)) - u_{i,t}) - \frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}\rho_f C_g^2\|h_{i,j}(\hat{\mathbf{w}}_t) - v_{i,j,t}\|^2 \\
&\quad - \frac{1}{n_1}\sum_{i\in S_1}\rho_f\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\|^2 \\
&\overset{(b)}{\geq} \frac{1}{n_1}\sum_{i\in S_1} \partial f_i(u_{i,t})^\top\Bigg[\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t} + \frac{1}{n_2}\sum_{j\in S_2}\partial g_i(v_{i,j,t})^\top(h_{i,j}(\hat{\mathbf{w}}_t) - v_{i,j,t}) \\
&\quad - \frac{1}{n_2}\sum_{j\in S_2}\frac{\rho_g}{2}\|h_{i,j}(\hat{\mathbf{w}}_t) - v_{i,j,t}\|^2\Bigg] - \frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2} 2\rho_f C_g^2\|h_{i,j}(\mathbf{w}_t) - v_{i,j,t}\|^2 \\
&\quad - 2\rho_f C_g^2\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 - \frac{1}{n_1}\sum_{i\in S_1}\rho_f\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\|^2 \\
&\overset{(c)}{\geq} \frac{1}{n_1}\sum_{i\in S_1} \partial f_i(u_{i,t})^\top\Bigg[\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\Bigg] \\
&\quad + \frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}\underbrace{\langle\partial f_i(u_{i,t})^\top\partial g_i(v_{i,j,t})^\top(h_{i,j}(\hat{\mathbf{w}}_t) - v_{i,j,t})}_{A_1} \\
&\quad - \frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}\frac{\rho_g C_f}{2}\|h_{i,j}(\hat{\mathbf{w}}_t) - v_{i,j,t}\|^2 - \frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2} 2\rho_f C_g^2\|h_{i,j}(\mathbf{w}_t) - v_{i,j,t}\|^2 \\
&\quad - 2\rho_f C_g^2\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 - \frac{1}{n_1}\sum_{i\in S_1}\rho_f\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\|^2 \\
&\geq \frac{1}{n_1}\sum_{i\in S_1} \partial f_i(u_{i,t})^\top\Bigg[\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\Bigg] \\
&\quad + \frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}\underbrace{\langle\partial f_i(u_{i,t})^\top\partial g_i(v_{i,j,t})^\top(h_{i,j}(\hat{\mathbf{w}}_t) - v_{i,j,t})}_{A_1} \\
&\quad - \frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}(2\rho_f C_g^2 + \rho_g C_f)\|h_{i,j}(\mathbf{w}_t) - v_{i,j,t}\|^2 \\
&\quad - (2\rho_f C_g^2 + \rho_g C_f C_h^2)\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 - \frac{1}{n_1}\sum_{i\in S_1}\rho_f\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\|^2
\end{aligned} \tag{19}
$$

where (a) follows from the convexity of f_i, (b) uses the assumption that $f_i(\cdot)$ is non-decreasing and g_i is weak convex, (c) is due to $0 \leq \partial f_i(u_{i,t}) \leq C_f$.

The L_h-smoothness assumption of $h_{i,j}(\mathbf{w})$ (or weakly-convexity of $h_{i,j}(\mathbf{w})$, then only the second inequality holds) for all $i, \mathbf{w}$ implies

$$\begin{aligned} h_{i,j}(\hat{\mathbf{w}}_t) &\leq h_{i,j}(\mathbf{w}_t) + \nabla h_{i,j}(\mathbf{w}_t)^\top(\hat{\mathbf{w}}_t - \mathbf{w}_t) + \frac{L_h}{2}\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2, \\ h_{i,j}(\hat{\mathbf{w}}_t) &\geq h_{i,j}(\mathbf{w}_t) + \nabla h_{i,j}(\mathbf{w}_t)^\top(\hat{\mathbf{w}}_t - \mathbf{w}_t) - \frac{L_h}{2}\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2. \end{aligned} \tag{20}$$

We first assume that $g_i(\cdot)$ is non-increasing. Since $\partial f_i(u_{i,t}) \geq 0$ and $\partial g_i(v_{i,j,t}) \leq 0$, we bound A_1 as following

$$\begin{aligned} A_1 &= \partial f_i(u_{i,t})\partial^\top g_i(v_{i,j,t})^\top(h_{i,j}(\hat{\mathbf{w}}_t) - v_{i,j,t}) \\ &\overset{(a)}{\geq} \langle \partial f_i(u_{i,t})^\top \partial g_i(v_{i,j,t})^\top(h_{i,j}(\mathbf{w}_t) - v_{i,j,t}) + \partial f_i(u_{i,t})^\top \partial g_i(v_{i,j,t})^\top \nabla h_{i,j}(\mathbf{w}_t)^\top(\hat{\mathbf{w}}_t - \mathbf{w}_t) \\ &\quad + \partial f_i(u_{i,t})^\top \partial g_i(v_{i,j,t})^\top \frac{L_h}{2}\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2\rangle \\ &\overset{(b)}{\geq} -C_f C_g\|h_{i,j}(\mathbf{w}_t) - v_{i,j,t}\| + \partial f_i(u_{i,t})^\top \partial g_i(v_{i,j,t})^\top \nabla h_{i,j}(\mathbf{w}_t)^\top(\hat{\mathbf{w}}_t - \mathbf{w}_t) \\ &\quad - \frac{C_f C_g L_h}{2}\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 \end{aligned} \tag{21}$$

where inequality (a) follows from the first inequality in (20), (b) follows from the Lipschitz continuity and monotone assumptions on $f_i, g_i, h_{i,j}$. On the other hand, if we assume $g_i(\cdot)$ is non-decreasing, we may use the second inequality in (20) and obtain the same result as (21). Now plugging the new formulation of A_1 back to inequality 19 yields

$$\begin{aligned} &\frac{1}{n_1}\sum_{i\in S_1} f_i(\frac{1}{n_2}\sum_{j\in S_2} g_i(h_{i,j}(\hat{\mathbf{w}}_t))) - \frac{1}{n_1}\sum_{i\in S_1} f_i(u_{i,t}) \\ &\geq \frac{1}{n_1}\sum_{i\in S_1}\partial f_i(u_{i,t})^\top\left[\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\right] + \frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2} -C_f C_g\|h_{i,j}(\mathbf{w}_t) - v_{i,j,t}\| \\ &\quad + \frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}\partial f_i(u_{i,t})^\top \partial g_i(v_{i,j,t})^\top \nabla h_{i,j}(\mathbf{w}_t)^\top(\hat{\mathbf{w}}_t - \mathbf{w}_t) - \frac{C_f C_g L_h}{2}\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 \\ &\quad - \frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}(2\rho_f C_g^2 + \rho_g C_f)\|h_{i,j}(\mathbf{w}_t) - v_{i,j,t}\|^2 \\ &\quad - (2\rho_f C_g^2 + \rho_g C_f C_h^2)\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 - \frac{1}{n_1}\sum_{i\in S_1}\rho_f\left\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\right\|^2 \\ &\geq \frac{1}{n_1}\sum_{i\in S_1} -C_f\left\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\right\| + \frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2} -C_f C_g\|h_{i,j}(\mathbf{w}_t) - v_{i,j,t}\| \\ &\quad + \langle \mathbb{E}_t[G_t], \hat{\mathbf{w}}_t - \mathbf{w}_t\rangle - \frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}(2\rho_f C_g^2 + \rho_g C_f)\|h_{i,j}(\mathbf{w}_t) - v_{i,j,t}\|^2 \\ &\quad - (2\rho_f C_g^2 + \rho_g C_f C_h^2 + \frac{C_f C_g L_h}{2})\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 - \frac{1}{n_1}\sum_{i\in S_1}\rho_f\left\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\right\|^2 \end{aligned}$$

It follows

$$\begin{aligned} &\langle \mathbb{E}_t[G_t], \hat{\mathbf{w}}_t - \mathbf{w}_t\rangle \\ &\leq \frac{1}{n_1}\sum_{i\in S_1} f_i(\frac{1}{n_2}\sum_{j\in S_2} g_i(h_{i,j}(\hat{\mathbf{w}}_t))) - \frac{1}{n_1}\sum_{i\in S_1} f_i(u_{i,t}) + \frac{1}{n_1}\sum_{i\in S_1} C_f\left\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\right\| \\ &\quad + \frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2} C_f C_g\|h_{i,j}(\mathbf{w}_t) - v_{i,j,t}\| + \frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}(2\rho_f C_g^2 + \rho_g C_f)\|h_{i,j}(\mathbf{w}_t) - v_{i,j,t}\|^2 \\ &\quad + (2\rho_f C_g^2 + \rho_g C_f C_h^2 + \frac{C_f C_g L_h}{2})\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 + \frac{1}{n_1}\sum_{i\in S_1}\rho_f\left\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\right\|^2 \end{aligned} \tag{22}$$

Combining inequality 22 and 18 yields

$$\begin{aligned}
&\mathbb{E}_t[F_{1/\bar\rho}(\mathbf{w}_{t+1})]\\
&\le F_{1/\bar\rho}(\mathbf{w}_t) + \frac{\eta^2\bar\rho M^2}{2} + \bar\rho\eta\Bigg\{\frac{1}{n_1}\sum_{i\in S_1}\Bigg[f_i(\frac{1}{n_2}\sum_{j\in S_2} g_i(h_{i,j}(\hat{\mathbf{w}}_t))) - f_i(u_{i,t})\\
&\quad + C_f\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\Bigg\| + \frac{1}{n_2}\sum_{j\in S_2} C_fC_g\|h_{i,j}(\mathbf{w}_t) - v_{i,j,t}\|\\
&\quad + \frac{1}{n_2}\sum_{j\in S_2}(2\rho_fC_g^2 + \rho_gC_f)\|h_{i,j}(\mathbf{w}_t) - v_{i,j,t}\|^2\\
&\quad + (2\rho_fC_g^2 + \rho_gC_fC_h^2 + \frac{C_fC_gL_h}{2})\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 + \rho_f\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\Bigg\|^2\Bigg]\Bigg\}\\
&\le F_{1/\bar\rho}(\mathbf{w}_t) + \frac{\eta^2\bar\rho M^2}{2} + \bar\rho\eta\Bigg\{\frac{1}{n_1}\sum_{i\in S_1}\Bigg[F_i(\hat{\mathbf{w}}_t) - F_i(\mathbf{w}_t) + F_i(\mathbf{w}_t) - f_i(\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}))\\
&\quad + f_i(\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t})) - f_i(u_{i,t}) + C_f\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\Bigg\|\\
&\quad + \frac{1}{n_2}\sum_{j\in S_2} C_fC_g\|h_{i,j}(\mathbf{w}_t) - v_{i,j,t}\| + \frac{1}{n_2}\sum_{j\in S_2}(2\rho_fC_g^2 + \rho_gC_f)\|h_{i,j}(\mathbf{w}_t) - v_{i,j,t}\|^2\\
&\quad + (2\rho_fC_g^2 + \rho_gC_fC_h^2 + \frac{C_fC_gL_h}{2})\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 + \rho_f\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\Bigg\|^2\Bigg]\Bigg\}\\
&\overset{(a)}{\le} F_{1/\bar\rho}(\mathbf{w}_t) + \frac{\eta^2\bar\rho M^2}{2} + \bar\rho\eta\Bigg\{\frac{1}{n_1}\sum_{i\in S_1}\Bigg[F_i(\hat{\mathbf{w}}_t) - F_i(\mathbf{w}_t) + 2C_f\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\Bigg\|\\
&\quad + \frac{1}{n_2}\sum_{j\in S_2} 2C_fC_g\|h_{i,j}(\mathbf{w}_t) - v_{i,j,t}\| + \frac{1}{n_2}\sum_{j\in S_2}(2\rho_fC_g^2 + \rho_gC_f)\|h_{i,j}(\mathbf{w}_t) - v_{i,j,t}\|^2\\
&\quad + (2\rho_fC_g^2 + \rho_gC_fC_h^2 + \frac{C_fC_gL_h}{2})\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 + \rho_f\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\Bigg\|^2\Bigg]\Bigg\}
\end{aligned}$$

where (a) follows from the Lipschitz continuity of $f_i, g_i, h_{i,j}$.

Due to the ρ_F-weak convexity of $F_i(\mathbf{w})$, we have $(\bar\rho - \rho_F)$-strong convexity of $\mathbf{w} \mapsto F_i(\mathbf{w}) + \frac{\bar\rho}{2}\|\mathbf{w}_t - \mathbf{w}\|^2$. Then it follows

$$\begin{aligned}
F_i(\hat{\mathbf{w}}_t) - F_i(\mathbf{w}_t) &= \Bigg[F_i(\hat{\mathbf{w}}_t) + \frac{\bar\rho}{2}\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2\Bigg] - \Bigg[F_i(\mathbf{w}_t) + \frac{\bar\rho}{2}\|\mathbf{w}_t - \mathbf{w}_t\|^2\Bigg] - \frac{\bar\rho}{2}\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2\\
&\le (\frac{\rho_F}{2} - \bar\rho)\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2
\end{aligned} \tag{23}$$

Plugging inequality 23 back into A.2, we obtain

$$\begin{aligned}
&\mathbb{E}_t[F_{1/\bar{\rho}}(\mathbf{w}_{t+1})] \\
&\leq F_{1/\bar{\rho}}(\mathbf{w}_t) + \frac{\eta^2\bar{\rho}M^2}{2} + \bar{\rho}\eta\Bigg\{\frac{1}{n_1}\sum_{i\in S_1}\Bigg[(\frac{\rho_F}{2}-\bar{\rho})\|\mathbf{w}_t-\hat{\mathbf{w}}_t\|^2 + 2C_f\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t}) - u_{i,t}\Bigg\| \\
&\quad + \frac{1}{n_2}\sum_{j\in S_2} 2C_fC_g\|h_{i,j}(\mathbf{w}_t)-v_{i,j,t}\| + \frac{1}{n_2}\sum_{j\in S_2}(2\rho_fC_g^2+\rho_gC_f)\|h_{i,j}(\mathbf{w}_t)-v_{i,j,t}\|^2 \\
&\quad + (2\rho_fC_g^2+\rho_gC_fC_h^2+\frac{C_fC_gL_h}{2})\|\hat{\mathbf{w}}_t-\mathbf{w}_t\|^2 + \rho_f\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t})-u_{i,t}\Bigg\|^2\Bigg]\Bigg\} \\
&\overset{(a)}{\leq} F_{1/\bar{\rho}}(\mathbf{w}_t) + \frac{\eta^2\bar{\rho}M^2}{2} + \bar{\rho}\eta\Bigg\{\frac{1}{n_1}\sum_{i\in S_1}\Bigg[-\frac{\bar{\rho}}{2}\|\mathbf{w}_t-\hat{\mathbf{w}}_t\|^2 + C_1\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t})-u_{i,t}\Bigg\| \\
&\quad + \frac{1}{n_2}\sum_{j\in S_2} C_1\|h_{i,j}(\mathbf{w}_t)-v_{i,j,t}\| + \frac{1}{n_2}\sum_{j\in S_2} C_1\|h_{i,j}(\mathbf{w}_t)-v_{i,j,t}\|^2 \\
&\quad + C_1\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t})-u_{i,t}\Bigg\|^2\Bigg]\Bigg\} \\
&\overset{(b)}{=} F_{1/\bar{\rho}}(\mathbf{w}_t) + \frac{\eta^2\bar{\rho}M^2}{2} - \frac{\eta}{2}\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t)\|^2 + C_1\bar{\rho}\eta\frac{1}{n_1}\sum_{i\in S_1}\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t})-u_{i,t}\Bigg\| \\
&\quad + C_1\bar{\rho}\eta\frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}\|h_{i,j}(\mathbf{w}_t)-v_{i,j,t}\| + C_1\bar{\rho}\eta\frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}\|h_{i,j}(\mathbf{w}_t)-v_{i,j,t}\|^2 \\
&\quad + C_1\bar{\rho}\eta\frac{1}{n_1}\sum_{i\in S_1}\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t})-u_{i,t}\Bigg\|^2
\end{aligned}$$

where in inequality (a) we use $\bar{\rho} = \rho_F + 4\rho_fC_g^2 + 2\rho_gC_fC_h^2 + C_fC_gL_h$ and $C_1 = \max\{2C_fC_g, 2C_f, (2\rho_fC_g^2+\rho_gC_f), \rho_f\}$, and equality (b) uses Lemma 3.2.

With general error bounds

$$\begin{aligned}
&\frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}\mathbb{E}[\|h_{i,j}(\mathbf{w}_t)-v_{i,j,t}\|] \leq (1-\mu_1)^t\frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}\|h_{i,j}(\mathbf{w}_0)-v_{i,j,0}\| + R_1, \\
&\frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}\mathbb{E}[\|h_{i,j}(\mathbf{w}_t)-v_{i,j,t}\|^2] \leq (1-\mu_1)^t\frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}\|h_{i,j}(\mathbf{w}_0)-v_{i,j,0}\|^2 + R_2, \\
&\frac{1}{n_1}\sum_{i\in S_1}\mathbb{E}\Bigg[\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t})-u_{i,t}\Bigg\|\Bigg] \leq (1-\mu_2)^t\frac{1}{n_+}\sum_{i\in S_+}\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,0})-u_{i,0}\Bigg\| + R_3, \\
&\frac{1}{n_1}\sum_{i\in S_1}\mathbb{E}\Bigg[\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,t})-u_{i,t}\Bigg\|^2\Bigg] \leq (1-\mu_2)^t\frac{1}{n_+}\sum_{i\in S_+}\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,0})-u_{i,0}\Bigg\|^2 + R_4,
\end{aligned}$$

we have

$$\begin{aligned}
&\mathbb{E}[F_{1/\bar{\rho}}(\mathbf{w}_{t+1})] \\
&\leq \mathbb{E}[F_{1/\bar{\rho}}(\mathbf{w}_t)] + \frac{\eta^2\bar{\rho}M^2}{2} - \frac{\eta}{2}\mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t)\|^2] + C_1\bar{\rho}\eta(1-\mu_{min})^t\Bigg[\frac{1}{n_1}\sum_{i\in S_1}\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,0})-u_{i,0}\Bigg\| \\
&\quad + \frac{1}{n_1}\sum_{i\in S_1}\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,0})-u_{i,0}\Bigg\|^2 + \frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}\|h_{i,j}(\mathbf{w}_0)-v_{i,j,0}\| \\
&\quad + \frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}\|h_{i,j}(\mathbf{w}_0)-v_{i,j,0}\|^2\Bigg] + C_1\bar{\rho}\eta(R_1+R_2+R_3+R_4),
\end{aligned}$$

where $\mu_{min} = \min\{\mu_1, \mu_2\}$.

Taking summation from $t = 0$ to $T - 1$ yields

$$\begin{aligned}
&\mathbb{E}[F_{1/\bar{\rho}}(\mathbf{w}_T)] \\
&\leq \mathbb{E}[F_{1/\bar{\rho}}(\mathbf{w}_0)] + \frac{\eta^2 \bar{\rho} M^2 T}{2} - \frac{\eta}{2}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t)\|^2] + C_1 \bar{\rho}\eta \sum_{t=0}^{T-1}(1-\mu_{min})^t \Delta_0 \\
&\quad + TC_1\bar{\rho}\eta(R_1 + R_2 + R_3 + R_4) \\
&\leq \mathbb{E}[F_{1/\bar{\rho}}(\mathbf{w}_0)] + \frac{\eta^2 \bar{\rho} M^2 T}{2} - \frac{\eta}{2}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t)\|^2] + \frac{C_1\bar{\rho}\eta\Delta_0}{\mu_{min}} + TC_1\bar{\rho}\eta(R_1 + R_2 + R_3 + R_4)
\end{aligned}$$

where we use $\sum_{t=0}^{T-1}(1-\mu_{min})^t \leq \frac{1}{\mu_{min}}$ and define constant Δ_0 such that

$$\begin{aligned}
&\Bigg[\frac{1}{n_1}\sum_{i\in S_1}\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,0}) - u_{i,0}\Bigg\| + \frac{1}{n_1}\sum_{i\in S_1}\Bigg\|\frac{1}{n_2}\sum_{j\in S_2} g_i(v_{i,j,0}) - u_{i,0}\Bigg\|^2 \\
&\quad + \frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}\|h_{i,j}(\mathbf{w}_0) - v_{i,j,0}\| + \frac{1}{n_1}\sum_{i\in S_1}\frac{1}{n_2}\sum_{j\in S_2}\|h_{i,j}(\mathbf{w}_0) - v_{i,j,0}\|^2\Bigg] \leq \Delta_0.
\end{aligned}$$

Then it follows

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t)\|^2] \\
&\leq \frac{2}{\eta T}\Bigg[F_{1/\bar{\rho}}(\mathbf{w}_0) - \mathbb{E}[F_{1/\bar{\rho}}(\mathbf{w}_T)] + \frac{\eta^2\bar{\rho}M^2 T}{2} + \frac{C_1\bar{\rho}\eta\Delta_0}{\mu_{min}} + TC_1\bar{\rho}\eta(R_1 + R_2 + R_3 + R_4)\Bigg] \\
&\leq \frac{2\Delta}{\eta T} + \eta\bar{\rho}M^2 + \frac{2C_1\bar{\rho}\Delta_0}{\mu_{min}T} + 2C_1\bar{\rho}(R_1 + R_2 + R_3) \\
&= \mathcal{O}(\frac{1}{T}(\frac{1}{\eta} + \frac{1}{\mu_{min}}) + \eta + R_1 + R_2 + R_3 + R_4)
\end{aligned}$$

where we define constant Δ such that $F_{1/\bar{\rho}}(\mathbf{w}_0, \mathbf{s}_0, s'_0) - \mathbb{E}[F_{1/\bar{\rho}}(\mathbf{w}_T, \mathbf{s}_T, s'_T)] \leq \Delta$.

With MSVR updates for $v_{i,j,t}$ and $u_{i,t}$, following from Lemma A.3 and Lemma A.4, we have

$$\begin{aligned}
&\mu_1 = \frac{B_1 B_2 \tau_1}{2n_1 n_2}, \quad \mu_2 = \frac{B_1\tau_2}{2n_1}, \quad R_1 = \frac{2\tau_1^{1/2}\sigma}{B_3^{1/2}} + \frac{4n_1 n_2\sqrt{C_h}M\eta}{B_1 B_2 \tau_1^{1/2}} \\
&R_2 = \frac{4\tau_1\sigma^2}{B_3} + \frac{16 n_1^2 n_2^2 C_h M^2\eta^2}{B_1^2 B_2^2 \tau_1}, \quad R_3 = \frac{2\tau_2^{1/2}\sigma}{B_2^{1/2}} + \frac{C_2 n_1^{1/2} B_2^{1/2}\tau_1}{B_1^{1/2} n_2^{1/2}\tau_2^{1/2}} + \frac{C_2 n_1^{3/2} n_2^{1/2}\eta}{B_1^{3/2} B_2^{1/2}\tau_2^{1/2}}, \\
&R_4 = \frac{4\tau_2\sigma^2}{B_2} + \frac{C_2^2 n_1 B_2 \tau_1^2}{B_1 n_2 \tau_2} + \frac{C_2^2 n_1^3 n_2 \eta^2}{B_1^3 B_2 \tau_2}.
\end{aligned}$$

Then

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t)\|^2] \\
&\leq \mathcal{O}\Bigg(\frac{1}{T}(\frac{1}{\eta} + \frac{1}{\mu_{min}}) + \eta + \frac{\tau_1^{1/2}}{B_3^{1/2}} + \frac{\tau_1}{B_3} + \frac{\tau_2^{1/2}}{B_2^{1/2}} + \frac{\tau_2}{B_2} \\
&\quad + \frac{n_1 n_2 \eta}{B_1 B_2 \tau_1^{1/2}} + \frac{n_1^2 n_2^2 \eta^2}{B_1^2 B_2^2 \tau_1} + \frac{n_1^{1/2} B_2^{1/2}\tau_1}{B_1^{1/2} n_2^{1/2}\tau_2^{1/2}} + \frac{n_1^{3/2} n_2^{1/2}\eta}{B_1^{3/2} B_2^{1/2}\tau_2^{1/2}} + \frac{n_1 B_2 \tau_1^2}{B_1 n_2 \tau_2} + \frac{n_1^3 n_2 \eta^2}{B_1^3 B_2 \tau_2}\Bigg) \\
&\leq \mathcal{O}\Bigg(\frac{1}{T}(\frac{1}{\eta} + \frac{1}{\mu_{min}}) + \frac{\tau_1^{1/2}}{B_3^{1/2}} + \frac{\tau_2^{1/2}}{B_2^{1/2}} + \frac{n_1 n_2 \eta}{B_1 B_2 \tau_1^{1/2}} + \frac{n_1^{1/2} B_2^{1/2}\tau_1}{B_1^{1/2} n_2^{1/2}\tau_2^{1/2}} + \frac{n_1^{3/2} n_2^{1/2}\eta}{B_1^{3/2} B_2^{1/2}\tau_2^{1/2}}\Bigg).
\end{aligned}$$

Algorithm 3 Stochastic Optimization algorithm for Non-smooth FCCO with coordinate moving average

```
Initialization: $\mathbf{w}_0$, $\{u_{i,0} : i \in \mathcal{S}\}$.
for $t = 0, \dots, T-1$ do
    Draw sample batches $\mathcal{B}_1^t \sim \mathcal{S}$, and $\mathcal{B}_{2,i}^t \sim \mathcal{D}_i$ for each $i \in \mathcal{B}_1^t$.
    $u_{i,t+1} = \begin{cases} (1-\tau)u_{i,t} + \tau g_i(\mathbf{w}_t; \mathcal{B}_{2,i}^t), & i \in \mathcal{B}_1^t \\ u_{i,t}, & i \notin \mathcal{B}_1^t \end{cases}$
    Compute $G_t = \frac{1}{B_1} \sum_{i \in \mathcal{B}_1^t} \partial g_i(\mathbf{w}_t; \mathcal{B}_{2,i}^t) \partial f_i(u_{i,t})$
    Update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta G_t$
end for
return $\mathbf{w}_{\bar{t}}$ with uniformly sampled $\bar{t} \in \{0, T-1\}$.
```

Setting

$$\tau_1 = \mathcal{O}\left(\min\{B_3, \frac{B_1^{1/2} n_2^{1/2}}{n_1^{1/2}}\}\epsilon^4\right), \quad \tau_2 = \mathcal{O}(B_2 \epsilon^4),$$

$$\eta = \mathcal{O}\left(\min\left\{\frac{B_1 B_2}{n_1 n_2}\tau_1^{1/2}\epsilon^2, \frac{B_1^{3/2} B_2^{1/2}}{n_1^{3/2} n_2^{1/2}}\tau_2^{1/2}\right\}\right)$$

$$= \mathcal{O}\left(\min\left\{B_3^{1/2}, \frac{B_1^{1/4} n_2^{1/4}}{n_1^{1/4}}, \frac{B_1^{1/2} n_2^{1/2}}{n_1^{1/2}}\right\}\frac{B_1 B_2}{n_1 n_2}\epsilon^4\right),$$

then with

$$T = \mathcal{O}\left(\max\left\{\frac{1}{B_3^{1/2}}, \frac{n_1^{1/4}}{B_1^{1/4} n_2^{1/4}}, \frac{n_1^{1/2}}{B_1^{1/2} n_2^{1/2}}\right\}\frac{n_1 n_2}{B_1 B_2}\epsilon^{-6}\right),$$

we have

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t)\|^2] \leq \epsilon^2$$

□

B Solving Non-smooth FCCO and TCCO with Coordinate Moving Average

In this section we consider solving non-smooth weakly-convex FCCO and TCCO without variance reduction method. To be specific, we use coordinate moving average updates for function values estimations instead of MSVR. This allows us to weaken the assumption on the Lipschitz continuity, i.e. the Lipschitz continuity of the stochastic function value estimation is not required, and can be replaced by the Lipschitz continuity of the function value. Moreover, compared with MSVR, coordinate moving average update does not need the stochastic evaluation from the previous iteration, and thus has a simpler implementation. However, as a result of not using variance reduction technique, the algorithms suffer from worse convergence rates in terms of ϵ.

B.1 Solving Non-smooth FCCO with Coordinate Moving Average

We first assume the followings assumptions hold.

Assumption B.1. For all $i \in \mathcal{S}$, we assume that

- $f_i(\cdot)$ is ρ_f-weakly-convex, C_f-Lipschitz continuous and non-decreasing;
- $g_i(\cdot)$ is ρ_g-weakly-convex and C_g-Lipschitz continuous;
- Stochastic gradient estimators $g_i(\mathbf{w};\xi)$ and $\partial g_i(\mathbf{w};\xi)$ have bounded variance σ^2.

With coordinate moving average update, we present the following lemma of error bound.

Lemma B.2. *Consider the coordinate moving average update for* $\{u_{i,t} : i \in \mathcal{S}_1\}$ *in Algorithm 3, assume* $g_i(\mathbf{w})$ *is* C_g*-Lipschitz continuous for all* $i \in \mathcal{S}_1$ *and* $\tau \leq 1$*, then we have*

$$\mathbb{E}[\|u_{i,t+1} - g_i(\mathbf{w}_{t+1})\|] \leq (1 - \frac{B_1\tau}{4n_1})^{t+1}\|u_{i,0} - g_i(\mathbf{w}_0)\| + \frac{2\sqrt{2}\tau^{1/2}\sigma}{B_2^{1/2}} + \frac{4\sqrt{2}n_1 C_g M\eta}{B_1\tau},$$

$$\mathbb{E}[\|u_{i,t+1} - g_i(\mathbf{w}_{t+1})\|^2] \leq (1 - \frac{B_1\tau}{4n_1})^{2(t+1)}\|u_{i,0} - g_i(\mathbf{w}_0)\|^2 + \frac{8\tau\sigma^2}{B_2} + \frac{32n_1^2C_g^2M^2\eta^2}{B_1^2\tau^2}.$$

Then we have a convergence analysis similar to Theorem 4.6.

Theorem B.3. *Consider non-smooth weakly-convex FCCO problem, under Assumption B.1, setting* $\tau = \mathcal{O}(B_2\epsilon^4) \leq 1$, $\eta = \mathcal{O}(\frac{B_1B_2}{n_1}\epsilon^6)$*, Algorithm 3 converges to an* ϵ*-stationary point of the Moreau envelope* $F_{1/\bar{\rho}}$ *in* $T = \mathcal{O}(\frac{n_1}{B_1B_2}\epsilon^{-8})$ *iterations.*

Proof of Theorem B.3. Since the only difference between SONX and Algorithm 3 is the update for $\{u_{i,t} : i \in \mathcal{S}_1\}$, the proof of Theorem 4.6 still holds with the error bound replaced by Lemma B.2, i.e.,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i\in\mathcal{S}}\|u_{i,t+1} - g_i(\mathbf{w}_{t+1})\|\right] \leq (1-\mu)^{t+1}\frac{1}{n}\sum_{i\in\mathcal{S}}\|u_{i,0} - g_i(\mathbf{w}_0)\| + R_1,$$

$$\mathbb{E}\left[\frac{1}{n}\sum_{i\in\mathcal{S}}\|u_{i,t+1} - g_i(\mathbf{w}_{t+1})\|^2\right] \leq (1-\mu)^{t+1}\frac{1}{n}\sum_{i\in\mathcal{S}}\|u_{i,0} - g_i(\mathbf{w}_0)\|^2 + R_2,$$

$$\mu = \frac{B_1\tau}{4n_1}, \quad R_1 = \frac{2\sqrt{2}\tau^{1/2}\sigma}{B_2^{1/2}} + \frac{4\sqrt{2}n_1C_gM\eta}{B_1\tau}, \quad R_2 = \frac{8\tau\sigma^2}{B_2} + \frac{32n_1^2C_g^2M^2\eta^2}{B_1^2\tau^2}.$$

Then proof proceeds to

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t)\|^2] \leq \mathcal{O}\left(\frac{1}{T}(\frac{1}{\eta} + \frac{1}{\mu}) + \eta + R_1 + R_2\right)$$

$$= \mathcal{O}\left(\frac{1}{T}(\frac{1}{\eta} + \frac{n_1}{B_1\tau}) + \eta + \frac{\tau^{1/2}\sigma}{B_2^{1/2}} + \frac{n_1\eta}{B_1\tau} + \frac{\tau\sigma^2}{B_2} + \frac{n_1^2\eta^2}{B_1^2\tau^2}\right).$$

Setting

$$\tau = \mathcal{O}(B_2\epsilon^4), \quad \eta = \mathcal{O}(\frac{B_1B_2}{n_1}\epsilon^6),$$

then to reach a nearly ϵ-stationary point, Algorithm 3 needs

$$T = \mathcal{O}(\frac{n_1}{B_1B_2}\epsilon^{-8})$$

iterations. □

B.2 Solving Non-smooth TCCO with Coordinate Moving Average

We first assume the following assumptions hold.

Assumption B.4. For all $(i,j) \in \mathcal{S}_1 \times \mathcal{S}_2$, we assume that

- $f_i(\cdot)$ is ρ_f-weakly-convex, C_f-Lipschitz continuous and non-decreasing;
- $g_i(\cdot)$ is ρ_g-weakly-convex and C_g-Lipschitz continuous. $h_{i,j}(\cdot)$ is differentiable and C_h-Lipschitz continuous.
- Either g_i is monotone and $h_{i,j}(\cdot)$ is L_h-smooth, or g_i is non-decreasing and $h_{i,j}(\cdot)$ is L_h-weakly-convex.
- Stochastic estimators $h_{i,j}(\mathbf{w},\xi)$, $\partial h_{i,j}(\mathbf{w},\xi)$ and $g_i(v_{i,j})$ have bounded variance σ^2, and $\|h_{i,j}(\mathbf{w})\| \leq \tilde{C}_h$.

With coordinate moving average update, we present the following lemmas of error bounds.

Algorithm 4 Stochastic Optimization algorithm for Non-smooth TCCO with coordinate moving average

```
Initialization: $\mathbf{w}_0$, $\{u_{i,0} : i \in \mathcal{S}_1\}$, $v_{i,j,0} = h_{i,j}(\mathbf{w}_0; \mathcal{B}^0_{3,i,j})$ for all $(i,j) \in \mathcal{S}_1 \times \mathcal{S}_2$.
for $t = 0, \ldots, T-1$ do
    Sample batches $\mathcal{B}^t_1 \subset \mathcal{S}_1$, $\mathcal{B}^t_2 \subset \mathcal{S}_2$, and $\mathcal{B}^t_{3,i,j} \subset \mathcal{D}_{i,j}$ for $i \in \mathcal{B}^t_1$ and $j \in \mathcal{B}^t_2$.
    $v_{i,j,t+1} = \begin{cases} (1-\tau_1)v_{i,j,t} + \tau_1 h_{i,j}(\mathbf{w}_t; \mathcal{B}^t_{3,i,j}), & (i,j) \in \mathcal{B}^t_1 \times \mathcal{B}^t_2 \\ v_{i,j,t}, & (i,j) \notin \mathcal{B}^t_1 \times \mathcal{B}^t_2 \end{cases}$
    $u_{i,t+1} = \begin{cases} (1-\tau_2)u_{i,t} + \frac{1}{B_2}\sum_{j \in \mathcal{B}^t_2} \tau_2 g_i(v_{i,j,t}), & i \in \mathcal{B}^t_1 \\ u_{i,t}, & i \notin \mathcal{B}^t_1 \end{cases}$
    $G_t = \frac{1}{B_1}\sum_{i \in \mathcal{B}^t_1}\left[\left(\frac{1}{B_2}\sum_{i \in \mathcal{B}^t_2} \nabla h_{i,j}(\mathbf{w}_t; \mathcal{B}^t_{3,i,j}) \partial g_i(v_{i,j,t})\right) \partial f_i(u_{i,t})\right]$
    Update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta G_t$
end for
return $\mathbf{w}_{\bar{t}}$ with uniformly sampled $\bar{t} \in \{0, T-1\}$.
```

Lemma B.5. *Consider the coordinate moving average update for* $\{v_{i,j,t} : (i,j) \in \mathcal{S}_1 \times \mathcal{S}_2\}$ *in Algorithm 4, assume* $h_{i,j}(\mathbf{w})$ *is* C_h*-Lipschitz continuous for all* $(i,j) \in \mathcal{S}_1 \times \mathcal{S}_2$ *and* $\tau_1 \leq 1$*, then we have*

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n_1 n_2}\sum_{i \in \mathcal{S}_1}\sum_{j \in \mathcal{S}_2}\|v_{i,j,t+1} - h_{i,j}(\mathbf{w}_{t+1})\|\right] \\
&\leq (1 - \frac{B_1 B_2 \tau_1}{4 n_1 n_2})^{t+1}\frac{1}{n_1 n_2}\sum_{i \in \mathcal{S}_1}\sum_{j \in \mathcal{S}_2}\|v_{i,j,0} - h_{i,j}(\mathbf{w}_0)\| + \frac{2\sqrt{2}\tau_1^{1/2}\sigma}{B_3^{1/2}} + \frac{4\sqrt{2} n_1 n_2 C_h M \eta}{B_1 B_2 \tau_1}, \\
&\mathbb{E}\left[\frac{1}{n_1 n_2}\sum_{i \in \mathcal{S}_1}\sum_{j \in \mathcal{S}_2}\|v_{i,j,t+1} - h_{i,j}(\mathbf{w}_{t+1})\|^2\right] \\
&\leq (1 - \frac{B_1 B_2 \tau_1}{4 n_1 n_2})^{2(t+1)}\frac{1}{n_1 n_2}\sum_{i \in \mathcal{S}_1}\sum_{j \in \mathcal{S}_2}\|v_{i,j,0} - h_{i,j}(\mathbf{w}_0)\|^2 + \frac{8\tau_1\sigma^2}{B_3} + \frac{32 n_1^2 n_2^2 C_h^2 M^2 \eta^2}{B_1^2 B_2^2 \tau_1^2}.
\end{aligned}$$

Lemma B.6. *Consider the coordinate moving average update for* $\{u_{i,t} : i \in \mathcal{S}_1\}$ *in Algorithm 4, assume* $g_i(\cdot)$ *is* C_g*-Lipschitz continuous for all* $i \in \mathcal{S}_1$ *and* $\tau_2 \leq 1$*, then we have*

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n_1}\sum_{i \in \mathcal{S}_1}\|u_{i,t+1} - \frac{1}{n_2}\sum_{j \in \mathcal{S}_2} g_i(v_{i,j,t+1})\|\right] \\
&\leq (1 - \frac{B_1 \tau_2}{4 n_1})^{t+1}\frac{1}{n_1}\sum_{i \in \mathcal{S}_1}\|u_{i,0} - \frac{1}{n_2}\sum_{j \in \mathcal{S}_2} g_i(v_{i,j,0})\| + \frac{2\sqrt{2}\tau_2^{1/2}\sigma}{B_2^{1/2}} + \frac{4\sqrt{2} C_g M n_1^{1/2} B_2^{1/2} \tau_1}{B_1^{1/2} n_2^{1/2} \tau_2}, \\
&\mathbb{E}\left[\frac{1}{n_1}\sum_{i \in \mathcal{S}_1}\|u_{i,t+1} - \frac{1}{n_2}\sum_{j \in \mathcal{S}_2} g_i(v_{i,j,t+1})\|^2\right] \\
&\leq (1 - \frac{B_1 \tau_2}{4 n_1})^{2(t+1)}\frac{1}{n_1}\sum_{i \in \mathcal{S}_1}\|u_{i,0} - \frac{1}{n_2}\sum_{j \in \mathcal{S}_2} g_i(v_{i,j,0})\|^2 + \frac{8\tau_2\sigma^2}{B_2} + \frac{32 C_g^2 M^2 n_1 B_2 \tau_1^2}{B_1 n_2 \tau_2^2}.
\end{aligned}$$

Then we have a convergence analysis similar to Theorem A.2.

Theorem B.7. *Consider non-smooth weakly-convex TCCO problem, under Assumption B.4, setting* $\tau_1 = \mathcal{O}\left(\min\left\{B_3\epsilon^4, \frac{B_1^{1/2} n_2^{1/2}}{n_1^{1/2} B_2^{1/2}} B_2 \epsilon^6\right\}\right) \leq 1$*,* $\tau_2 = \mathcal{O}(B_2\epsilon^4) \leq 1$*,* $\eta = \mathcal{O}\left(\min\left\{B_3\epsilon^4, \frac{B_1^{1/2} n_2^{1/2}}{n_1^{1/2} B_2^{1/2}} B_2 \epsilon^6\right\}\frac{B_1 B_2}{n_1 n_2}\epsilon^2\right)$*, Algorithm 4 converges to an* ϵ*-stationary point of the Moreau envelope* $F_{1/\bar{\rho}}$ *in* $T = \mathcal{O}\left(\max\left\{\frac{1}{B_3}, \frac{n_1^{1/2}}{B_1^{1/2} B_2^{1/2} n_2^{1/2}}\epsilon^{-2}\right\}\frac{n_1 n_2}{B_1 B_2}\epsilon^{-8}\right)$ *iterations.*

Proof of Theorem B.7. Since the only difference between SONT and Algorithm 4 is the update for $\{u_{i,t} : i \in \mathcal{S}_1\}$ and $\{v_{i,j,t} : (i,j) \in \mathcal{S}_1 \times \mathcal{S}_2\}$, the proof of Theorem A.2 still holds with the error bound replaced by Lemma B.5 and Lemma B.6, i.e.,

$$\frac{1}{n_1 n_2} \sum_{i \in S_1} \sum_{j \in S_2} \mathbb{E}[\|h_{i,j}(\mathbf{w}_t) - v_{i,j,t}\|] \leq (1-\mu_1)^t \frac{1}{n_1 n_2} \sum_{i \in S_1} \sum_{j \in S_2} \|h_{i,j}(\mathbf{w}_0) - v_{i,j,0}\| + R_1,$$

$$\frac{1}{n_1 n_2} \sum_{i \in S_1} \sum_{j \in S_2} \mathbb{E}[\|h_{i,j}(\mathbf{w}_t) - v_{i,j,t}\|^2] \leq (1-\mu_1)^t \frac{1}{n_1 n_2} \sum_{i \in S_1} \sum_{j \in S_2} \|h_{i,j}(\mathbf{w}_0) - v_{i,j,0}\|^2 + R_2,$$

$$\frac{1}{n_1} \sum_{i \in S_1} \mathbb{E}\left[\left\| \frac{1}{n_2} \sum_{j \in S_2} g_i(v_{i,j,t}) - u_{i,t} \right\|\right] \leq (1-\mu_2)^t \frac{1}{n_+} \sum_{i \in S_+} \left\| \frac{1}{n_2} \sum_{j \in S_2} g_i(v_{i,j,0}) - u_{i,0} \right\| + R_3,$$

$$\frac{1}{n_1} \sum_{i \in S_1} \mathbb{E}\left[\left\| \frac{1}{n_2} \sum_{j \in S_2} g_i(v_{i,j,t}) - u_{i,t} \right\|^2\right] \leq (1-\mu_2)^t \frac{1}{n_+} \sum_{i \in S_+} \left\| \frac{1}{n_2} \sum_{j \in S_2} g_i(v_{i,j,0}) - u_{i,0} \right\|^2 + R_4,$$

with

$$\begin{aligned}
&\mu_1 = \frac{B_1 B_2 \tau_1}{4 n_1 n_2}, \quad \mu_2 = \frac{B_1 \tau_2}{4 n_1}, \quad R_1 = \frac{2\sqrt{2}\tau_1^{1/2}\sigma}{B_3^{1/2}} + \frac{4\sqrt{2} n_1 n_2 C_h M \eta}{B_1 B_2 \tau_1} \\
&R_2 = \frac{8\tau_1\sigma^2}{B_3} + \frac{32 n_1^2 n_2^2 C_h^2 M^2 \eta^2}{B_1^2 B_2^2 \tau_1^2}, \quad R_3 = \frac{2\sqrt{2}\tau_2^{1/2}\sigma}{B_2^{1/2}} + \frac{4\sqrt{2} C_g M n_1^{1/2} B_2^{1/2} \tau_1}{B_1^{1/2} n_2^{1/2} \tau_2}, \\
&R_4 = \frac{8\tau_2\sigma^2}{B_2} + \frac{32 C_g^2 M^2 n_1 B_2 \tau_1^2}{B_1 n_2 \tau_2^2}
\end{aligned}$$

Then the proof proceeds to

$$\begin{aligned}
&\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla F_{1/\bar\rho}(\mathbf{w}_t)\|^2] \\
&\leq \mathcal{O}\left(\frac{1}{T}(\frac{1}{\eta} + \frac{1}{\mu_{min}}) + \eta + R_1 + R_2 + R_3 + R_4\right) \\
&\leq \mathcal{O}\bigg(\frac{1}{T}(\frac{1}{\eta} + \frac{1}{\mu_{min}}) + \eta + \frac{\tau_1^{1/2}\sigma}{B_3^{1/2}} + \frac{\tau_1\sigma^2}{B_3} + \frac{\tau_2^{1/2}\sigma}{B_2^{1/2}} + \frac{\tau_2\sigma^2}{B_2} + \frac{n_1 n_2 \eta}{B_1 B_2 \tau_1} + \frac{n_1^2 n_2^2 \eta^2}{B_1^2 B_2^2 \tau_1^2} \\
&\quad + \frac{n_1^{1/2} B_2^{1/2} \tau_1}{B_1^{1/2} n_2^{1/2} \tau_2} + \frac{n_1 B_2 \tau_1^2}{B_1 n_2 \tau_2^2}\bigg) \\
&\leq \mathcal{O}\bigg(\frac{1}{T}(\frac{1}{\eta} + \frac{1}{\mu_{min}}) + \frac{\tau_1^{1/2}\sigma}{B_3^{1/2}} + \frac{\tau_2^{1/2}\sigma}{B_2^{1/2}} + \frac{n_1 n_2 \eta}{B_1 B_2 \tau_1} + \frac{n_1^{1/2} B_2^{1/2} \tau_1}{B_1^{1/2} n_2^{1/2} \tau_2}\bigg).
\end{aligned}$$

Setting

$$\begin{aligned}
&\tau_1 = \mathcal{O}\left(\min\left\{B_3\epsilon^4, \frac{B_1^{1/2} n_2^{1/2}}{n_1^{1/2} B_2^{1/2}} B_2 \epsilon^6\right\}\right), \quad \tau_2 = \mathcal{O}(B_2\epsilon^4), \\
&\eta = \mathcal{O}\left(\min\left\{B_3\epsilon^4, \frac{B_1^{1/2} n_2^{1/2}}{n_1^{1/2} B_2^{1/2}} B_2 \epsilon^6\right\} \frac{B_1 B_2}{n_1 n_2} \epsilon^2\right),
\end{aligned}$$

then to reach a nearly ϵ-stationary point, Algorithm 4 need

$$T = \mathcal{O}\left(\max\left\{\frac{1}{B_3}, \frac{n_1^{1/2}}{B_1^{1/2} B_2^{1/2} n_2^{1/2}} \epsilon^{-2}\right\} \frac{n_1 n_2}{B_1 B_2} \epsilon^{-10}\right)$$

iterations. □

C Details for TPAUC Maximization

C.1 Assumption Verification

We first present two lemmas about the weak convexity of the objective in the regular learning setting and in the multi-instance learning setting with mean pooling.

Lemma C.1. *Consider the formulation in problem (9) in the regular learning setting and assume that function $\ell(\cdot)$ is non-decreasing, C_ℓ Lipschitz continuous and ρ_ℓ-weakly-convex, and function $h_{\mathbf{w}}(X_i)$ is C_h Lipschitz continuous and ρ_h-weakly-convex. then the following statements are true:*

- *$f_i(g, s')$ is convex and C_f-Lipschitz continuous w.r.t. (g, s'), and non-decreasing w.r.t. g.*
- *$\psi_i(\mathbf{w}, s_i)$ is ρ_ψ-weakly-convex w.r.t. $(\mathbf{w}, s_i)$, and the stochastic estimator of the finite sum function value $\psi_i(\mathbf{w}, s_i)$ is C_ψ-Lipschitz continuous w.r.t. $(\mathbf{w}, s_i)$.*
- *$\frac{1}{n_+}\sum_{i\in\mathcal{S}_+} f_i(\psi_i(\mathbf{w}, s_i), s')$ is ρ_F-weakly-convex w.r.t. $(\mathbf{w}, \mathbf{s}, s')$.*

Lemma C.2. *Consider the formulation in problem (9) in the multi-instance learning setting with mean pooling, and assume that function $h_i(\mathbf{w}) = \frac{1}{|X_i|}\sum_{\mathbf{x}\in X_i} e(\mathbf{w}_e; \mathbf{x})^\top \mathbf{w}_c$ is $\tilde{L}_h$-smooth and is bounded by $\tilde{C}_h$, and $h_i(\mathbf{w}; \xi) = e(\mathbf{w}_e; \xi)^\top \mathbf{w}_c$ is C_h-Lipschitz continuous and has bounded variance σ^2, ℓ is non-decreasing and L_ℓ-weakly-convex, then the followings are true:*

- *$f_i(g, s')$ is convex and C_f-Lipschitz-continuous w.r.t. (g, s'), and non-decreasing w.r.t. g;*
- *$g_i(v, s_i) = s_i + \frac{(\ell(v)-s_i)_+}{\beta}$ is ρ_g-weakly convex and non-decreasing w.r.t. v, convex w.r.t. s_i, and C_g-Lipschitz continuous w.r.t. (v, s_i);*
- *$h_{i,j}(\mathbf{w}) = h_j(\mathbf{w}) - h_i(\mathbf{w})$ is L_h-weakly-convex, and $h_{i,j}(\mathbf{w}; \xi, \zeta)$ is C_h-Lipschitz continuous;*
- *$\frac{1}{n_+}\sum_{X_i\in\mathcal{S}_+} f_i(g_i(h_{i,j}(\mathbf{w}), s_i), s')$ is ρ_F-weakly-convex w.r.t. $(\mathbf{w}, \mathbf{s}, s')$.*

C.1.1 Proof of Lemma C.1

Proof of Lemma C.1. The convexity of $f_i(g, s')$ with respect to (g, s') follows from the convexity definition. With subgradients $\partial_{s'} f_i(g, s') \in [1 - \frac{1}{\alpha}, 1]$, $\partial_g f_i(g, s') \in [0, \frac{1}{\alpha}]$, we can see that $f_i(g, s')$ is $\frac{1}{\alpha}$-Lipschitz continuous w.r.t. (g, s'), and non-decreasing w.r.t. u.

We first show that $\ell(h_{\mathbf{w}}(X_j) - h_{\mathbf{w}}(X_i))$ is weakly-convex w.r.t. $\mathbf{w}$.

$$
\begin{aligned}
&\ell(h_{\tilde{\mathbf{w}}}(X_j) - h_{\tilde{\mathbf{w}}}(X_i)) \\
&\geq \ell(h_{\mathbf{w}}(X_j) - h_{\mathbf{w}}(X_i)) + \langle \partial\ell(h_{\mathbf{w}}(X_j) - h_{\mathbf{w}}(X_i)), (h_{\tilde{\mathbf{w}}}(X_j) - h_{\tilde{\mathbf{w}}}(X_i)) - (h_{\mathbf{w}}(X_j) - h_{\mathbf{w}}(X_i))\rangle \\
&\quad + \frac{\rho_\ell}{2}\|(h_{\tilde{\mathbf{w}}}(X_j) - h_{\tilde{\mathbf{w}}}(X_i)) - (h_{\mathbf{w}}(X_j) - h_{\mathbf{w}}(X_i))\|^2 \\
&\overset{(a)}{\geq} \ell(h_{\mathbf{w}}(X_j) - h_{\mathbf{w}}(X_i)) + \langle \partial\ell(h_{\mathbf{w}}(X_j) - h_{\mathbf{w}}(X_i)), \langle \nabla h_{\mathbf{w}}(X_j) - \nabla h_{\mathbf{w}}(X_i), \tilde{\mathbf{w}} - \mathbf{w}\rangle\rangle \\
&\quad + 2\rho_\ell C_h^2\|\tilde{\mathbf{w}} - \mathbf{w}\|^2
\end{aligned}
$$

where (a) uses the weak-convexity of $h_{\mathbf{w}}(X_i)$ and $h_{\mathbf{w}}(X_j)$,

$$
\begin{aligned}
h_{\tilde{\mathbf{w}}}(X_j) - h_{\mathbf{w}}(X_j) &\geq \langle \nabla h_{\mathbf{w}}(X_j), \tilde{\mathbf{w}} - \mathbf{w}\rangle - \frac{\rho_h}{2}\|\tilde{\mathbf{w}} - \mathbf{w}\|^2, \\
-h_{\tilde{\mathbf{w}}}(X_i) + h_{\mathbf{w}}(X_i) &\geq -\langle \nabla h_{\mathbf{w}}(X_i), \tilde{\mathbf{w}} - \mathbf{w}\rangle + \frac{\rho_h}{2}\|\tilde{\mathbf{w}} - \mathbf{w}\|^2.
\end{aligned}
$$

Thus $\ell(h_{\mathbf{w}}(X_j) - h_{\mathbf{w}}(X_i))$ is $4\rho_\ell C_h^2$-weakly-convex w.r.t. $\mathbf{w}$.

By convexity of $(\ell, s_i) \mapsto s_i + \frac{(\ell - s_i)_+}{\beta}$, we have

$$\begin{aligned}
&\psi_i(\tilde{\mathbf{w}}, \tilde{s}_i) \\
&\geq \psi_i(\mathbf{w}, s_i) + \langle \partial_\ell \psi_i(\mathbf{w}, s_i), \ell(h_{\tilde{\mathbf{w}}}(X_j) - h_{\tilde{\mathbf{w}}}(X_i)) - \ell(h_{\mathbf{w}}(X_j) - h_{\mathbf{w}}(X_i)) \rangle + \langle \partial_{s_i} \psi_i(\mathbf{w}, s_i), \tilde{s}_i - s_i \rangle \\
&\overset{(a)}{\geq} \psi_i(\mathbf{w}, s_i) + \partial_\ell \psi_i(\mathbf{w}, s_i) \Big[\partial \ell(h_{\mathbf{w}}(X_j) - h_{\mathbf{w}}(X_i)) \langle \nabla h_{\mathbf{w}}(X_j) - \nabla h_{\mathbf{w}}(X_i), \tilde{\mathbf{w}} - \mathbf{w} \rangle \rangle - 2\rho_\ell C_h^2 \|\tilde{\mathbf{w}} - \mathbf{w}\|^2 \Big] \\
&\quad + \langle \partial_{s_i} \psi_i(\mathbf{w}, s_i), \tilde{s}_i - s_i \rangle \\
&\overset{(b)}{\geq} \psi_i(\mathbf{w}, s_i) + \partial_\ell \psi_i(\mathbf{w}, s_i) \partial \ell(h_{\mathbf{w}}(X_j) - h_{\mathbf{w}}(X_i)) \langle \nabla h_{\mathbf{w}}(X_j) - \nabla h_{\mathbf{w}}(X_i), \tilde{\mathbf{w}} - \mathbf{w} \rangle \rangle \\
&\quad + \langle \partial_{s_i} \psi_i(\mathbf{w}, s_i), \tilde{s}_i - s_i \rangle - \frac{2\rho_\ell C_h^2}{\beta} \|\tilde{\mathbf{w}} - \mathbf{w}\|^2
\end{aligned}$$

where (a) follows from the monotonicity of ψ_i w.r.t. ℓ and weak-convexity of $\ell(h_{\mathbf{w}}(X_j) - h_{\mathbf{w}}(X_i))$, and (b) is due to the Lipschitz continuity of $(\ell, s_i) \mapsto s_i + \frac{(\ell - s_i)_+}{\beta}$ w.r.t. ℓ. Thus ψ_i is $\frac{4\rho_\ell C_h^2}{\beta}$-weakly-convex w.r.t. $(\mathbf{w}, s_i)$.

With a similar argument using the convexity and Lipschitz continuity of $f_i(g, s')$ w.r.t. (g, s') and the weak-convexity of $\psi_i(\mathbf{w}, s_i)$, we can show that $f_i(\psi_i(\mathbf{w}, s_i), s')$ is $\frac{4\rho_\ell C_h^2}{\beta}$-weakly-convex w.r.t. $(\mathbf{w}, s_i, s')$. Thus, $F(\mathbf{w}, s_i, s')$ is $\rho_F = \frac{4\rho_\ell C_h^2}{\beta}$-weakly-convex w.r.t. $(\mathbf{w}, \mathbf{s}, s')$.

Now we show the Lipschitz continuity of $\psi_i(\mathbf{w}, s_i; X_j)$, i.e. an unbiased stochastic estimator of $\psi_i(\mathbf{w}, s_i)$. We have

$$\begin{aligned}
&\|\psi_i(\mathbf{w}, s_i; X_j) - \psi_i(\tilde{\mathbf{w}}, \tilde{s}_i; X_j)\|^2 \\
&= \left\| (s_i + \frac{(\ell(h_{\mathbf{w}}(X_j) - h_{\mathbf{w}}(X_i)) - s_i)_+}{\beta}) - (\tilde{s}_i + \frac{(\ell(h_{\tilde{\mathbf{w}}}(X_j) - h_{\tilde{\mathbf{w}}}(X_i)) - \tilde{s}_i)_+}{\beta}) \right\|^2 \\
&\leq 2\|s_i - \tilde{s}_i\|^2 + 2 \left\| \frac{(\ell(h_{\mathbf{w}}(X_j) - h_{\mathbf{w}}(X_i)) - s_i)_+}{\beta} - \frac{(\ell(h_{\tilde{\mathbf{w}}}(X_j) - h_{\tilde{\mathbf{w}}}(X_i)) - \tilde{s}_i)_+}{\beta} \right\|^2 \\
&\leq 2\|s_i - \tilde{s}_i\|^2 + \frac{2}{\beta^2}(8C_\ell^2 C_h^2 \|\tilde{\mathbf{w}} - \mathbf{w}\|^2 + 2\|\tilde{s}_i - s_i\|^2) \\
&\leq (2 + \frac{4 + 16C_\ell^2 C_h^2}{\beta^2})(\|\tilde{\mathbf{w}} - \mathbf{w}\|^2 + \|\tilde{s}_i - s_i\|^2).
\end{aligned}$$

Thus $\psi_i(\mathbf{w}, s_i; X_j)$ is $(2 + \frac{4 + 16C_\ell^2 C_h^2}{\beta^2})^{1/2}$-Lipschitz continuous w.r.t. $(\mathbf{w}, s_i)$. □

C.1.2 Proof of Lemma C.2

Proof of Lemma C.2. First of all, the convexity of $f_i(u, s')$ w.r.t. (u, s') and the convexity of $g_i(v_{ij}, s_i)$ w.r.t. (ℓ, s_i) directly follows from the convexity definition. Moreover, one can see from the formulation that $\partial_{s'} f_i(g, s') \in [1 - \frac{1}{\alpha}, 1]$, $\partial_u f_i(g, s') \in [0, \frac{1}{\alpha}]$, $\partial_\ell g_i(v_{ij}, s_i) \in [1 - \frac{1}{\beta}, 1]$, $\partial_{s_i} g_i(v_{ij}, s_i) \in [0, \frac{1}{\beta}]$. Thus f_i is $C_f = \frac{1}{\alpha}$-Lipschitz continuous w.r.t. (u, s') and non-decreasing w.r.t. u, g_i is $\frac{1}{\beta}$-Lipschitz continuous w.r.t. (ℓ, s_i) and non-decreasing w.r.t. ℓ. Since $\ell(\cdot)$ is non-decreasing, $g_i(v_{ij}, s_i)$ is non-decreasing w.r.t. v_{ij}. As a result of Proposition 4.2, $g_i(v_{ij}, s_i)$ is $\rho_g = \frac{1}{\beta} L_\ell$-weakly-convex w.r.t. v_{ij}. Due to the composition structure and the Lipschitz continuity of g_i and ℓ, one can see that $g_i(v_{ij}, s_i)$ is $C_g = \frac{1}{\beta} C_\ell$-Lipschitz continuous w.r.t. (v_{ij}, s_i).

The $L_h = 2\tilde{L}_h$-weakly-convexity of $h_{i,j}(\mathbf{w})$ and $C_h = 2\tilde{C}_h$-Lipschitz continuity of $h_{i,j}(\mathbf{w}; \xi, \zeta)$ directly follows from the $\tilde{L}_h$-smoothness of $h_i(\mathbf{w})$ and $\tilde{C}_h$-Lipschitz continuity of $h_i(\mathbf{w}; \xi)$. Finally, we show the weakly-convexity of $f_i(g_i(h_{i,j}(\mathbf{w}), s_i), s')$:

$$
\begin{aligned}
& f_i(g_i(h_{i,j}(\tilde{\mathbf{w}}), \tilde{s}_i)\tilde{s}') \\
& \overset{(a)}{\geq} f_i(g_i(h_{i,j}(\mathbf{w}), s_i), s') + \langle \partial_{s'} f_i(g_i(h_{i,j}(\mathbf{w}), s_i), s'), \tilde{s}' - s' \rangle \\
& \quad + \langle \partial_u f_i(g_i(h_{i,j}(\mathbf{w}), s_i), s'), g_i(h_{i,j}(\tilde{\mathbf{w}}), \tilde{s}_i) - g_i(h_{i,j}(\mathbf{w}), s_i) \rangle \\
& \overset{(b)}{\geq} f_i(g_i(h_{i,j}(\mathbf{w}), s_i), s') + \langle \partial_{s'} f_i(g_i(h_{i,j}(\mathbf{w}), s_i), s'), \tilde{s}' - s' \rangle \\
& \quad + \langle \partial_u f_i(g_i(h_{i,j}(\mathbf{w}), s_i), s'), \langle \partial_\ell g_i(h_{i,j}(\mathbf{w}), s_i), \ell(h_{i,j}(\tilde{\mathbf{w}})) - \ell(h_{i,j}(\mathbf{w})) \rangle \rangle \\
& \quad + \langle \partial_u f_i(g_i(h_{i,j}(\mathbf{w}), s_i)s'), \langle \partial_{s_i} g_i(h_{i,j}(\mathbf{w}), s_i), \tilde{s}_i - s_i \rangle \rangle \\
& \overset{(c)}{\geq} f_i(g_i(h_{i,j}(\mathbf{w}), s_i), s') + \langle \partial_{s'} f_i(g_i(h_{i,j}(\mathbf{w}), s_i), s'), \tilde{s}' - s' \rangle \\
& \quad + \langle \partial_u f_i(g_i(h_{i,j}(\mathbf{w}), s_i), s') \partial_\ell g_i(h_{i,j}(\mathbf{w}), s_i) \partial \ell(h_{i,j}(\mathbf{w})), h_{i,j}(\tilde{\mathbf{w}}) - h_{i,j}(\mathbf{w}) \rangle \\
& \quad - \frac{C_f C_g L_\ell}{2} \|h_{i,j}(\tilde{\mathbf{w}}) - h_{i,j}(\mathbf{w})\|^2 + \langle \partial_u f_i(s', g_i(h_{i,j}(\mathbf{w}), s_i)) \partial_{s_i} g_i(h_{i,j}(\mathbf{w}), s_i), \tilde{s}_i - s_i \rangle \\
& \overset{(d)}{\geq} f_i(g_i(h_{i,j}(\mathbf{w}), s_i), s') + \langle \partial_{s'} f_i(g_i(h_{i,j}(\mathbf{w}), s_i), s'), \tilde{s}' - s' \rangle \\
& \quad + \langle \partial_u f_i(g_i(h_{i,j}(\mathbf{w}), s_i), s') \partial_\ell g_i(h_{i,j}(\mathbf{w}), s_i) \partial \ell(h_{i,j}(\mathbf{w})) \nabla h_{i,j}(\mathbf{w}), \tilde{\mathbf{w}} - \mathbf{w} \rangle \\
& \quad + \langle \partial_u f_i(g_i(h_{i,j}(\mathbf{w}), s_i), s') \partial_{s_i} g_i(h_{i,j}(\mathbf{w}), s_i), \tilde{s}_i - s_i \rangle - (\frac{C_f C_g C_h^2 L_\ell}{2} + \frac{C_f C_g L_h}{2}) \|\tilde{\mathbf{w}} - \mathbf{w}\|^2
\end{aligned}
$$

where (a) uses the convexity of f_i, (b) uses the monotonicity of f_i w.r.t. u and convexity of $g_i(\ell, s_i)$ w.r.t. (ℓ, s_i), (c) uses monotonicity of f_i w.r.t. u, monotonicity of g_i w.r.t. ℓ and L_ℓ-weak-convexity of ℓ, (d) uses the smoothness of $h_{i,j}$. Thus $f_i(g_i(h_{i,j}(\mathbf{w}), s_i), s')$ is $\rho_F = (C_f C_g C_h^2 L_\ell + C_f C_g L_h)$-weakly-convex w.r.t. $(\mathbf{w}, s_i, s')$. Therefore, $\frac{1}{n_+} \sum_{i \in \mathcal{S}_+} f_i(g_i(h_{i,j}(\mathbf{w}), s_i), s')$ is ρ_F-weakly-convex w.r.t. $(\mathbf{w}, \mathbf{s}, s')$. □

C.2 Algorithms for TPAUC and Multi-instance TPAUC Maximization

Algorithm 5 SONX for TPAUC

```
Initialization: $\mathbf{w}_0, \{u_{i,0} : i \in \mathcal{S}_+\}, \{s_{i,0} : i \in \mathcal{S}_+\}, s'_0$
for $t = 0, \ldots, T-1$ do
    Sample batches $\mathcal{B}_1^t \subset S_+$ and $\mathcal{B}_2^t \subset S_-$.
    $u_{i,t+1} = \begin{cases} (1-\tau)u_{i,t} + \tau\psi_i(\mathbf{w}_t, s_{i,t}; \mathcal{B}_2^t) + \gamma(\psi_i(\mathbf{w}_t, s_{i,t}; \mathcal{B}_2^t) - \psi_i(\mathbf{w}_{t-1}, s_{i,t-1}; \mathcal{B}_2^t)), \ i \in \mathcal{B}_1^t \\ u_{i,t}, \quad i \notin \mathcal{B}_1^t \end{cases}$
    $s_{i,t+1} = \begin{cases} s_{i,t} - \eta\frac{1}{B_1}\partial_s\psi_i(\mathbf{w}_t, s_{i,t}; \mathcal{B}_2^t)\partial_u f(u_{i,t}, s'_t), \quad i \in \mathcal{B}_1^t \\ s_{i,t}, \quad i \notin \mathcal{B}_1^t \end{cases}$
    $s'_{t+1} = s'_t - \eta\frac{1}{B_1}\sum_{i \in \mathcal{B}_1^t} \partial_{s'} f(u_{i,t}, s'_t)$
    Compute $G_t = \frac{1}{B_1}\sum_{i \in \mathcal{B}_1^t} \partial_w\psi_i(\mathbf{w}_t, s_{i,t}; \mathcal{B}_2^t)\partial_u f(u_{i,t}, s'_t)$
    Update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta G_t$
end for
return $\mathbf{w}_{\bar{t}}$ with $\bar{t}$ uniformly sampled from $\{0, \ldots, T-1\}$.
```

Algorithm 6 SONT for Multi-instance TPAUC

```
Initialization: $\mathbf{w}_0, \{u_{i,0} : i \in \mathcal{S}_+\}, \{s_{i,0} : i \in \mathcal{S}_+\}, s'_0, \{v_{i,j,0} : (i,j) \in \mathcal{S}_+ \times \mathcal{S}_-\}$
for $t = 0, \dots, T-1$ do
    Sample batches $\mathcal{B}_1^t \subset S_+$, $\mathcal{B}_2^t \subset S_-$, and $\mathcal{B}_{3,i}^t \subset X_i$ for $i \in \mathcal{B}_1^t \cup \mathcal{B}_2^t$.
    $v_{i,t+1} = \begin{cases} \Pi_{\tilde{C}_h}[(1-\tau_1)v_{i,t} + \tau_1 h_i(\mathbf{w}_t; \mathcal{B}_{3,i}^t) + \gamma_1(h_i(\mathbf{w}_t; \mathcal{B}_{3,i}^t) - h_i(\mathbf{w}_{t-1}; \mathcal{B}_{3,i}^t))], & i \in \mathcal{B}_1^t \\ \Pi_{\tilde{C}_h}[(1-\tau_1)v_{i,t} + \tau_1 h_i(\mathbf{w}_t; \mathcal{B}_{3,i}^t) + \gamma_2(h_i(\mathbf{w}_t; \mathcal{B}_{3,i}^t) - h_i(\mathbf{w}_{t-1}; \mathcal{B}_{3,i}^t))], & i \in \mathcal{B}_2^t \\ v_{i,t}, \quad i \notin \mathcal{B}_1^t \text{ and } i \notin \mathcal{B}_2^t \end{cases}$
    $u_{i,t+1} = \begin{cases} (1-\tau_2)u_{i,t} + \frac{1}{B_2}\sum_{j\in\mathcal{B}_2^t}[\tau_2 g(v_{j,t} - v_{i,t}, s_{i,t}) \\ \qquad +\gamma_3(g(v_{j,t} - v_{i,t}, s_{i,t}) - g(v_{j,t-1} - v_{i,t-1}, s_{i,t-1}))], & i \in \mathcal{B}_1^t \\ u_{i,t}, & i \notin \mathcal{B}_1^t \end{cases}$
    $s_{i,t+1} = \begin{cases} s_{i,t} - \eta_1 \frac{1}{B_1}\left[\frac{1}{B_2}\sum_{j\in\mathcal{B}_2^t} \partial_{s_i} g(v_{j,t} - v_{i,t}, s_{i,t})\right] \partial_u f(s'_t, u_{i,t}), & i \in \mathcal{B}_1^t \\ s_{i,t}, & i \notin \mathcal{B}_1^t \end{cases}$
    $s'_{t+1} = s'_t - \eta_2 \frac{1}{B_1}\sum_{i\in\mathcal{B}_1^t} \partial_{s'} f(u_{i,t}, s'_t)$
    $G_t = \frac{1}{B_1}\sum_{i\in\mathcal{B}_1^t} \partial_u f(u_{i,t}, s'_t)$
        $\left[\frac{1}{B_2}\sum_{j\in\mathcal{B}_2^t}\left(\nabla h_j(\mathbf{w}_t; \mathcal{B}_{3,j}^t) - \nabla h_i(\mathbf{w}_t; \mathcal{B}_{3,i}^t)\right)\partial_v g(v_{j,t} - v_{i,t}, s_{i,t})\right]$
    Update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta G_t$
end for
return $\mathbf{w}_{\bar{t}}$ with $\bar{t}$ uniformly sampled from $\{0, \dots, T-1\}$.
```

C.3 TPAUC in MIL with smoothed-max pooling and attention-based pooling

We can extend our results to smoothed-max pooling and attention-based pooling.

Smoothed-max Pooling. The smoothed-max pooling can be written as [45]:

$$h_{\mathbf{w}}(X) = \tau \log\left(\frac{1}{|X|}\sum_{\mathbf{x}\in X} \exp(\phi(\mathbf{w};\mathbf{x})/\tau)\right), \tag{24}$$

where $\tau > 0$ is a hyperparameter and $\phi(\mathbf{w};\mathbf{x}) = e(\mathbf{w}_e, \mathbf{x})^\top \mathbf{w}_c$ is the prediction score for instance $\mathbf{x}$.

We can see that $h_{\mathbf{w}}(X)$ itself is a compositional function. To map the problem into TCCO, we define $h_i(\mathbf{w}) = \frac{1}{|X_i|}\sum_{\mathbf{x}\in X_i} \exp(\phi(\mathbf{w};\mathbf{x})/\tau) + C$, where $C > 0$ is a constant. Then the objective function becomes

$$\begin{aligned} &\min_{\mathbf{w},\mathbf{s}',\mathbf{s}} \frac{1}{n_+}\sum_{X_i\in\mathcal{S}_+} f_i(\psi_i(\mathbf{w}, s_i), s'), \\ &\text{where } f_i(g, s') = s' + \frac{(g - s')_+}{\alpha}, \\ &\qquad \psi_i(\mathbf{w}, s_i) = \frac{1}{n_-}\sum_{X_j\in\mathcal{S}_-} s_i + \frac{(\ell(\tau\log h_j(\mathbf{w}) - \tau\log h_i(\mathbf{w})) - s_i)_+}{\beta}, \end{aligned} \tag{25}$$

In this case we define $g_i(\ell(\mathbf{v}), s_i) = s_i + \frac{(\ell(\tau\log v_1 - \tau\log v_2) - s_i)_+}{\beta}$ and $h_{i,j}(\mathbf{w}) = [h_i(\mathbf{w}), h_j(\mathbf{w})]$. We can still prove that $g_i(\ell(\mathbf{v}), s_i)$ is monotone w.r.t to each component of $\mathbf{v}$. It is not difficult to prove that $\ell(\tau\log v_1 - \tau\log v_2)$ is weakly convex w.r.t $\mathbf{v}$ because $\tau\log v_1 - \tau\log v_2$ is a smooth mapping of $\mathbf{v}$ due to $\mathbf{v} \geq C$ and ℓ is a convex function [8]. As a result, since $g_i(\ell, s_i)$ is non-decreasing and convex w.r.t to ℓ, it is easy to prove that $g_i(\ell(\mathbf{v}), s_i)$ is weakly convex w.r.t $\mathbf{v}$ and is monotone (either non-decreasing or non-increasing) w.r.t to each component of $\mathbf{v}$. Hence, assuming $h_i(\mathbf{w})$ is a smooth and Lipchitz continuous function, we can prove that $g_i(h_{i,j}(\mathbf{w}), s_i)$ is weakly convex w.r.t. to $\mathbf{w}$.

Attention-based Pooling. Attention-based pooling was recently introduced for deep MIL [14], which aggregates the feature representations using attention, i.e.,

$$E(\mathbf{w}; X) = \sum_{\mathbf{x}\in X} \frac{\exp(g(\mathbf{w};\mathbf{x}))}{\sum_{\mathbf{x}'\in X}\exp(g(\mathbf{w};\mathbf{x}'))} e(\mathbf{w}_e; \mathbf{x}) \tag{26}$$

where $g(\mathbf{w};\mathbf{x})$ is a parametric function, e.g., $g(\mathbf{w};\mathbf{x}) = \mathbf{w}_a^\top \tanh(Ve(\mathbf{w}_e;\mathbf{x})) + C$, where $V \in \mathbb{R}^{m\times d_o}$ and $\mathbf{w}_a \in \mathbb{R}^m$. Based on the aggregated feature representation, the bag level prediction can be computed by

$$\begin{aligned} h_{\mathbf{w}}(\mathbf{w}, X) &= (\mathbf{w}_c^\top E(\mathbf{w}; X)) \\ &= \left(\sum_{\mathbf{x}\in X} \frac{\exp(g(\mathbf{w};\mathbf{x}))\delta(\mathbf{w};\mathbf{x})}{\sum_{\mathbf{x}'\in X}\exp(g(\mathbf{w};\mathbf{x}'))}\right), \end{aligned} \tag{27}$$

where $\delta(\mathbf{w};\mathbf{x}) = \mathbf{w}_c^\top e(\mathbf{w}_e;\mathbf{x})$.

We can see that $h_{\mathbf{w}}(X)$ itself is a compositional function. To map the problem into TCCO, we define $h_i^1(\mathbf{w}) = \frac{1}{|X_i|}\sum_{\mathbf{x}\in X_i}\exp(g(\mathbf{w};\mathbf{x}))\delta(\mathbf{w};\mathbf{x})$, and $h_i^2(\mathbf{w}) = \frac{1}{|X_i|}\sum_{\mathbf{x}'\in X_i}\exp(g(\mathbf{w};\mathbf{x}'))$. Assume $|\mathbf{w}_a^\top \tanh(Ve(\mathbf{w}_e;\mathbf{x}))| \leq C_b$ then $h_i^2(\mathbf{w}) \geq \exp(C - C_b)$. Then the objective function becomes

$$\min_{\mathbf{w},\mathbf{s}',\mathbf{s}} \frac{1}{n_+}\sum_{X_i\in\mathcal{S}_+} f_i(\psi_i(\mathbf{w}, s_i), s'),$$

$$\text{where } f_i(g, s') = s' + \frac{(g-s')_+}{\alpha}, \quad \psi_i(\mathbf{w}, s_i) = \frac{1}{n_-}\sum_{X_j\in\mathcal{S}_-} s_i + \frac{(\ell(\frac{h_j^1(\mathbf{w})}{h_j^2(\mathbf{w})} - \frac{h_i^1(\mathbf{w})}{h_i^2(\mathbf{w})}) - s_i)_+}{\beta}, \tag{28}$$

In this case we define $g_i(\ell(\mathbf{v}), s_i) = s_i + \frac{(\ell(\frac{v_3}{v_4} - \frac{v_1}{v_2}) - s_i)_+}{\beta}$ and $h_{i,j}(\mathbf{w}) = [h_i^1(\mathbf{w}), h_i^2(\mathbf{w}), h_j^1(\mathbf{w}), h_j^2(\mathbf{w})]$. We can still prove that $g_i(\ell(\mathbf{v}), s_i)$ is monotone w.r.t to each component of $\mathbf{v}$. It is not difficult to prove that $\ell(\frac{v_3}{v_4} - \frac{v_1}{v_2})$ is weakly convex w.r.t $\mathbf{v}$ because $\frac{v_3}{v_4} - \frac{v_1}{v_2}$ is a smooth mapping of $\mathbf{v}$ when v_2, v_4 are lower bounded and ℓ is a convex function [8]. As a result, since $g_i(\ell, s_i)$ is non-decreasing and convex w.r.t to ℓ, it is easy to prove that $g_i(\ell(\mathbf{v}), s_i)$ is weakly convex w.r.t $\mathbf{v}$ and is monotone (either non-decreasing or non-increasing) w.r.t to each component of $\mathbf{v}$. Hence, assuming $h_i^1(\mathbf{w}), h_i^2(\mathbf{w})$ are smooth and Lipchitz continuous, we can prove that $g_i(h_{i,j}(\mathbf{w}), s_i)$ is weakly convex w.r.t. to $\mathbf{w}$.

C.4 Convergence Analysis of TPAUC Maximization

C.4.1 Convergence analysis for Algorithm 5

We first consider TPAUC maximization in the regular learning setting. Define $F(\mathbf{w}, \mathbf{s}, s') := \frac{1}{n_+}\sum_{X_i\in\mathcal{S}_+} f_i(\psi_i(\mathbf{w}, s_i), s')$. Due to the weak-convexity of $F(\mathbf{w}, \mathbf{s}, s')$ w.r.t. $(\mathbf{w}, \mathbf{s}, s')$, we consider the following Moreau envelope and proximal map defined as

$$\begin{aligned} F_\lambda(\mathbf{w}, \mathbf{s}, s') &= \min_{\tilde{\mathbf{w}},\tilde{\mathbf{s}},\tilde{s}'} F(\tilde{\mathbf{w}}, \tilde{\mathbf{s}}, \tilde{s}') + \frac{1}{2\lambda}\left(\|\tilde{\mathbf{w}} - \mathbf{w}\|^2 + \|\tilde{\mathbf{s}} - \mathbf{s}\|^2 + \|\tilde{s}' - s'\|^2\right), \\ \text{prox}_{\lambda F}(\mathbf{w}, \mathbf{s}, s') &= \arg\min_{\tilde{\mathbf{w}},\tilde{\mathbf{s}},\tilde{s}'} F(\tilde{\mathbf{w}}, \tilde{\mathbf{s}}, \tilde{s}') + \frac{1}{2\lambda}\left(\|\tilde{\mathbf{w}} - \mathbf{w}\|^2 + \|\tilde{\mathbf{s}} - \mathbf{s}\|^2 + \|\tilde{s}' - s'\|^2\right). \end{aligned}$$

Following the same proof of Lemma 4.5, we have the following error bound

Lemma C.3. *Consider the update for $\{u_{i,t} : X_i \in \mathcal{S}_+\}$ in Algorithm 5. Assume $\psi_i(\mathbf{w}, s_i)$ is C_ψ-Lipshitz continuous for all $X_i \in \mathcal{S}_+$. Assume $\mathbb{E}_t[\|G_t\|^2] \leq M^2$ and $\mathbb{E}_t[\|\frac{1}{B_1}\sum_{X_i\in\mathcal{B}_1^t}\partial_s\psi_i(\mathbf{w}_t, s_{i,t}; \mathcal{B}_2^t)\partial_u f(u_{i,t}, s_t')e_i\|^2] \leq M^2$, where e_i is the n_+-dimensional vector with 1 at the i-th entry and 0 everywhere else. With $\gamma = \frac{n_+ - B_1}{B_1(1-\tau)} + (1-\tau)$ and $\tau \leq \frac{1}{2}$, we have*

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{n_+}\sum_{X_i\in\mathcal{S}_+}\|u_{i,t+1} - \psi_i(\mathbf{w}_{t+1}, s_{i,t+1})\|\right] \\ &\leq (1 - \frac{B_1\tau}{2n_1})^{t+1}\frac{1}{n}\sum_{X_i\in\mathcal{S}_+}\|u_{i,0} - \psi_i(\mathbf{w}_0, s_{i,0})\| + \frac{2\tau^{1/2}\sigma}{B_2^{1/2}} + \frac{8n_+C_\psi M\eta}{B_1\tau^{1/2}}. \end{aligned}$$

Then we have following convergence guarantee.

Theorem C.4. *Under the assumptions given in Lemma C.1, with* $\gamma = \frac{n_+ - B_1}{B_1(1-\tau)} + (1-\tau)$*,* $\tau = \mathcal{O}(B_2\epsilon^4) \leq \frac{1}{2}$*,* $\eta = \mathcal{O}(\frac{B_1 B_2^{1/2}\epsilon^4}{n_+})$*, and* $\bar{\rho} = \rho_F + \rho_\psi C_f$*, Algorithm 5 converges to an* ϵ*-stationary point of the Moreau envelope* $F_{1/\bar{\rho}}$ *in* $T = \mathcal{O}(\frac{n_+}{B_1 B_2^{1/2}}\epsilon^{-6})$ *iterations.*

Proof of Theorem C.4. Define $(\hat{\mathbf{w}}_t, \hat{\mathbf{s}}_t, \hat{s}'_t) := \text{prox}_{F/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t)$. For a given $X_i \in \mathcal{S}_+$, we have

$$
\begin{aligned}
& f_i(\psi_i(\hat{\mathbf{w}}_t, \hat{s}_{i,t}), \hat{s}'_t) - f_i(u_{i,t}, s'_t) \\
& \overset{(a)}{\geq} \partial_{s'} f_i(u_{i,t}, s'_t)(\hat{s}'_t - s'_t) + \partial_u f_i(u_{i,t}, s'_t)(\psi_i(\hat{\mathbf{w}}_t, \hat{s}_{i,t}) - u_{i,t}) \\
& \overset{(b)}{\geq} \partial_{s'} f_i(u_{i,t}, s'_t)(\hat{s}'_t - s'_t) + \partial_u f_i(u_{i,t}, s'_t)\Big[\psi_i(\mathbf{w}_t, s_{i,t}) - u_{i,t} + \langle \partial_w \psi_i(\mathbf{w}_t, s_{i,t}), \hat{\mathbf{w}}_t - \mathbf{w}_t\rangle \\
& \quad - \frac{\rho_\psi}{2}\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 + \langle \partial_{s_i}\psi_i(\mathbf{w}_t, s_{i,t}), \hat{s}_{i,t} - s_{i,t}\rangle - \frac{\rho_\psi}{2}\|\hat{s}_{i,t} - s_{i,t}\|^2\Big] \\
& \overset{(c)}{\geq} \partial_{s'} f_i(u_{i,t}, s'_t)(\hat{s}'_t - s'_t) + \partial_u f_i(u_{i,t}, s'_t)\big[\psi_i(\mathbf{w}_t, s_{i,t}) - u_{i,t}\big] + \langle \partial_u f_i(u_{i,t}, s'_t)\partial_w \psi_i(\mathbf{w}_t, s_{i,t}), \hat{\mathbf{w}}_t - \mathbf{w}_t\rangle \\
& \quad + \langle \partial_u f_i(u_{i,t}, s'_t)\partial_{s_i}\psi_i(\mathbf{w}_t, s_{i,t}), \hat{s}_{i,t} - s_{i,t}\rangle - \frac{\rho_\psi C_f}{2}\left(\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 + \|\hat{s}_{i,t} - s_{i,t}\|^2\right)
\end{aligned}
$$

where (a) follows from the convexity of f_i, (b) follows from the monotonicity of $f_i(\cdot, s')$ and weak convexity of ψ_i, (c) is due to $0 \leq \partial_u f_i(u_{i,t}, s'_t) \leq C_f$. Then it follows

$$
\begin{aligned}
& \frac{1}{n_+}\sum_{X_i \in S_+}\Big[\partial_{s'} f_i(u_{i,t}, s'_t)(\hat{s}'_t - s'_t) + \langle \partial_u f_i(u_{i,t}, s'_t)\partial_w \psi_i(\mathbf{w}_t, s_{i,t}), \hat{\mathbf{w}}_t - \mathbf{w}_t\rangle \\
& \quad + \langle \partial_u f_i(u_{i,t}, s'_t)\partial_{s_i}\psi_i(\mathbf{w}_t, s_{i,t}), \hat{s}_{i,t} - s_{i,t}\rangle\Big] \\
& \leq \frac{1}{n_+}\sum_{X_i \in S_+}\Big[f_i(\psi_i(\hat{\mathbf{w}}_t, \hat{s}_{i,t}), \hat{s}'_t) - f_i(u_{i,t}, s'_t) - \partial_u f_i(u_{i,t}, s'_t)\big[\psi_i(\mathbf{w}_t, s_{i,t}) - u_{i,t}\big] \\
& \quad + \frac{\rho_\psi C_f}{2}\left(\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 + \|\hat{s}_{i,t} - s_{i,t}\|^2\right)\Big]
\end{aligned}
\tag{29}
$$

Now we consider the change in the Moreau envelope:

$$
\begin{aligned}
& \mathbb{E}_t[F_{1/\bar{\rho}}(\mathbf{w}_{t+1}, \mathbf{s}_{t+1}, s'_{t+1})] \\
& = \mathbb{E}_t\left[\min_{\tilde{\mathbf{w}}, \tilde{\mathbf{s}}, \tilde{s}'} F(\tilde{\mathbf{w}}, \tilde{\mathbf{s}}_t, \tilde{s}'_t) + \frac{\bar{\rho}}{2}\left(\|\tilde{\mathbf{w}} - \mathbf{w}_{t+1}\|^2 + \|\tilde{\mathbf{s}} - \mathbf{s}_{t+1}\|^2 + \|\tilde{s}' - \mathbf{s}'_{t+1}\|^2\right)\right] \\
& \leq \mathbb{E}_t\left[F(\hat{\mathbf{w}}_t, \hat{\mathbf{s}}_t, \hat{s}'_t) + \frac{\bar{\rho}}{2}\left(\|\hat{\mathbf{w}}_t - \mathbf{w}_{t+1}\|^2 + \|\hat{\mathbf{s}}_t - \mathbf{s}_{t+1}\|^2 + \|\hat{s}'_t - s'_{t+1}\|^2\right)\right] \\
& = F(\hat{\mathbf{w}}_t, \hat{\mathbf{s}}_t, \hat{s}'_t) + \mathbb{E}_t\Big[\frac{\bar{\rho}}{2}\big(\|\hat{\mathbf{w}}_t - (\mathbf{w}_t - \eta G_t)\|^2 + \|\hat{\mathbf{s}}_t - (\mathbf{s}_t - \eta G_t^1)\|^2 \\
& \quad + \|\hat{s}'_t - (s'_t - \eta G_t^2)\|^2\big)\Big] \\
& \leq F(\hat{\mathbf{w}}_t, \hat{\mathbf{s}}_{\mathbf{t}}, \hat{s}'_t) + \frac{\bar{\rho}}{2}\left(\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 + \|\hat{\mathbf{s}}_t - \mathbf{s}_t\|^2 + \|\hat{s}'_t - s'_t\|^2\right) \\
& \quad + \bar{\rho}\mathbb{E}_t[\eta\langle \hat{\mathbf{w}}_t - \mathbf{w}_t, G_t\rangle + \eta\langle \hat{\mathbf{s}}_t - \mathbf{s}_t, G_t^1\rangle + \eta\langle \hat{s}'_t - s'_t, G_t^2\rangle] + \frac{3\eta^2\bar{\rho}M^2}{2} \\
& = F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t) + \bar{\rho}\mathbb{E}_t[\eta\langle \hat{\mathbf{w}}_t - \mathbf{w}_t, G_t\rangle + \eta\langle \hat{\mathbf{s}}_t - \mathbf{s}_t, G_t^1\rangle + \eta\langle \hat{s}'_t - s'_t, G_t^2\rangle] \\
& \quad + \frac{3\eta^2\bar{\rho}M^2}{2}
\end{aligned}
\tag{30}
$$

where for simplicity we denote $G_t^1 = \frac{1}{B_1}\sum_{X_i \in \mathcal{B}_1^t}\partial_u f_i(u_{i,t}, s'_t)\partial_{\mathbf{s}}\psi_i(\mathbf{w}_t, s_{i,t}; \mathcal{B}_2^t)$ and $G_t^2 = \frac{1}{B_1}\sum_{X_i \in \mathcal{B}_1^t}\partial_{s'} f_i(u_{i,t}, s'_t)$. The second inequality in the above derivation uses the bounds of $\mathbb{E}[\|G_t\|^2]$, $\mathbb{E}[\|G_t^1\|^2]$ and $\mathbb{E}[\|G_t^2\|^2]$, which follow from the Lipschitz continuity and bounded variance assumptions and are denoted by M. Moreover, we have

$$\begin{aligned}&\mathbb{E}_t[\eta\langle\hat{\mathbf{w}}_t-\mathbf{w}_t,G_t\rangle+\eta\langle\hat{\mathbf{s}}_t-\mathbf{s}_t,G_t^1\rangle+\eta\langle\hat{s}'_t-s'_t,G_t^2\rangle]\\&=\eta\langle\hat{\mathbf{w}}_t-\mathbf{w}_t,\mathbb{E}_t[G_t]\rangle+\eta\langle\hat{\mathbf{s}}_t-\mathbf{s}_t,\mathbb{E}_t[G_t^1]\rangle+\eta\langle\hat{s}'_t-s'_t,\mathbb{E}_t[G_t^2]\rangle,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_t[G_t]&=\frac{1}{n_+}\sum_{X_i\in\mathcal{S}_+}\partial_u f_i(u_{i,t},s'_t)\partial_w\psi_i(\mathbf{w}_t,s_{i,t})\\\mathbb{E}_t[G_t^1]&=\frac{1}{n_+}\sum_{X_i\in\mathcal{S}_+}\partial_u f_i(u_{i,t},s'_t)\partial_{\mathbf{s}}\psi_i(\mathbf{w}_t,s_{i,t})\\\mathbb{E}_t[G_t^2]&=\frac{1}{n_+}\sum_{X_i\in\mathcal{S}_+}\partial_{s'} f_i(u_{i,t},s'_t).\end{aligned}$$

Combining inequality 29 and 30 yields

$$\begin{aligned}&\mathbb{E}_t[F_{1/\bar{\rho}}(\mathbf{w}_{t+1},\mathbf{s}_{t+1},s'_{t+1})]\\&\leq F_{1/\bar{\rho}}(\mathbf{w}_t,\mathbf{s}_t,s'_t)+\frac{3\eta^2\bar{\rho}M^2}{2}+\frac{\bar{\rho}\eta}{n_+}\sum_{X_i\in S_+}\Bigg[f_i(\psi_i(\hat{\mathbf{w}}_t,\hat{s}_{i,t}),\hat{s}'_t)-f_i(u_{i,t},s'_t)\\&\quad-\partial_u f_i(u_{i,t},s'_t)\big[\psi_i(\mathbf{w}_t,s_{i,t})-u_{i,t}\big]+\frac{\rho_\psi C_f}{2}\left(\|\hat{\mathbf{w}}_t-\mathbf{w}_t\|^2+\|\hat{s}_{i,t}-s_{i,t}\|^2\right)\Bigg]\\&\leq F_{1/\bar{\rho}}(\mathbf{w}_t,\mathbf{s}_t,s'_t)+\frac{3\eta^2\bar{\rho}M^2}{2}+\bar{\rho}\eta(F(\hat{\mathbf{w}}_t,\hat{\mathbf{s}}_t,\hat{s}'_t)-F(\mathbf{w}_t,\mathbf{s}_t,s'_t))\\&\quad+\frac{\bar{\rho}\eta}{n_+}\sum_{X_i\in S_+}\Bigg[f_i(\psi_i(\mathbf{w}_t,s_{i,t}),s'_t)-f_i(u_{i,t},s'_t)-\partial_u f_i(u_{i,t},s'_t)\big[\psi_i(\mathbf{w}_t,s_{i,t})-u_{i,t}\big]\\&\quad+\frac{\rho_\psi C_f}{2}\left(\|\hat{\mathbf{w}}_t-\mathbf{w}_t\|^2+\|\hat{s}_{i,t}-s_{i,t}\|^2\right)\Bigg]\end{aligned}\tag{31}$$

Due to the ρ_F-weak convexity of $F(\mathbf{w},\mathbf{s},s')$, we have $(\bar{\rho}-\rho_F)$-strong convexity of $(\mathbf{w},\mathbf{s},s')\mapsto F(\mathbf{w},\mathbf{s},s')+\frac{\bar{\rho}}{2}\|(\mathbf{w}_t,\mathbf{s}_t,s'_t)-(\mathbf{w},\mathbf{s},s')\|^2$. Then it follows

$$\begin{aligned}F(\hat{\mathbf{w}}_t,\hat{\mathbf{s}}_t,\hat{s}'_t)-F(\mathbf{w}_t,\mathbf{s}_t,s'_t)&=\left[F(\hat{\mathbf{w}}_t,\hat{\mathbf{s}}_t,\hat{s}'_t)+\frac{\bar{\rho}}{2}\|(\mathbf{w}_t,\mathbf{s}_t,s'_t)-(\hat{\mathbf{w}}_t,\hat{\mathbf{s}}_t,\hat{s}'_t)\|^2\right]\\&\quad-\left[F(\mathbf{w}_t,\mathbf{s}_t,s'_t)+\frac{\bar{\rho}}{2}\|(\mathbf{w}_t,\mathbf{s}_t,s'_t)-(\mathbf{w}_t,\mathbf{s}_t,s'_t)\|^2\right]\\&\quad-\frac{\bar{\rho}}{2}\|(\mathbf{w}_t,\mathbf{s}_t,s'_t)-(\hat{\mathbf{w}}_t,\hat{\mathbf{s}}_t,\hat{s}'_t)\|^2\\&\leq(\frac{\rho_F}{2}-\bar{\rho})\|(\mathbf{w}_t,\mathbf{s}_t,s'_t)-(\hat{\mathbf{w}}_t,\hat{\mathbf{s}}_t,\hat{s}'_t)\|^2\end{aligned}\tag{32}$$

Plugging inequality 32 into inequality 31 yields

$$\begin{aligned}&\mathbb{E}_t[F_{1/\bar{\rho}}(\mathbf{w}_{t+1},\mathbf{s}_{t+1},s'_{t+1})]\\&\leq\mathbb{E}[F_{1/\bar{\rho}}(\mathbf{w}_t,\mathbf{s}_t,s'_t)]+\frac{3\eta^2\bar{\rho}M^2}{2}+\bar{\rho}\eta(\frac{\rho_F}{2}-\bar{\rho})\|(\mathbf{w}_t,\mathbf{s}_t,s'_t)-(\hat{\mathbf{w}}_t,\hat{\mathbf{s}}_t,\hat{s}'_t)\|^2\\&\quad+\frac{\bar{\rho}\eta}{n_+}\sum_{X_i\in S_+}\Bigg[f_i(\psi_i(\mathbf{w}_t,s_{i,t}),s'_t)-f_i(u_{i,t},s'_t)-\partial_u f_i(u_{i,t},s'_t)\big[\psi_i(\mathbf{w}_t,s_{i,t})-u_{i,t}\big]\\&\quad+\frac{\rho_\psi C_f}{2}\left(\|\hat{\mathbf{w}}_t-\mathbf{w}_t\|^2+\|\hat{s}_{i,t}-s_{i,t}\|^2\right)\Bigg]\end{aligned}\tag{33}$$

Set $\bar{\rho}=\rho_F+\rho_\psi C_f$. We have

$$\begin{aligned}
&\mathbb{E}_t[F_{1/\bar{\rho}}(\mathbf{w}_{t+1}, \mathbf{s}_{t+1}, s'_{t+1})] \\
&\leq F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t) + \frac{3\eta^2\bar{\rho}M^2}{2} - \frac{\bar{\rho}^2\eta}{2}\|(\mathbf{w}_t, \mathbf{s}_t, s'_t) - (\hat{\mathbf{w}}_t, \hat{\mathbf{s}}_t, \hat{s}'_t)\|^2 \\
&\quad + \frac{\bar{\rho}\eta}{n_+}\sum_{X_i\in S_+}\Big[f_i(\psi_i(\mathbf{w}_t, s_{i,t}), s'_t) - f_i(u_{i,t}, s'_t) - \partial_u f_i(u_{i,t}, s'_t)\big[\psi_i(\mathbf{w}_t, s_{i,t}) - u_{i,t}\big]\Big] \\
&\overset{(a)}{\leq} F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t) + \frac{3\eta^2\bar{\rho}M^2}{2} - \frac{\eta}{2}\|\nabla\varphi_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t)\|^2 \\
&\quad + \frac{\bar{\rho}\eta}{n_+}\sum_{X_i\in S_+}\Big[f_i(\psi_i(\mathbf{w}_t, s_{i,t}), s'_t) - f_i(u_{i,t}, s'_t) - \partial_u f_i(u_{i,t}, s'_t)\big[\psi_i(\mathbf{w}_t, s_{i,t}) - u_{i,t}\big]\Big]
\end{aligned}$$

where inequality (a) follows from Lemma 3.2.

Using the Lipschitz continuity of f, we have

$$\begin{aligned}
&\mathbb{E}_t[F_{1/\bar{\rho}}(\mathbf{w}_{t+1}, \mathbf{s}_{t+1}, s'_{t+1})] \\
&\leq F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t) + \frac{3\eta^2\bar{\rho}M^2}{2} - \frac{\eta}{2}\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t)\|^2 \\
&\quad + \frac{\bar{\rho}\eta}{n_+}\sum_{X_i\in S_+} 2C_f\|\psi_i(\mathbf{w}_t, s_{i,t}) - u_{i,t}\|
\end{aligned} \tag{34}$$

With the error bound from Lemma C.3, we have

$$\mathbb{E}\left[\frac{1}{n_+}\sum_{X_i\in S_+}\|\psi_i(\mathbf{w}_t, s_{i,t}) - u_{i,t}\|\right] \leq (1-\mu)^t\frac{1}{n_+}\sum_{X_i\in S_+}\|\psi_i(\mathbf{w}_0, s_{i,0}) - u_{i,0}\| + R$$

with $\mu = \frac{B_1\tau}{2n_+}$, $R = \frac{2\tau^{1/2}\sigma}{B_2^{1/2}} + \frac{4n_+C_\psi M\eta}{B_1\tau^{1/2}} + \frac{4n_+^{1/2}C_\psi M\eta}{B_1\tau^{1/2}}$. Then

$$\begin{aligned}
&\mathbb{E}[F_{1/\bar{\rho}}(\mathbf{w}_{t+1}, \mathbf{s}_{t+1}, s'_{t+1})] \\
&\leq F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t) + \frac{3\eta^2\bar{\rho}M^2}{2} - \frac{\eta}{2}\mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t)\|^2] \\
&\quad + 2C_f\bar{\rho}\eta\left((1-\mu)^t\frac{1}{n_+}\sum_{X_i\in S_+}\|\psi_i(\mathbf{w}_0, s_{i,0}) - u_{i,0}\| + R\right)
\end{aligned} \tag{35}$$

Taking summation from $t = 0$ to $T - 1$ yields

$$\begin{aligned}
&\mathbb{E}[F_{1/\bar{\rho}}(\mathbf{w}_T, \mathbf{s}_T, s'_T)] \\
&\leq F_{1/\bar{\rho}}(\mathbf{w}_0, \mathbf{s}_0, s'_0) + \frac{3\eta^2\bar{\rho}M^2T}{2} - \frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t)\|^2] \\
&\quad + 2C_f\bar{\rho}\eta\left(\sum_{t=0}^{T-1}(1-\mu)^t\frac{1}{n_+}\sum_{X_i\in S_+}\|\psi_i(\mathbf{w}_0, s_{i,0}) - u_{i,0}\| + RT\right) \\
&\overset{(a)}{\leq} F_{1/\bar{\rho}}(\mathbf{w}_0, \mathbf{s}_0, s'_0) + \frac{3\eta^2\bar{\rho}M^2T}{2} - \frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t)\|^2] \\
&\quad + \frac{4C_f\bar{\rho}\eta}{\mu}\sum_{X_i\in S_+}\frac{1}{n_+}\|\psi_i(\mathbf{w}_0, s_{i,0}) - u_{i,0}\| + 2C_f\bar{\rho}\eta RT
\end{aligned} \tag{36}$$

where (a) uses $\sum_{t=0}^{T-1}(1-\mu)^t \leq \frac{1}{\mu}$.

Lower bounding the left-hand-side by $\min_{\mathbf{w},\mathbf{s},s'} F_{1/\bar{\rho}}(\mathbf{w},\mathbf{s},s')$, we obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t,\mathbf{s}_t,s'_t)\|^2] \\
&\leq \frac{2}{\eta T}\Bigg[F_{1/\bar{\rho}}(\mathbf{w}_0,\mathbf{s}_0,s'_0) - \min_{\mathbf{w},\mathbf{s},s'} F_{1/\bar{\rho}}(\mathbf{w},\mathbf{s},s') + \frac{3\eta^2\bar{\rho}M^2T}{2} \\
&\quad + \frac{4C_f\bar{\rho}\eta}{n_+}\sum_{X_i\in S_+}\|\psi_i(\mathbf{w}_0,s_{i,0}) - u_{i,0}\| + 2C_f\bar{\rho}\eta RT\Bigg] \\
&\leq \frac{2\Delta}{\eta T} + 3\eta\bar{\rho}M^2 + \frac{8C_f\bar{\rho}}{\mu T n_+}\sum_{X_i\in S_+}\|\psi_i(\mathbf{w}_0,s_{i,0}) - u_{i,0}\| + 4C_f\bar{\rho}R \\
&\leq \frac{C}{T}(\frac{1}{\eta}+\frac{1}{\mu}) + C(\eta+R)
\end{aligned}$$

where we assume $F_{1/\bar{\rho}}(\mathbf{w}_0,\mathbf{s}_0,s'_0) - \min_{\mathbf{w},\mathbf{s},s'} F_{1/\bar{\rho}}(\mathbf{w},\mathbf{s},s') \leq \Delta$ and

$$C = \max\{8\Delta, 12\bar{\rho}M^2, 32C_f\bar{\rho}\sum_{X_i\in S_+}\|\psi_i(\mathbf{w}_0,s_{i,0}) - u_{i,0}\|, 16C_f\bar{\rho}\}.$$

Plugging the expression of μ and R yields

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t,\mathbf{s}_t,s'_t)\|^2] \\
&\leq \mathcal{O}\left(\frac{1}{T}(\frac{1}{\eta}+\frac{n_+}{B_1\tau}) + (\frac{\tau^{1/2}\sigma}{B_2^{1/2}} + \frac{n_+\eta}{B_1\tau^{1/2}})\right)
\end{aligned}$$

Setting $\tau = \mathcal{O}(B_2\epsilon^4)$ and $\eta = \mathcal{O}(\frac{B_1B_2^{1/2}}{n_+}\epsilon^4)$, with $T = \mathcal{O}(\frac{n_+}{B_1B_2^{1/2}}\epsilon^{-6})$ iterations, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t,\mathbf{s}_t,s'_t)\|^2] \leq \epsilon^2.$$

□

C.4.2 Convergence analysis for Algorithm 6

We now consider MIL TPAUC maximization with mean pooling. Define $F(\mathbf{w},\mathbf{s},s') := \frac{1}{n_+}\sum_{X_i\in\mathcal{S}_+} f_i(g_i(h_j(\mathbf{w})-h_i(\mathbf{w}),s_i),s')$. Due to the weak-convexity of $F(\mathbf{w},\mathbf{s},s')$ w.r.t. $(\mathbf{w},\mathbf{s},s')$, we consider the following Moreau envelope and proximal map defined as

$$\begin{aligned}
F_\lambda(\mathbf{w},\mathbf{s},s') &= \min_{\tilde{\mathbf{w}},\tilde{\mathbf{s}},\tilde{s}'} F(\tilde{\mathbf{w}},\tilde{\mathbf{s}},\tilde{s}') + \frac{1}{2\lambda}\left(\|\tilde{\mathbf{w}}-\mathbf{w}\|^2 + \|\tilde{\mathbf{s}}-\mathbf{s}\|^2 + \|\tilde{s}'-s'\|^2\right), \\
\text{prox}_{\lambda F}(\mathbf{w},\mathbf{s},s') &= \arg\min_{\tilde{\mathbf{w}},\tilde{\mathbf{s}},\tilde{s}'} F(\tilde{\mathbf{w}},\tilde{\mathbf{s}},\tilde{s}') + \frac{1}{2\lambda}\left(\|\tilde{\mathbf{w}}-\mathbf{w}\|^2 + \|\tilde{\mathbf{s}}-\mathbf{s}\|^2 + \|\tilde{s}'-s'\|^2\right).
\end{aligned}$$

Following the same proofs of Lemma A.3 and Lemma A.4, we have the following error bounds

Lemma C.5. *Consider the update for $\{v_{i,t} : X_i \in \mathcal{S}_+\cup\mathcal{S}_-\}$ in Algorithm 6. Assume $h_i(\mathbf{w};\xi)$ is C_h-Lipshitz for all $X_i \in S_+\cup S_-$, and $\mathbb{E}[\|G_t\|^2] \leq M^2$. With $\gamma_1 = \frac{n_+-B_1}{B_1(1-\tau_1)} + (1-\tau_1)$,*

$\gamma_2 = \frac{n_- - B_2}{B_2(1-\tau_1)} + (1 - \tau_1)$ *and* $\tau_1 \leq \frac{1}{2}$, *we have*

$$\mathbb{E}\left[\frac{1}{n_+}\sum_{X_i\in\mathcal{S}_+}\|v_{i,t+1} - h_i(\mathbf{w}_{t+1})\|\right] \leq (1 - \frac{B_1\tau_1}{2n_+})^{t+1}\sum_{X_i\in\mathcal{S}_+}\|v_{i,0} - h_i(\mathbf{w}_t)\| + 2\tau_1^{1/2}\sigma + \frac{4n_+C_hM\eta}{B_1\tau_1^{1/2}}$$

$$\mathbb{E}\left[\frac{1}{n_-}\sum_{X_j\in\mathcal{S}_-}\|v_{j,t+1} - h_j(\mathbf{w}_{t+1})\|\right] \leq (1 - \frac{B_1\tau_1}{2n_-})^{t+1}\frac{1}{n_-}\sum_{X_j\in\mathcal{S}_-}\|v_{j,0} - h_j(\mathbf{w}_t)\| + 2\tau_1^{1/2}\sigma + \frac{4n_-C_hM\eta}{B_1\tau_1^{1/2}}$$

$$\mathbb{E}\left[\frac{1}{n_+}\sum_{X_i\in\mathcal{S}_+}\|v_{i,t+1} - h_i(\mathbf{w}_{t+1})\|^2\right] \leq (1 - \frac{B_1\tau_1}{2n_+})^{2(t+1)}\frac{1}{n_+}\sum_{X_i\in\mathcal{S}_+}\|v_{i,0} - h_i(\mathbf{w}_t)\|^2 + 4\tau_1\sigma^2 + \frac{16n_+^2C_h^2M^2\eta^2}{B_1^2\tau_1}$$

$$\mathbb{E}\left[\frac{1}{n_-}\sum_{X_j\in\mathcal{S}_-}\|v_{j,t+1} - h_j(\mathbf{w}_{t+1})\|^2\right] \leq (1 - \frac{B_1\tau_1}{2n_-})^{2(t+1)}\frac{1}{n_-}\sum_{X_j\in\mathcal{S}_-}\|v_{j,0} - h_j(\mathbf{w}_t)\|^2 + 4\tau_1\sigma^2 + \frac{16n_-^2C_h^2M^2\eta^2}{B_1^2\tau_1}$$

Lemma C.6. *Consider update for* $\{u_{i,t} : X_i \in \mathcal{S}_+\}$ *in Algorithm 6. Assume* $g_i(v_{ij}, s_i)$ *is* C_g*-Lipshitz w.r.t.* (v_{ij}, s_i) *for all* $X_i \in S_+$ *and* $X_j \in S_-$*. With* $\gamma_3 = \frac{n_+ - B_1}{B_1(1-\tau_2)} + (1 - \tau_2)$ *and* $\tau_2 \leq \frac{1}{2}$, *we have*

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{n_+}\sum_{X_i\in S_+}\|u_{i,t+1} - \frac{1}{n_-}\sum_{X_j\in S_-}g_i(v_{j,t+1} - v_{i,t+1}, s_{i,t+1})\|\right]\\ &\leq (1 - \frac{B_1\tau_2}{2n_+})^{t+1}\frac{1}{n_+}\sum_{X_i\in S_+}\|u_{i,0} - \frac{1}{n_-}\sum_{X_j\in S_-}g_i(v_{j,0} - v_{i,0}, s_{i,0})\| + 2\tau_2^{1/2}\sigma\\ &\quad + C_2\frac{n_+}{B_1}(\frac{B_1^{1/2}}{n_+^{1/2}} + \frac{B_2^{1/2}}{n_-^{1/2}})\frac{\tau_1}{\tau_2^{1/2}} + C_2\frac{n_+}{B_1}(\frac{n_+^{1/2}}{B_1^{1/2}} + \frac{n_-^{1/2}}{B_2^{1/2}})\frac{\eta}{\tau_2^{1/2}} + C_2\frac{n_+^{1/2}\eta}{B_1\tau_2^{1/2}}\end{aligned}$$

where C_2 *is a constant defined in the proof.*

Then we have the following covnergence guarantee.

Theorem C.7. *Under assumptions given in Lemma C.2, with* $\gamma_1 = \frac{n_1 - B_1}{B_1(1-\tau_1)} + (1 - \tau_1)$, $\gamma_2 = \frac{n_2 - B_2}{B_2(1-\tau_1)} + (1 - \tau_1)$, $\gamma_3 = \frac{n_1 - B_1}{B_1(1-\tau_2)} + (1 - \tau_2)$, $\tau_1 = \mathcal{O}\left(\min\left\{B_3, \frac{B_1}{n_+}\min\{\frac{n_+^{1/2}}{B_1^{1/2}}, \frac{n_-^{1/2}}{B_2^{1/2}}\}B_2^{1/2}\right\}\epsilon^4\right) \leq 1/2$, $\tau_2 = \mathcal{O}(B_2\epsilon^4) \leq 1/2$, $\eta = \mathcal{O}\left(\min\left\{\min\{\frac{B_1}{n_+}, \frac{B_2}{n_-}\}\min\{B_3^{1/2}, \frac{B_1^{1/2}}{n_+^{1/2}}\min\{\frac{n_+^{1/4}}{B_1^{1/4}}, \frac{n_-^{1/4}}{B_2^{1/4}}\}B_2^{1/4}\}, \frac{B_1}{n_+}\min\{\frac{B_1^{1/2}}{n_+^{1/2}}, \frac{B_2^{1/2}}{n_-^{1/2}}\}B_3^{1/2}\right\}\epsilon^4\right)$, *then after*

$$T \geq \mathcal{O}\left(\max\left\{\max\{\frac{n_+}{B_1}, \frac{n_-}{B_2}\}\max\{\frac{1}{B_3^{1/2}}, \frac{n_+^{1/2}}{B_1^{1/2}}\max\{\frac{B_1^{1/4}}{n_+^{1/4}}, \frac{B_2^{1/4}}{n_-^{1/4}}\}\frac{1}{B_2^{1/4}}\}, \frac{n_+}{B_1}\max\{\frac{n_+^{1/2}}{B_1^{1/2}}, \frac{n_-^{1/2}}{B_2^{1/2}}\}\frac{1}{B_2^{1/2}}\right\}\epsilon^{-6}\right)$$

iterations, Algorithm 6 gives ϵ*-stationary point to the Moreau envelope, i.e.,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s_t')\|^2 \leq \epsilon^2.$$

where $\bar{\rho} = \rho_F + \rho_g C_f + 8\rho_g C_f C_h + C_f C_g L_h$.

Proof of Theorem C.7. Consider the change in the Moreau envelope:

$$\begin{aligned}
&\mathbb{E}_t[F_{1/\bar{\rho}}(\mathbf{w}_{t+1}, \mathbf{s}_{t+1}, s'_{t+1})] \\
&= \mathbb{E}_t\left[\min_{\tilde{\mathbf{w}}, \tilde{\mathbf{s}}, \tilde{s}'} F(\tilde{\mathbf{w}}, \tilde{\mathbf{s}}_t, \tilde{s}'_t) + \frac{\bar{\rho}}{2}\left(\|\tilde{\mathbf{w}} - \mathbf{w}_{t+1}\|^2 + \|\tilde{\mathbf{s}} - \mathbf{s}_{t+1}\|^2 + \|\tilde{s}' - \mathbf{s}'_{t+1}\|^2\right)\right] \\
&\le \mathbb{E}_t\left[F(\hat{\mathbf{w}}_t, \hat{\mathbf{s}}_t, \hat{s}'_t) + \frac{\bar{\rho}}{2}\left(\|\hat{\mathbf{w}}_t - \mathbf{w}_{t+1}\|^2 + \|\hat{\mathbf{s}}_t - \mathbf{s}_{t+1}\|^2 + \|\hat{s}'_t - s'_{t+1}\|^2\right)\right] \\
&= F(\hat{\mathbf{w}}_t, \hat{\mathbf{s}}_t, \hat{s}'_t) + \mathbb{E}_t\left[\frac{\bar{\rho}}{2}\big(\|\hat{\mathbf{w}}_t - (\mathbf{w}_t - \eta G_t)\|^2 + \|\hat{\mathbf{s}}_t - (\mathbf{s}_t - \eta G_t^1)\|^2\right. \\
&\quad \left. + \|\hat{s}'_t - (s'_t - \eta G_t^2)\|^2\big)\right] \qquad (37) \\
&\le F(\hat{\mathbf{w}}_t, \hat{\mathbf{s}}_{\mathbf{t}}, \hat{s}'_t) + \frac{\bar{\rho}}{2}\left(\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 + \|\hat{\mathbf{s}}_t - \mathbf{s}_t\|^2 + \|\hat{s}'_t - s'_t\|^2\right) \\
&\quad + \bar{\rho}\mathbb{E}_t[\eta\langle\hat{\mathbf{w}}_t - \mathbf{w}_t, G_t\rangle + \eta\langle\hat{\mathbf{s}}_t - \mathbf{s}_t, G_t^1\rangle + \eta\langle\hat{s}'_t - s'_t, G_t^2\rangle] + \frac{3\eta^2\bar{\rho}M^2}{2} \\
&= F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t) + \bar{\rho}\mathbb{E}_t[\eta\langle\hat{\mathbf{w}}_t - \mathbf{w}_t, G_t\rangle + \eta\langle\hat{\mathbf{s}}_t - \mathbf{s}_t, G_t^1\rangle + \eta\langle\hat{s}'_t - s'_t, G_t^2\rangle] \\
&\quad + \frac{3\eta^2\bar{\rho}M^2}{2}
\end{aligned}$$

where for simplicity we denote $G_t^2 = \frac{1}{B_1}\sum_{i\in\mathcal{B}_1^t} \partial_{s'} f_i(u_{i,t}, s'_t)$, and G_t^1 is a n_+-dimensional vector whose i-th coordinate is defined as

$$\begin{cases} \frac{1}{B_1}\partial_u f_i(u_{i,t}, s'_t)\left[\frac{1}{B_2}\sum_{X_j\in\mathcal{B}_2^t} \partial_{s_i} g_i(v_{j,t} - v_{i,t}, s_{i,t})\right], & X_i \in \mathcal{B}_1^t \\ 0, & X_i \notin \mathcal{B}_1^t \end{cases}.$$

The second inequality in the above derivation uses the bounds of $\mathbb{E}[\|G_t\|^2], \mathbb{E}[\|G_t^1\|^2]$ and $\mathbb{E}[\|G_t^2\|^2]$, which follow from the Lipschitz continuity and bounded variance assumptions and are denoted by M.

Note that

$$\begin{aligned}
&\mathbb{E}_t[G_t] \\
&\quad = \frac{1}{n_+}\sum_{X_i\in S_+} \partial_u f_i(u_{i,t}, s'_t)\left[\frac{1}{n_-}\sum_{X_j\in S_-} \partial_v g_i(v_{j,t} - v_{i,t}, s_{i,t})\left(\nabla h_i(\mathbf{w}) - \nabla h_j(\mathbf{w})\right)\right] \\
&\mathbb{E}_t[G_t^1] = \frac{1}{n_+}\sum_{X_i\in\mathcal{S}_+} \partial_u f_i(u_{i,t}, s'_t)\left[\frac{1}{n_-}\sum_{X_j\in S_-} \partial_{\mathbf{s}} g_i(v_{j,t} - v_{i,t}, s_{i,t})\right] \\
&\mathbb{E}_t[G_t^2] = \frac{1}{n_+}\sum_{X_i\in S_+} \partial_{s'} f_i(u_{i,t}, s'_t)
\end{aligned}$$

Define $(\hat{\mathbf{w}}_t, \hat{\mathbf{s}}_t, \hat{s}'_t) := \text{prox}_{F/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t)$. For a given $i \in \{1, \dots, m\}$, we have

$$
\begin{aligned}
& f_i(\frac{1}{n_-} \sum_{X_j \in S_-} g_i(h_j(\hat{\mathbf{w}}_t) - h_i(\hat{\mathbf{w}}_t), \hat{s}_{i,t}), \hat{s}'_t) - f_i(u_{i,t}, s'_t) \\
& \overset{(a)}{\geq} \partial_{s'} f_i(u_{i,t}, s'_t)(\hat{s}'_t - s'_t) + \partial_u f_i(u_{i,t}, s'_t)(\frac{1}{n_-} \sum_{X_j \in S_-} g_i(h_j(\hat{\mathbf{w}}_t) - h_i(\hat{\mathbf{w}}_t), \hat{s}_{i,t}) - u_{i,t}) \\
& \overset{(b)}{\geq} \partial_{s'} f_i(u_{i,t}, s'_t)(\hat{s}'_t - s'_t) + \partial_u f_i(u_{i,t}, s'_t)\Bigg[\frac{1}{n_-} \sum_{X_j \in S_-} g_i(v_{j,t} - v_{i,t}, s_{i,t}) - u_{i,t} \\
& \quad + \frac{1}{n_-} \sum_{X_j \in S_-} \langle \partial_v g_i(v_{j,t} - v_{i,t}, s_{i,t}), (h_j(\hat{\mathbf{w}}_t) - h_i(\hat{\mathbf{w}}_t)) - (v_{j,t} - v_{i,t})) \rangle \\
& \quad - \frac{1}{n_-} \sum_{X_j \in S_-} \frac{\rho_g}{2} \|(h_j(\hat{\mathbf{w}}_t) - h_i(\hat{\mathbf{w}}_t)) - (v_{j,t} - v_{i,t})\|^2 \\
& \quad + \langle \frac{1}{n_-} \sum_{X_j \in S_-} \partial_{s_i} g_i(v_{j,t} - v_{i,t}), s_{i,t}, \hat{s}_{i,t} - s_{i,t} \rangle - \frac{\rho_g}{2} \|\hat{s}_{i,t} - s_{i,t}\|^2 \Bigg] \\
& \overset{(c)}{\geq} \partial_{s'} f_i(u_{i,t}, s'_t)(\hat{s}'_t - s'_t) + \partial_u f_i(u_{i,t}, s'_t)\Bigg[\frac{1}{n_-} \sum_{X_j \in S_-} g_i(v_{j,t} - v_{i,t}, s_{i,t}) - u_{i,t}\Bigg] \\
& \quad + \frac{1}{n_-} \sum_{X_j \in S_-} \underbrace{\langle \partial_u f_i(u_{i,t}, s'_t) \partial_v g_i(v_{j,t} - v_{i,t}, s_{i,t}), (h_j(\hat{\mathbf{w}}_t) - h_i(\hat{\mathbf{w}}_t)) - (v_{j,t} - v_{i,t})) \rangle}_{A_1} \\
& \quad + \frac{1}{n_-} \sum_{X_j \in S_-} \langle \partial_u f_i(u_{i,t}, s'_t) \partial_{s_i} g_i(v_{j,t} - v_{i,t}, s_{i,t}), \hat{s}_{i,t} - s_{i,t} \rangle \\
& \quad - \frac{1}{n_-} \sum_{X_j \in S_-} \frac{\rho_g C_f}{2} \|(h_j(\hat{\mathbf{w}}_t) - h_i(\hat{\mathbf{w}}_t)) - (v_{j,t} - v_{i,t})\|^2 - \frac{\rho_g C_f}{2} \|\hat{s}_{i,t} - s_{i,t}\|^2
\end{aligned}
\tag{38}
$$

where (a) follows from the convexity of f_i, (b) follows from the monotonicity of $f_i(\cdot, s')$ and weak convexity of g_i, (c) is due to $0 \leq \partial_u f_i(u_{i,t}, s'_t) \leq C_f$.

The L_h-smoothness assumption of $h_i(\mathbf{w}) - h_j(\mathbf{w})$ for all $i, \mathbf{w}$ implies

$$
\begin{aligned}
& h_i(\hat{\mathbf{w}}_t) - h_j(\hat{\mathbf{w}}_t) \\
& \geq h_i(\mathbf{w}_t) - h_j(\mathbf{w}_t) + \langle (\nabla h_i(\mathbf{w}_t) - \nabla h_j(\mathbf{w}_t)), \hat{\mathbf{w}}_t - \mathbf{w}_t \rangle - \frac{L_h}{2} \|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2
\end{aligned}
\tag{39}
$$

Since $\partial_u f_i(u_{i,t}, s'_t) \partial_v g_i(v_{j,t} - v_{i,t}, s_{i,t}) \geq 0$, we bound A_1 as following

$$
\begin{aligned}
A_1 &= \langle \partial_u f_i(u_{i,t}, s'_t) \partial_v g_i(v_{j,t} - v_{i,t}, s_{i,t}), (h_j(\hat{\mathbf{w}}_t) - h_i(\hat{\mathbf{w}}_t)) - (v_{j,t} - v_{i,t})) \rangle \\
& \overset{(a)}{\geq} \langle \partial_u f_i(u_{i,t}, s'_t) \partial_v g_i(v_{j,t} - v_{i,t}, s_{i,t}), (h_i(\mathbf{w}_t) - h_j(\mathbf{w}_t)) - (v_{j,t} - v_{i,t})) \rangle \\
& \quad - \langle \partial_u f_i(u_{i,t}, s'_t) \partial_v g_i(v_{j,t} - v_{i,t}, s_{i,t}), \frac{L_h}{2} \|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 \rangle \\
& \quad + \langle \partial_u f_i(u_{i,t}, s'_t) \partial_v g_i(v_{j,t} - v_{i,t}, s_{i,t})(\nabla h_i(\mathbf{w}_t) - \nabla h_j(\mathbf{w}_t)), \hat{\mathbf{w}}_t - \mathbf{w}_t \rangle \\
& \overset{(b)}{\geq} -C_f C_g [\|h_i(\mathbf{w}_t) - v_{i,t}\| + \|h_j(\mathbf{w}_t) - v_{j,t}\|] - \frac{C_f C_g L_h}{2} \|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 \\
& \quad + \langle \partial_u f_i(u_{i,t}, s'_t) \partial_\ell g_i(v_{j,t} - v_{i,t}, s_{i,t})(\nabla h_i(\mathbf{w}_t) - \nabla h_j(\mathbf{w}_t)), \hat{\mathbf{w}}_t - \mathbf{w}_t \rangle
\end{aligned}
$$

where inequality (a) follows from inequality 39, (b) follows from the Lipschitz continuity and monotone assumptions on f_i, g_i, h_i, h_j. Then plugging the new formulation of A_1 back to inequality 38 yields

$$
\begin{aligned}
&f_i(\frac{1}{n_-}\sum_{X_j\in S_-} g_i(h_j(\hat{\mathbf{w}}_t)-h_i(\hat{\mathbf{w}}_t),\hat{s}_{i,t}),\hat{s}'_t)-f_i(u_{i,t},s'_t)\\
&\geq \partial_{s'}f_i(u_{i,t},s'_t)(\hat{s}'_t-s'_t)+\partial_u f_i(u_{i,t},s'_t)\left[\frac{1}{n_-}\sum_{X_j\in S_-} g_i(v_{j,t}-v_{i,t},s_{i,t})-u_{i,t}\right]\\
&\quad+\frac{1}{n_-}\sum_{X_j\in S_-}[-C_fC_g[\|h_i(\mathbf{w}_t)-v_{i,t}\|+\|h_j(\mathbf{w}_t)-v_{j,t}\|]]-\frac{C_fC_gL_h}{2}\|\hat{\mathbf{w}}_t-\mathbf{w}_t\|^2\\
&\quad+\frac{1}{n_-}\sum_{X_j\in S_-}\langle \partial_u f_i(u_{i,t},s'_t)\partial_v g_i(v_{j,t}-v_{i,t},s_{i,t})(\nabla h_i(\mathbf{w}_t)-\nabla h_j(\mathbf{w}_t)),\hat{\mathbf{w}}_t-\mathbf{w}_t\rangle\\
&\quad+\frac{1}{n_-}\sum_{X_j\in S_-}\langle \partial_u f_i(u_{i,t},s'_t)\partial_{s_i} g_i(v_{j,t}-v_{i,t},s_{i,t}),\hat{s}_{i,t}-s_{i,t}\rangle\\
&\quad-\frac{1}{n_-}\sum_{X_j\in S_-}\frac{\rho_gC_f}{2}\|(h_j(\hat{\mathbf{w}}_t)-h_i(\hat{\mathbf{w}}_t))-(v_{j,t}-v_{i,t})\|^2-\frac{\rho_gC_f}{2}\|\hat{s}_{i,t}-s_{i,t}\|^2
\end{aligned}
$$

Taking average over $i\in S_+$ gives

$$
\begin{aligned}
&\frac{1}{n_+}\sum_{X_i\in S_+} f_i(\frac{1}{n_-}\sum_{X_j\in S_-} g_i(h_j(\hat{\mathbf{w}}_t)-h_i(\hat{\mathbf{w}}_t),\hat{s}_{i,t}),\hat{s}'_t)-f_i(u_{i,t},s'_t)\\
&\geq \langle \mathbb{E}_t[G_t^2],\hat{s}'_t-s'_t\rangle+\langle \mathbb{E}_t[G_t],\hat{\mathbf{w}}_t-\mathbf{w}_t\rangle+\langle \mathbb{E}_t[G_t^1],\hat{\mathbf{s}}_t-\mathbf{s}_t\rangle\\
&\quad+\frac{1}{n_+}\sum_{X_i\in S_+}\partial_u f_i(u_{i,t},s'_t)\left[\frac{1}{n_-}\sum_{X_j\in S_-} g_i(v_{j,t}-v_{i,t},s_{i,t})-u_{i,t}\right]\\
&\quad-C_fC_g\left[\frac{1}{n_+}\sum_{X_i\in S_+}\|h_i(\mathbf{w}_t)-v_{i,t}\|+\frac{1}{n_-}\sum_{X_j\in S_-}\|h_j(\mathbf{w}_t)-v_{j,t}\|\right]-\frac{C_fC_gL_h}{2}\|\hat{\mathbf{w}}_t-\mathbf{w}_t\|^2\\
&\quad-\frac{1}{n_+}\sum_{X_i\in S_+}\frac{1}{n_-}\sum_{X_j\in S_-}\frac{\rho_gC_f}{2}\|(h_j(\hat{\mathbf{w}}_t)-h_i(\hat{\mathbf{w}}_t))-(v_{j,t}-v_{i,t})\|^2-\frac{1}{n_+}\sum_{X_i\in S_+}\frac{\rho_gC_f}{2}\|\hat{s}_{i,t}-s_{i,t}\|^2
\end{aligned}
$$

It follows

$$
\begin{aligned}
&\langle \mathbb{E}_t[G_t^2],\hat{s}'_t-s'_t\rangle+\langle \mathbb{E}_t[G_t],\hat{\mathbf{w}}_t-\mathbf{w}_t\rangle+\langle \mathbb{E}_t[G_t^1],\hat{\mathbf{s}}_t-\mathbf{s}_t\rangle\\
&\leq \frac{1}{n_+}\sum_{X_i\in S_+}\Bigg[f_i(\frac{1}{n_-}\sum_{X_j\in S_-} g_i(h_j(\hat{\mathbf{w}}_t)-h_i(\hat{\mathbf{w}}_t),\hat{s}_{i,t}),\hat{s}'_t)-f_i(u_{i,t},s'_t)\\
&\quad-\partial_u f_i(u_{i,t},s'_t)\left[\frac{1}{n_-}\sum_{X_j\in S_-} g_i(v_{j,t}-v_{i,t},s_{i,t})-u_{i,t}\right]\\
&\quad+\frac{1}{n_-}\sum_{X_j\in S_-}\frac{\rho_gC_f}{2}\|(h_j(\hat{\mathbf{w}}_t)-h_i(\hat{\mathbf{w}}_t))-(v_{j,t}-v_{i,t})\|^2+\frac{\rho_gC_f}{2}\|\hat{s}_{i,t}-s_{i,t}\|^2\Bigg]\\
&\quad+C_fC_g\left[\frac{1}{n_+}\sum_{X_i\in S_+}\|h_i(\mathbf{w}_t)-v_{i,t}\|+\frac{1}{n_-}\sum_{X_j\in S_-}\|h_j(\mathbf{w}_t)-v_{j,t}\|\right]+\frac{C_fC_gL_h}{2}\|\hat{\mathbf{w}}_t-\mathbf{w}_t\|^2
\end{aligned}
\tag{40}
$$

Combining inequality 37 and 40 yields

$$\begin{aligned}
&\mathbb{E}_t[F_{1/\bar{\rho}}(\mathbf{w}_{t+1}, \mathbf{s}_{t+1}, s'_{t+1})] \\
&= F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t) + \bar{\rho}\eta \left[\langle \hat{\mathbf{w}}_t - \mathbf{w}_t, \mathbb{E}_t[G_t]\rangle + \langle \hat{\mathbf{s}}_t - \mathbf{s}_t, \mathbb{E}_t[G^1_t]\rangle + \langle \hat{s}'_t - s'_t, \mathbb{E}_t[G^2_t]\rangle\right] \\
&\quad + \frac{3\eta^2\bar{\rho}M^2}{2} \\
&\overset{(a)}{\leq} F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t) + \frac{3\eta^2\bar{\rho}M^2}{2} + \bar{\rho}\eta \Bigg\{ \frac{1}{n_+} \sum_{X_i \in S_+} \Bigg[F_i(\hat{s}'_t, \hat{\mathbf{w}}_t, \hat{s}_{i,t}) - F_i(s'_t, \mathbf{w}_t, s_{i,t}) \\
&\quad + C_f C_g \frac{1}{n_-} \sum_{X_j \in S_-} \left[\|h_i(\mathbf{w}_t) - v_{i,t}\| + \|h_j(\mathbf{w}_t)) - v_{j,t}\|\right] \\
&\quad + C_f \left\| \frac{1}{n_-} \sum_{X_j \in S_-} g_i(v_{j,t} - v_{i,t}, s_{i,t}) - u_{i,t} \right\| + C_f \left\| \frac{1}{n_-} \sum_{X_j \in S_-} g_i(v_{j,t} - v_{i,t}, s_{i,t}) - u_{i,t} \right\| \\
&\quad + \frac{1}{n_-} \sum_{X_j \in S_-} \rho_g C_f \left[\|(h_i(\hat{\mathbf{w}}_t) - v_{i,t}\|^2 + \|h_j(\hat{\mathbf{w}}_t)) - v_{j,t}\|^2\right] + \frac{\rho_g C_f}{2}\|\hat{s}_{i,t} - s_{i,t}\|^2 \Bigg] \\
&\quad + C_f C_g \left[\frac{1}{n_+} \sum_{X_i \in S_+} \|h_i(\mathbf{w}_t) - v_{i,t}\| + \frac{1}{n_-} \sum_{X_j \in S_-} \|h_j(\mathbf{w}_t) - v_{j,t}\| \right] + \frac{C_f C_g L_h}{2}\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 \Bigg\} \\
&= F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t) + \frac{3\eta^2\bar{\rho}M^2}{2} + \bar{\rho}\eta(F(\hat{\mathbf{w}}_t, \hat{\mathbf{s}}_t, \hat{s}'_t) - F(\mathbf{w}_t, \mathbf{s}_t, s'_t)) \\
&\quad + \bar{\rho}\eta \Bigg\{ \frac{1}{n_+} \sum_{X_i \in S_+} \Bigg[\frac{2C_f C_g}{n_-} \sum_{X_j \in S_-} \left[\|h_i(\mathbf{w}_t) - v_{i,t}\| + \|h_j(\mathbf{w}_t)) - v_{j,t}\|\right] \\
&\quad + \frac{2\rho_g C_f}{n_-} \sum_{X_j \in S_-} \left[\|h_i(\mathbf{w}_t) - v_{i,t}\|^2 + \|h_j(\mathbf{w}_t)) - v_{j,t}\|^2\right] + 4\rho_g C_f C_h \|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 \\
&\quad + 2C_f \left\| \frac{1}{n_-} \sum_{X_j \in S_-} g_i(v_{j,t} - v_{i,t}, s_{i,t}) - u_{i,t} \right\| + \frac{\rho_g C_f}{2}\|\hat{s}_{i,t} - s_{i,t}\|^2 \Bigg] + \frac{C_f C_g L_h}{2}\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 \Bigg\}
\end{aligned} \tag{41}$$

where (a) follows from the Lipschitz continuity of f_i, g_i, h_i, h_j and inequality 40.

Due to the ρ_F-weak convexity of $F(\mathbf{w}, \mathbf{s}_i, s')$, we have $(\bar{\rho} - \rho_F)$-strong convexity of $(\mathbf{w}, s_i, s') \mapsto F(\mathbf{w}, \mathbf{s}, s') + \frac{\bar{\rho}}{2}\|(\mathbf{w}_t, \mathbf{s}_t, s'_t) - (\mathbf{w}, \mathbf{s}, s')\|^2$. Then it follows

$$\begin{aligned}
F(\hat{\mathbf{w}}_t, \hat{\mathbf{s}}_t, \hat{s}'_t) - F_i(\mathbf{w}_t, \mathbf{s}_t, s'_t) &= \left[F_i(\hat{\mathbf{w}}_t, \hat{\mathbf{s}}_t, \hat{s}'_t) + \frac{\bar{\rho}}{2}\|(\mathbf{w}_t, \mathbf{s}_t, s'_t) - (\hat{\mathbf{w}}_t, \hat{\mathbf{s}}_t, \hat{s}'_t)\|^2\right] \\
&\quad - \left[F_i(\mathbf{w}_t, \mathbf{s}_t, s'_t) + \frac{\bar{\rho}}{2}\|(\mathbf{w}_t, \mathbf{s}_t, s'_t) - (\mathbf{w}_t, \mathbf{s}_t, s'_t)\|^2\right] \\
&\quad - \frac{\bar{\rho}}{2}\|(\mathbf{w}_t, \mathbf{s}_t, s'_t) - (\hat{\mathbf{w}}_t, \hat{\mathbf{s}}_t, \hat{s}'_t)\|^2 \\
&\leq (\frac{\rho_F}{2} - \bar{\rho})\|(\mathbf{w}_t, \mathbf{s}_t, s'_t) - (\hat{\mathbf{w}}_t, \hat{\mathbf{s}}_t, \hat{s}'_t)\|^2
\end{aligned} \tag{42}$$

Plugging inequality 42 back into 41, we obtain

$$\begin{aligned}
&\mathbb{E}_t[F_{1/\bar{\rho}}(\mathbf{w}_{t+1}, \mathbf{s}_{t+1}, s'_{t+1})] \\
&\leq F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t) + \frac{3\eta^2\bar{\rho}M^2}{2} + \bar{\rho}\eta\Bigg\{\frac{1}{n_+}\sum_{X_i\in S_+}\Bigg[(\frac{\rho_F}{2} - \bar{\rho})\|(\mathbf{w}_t, \mathbf{s}_t, s'_t) - (\hat{\mathbf{w}}_t, \hat{\mathbf{s}}_t, \hat{s}'_t)\|^2 \\
&\quad + \frac{2C_fC_g}{n_-}\sum_{X_j\in S_-}\big[\|h_i(\mathbf{w}_t) - v_{i,t}\| + \|h_j(\mathbf{w}_t) - v_{j,t}\|\big] \\
&\quad + \frac{2\rho_gC_f}{n_-}\sum_{X_j\in S_-}\big[\|h_i(\mathbf{w}_t) - v_{i,t}\|^2 + \|h_j(\mathbf{w}_t) - v_{j,t}\|^2\big] \\
&\quad + 2C_f\Bigg\|\frac{1}{n_-}\sum_{X_j\in S_-} g_i(v_{j,t} - v_{i,t}, s_{i,t}) - u_{i,t}\Bigg\| + \frac{\rho_gC_f}{2}\|\hat{s}_{i,t} - s_{i,t}\|^2\Bigg] + (4\rho_gC_fC_h + \frac{C_fC_gL_h}{2})\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2\Bigg\} \\
&\leq F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t) + \frac{3\eta^2\bar{\rho}M^2}{2} + \bar{\rho}\eta\Bigg\{\frac{1}{n_+}\sum_{X_i\in S_+}\Bigg[-\frac{\bar{\rho}}{2}\|(\mathbf{w}_t, \mathbf{s}_t, s'_t) - (\hat{\mathbf{w}}_t, \hat{\mathbf{s}}_t, \hat{s}'_t)\|^2 \\
&\quad + \frac{C_1}{n_-}\sum_{X_j\in S_-}\big[\|h_i(\mathbf{w}_t) - v_{i,t}\| + \|h_j(\mathbf{w}_t) - v_{j,t}\| + \|h_i(\mathbf{w}_t) - v_{i,t}\|^2 + \|h_j(\mathbf{w}_t) - v_{j,t}\|^2\big] \\
&\quad + C_1\Bigg\|\frac{1}{n_-}\sum_{X_j\in S_-} g_i(v_{j,t} - v_{i,t}, s_{i,t}) - u_{i,t}\Bigg\|\Bigg]\Bigg\} \\
&\overset{(b)}{\leq} F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t) + \frac{3\eta^2\bar{\rho}M^2}{2} - \frac{\eta}{2}\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t)\|^2 \\
&\quad + \frac{\bar{\rho}\eta C_1}{n_+n_-}\sum_{X_i\in S_+}\sum_{X_j\in S_-}\big[\|h_i(\mathbf{w}_t) - v_{i,t}\| + \|h_j(\mathbf{w}_t) - v_{j,t}\| + \|h_i(\mathbf{w}_t) - v_{i,t}\|^2 + \|h_j(\mathbf{w}_t) - v_{j,t}\|^2\big] \\
&\quad + \frac{\bar{\rho}\eta C_1}{n_+}\sum_{X_i\in S_+}\Bigg\|\frac{1}{n_-}\sum_{X_j\in S_-} g_i(v_{j,t} - v_{i,t}, s_{i,t}) - u_{i,t}\Bigg\|
\end{aligned}$$

where in inequality (a) we use $\bar{\rho} = \rho_F + \rho_gC_f + 8\rho_gC_fC_h + C_fC_gL_h$ and $C_1 = \max\{2C_fC_g, 2\rho_gC_f, 2C_f\}$, and inequality (b) uses Lemma 3.2.

With general error bounds

$$\begin{aligned}
&\frac{1}{n_+}\sum_{X_i\in S_+}\mathbb{E}[\|h_i(\mathbf{w}_t) - v_{i,t}\|] \leq (1-\mu_1)^t\frac{1}{n_+}\sum_{X_i\in S_+}\|h_i(\mathbf{w}_0) - v_{i,0}\| + R_1, \\
&\frac{1}{n_-}\sum_{X_j\in S_-}\mathbb{E}[\|h_j(\mathbf{w}_t) - v_{j,t}\|] \leq (1-\mu_2)^t\frac{1}{n_-}\sum_{X_j\in S_-}\|h_j(\mathbf{w}_0) - v_{j,0}\| + R_2, \\
&\frac{1}{n_+}\sum_{X_i\in S_+}\mathbb{E}[\|h_i(\mathbf{w}_t) - v_{i,t}\|^2] \leq (1-\mu_1)^t\frac{1}{n_+}\sum_{X_i\in S_+}\|h_i(\mathbf{w}_0) - v_{i,0}\|^2 + R_3, \\
&\frac{1}{n_-}\sum_{X_j\in S_-}\mathbb{E}[\|h_j(\mathbf{w}_t) - v_{j,t}\|^2] \leq (1-\mu_2)^t\frac{1}{n_-}\sum_{X_j\in S_-}\|h_j(\mathbf{w}_0) - v_{j,0}\|^2 + R_4, \\
&\frac{1}{n_+}\sum_{X_i\in S_+}\mathbb{E}\Bigg[\Bigg\|\frac{1}{n_-}\sum_{X_j\in S_-} g_i(v_{j,t} - v_{i,t}, s_{i,t}) - u_{i,t}\Bigg\|\Bigg] \\
&\quad \leq (1-\mu_3)^t\frac{1}{n_+}\sum_{X_i\in S_+}\Bigg\|\frac{1}{n_-}\sum_{X_j\in S_-} g_i(v_{i,0} - v_{j,0}, s_{i,0}) - u_{i,0}\Bigg\| + R_5,
\end{aligned}$$

we have

$$\begin{aligned}
&\mathbb{E}[F_{1/\bar{\rho}}(\mathbf{w}_{t+1}, \mathbf{s}_{t+1}, s'_{t+1})] \\
&\leq \mathbb{E}[F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t)] + \frac{3\eta^2\bar{\rho}M^2}{2} - \frac{\eta}{2}\mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t)\|^2] \\
&\quad + \bar{\rho}\eta C_1\bigg[(1-\mu_1)^t \frac{1}{n_+}\sum_{X_i \in S_+} \|h_i(\mathbf{w}_0) - v_{i,0}\| + (1-\mu_2)^t \frac{1}{n_-}\sum_{X_j \in S_-} \|h_j(\mathbf{w}_0) - v_{j,0}\| \\
&\quad + (1-\mu_1)^t \frac{1}{n_+}\sum_{X_i \in S_+} \|h_i(\mathbf{w}_0) - v_{i,0}\|^2 + (1-\mu_2)^t \frac{1}{n_-}\sum_{X_j \in S_-} \|h_j(\mathbf{w}_0) - v_{j,0}\|^2 \\
&\quad + (1-\mu_3)^t \frac{1}{n_+}\sum_{X_i \in S_+} \bigg\|\frac{1}{n_-}\sum_{X_j \in S_-} g_i(v_{i,0} - v_{j,0}, s_{i,0}) - u_{i,0}\bigg\| + R_1 + R_2 + R_3 + R_4 + R_5\bigg] \\
&\leq \mathbb{E}[F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t)] + \frac{3\eta^2\bar{\rho}M^2}{2} - \frac{\eta}{2}\mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t)\|^2] \\
&\quad + \bar{\rho}\eta C_1\bigg[(1-\mu_{min})^t \bigg(\frac{1}{n_+}\sum_{X_i \in S_+} \|h_i(\mathbf{w}_0) - v_{i,0}\| + \frac{1}{n_-}\sum_{X_j \in S_-} \|h_j(\mathbf{w}_0) - v_{j,0}\| \\
&\quad + \frac{1}{n_+}\sum_{X_i \in S_+} \|h_i(\mathbf{w}_0) - v_{i,0}\|^2 + \frac{1}{n_-}\sum_{X_j \in S_-} \|h_j(\mathbf{w}_0) - v_{j,0}\|^2 \\
&\quad + \frac{1}{n_+}\sum_{X_i \in S_+} \bigg\|\frac{1}{n_-}\sum_{X_j \in S_-} g_i(v_{i,0} - v_{j,0}, s_{i,0}) - u_{i,0}\bigg\|\bigg) + R_1 + R_2 + R_3 + R_4 + R_5\bigg]
\end{aligned}$$

where $\mu_{min} = \min\{\mu_1, \mu_2, \mu_3\}$.

Taking summation from $t = 0$ to $T - 1$ yields

$$\begin{aligned}
&\mathbb{E}[F_{1/\bar{\rho}}(\mathbf{w}_T, \mathbf{s}_T, s'_T)] \\
&\leq F_{1/\bar{\rho}}(\mathbf{w}_0, \mathbf{s}_0, s'_0) + \frac{3\eta^2\bar{\rho}M^2 T}{2} - \frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t)\|^2] \\
&\quad + \bar{\rho}\eta C_1\bigg[\sum_{t=0}^{T-1}(1-\mu_{min})^t \bigg(\frac{1}{n_+}\sum_{X_i \in S_+} \|h_i(\mathbf{w}_0) - v_{i,0}\| + \frac{1}{n_-}\sum_{X_j \in S_-} \|h_j(\mathbf{w}_0) - v_{j,0}\| \\
&\quad + \frac{1}{n_+}\sum_{X_i \in S_+} \|h_i(\mathbf{w}_0) - v_{i,0}\|^2 + \frac{1}{n_-}\sum_{X_j \in S_-} \|h_j(\mathbf{w}_0) - v_{j,0}\|^2 \\
&\quad + \frac{1}{n_+}\sum_{X_i \in S_+} \bigg\|\frac{1}{n_-}\sum_{X_j \in S_-} g_i(v_{i,0} - v_{j,0}, s_{i,0}) - u_{i,0}\bigg\|\bigg) + T(R_1 + R_2 + R_3 + R_4 + R_5)\bigg] \\
&\leq F_{1/\bar{\rho}}(\mathbf{w}_0, \mathbf{s}_0, s'_0) + \frac{3\eta^2\bar{\rho}M^2 T}{2} - \frac{\eta}{2}\sum_{t=0}^{T-1}\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t)\|^2 \\
&\quad + \bar{\rho}\eta C_1\bigg[\frac{\Delta_0}{\mu_{min}} + T(R_1 + R_2 + R_3 + R_4 + R_5)\bigg]
\end{aligned}$$

where we use $\sum_{t=0}^{T-1}(1-\mu_{min})^t \leq \frac{1}{\mu_{min}}$ and define constant Δ_0 such that

$$\begin{aligned}
&\bigg(\frac{1}{n_+}\sum_{X_i \in S_+} \|h_i(\mathbf{w}_0) - v_{i,0}\| + \frac{1}{n_-}\sum_{X_j \in S_-} \|h_j(\mathbf{w}_0) - v_{j,0}\| \\
&\quad + \frac{1}{n_+}\sum_{X_i \in S_+} \|h_i(\mathbf{w}_0) - v_{i,0}\|^2 + \frac{1}{n_-}\sum_{X_j \in S_-} \|h_j(\mathbf{w}_0) - v_{j,0}\|^2 \\
&\quad + \frac{1}{n_+}\sum_{X_i \in S_+} \bigg\|\frac{1}{n_-}\sum_{X_j \in S_-} g_i(v_{i,0} - v_{j,0}, s_{i,0}) - u_{i,0}\bigg\|\bigg) \leq \Delta_0.
\end{aligned}$$

Then it follows

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t)\|^2 \\
&\le \frac{2}{\eta T}\Bigg[F_{1/\bar{\rho}}(\mathbf{w}_0, \mathbf{s}_0, s'_0) - \mathbb{E}[F_{1/\bar{\rho}}(\mathbf{w}_T, \mathbf{s}_T, s'_T)] + \frac{3\eta^2\bar{\rho}M^2T}{2} \\
&\quad + \bar{\rho}\eta C_1\left[\frac{\Delta_0}{\mu_{min}} + T(R_1 + R_2 + R_3 + R_4 + R_5)\right]\Bigg] \\
&\le \frac{2\Delta}{\eta T} + (2 + \frac{n_+}{B_1})\eta\bar{\rho}M^2 + \frac{2\bar{\rho}C_1\Delta_0}{\mu_{min}T} + 2\bar{\rho}C_1(R_1 + R_2 + R_3 + R_4 + R_5) \\
&= \mathcal{O}\left(\frac{1}{T}(\frac{1}{\eta} + \frac{1}{\mu_{min}}) + \eta + R_1 + R_2 + R_3 + R_4 + R_5\right)
\end{aligned}
$$

where we define constant Δ such that $F_{1/\bar{\rho}}(\mathbf{w}_0, \mathbf{s}_0, s'_0) - \mathbb{E}[F_{1/\bar{\rho}}(\mathbf{w}_T, \mathbf{s}_T, s'_T)] \le \Delta$.

With MSVR updates for $v_{i,t}$ and $u_{i,t}$, following from Lemma C.5 and Lemma C.6, we have

$$
\begin{aligned}
&\mu_1 = \frac{B_1\tau_1}{2n_+}, \quad \mu_2 = \frac{B_1\tau_1}{2n_-}, \quad \mu_3 = \frac{B_1\tau_2}{2n_+} \\
&R_1 = \frac{2\tau_1^{1/2}\sigma}{B_3^{1/2}} + \frac{4n_+C_hM\eta}{B_1\tau_1^{1/2}}, \quad R_2 = \frac{2\tau_1^{1/2}\sigma}{B_3^{1/2}} + \frac{4n_-C_hM\eta}{B_2\tau_1^{1/2}} \\
&R_3 = \frac{4\tau_1\sigma^2}{B_3} + \frac{16n_+^2C_h^2M^2\eta^2}{B_1^2\tau_1}, \quad R_4 = \frac{4\tau_1\sigma^2}{B_3} + \frac{16n_-^2C_h^2M^2\eta^2}{B_2^2\tau_1} \\
&R_5 = \frac{2\tau_2^{1/2}\sigma}{B_2^{1/2}} + C_2\frac{n_+}{B_1}(\frac{B_1^{1/2}}{n_+^{1/2}} + \frac{B_2^{1/2}}{n_-^{1/2}})\frac{\tau_1}{\tau_2^{1/2}} + C_2\frac{n_+}{B_1}(\frac{n_+^{1/2}}{B_1^{1/2}} + \frac{n_-^{1/2}}{B_2^{1/2}})\frac{\eta}{\tau_2^{1/2}} + C_2\frac{n_+^{1/2}\eta}{B_1\tau_2^{1/2}}
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t)\|^2 \\
&\le \mathcal{O}\Bigg(\frac{1}{T}(\frac{1}{\eta} + \frac{1}{\mu_{min}}) + (\frac{\tau_1^{1/2}}{B_3^{1/2}} + \frac{\tau_2^{1/2}}{B_2^{1/2}})\sigma \\
&\quad + \frac{n_+\eta}{B_1\tau_1^{1/2}} + \frac{n_-\eta}{B_2\tau_1^{1/2}} + \frac{n_+}{B_1}\max\{\frac{B_1^{1/2}}{n_+^{1/2}}, \frac{B_2^{1/2}}{n_-^{1/2}}\}\frac{\tau_1}{\tau_2^{1/2}} + \frac{n_+}{B_1}\max\{\frac{n_+^{1/2}}{B_1^{1/2}}, \frac{n_-^{1/2}}{B_2^{1/2}}\}\frac{\eta}{\tau_2^{1/2}}\Bigg)
\end{aligned}
$$

Setting

$$
\tau_1 = \mathcal{O}\left(\min\left\{B_3, \frac{B_1}{n_+}\min\{\frac{n_+^{1/2}}{B_1^{1/2}}, \frac{n_-^{1/2}}{B_2^{1/2}}\}B_2^{1/2}\right\}\epsilon^4\right), \quad \tau_2 = \mathcal{O}(B_2\epsilon^4),
$$

$$
\eta = \mathcal{O}\left(\min\left\{\min\{\frac{B_1}{n_+}, \frac{B_2}{n_-}\}\min\{B_3^{1/2}, \frac{B_1^{1/2}}{n_+^{1/2}}\min\{\frac{n_+^{1/4}}{B_1^{1/4}}, \frac{n_-^{1/4}}{B_2^{1/4}}\}B_2^{1/4}\}, \frac{B_1}{n_+}\min\{\frac{B_1^{1/2}}{n_+^{1/2}}, \frac{B_2^{1/2}}{n_-^{1/2}}\}B_3^{1/2}\right\}\epsilon^4\right),
$$

Then with

$$
T \ge \mathcal{O}\left(\max\left\{\max\{\frac{n_+}{B_1}, \frac{n_-}{B_2}\}\max\{\frac{1}{B_3^{1/2}}, \frac{n_+^{1/2}}{B_1^{1/2}}\max\{\frac{B_1^{1/4}}{n_+^{1/4}}, \frac{B_2^{1/4}}{n_-^{1/4}}\}\frac{1}{B_2^{1/4}}\}, \frac{n_+}{B_1}\max\{\frac{n_+^{1/2}}{B_1^{1/2}}, \frac{n_-^{1/2}}{B_2^{1/2}}\}\frac{1}{B_2^{1/2}}\right\}\epsilon^{-6}\right)
$$

iterations, we have

$$
\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t, \mathbf{s}_t, s'_t)\|^2 \le \epsilon^2.
$$

□

D Proofs of Lemmas and Propositions

D.1 Additional Proposition

Proposition D.1. *Consider a Lipschitz continuous function $f : O \to \mathbb{R}$ where $O \subset \mathbb{R}^d$ is an open set. Assume f to be non-increasing (resp. non-decreasing) with respect to each element in the input, then all subgradients of f are element-wise non-positive (resp. non-negative).*

Proof of Proposition D.1. Let D be the subset of O where f is differentiable. By Theorem 9.60 in [24], a Lipschitz continuous function $f : O \to \mathbb{R}$, where $O \subset \mathbb{R}^d$ is an open set, is differentiable almost everywhere, i.e., D is dense in O. Then by Theorem 9.61 in [24], the subdifferential of f at x is defined as

$$\partial f(x) = \text{con}\{v | \exists x_k \to x \text{ with } x_k \in D, \nabla f(x_k) \to v\},$$

where con denotes the convex hull. If we assume that f is non-increasing with respect to each element in the input, then $\nabla f(x) \leq 0$ (element-wise) for all differentiable points $x \in D$. It implies that the all vectors in $\{v | \exists x_k \to x \text{ with } x_k \in D, \nabla f(x_k) \to v\}$ are element-wise non-positive. Therefore, all subgradients of f are element-wise non-positive. On the other hand, if we assume that f is non-decreasing, one may follow the same argument and conclude that all subgradients of f are element-wise non-negative. ☐

For functions $f : O \to \mathbb{R}^m$ where $O \subset \mathbb{R}^d$ is an open set, one may write $f = (f_1, \ldots, f_m)$ and apply the above proposition for each $f_k, k = 1, \ldots, m$.

D.2 Proofs of Proposition 4.2 and Proposition 4.4

To prove Proposition 4.2 and Proposition 4.4, we first present the following proposition on the weak-convexity of composition functions.

Proposition D.2. *Assume $f : \mathbb{R}^d \to \mathbb{R}$ is ρ_1-weakly-convex and C_1-Lipschitz continuous, $g : \mathbb{R}^{\bar{d}} \to \mathbb{R}^d$ is C_2-Lipschitz continuous, and either of the followings holds:*

1. *$f(\cdot)$ is monotone and $g(\cdot)$ is L_2-smooth;*
2. *$f(\cdot)$ is non-decreasing and $g(\cdot)$ is L_2-weakly-convex,*

then $f \circ g$ is $\tilde{\rho}$-weakly-convex with $\tilde{\rho} = \sqrt{d} L_2 C_1 + \rho_1 C_2^2$.

Proof of Proposition D.2. The weak convexity of f implies

$$\begin{aligned} f(g(y)) &\geq f(g(x)) + v^\top (g(y) - g(x)) - \frac{\rho_1}{2} \|g(y) - g(x)\|^2 \\ &\geq f(g(x)) + v^\top (g(y) - g(x)) - \frac{\rho_1 C_2^2}{2} \|x - y\|^2 \end{aligned}$$

where $v \in \partial f(g(x))$. Moreover, due to the smoothness of $g(\cdot)$ (or weakly-convexity of $g(\cdot)$, then only the second inequality holds), we have

$$\begin{aligned} g(y) - g(x) &\leq \nabla g(x)^\top (y - x) + \mathbf{v}\left(\frac{L_2}{2}\|x - y\|^2\right), \\ g(y) - g(x) &\geq \nabla g(x)^\top (y - x) - \mathbf{v}\left(\frac{L_2}{2}\|x - y\|^2\right). \end{aligned} \tag{43}$$

where $\mathbf{v}(e)$ denotes a d-dimensional vector with value e on each dimensions. We first assume that f is non-increasing, then we may use the first inequality in (43) and the Lipschitz continuity of g to get

$$\begin{aligned} f(g(y)) &\geq f(g(x)) + v^\top \left[\nabla g(x)^\top (y - x) + \mathbf{v}\left(\frac{L_2}{2}\|x - y\|^2\right)\right] - \frac{\rho_1 C_2^2}{2}\|x - y\|^2 \\ &\geq f(g(x)) + v^\top \nabla g(x)^\top (y - x) + v^\top \mathbf{v}\left(\frac{L_2}{2}\|x - y\|^2\right) - \frac{\rho_1 C_2^2}{2}\|x - y\|^2 \\ &\geq f(g(x)) + \langle v^\top \nabla g(x)^\top (y - x) - \frac{\sqrt{d} L_2 C_1 + \rho_1 C_2^2}{2}\|x - y\|^2. \end{aligned}$$

On the other hand, if we assume f is non-decreasing, the same result follows from the second inequality in (43). Thus $f \circ g$ is $\tilde{\rho}$-weakly-convex with $\tilde{\rho}_g = \sqrt{d}L_2C_1 + \rho_1 C_2^2$. $\square$

Proof of Proposition 4.2. Under Assumption 4.1, Proposition D.2 directly implies the ρ_F-weak-convexity of $F(\mathbf{w})$ with $\rho_F = \sqrt{d_1}\rho_g C_f + \rho_f C_g^2$. $\square$

Proof of Proposition 4.4. Under Assumption 4.3, we first apply Proposition D.2 to the composite function $g_i(h_{i,j}(\cdot))$ and obtain its $\rho_{\tilde{g}} = \sqrt{d_2}L_h C_g + \rho_g C_h^2$-weak-convexity. To show it Lipschitz continuity, we use the Lipschitz continuity of g_i and $h_{i,j}$ to obtain

$$\|g_i(h_{i,j}(\mathbf{w})) - g_i(h_{i,j}(\tilde{\mathbf{w}}))\|^2 \leq C_g^2 C_h^2 \|\mathbf{w} - \tilde{\mathbf{w}}\|^2.$$

Thus $g_i(h_{i,j}(\mathbf{w}))$ is $C_{\tilde{g}} = C_g C_h$-Lipschitz-continuous

Since we assume $f_i(\cdot)$ is non-decreasing, ρ_f-weakly-convex and C_f-Lipschitz continuous, and $g_i(h_{i,j}(\cdot))$ is $\rho_{\tilde{g}}$-weakly-convex and $C_{\tilde{g}}$-Lipschitz-continuous, we apply Proposition D.2 again to conclude that $F(\cdot)$ is $\rho_F = \sqrt{d_1}\rho_{\tilde{g}} C_f + \rho_f C_{\tilde{g}}^2$-weakly-convex. $\square$

D.3 Proof of Lemma 4.5

Proof of Lemma 4.5. With $\gamma = \frac{n_1 - B_1}{B_1(1-\tau)} + (1-\tau)$, $\tau \leq \frac{1}{2}$, MSVR update gives recursive error bound [15]

$$\begin{aligned}
&\mathbb{E}[\|u_{i,t+1} - g_i(\mathbf{w}_{t+1})\|^2] \\
&\leq (1 - \frac{B_1\tau}{n_1})\mathbb{E}[\|u_{i,t} - g_i(\mathbf{w}_t)\|^2] + \frac{2\tau^2 B_1 \sigma^2}{n_1 B_2} + \frac{8n_1 C_g^2}{B_1}\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2] \\
&\leq (1 - \frac{B_1\tau}{n_1})\mathbb{E}[\|u_{i,t} - g_i(\mathbf{w}_t)\|^2] + \frac{2\tau^2 B_1 \sigma^2}{n_1 B_2} + \frac{8n_1 C_g^2}{B_1}\eta^2\mathbb{E}[\|G_t\|^2] \\
&\leq (1 - \frac{B_1\tau}{2n_1})^2\mathbb{E}[\|u_{i,t} - g_i(\mathbf{w}_t)\|^2] + \frac{2\tau^2 B_1 \sigma^2}{n_1 B_2} + \frac{8n_1 C_g^2 M^2 \eta^2}{B_1}
\end{aligned}$$

Applying this inequality recursively, we obtain

$$\begin{aligned}
&\mathbb{E}[\|u_{i,t+1} - g_i(\mathbf{w}_{t+1})\|^2] \\
&\leq (1 - \frac{B_1\tau}{2n_1})^{2(t+1)}\|u_{i,0} - g_i(\mathbf{w}_0)\|^2 + \sum_{j=0}^{t}(1 - \frac{B_1\tau}{2n_1})^{2(t-j)}\left(\frac{2\tau^2 B_1\sigma^2}{n_1 B_2} + \frac{8n_+ C_g^2 M^2 \eta^2}{B_1}\right) \\
&\leq (1 - \frac{B_1\tau}{2n_1})^{2(t+1)}\|u_{i,0} - g_i(\mathbf{w}_0)\|^2 + \frac{4\tau\sigma^2}{B_2} + \frac{16n_1^2 C_g^2 M^2 \eta^2}{B_1^2 \tau}
\end{aligned}$$

where we use $\sum_{j=0}^{t}(1 - \frac{B_1\tau}{2n_1})^{2(t-j)} \leq \frac{2n_1}{B_1\tau}$.

It follows

$$\begin{aligned}
&\mathbb{E}\left[\|u_{i,t+1} - g_i(\mathbf{w}_{t+1})\|\right]^2 \\
&\leq \mathbb{E}[\|u_{i,t+1} - g_i(\mathbf{w}_{t+1})\|^2] \\
&\leq (1 - \frac{B_1\tau}{2n_1})^{2(t+1)}\|u_{i,0} - g_i(\mathbf{w}_0)\|^2 + \frac{4\tau\sigma^2}{B_2} + \frac{16n_1^2 C_g^2 M^2 \eta^2}{B_1^2 \tau} \\
&\leq \left[(1 - \frac{B_1\tau}{2n_1})^{t+1}\|u_{i,0} - g_i(\mathbf{w}_0)\| + \frac{2\tau^{1/2}\sigma}{B_2^{1/2}} + \frac{4n_1 C_g M \eta}{B_1 \tau^{1/2}}\right]^2
\end{aligned}$$

Thus

$$\begin{aligned}
&\mathbb{E}\Big[\|u_{i,t+1} - g_i(\mathbf{w}_{t+1})\|\Big] \\
&\leq (1 - \frac{B_1\tau}{2n_1})^{t+1}\|u_{i,0} - g_i(\mathbf{w}_0)\| + \frac{2\tau^{1/2}\sigma}{B_2^{1/2}} + \frac{4n_1 C_g M \eta}{B_1 \tau^{1/2}}
\end{aligned}$$

Taking summation over $i \in \mathcal{S}$, we obtain the desired result

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i\in\mathcal{S}}\|u_{i,t+1} - g_i(\mathbf{w}_{t+1})\|\right] \\
&\leq (1 - \frac{B_1\tau}{2n_1})^{t+1}\frac{1}{n}\sum_{i\in\mathcal{S}}\|u_{i,0} - g_i(\mathbf{w}_0)\| + \frac{2\tau^{1/2}\sigma}{B_2^{1/2}} + \frac{4nC_gM\eta}{B_1\tau^{1/2}}
\end{aligned}$$

□

D.4 Proof of Lemma C.6

Proof of Lemma C.6. With $\gamma_3 = \frac{n_+ - B_1}{B_1(1-\tau_2)} + (1-\tau_2)$ and $\tau_2 \leq \frac{1}{2}$, MSVR update gives the following recursive error bound [15]

$$\begin{aligned}
&\mathbb{E}[\|u_{i,t+1} - \frac{1}{n_-}\sum_{j\in S_-} g_i(v_{j,t+1} - v_{i,t+1}, s_{i,t+1})\|^2] \\
&\leq (1 - \frac{B_1\tau_2}{n_+})\mathbb{E}[\|u_{i,t} - \frac{1}{n_-}\sum_{j\in S_-} g_i(v_{j,t} - v_{i,t}, s_{i,t})\|^2] + \frac{2\tau_2^2B_1\sigma^2}{n_+B_2} \\
&\quad + \frac{8n_+C_g^2}{B_1}\mathbb{E}[\|(v_{j,t} - v_{i,t}, s_{i,t}) - (v_{j,t+1} - v_{i,t+1}, s_{i,t+1})\|^2] \qquad (44) \\
&\leq (1 - \frac{B_1\tau_2}{n_+})\mathbb{E}[\|u_{i,t} - \frac{1}{n_-}\sum_{j\in S_-} g_i(v_{j,t} - v_{i,t}, s_{i,t})\|^2] + \frac{2\tau_2^2B_1\sigma^2}{n_+B_2} \\
&\quad + \frac{16n_+C_g^2}{B_1}\mathbb{E}[\|v_{i,t} - v_{i,t+1}\|^2 + \|v_{j,t} - v_{j,t+1}\|^2] + \frac{8C_g^2M^2\eta^2}{B_1}
\end{aligned}$$

It remains to bound $\mathbb{E}[\|v_{i,t} - v_{i,t+1}\|^2]$ and $\mathbb{E}[\|v_{j,t} - v_{j,t+1}\|^2]$. We bound the former, and the latter's bound naturally follows. Consider the update of $v_{i,t+1}$ and we have

$$\begin{aligned}
&\mathbb{E}[\|v_{i,t} - v_{i,t+1}\|^2] \\
&\leq \mathbb{E}\left[\frac{B_1}{n_+}\|\tau_1 v_{i,t} - \tau_1 h^{(i)}(\mathbf{w}_t;\mathcal{B}_{3,i}^t) - \gamma_1(h^{(i)}(\mathbf{w}_t;\mathcal{B}_{3,i}^t) - h^{(i)}(\mathbf{w}_{t-1};\mathcal{B}_{3,i}^t))\|^2\right] \\
&\leq \mathbb{E}\left[\frac{2B_1\tau_1^2}{n_+}\|v_{i,t} - h^{(i)}(\mathbf{w}_t;\mathcal{B}_{3,i}^t)\|^2 + \frac{2B_1\gamma_1^2}{n_+}\|h^{(i)}(\mathbf{w}_t;\mathcal{B}_{3,i}^t) - h^{(i)}(\mathbf{w}_{t-1};\mathcal{B}_{3,i}^t)\|^2\right] \\
&\leq \mathbb{E}\left[\frac{2B_1\tau_1^2}{n_+}\|v_{i,t} - h^{(i)}(\mathbf{w}_t;\mathcal{B}_{3,i}^t)\|^2 + \frac{2B_1\gamma_1^2C_h}{n_+}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2\right] \\
&\overset{(a)}{\leq} \frac{8B_1\tau_1^2M^2}{n_+} + \frac{8n_+C_h^2\eta^2M^2}{B_1}
\end{aligned}$$

where inequality (a) uses $\tau_1 \leq 1/2$ and $\gamma_1 = \frac{n_+ - B_1}{B_1(1-\tau_1)} + (1-\tau_1) \leq \frac{2n_+}{B_1}$. Plugging the above inequality back into inequality 44 gives

$$\begin{aligned}
&\mathbb{E}[\|u_{i,t+1} - \frac{1}{n_-}\sum_{j\in S_-} g_i(v_{j,t+1} - v_{i,t+1}, s_{i,t+1})\|^2] \\
&\leq (1 - \frac{B_1\tau_2}{n_+})\mathbb{E}[\|u_{i,t} - \frac{1}{n_-}\sum_{j\in S_-} g_i(v_{j,t} - v_{i,t}, s_{i,t})\|^2] + \frac{2\tau_2^2B_1\sigma^2}{n_+B_2} \\
&\quad + \frac{16n_+C_g^2}{B_1}\left(8\tau_1^2M^2(\frac{B_1}{n_+} + \frac{B_2}{n_-}) + 8C_h^2\eta^2M^2(\frac{n_+}{B_1} + \frac{n_-}{B_2})\right) + \frac{8C_g^2M^2\eta^2}{B_1} \\
&\leq (1 - \frac{B_1\tau_2}{n_+})\mathbb{E}[\|u_{i,t} - \frac{1}{n_-}\sum_{j\in S_-} g_i(v_{j,t} - v_{i,t}, s_{i,t})\|^2] + \frac{2\tau_2^2B_1\sigma^2}{n_+B_2} \\
&\quad + 128C_g^2M^2\frac{n_+}{B_1}(\frac{B_1}{n_+} + \frac{B_2}{n_-})\tau_1^2 + 128C_g^2C_h^2M^2\frac{n_+}{B_1}(\frac{n_+}{B_1} + \frac{n_-}{B_2})\eta^2 + \frac{8C_g^2M^2\eta^2}{B_1}
\end{aligned}$$

Applying this inequality recursively, we obtain

$$\begin{aligned}
&\mathbb{E}[\|u_{i,t+1} - \frac{1}{n_-}\sum_{j\in S_-} g_i(v_{j,t+1} - v_{i,t+1}, s_{i,t+1})\|^2] \\
&\leq (1 - \frac{B_1\tau_2}{2n_+})^{2(t+1)}\|u_{i,0} - \frac{1}{n_-}\sum_{j\in S_-} g_i(v_{i,0} - v_{j,0}, s_{i,0})\|^2 + \sum_{j=0}^{t}(1 - \frac{B_1\tau_2}{2n_+})^{2(t-j)}\bigg(\frac{2\tau_2^2 B_1\sigma^2}{n_+ B_2} \\
&\quad + 128C_g^2M^2\frac{n_+}{B_1}(\frac{B_1}{n_+} + \frac{B_2}{n_-})\tau_1^2 + 128C_g^2C_h^2M^2\frac{n_+}{B_1}(\frac{n_+}{B_1} + \frac{n_-}{B_2})\eta^2 + \frac{8C_g^2M^2\eta^2}{B_1}\bigg) \\
&\leq (1 - \frac{B_1\tau_2}{2n_+})^{2(t+1)}\|u_{i,0} - \frac{1}{n_-}\sum_{j\in S_-} g_i(v_{i,0} - v_{j,0}, s_{i,0})\|^2 + \frac{4\tau_2\sigma^2}{B_2} \\
&\quad + 256C_g^2M^2\frac{n_+^2}{B_1^2}(\frac{B_1}{n_+} + \frac{B_2}{n_-})\frac{\tau_1^2}{\tau_2} + 256C_g^2C_h^2M^2\frac{n_+^2}{B_1^2}(\frac{n_+}{B_1} + \frac{n_-}{B_2})\frac{\eta^2}{\tau_2} + \frac{16n_+C_g^2M^2\eta^2}{B_1^2\tau_2} \\
&\leq (1 - \frac{B_1\tau_2}{2n_+})^{2(t+1)}\|u_{i,0} - \frac{1}{n_-}\sum_{j\in S_-} g_i(v_{i,0} - v_{j,0}, s_{i,0})\|^2 + \frac{4\tau_2\sigma^2}{B_2} + C_2^2\frac{n_+^2}{B_1^2}(\frac{B_1}{n_+} + \frac{B_2}{n_-})\frac{\tau_1^2}{\tau_2} \\
&\quad + C_2^2\frac{n_+^2}{B_1^2}(\frac{n_+}{B_1} + \frac{n_-}{B_2})\frac{\eta^2}{\tau_2} + C_2^2\frac{n_+\eta^2}{B_1^2\tau_2}
\end{aligned}$$

where we use $\sum_{j=0}^{t}(1 - \frac{B_1\tau_2}{2n_+})^{2(t-j)} \leq \frac{2n_+}{B_1\tau_1}$ and denotes $C_2^2 = 2\max\{256C_g^2M^2, 256C_g^2C_h^2M^2, 16C_g^2M^2\}$. Taking average over $i \in S_+$ gives the squared-norm error bound.

To derive the norm error bound, we derive

$$\begin{aligned}
&\mathbb{E}[\|u_{i,t+1} - \frac{1}{n_-}\sum_{j\in S_-} g_i(v_{j,t+1} - v_{i,t+1}, s_{i,t+1})\|]^2 \\
&\leq \mathbb{E}[\|u_{i,t+1} - \frac{1}{n_-}\sum_{j\in S_-} g_i(v_{j,t+1} - v_{i,t+1}, s_{i,t+1})\|^2] \\
&\leq (1 - \frac{B_1\tau_2}{2n_+})^{2(t+1)}\|u_{i,0} - \frac{1}{n_-}\sum_{j\in S_-} g_i(v_{i,0} - v_{j,0}, s_{i,0})\|^2 + \frac{4\tau_2\sigma^2}{B_2} + C_2^2\frac{n_+^2}{B_1^2}(\frac{B_1}{n_+} + \frac{B_2}{n_-})\frac{\tau_1^2}{\tau_2} \\
&\quad + C_2^2\frac{n_+^2}{B_1^2}(\frac{n_+}{B_1} + \frac{n_-}{B_2})\frac{\eta^2}{\tau_2} + C_2^2\frac{n_+\eta^2}{B_1^2\tau_2} \\
&\leq \bigg[(1 - \frac{B_1\tau_2}{2n_+})^{t+1}\|u_{i,0} - \frac{1}{n_-}\sum_{j\in S_-} g_i(v_{i,0} - v_{j,0}, s_{i,0})\| + \frac{2\tau_2^{1/2}\sigma}{B_2^{1/2}} + C_2\frac{n_+}{B_1}(\frac{B_1^{1/2}}{n_+^{1/2}} + \frac{B_2^{1/2}}{n_-^{1/2}})\frac{\tau_1}{\tau_2^{1/2}} \\
&\quad + C_2\frac{n_+}{B_1}(\frac{n_+^{1/2}}{B_1^{1/2}} + \frac{n_-^{1/2}}{B_2^{1/2}})\frac{\eta}{\tau_2^{1/2}} + C_2\frac{n_+^{1/2}\eta}{B_1\tau_2^{1/2}}\bigg]^2
\end{aligned}$$

Thus

$$\begin{aligned}
&\mathbb{E}[\|u_{i,t+1} - \frac{1}{n_-}\sum_{j\in S_-} g_i(v_{j,t+1} - v_{i,t+1}, s_{i,t+1})\|] \\
&\leq (1 - \frac{B_1\tau_2}{2n_+})^{t+1}\|u_{i,0} - \frac{1}{n_-}\sum_{j\in S_-} g_i(v_{i,0} - v_{j,0}, s_{i,0})\| + \frac{2\tau_2^{1/2}\sigma}{B_2^{1/2}} + C_2\frac{n_+}{B_1}(\frac{B_1^{1/2}}{n_+^{1/2}} + \frac{B_2^{1/2}}{n_-^{1/2}})\frac{\tau_1}{\tau_2^{1/2}} \\
&\quad + C_2\frac{n_+}{B_1}(\frac{n_+^{1/2}}{B_1^{1/2}} + \frac{n_-^{1/2}}{B_2^{1/2}})\frac{\eta}{\tau_2^{1/2}} + C_2\frac{n_+^{1/2}\eta}{B_1\tau_2^{1/2}}
\end{aligned}$$

Taking average over $i \in S_+$, we obtain the norm error bound

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n_+}\sum_{i\in S_+}\|u_{i,t+1}-\frac{1}{n_-}\sum_{j\in S_-}g_i(v_{j,t+1}-v_{i,t+1},s_{i,t+1})\|\right] \\
&\le (1-\frac{B_1\tau_2}{2n_+})^{t+1}\frac{1}{n_+}\sum_{i\in S_+}\|u_{i,0}-\frac{1}{n_-}\sum_{j\in S_-}g_i(v_{i,0}-v_{j,0},s_{i,0})\|+\frac{2{\tau_2}^{1/2}\sigma}{B_2^{1/2}} \\
&\quad + C_2\frac{n_+}{B_1}(\frac{B_1^{1/2}}{n_+^{1/2}}+\frac{B_2^{1/2}}{n_-^{1/2}})\frac{\tau_1}{\tau_2^{1/2}}+C_2\frac{n_+}{B_1}(\frac{n_+^{1/2}}{B_1^{1/2}}+\frac{n_-^{1/2}}{B_2^{1/2}})\frac{\eta}{\tau_2^{1/2}}+C_2\frac{n_+^{1/2}\eta}{B_1\tau_2^{1/2}}
\end{aligned}$$

□

D.5 Proof of Lemma A.3

Proof of Lemma A.3. With $\gamma_1 = \frac{n_1 n_2 - B_1 B_2}{B_1 B_2 (1-\tau_1)} + (1-\tau_1)$ and $\tau_1 \leq \frac{1}{2}$, MSVR update has the following recursive error bound [15][15]

$$\begin{aligned}
&\mathbb{E}[\|v_{i,j,t+1} - h_{i,j}(\mathbf{w}_{t+1})\|^2] \\
&\leq (1 - \frac{B_1 B_2 \tau_1}{n_1 n_2})\mathbb{E}[\|v_{i,j,t} - h_{i,j}(\mathbf{w}_t)\|^2] + \frac{2\tau_1^2 B_1 B_2 \sigma^2}{n_1 n_2 B_3} + \frac{8 n_1 n_2 C_h^2}{B_1 B_2}\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2] \\
&\leq (1 - \frac{B_1 B_2 \tau_1}{2 n_1 n_2})^2\mathbb{E}[\|v_{i,j,t} - h_{i,j}(\mathbf{w}_t)\|^2] + \frac{2\tau_1^2 B_1 B_2 \sigma^2}{n_1 n_2 B_3} + \frac{8 n_1 n_2 C_h^2 M^2 \eta^2}{B_1 B_2}
\end{aligned}$$

Applying this inequality recursively, we obtain

$$\begin{aligned}
&\mathbb{E}[\|v_{i,j,t+1} - h_{i,j}(\mathbf{w}_{t+1})\|^2] \\
&\leq (1 - \frac{B_1 B_2 \tau_1}{2 n_1 n_2})^{2(t+1)}\|v_{i,j,0} - h_{i,j}(\mathbf{w}_0)\|^2 + \sum_{j=0}^{t}(1 - \frac{B_1 B_2 \tau_1}{2 n_1 n_2})^{2(t-j)}(\frac{2\tau_1^2 B_1 B_2 \sigma^2}{n_1 n_2 B_3} + \frac{8 n_1 n_2 C_h^2 M^2 \eta^2}{B_1 B_2}) \\
&\leq (1 - \frac{B_1 B_2 \tau_1}{2 n_1 n_2})^{2(t+1)}\|v_{i,j,0} - h_{i,j}(\mathbf{w}_0)\|^2 + \frac{4\tau_1 \sigma^2}{B_3} + \frac{16 n_1^2 n_2^2 C_h^2 M^2 \eta^2}{B_1^2 B_2^2 \tau_1}
\end{aligned}$$

where we use $\sum_{j=0}^{t}(1 - \frac{B_1 B_2 \tau_1}{2 n_1 n_2})^{2(t-j)} \leq \frac{2 n_1 n_2}{B_1 B_2 \tau_1}$. Taking average over $(i,j) \in S_1 \times S_2$ gives the squared-norm error bound.

To derive the norm error bound, we derive

$$\begin{aligned}
&\mathbb{E}[\|v_{i,j,t+1} - h_{i,j}(\mathbf{w}_{t+1})\|]^2 \\
&\leq \mathbb{E}[\|v_{i,j,t+1} - h_{i,j}(\mathbf{w}_{t+1})\|^2] \\
&\leq (1 - \frac{B_1 B_2 \tau_1}{2 n_1 n_2})^{2(t+1)}\|v_{i,j,0} - h_{i,j}(\mathbf{w}_0)\|^2 + \frac{4\tau_1 \sigma^2}{B_3} + \frac{16 n_1^2 n_2^2 C_h^2 M^2 \eta^2}{B_1^2 B_2^2 \tau_1} \\
&\leq \left[(1 - \frac{B_1 B_2 \tau_1}{2 n_1 n_2})^{t+1}\|v_{i,j,0} - h_{i,j}(\mathbf{w}_0)\| + \frac{2\tau_1^{1/2}\sigma}{B_3^{1/2}} + \frac{4 n_1 n_2 C_h M \eta}{B_1 B_2 \tau_1^{1/2}}\right]^2
\end{aligned}$$

Thus

$$\begin{aligned}
&\mathbb{E}[\|v_{i,j,t+1} - h_{i,j}(\mathbf{w}_{t+1})\|] \\
&\leq (1 - \frac{B_1 B_2 \tau_1}{2 n_1 n_2})^{t+1}\|v_{i,j,0} - h_{i,j}(\mathbf{w}_0)\| + \frac{2\tau_1^{1/2}\sigma}{B_3^{1/2}} + \frac{4 n_1 n_2 C_h M \eta}{B_1 B_2 \tau_1^{1/2}}
\end{aligned}$$

Taking average over $(i,j) \in \mathcal{S}_1 \times \mathcal{S}_2$, we obtain the norm error bound

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n_1}\sum_{i\in\mathcal{S}_1}\frac{1}{n_2}\sum_{j\in\mathcal{S}_2}\|v_{i,j,t+1} - h_{i,j}(\mathbf{w}_{t+1})\|\right] \\
&\leq (1 - \frac{B_1 B_2 \tau_1}{2 n_1 n_2})^{t+1}\frac{1}{n_1}\sum_{i\in\mathcal{S}_1}\frac{1}{n_2}\sum_{j\in\mathcal{S}_2}\|v_{i,j,0} - h_{i,j}(\mathbf{w}_0)\| + \frac{2\tau_1^{1/2}\sigma}{B_3^{1/2}} + \frac{4 n_1 n_2 C_h M \eta}{B_1 B_2 \tau_1^{1/2}}.
\end{aligned}$$

□

D.6 Proof of Lemma A.4

Proof of Lemma A.4. With $\gamma_2 = \frac{n_1 - B_1}{B_1(1-\tau_2)} + (1-\tau_2)$ and $\tau_2 \leq \frac{1}{2}$, MSVR update has the following recursive error bound [15]

$$\begin{aligned}
&\mathbb{E}[\|u_{i,t+1} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t+1})\|^2] \\
&\leq (1 - \frac{B_1 \tau_2}{n_1})\mathbb{E}[\|u_{i,t} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t})\|^2] + \frac{2\tau_2^2 B_1 \sigma^2}{n_1 B_2} + \frac{8 n_1 C_g^2}{B_1}\mathbb{E}[\|v_{i,j,t+1} - v_{i,j,t}\|^2]
\end{aligned} \tag{45}$$

It remains to bound $\mathbb{E}[\|v_{i,j,t+1} - v_{i,j,t}\|^2]$, which is done as following

$$
\begin{aligned}
&\mathbb{E}[\|v_{i,j,t+1} - v_{i,j,t}\|^2] \\
&\leq \mathbb{E}\left[\frac{B_1 B_2}{n_1 n_2}\|\tau_1 v_{i,j,t} - \tau_1 h_{i,j}(\mathbf{w}_t; \mathcal{B}^t_{3,i,j}) - \gamma_1(h_{i,j}(\mathbf{w}_t; \mathcal{B}^t_{3,i,j}) - h_{i,j}(\mathbf{w}_{t-1}; \mathcal{B}^t_{3,i,j}))\|^2\right] \\
&\leq \mathbb{E}\left[\frac{2B_1 B_2 \tau_1^2}{n_1 n_2}\|v_{i,j,t} - h_{i,j}(\mathbf{w}_t; \mathcal{B}^t_{3,i,j})\|^2 + \frac{2B_1 B_2 \gamma_1^2}{n_1 n_2}\|h_{i,j}(\mathbf{w}_t; \mathcal{B}^t_{3,i,j}) - h_{i,j}(\mathbf{w}_{t-1}; \mathcal{B}^t_{3,i,j})\|^2\right] \\
&\leq \mathbb{E}\left[\frac{2B_1 B_2 \tau_1^2}{n_1 n_2}\|v_{i,j,t} - h_{i,j}(\mathbf{w}_t; \mathcal{B}^t_{3,i,j})\|^2 + \frac{2B_1 B_2 \gamma_1^2 C_h}{n_1 n_2}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2\right] \\
&\overset{(a)}{\leq} \frac{8B_1 B_2 \tau_1^2 M^2}{n_1 n_2} + \frac{8 n_1 n_2 C_h^2 \eta^2 M^2}{B_1 B_2}
\end{aligned}
$$

where inequality (a) uses $\tau_1 \leq 1/2$ and $\gamma_1 = \frac{n_1 n_2 - B_1 B_2}{B_1 B_2 (1-\tau_1)} + (1-\tau_1) \leq \frac{2 n_1 n_2}{B_1 B_2}$. Plugging the above inequality back into inequality 45 gives

$$
\begin{aligned}
&\mathbb{E}[\|u_{i,t+1} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t+1})\|^2] \\
&\leq (1 - \frac{B_1 \tau_2}{n_1})\mathbb{E}[\|u_{i,t} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t})\|^2] + \frac{2\tau_2^2 B_1 \sigma^2}{n_1 B_2} + \frac{8 n_1 C_g^2}{B_1}\left(\frac{8B_1 B_2 \tau_1^2 M^2}{n_1 n_2} + \frac{8 n_1 n_2 C_h^2 \eta^2 M^2}{B_1 B_2}\right) \\
&\leq (1 - \frac{B_1 \tau_2}{n_1})\mathbb{E}[\|u_{i,t} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t})\|^2] + \frac{2\tau_2^2 B_1 \sigma^2}{n_1 B_2} + \frac{64 B_2 \tau_1^2 M^2 C_g^2}{n_2} + \frac{64 n_1^2 n_2 C_h^2 \eta^2 M^2 C_g^2}{B_1^2 B_2}
\end{aligned}
$$

Applying this inequality recursively, we obtain

$$
\begin{aligned}
&\mathbb{E}[\|u_{i,t+1} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t+1})\|^2] \\
&\leq (1 - \frac{B_1 \tau_2}{2n_1})^{2(t+1)}\|u_{i,0} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,0})\|^2 + \sum_{j=0}^{t}(1 - \frac{B_1\tau_2}{2n_1})^{t-j}\left(\frac{2\tau_2^2 B_1 \sigma^2}{n_1 B_2} + \frac{64 B_2 \tau_1^2 M^2 C_g^2}{n_2}\right. \\
&\quad \left. + \frac{64 n_1^2 n_2 C_h^2 \eta^2 M^2 C_g^2}{B_1^2 B_2}\right) \\
&\leq (1 - \frac{B_1 \tau_2}{2n_1})^{2(t+1)}\|u_{i,0} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,0})\|^2 + \frac{4\tau_2 \sigma^2}{B_2} + \frac{128 n_1 B_2 \tau_1^2 M^2 C_g^2}{B_1 n_2 \tau_2} + \frac{128 n_1^3 n_2 C_h^2 \eta^2 M^2 C_g^2}{B_1^3 B_2 \tau_2}
\end{aligned}
$$

where we use $\sum_{j=0}^{t}(1 - \frac{B_1\tau_2}{2n_1})^{2(t-j)} \leq \frac{2n_1}{B_1 \tau_1}$. Taking average over $i \in \mathcal{S}_1$ gives the squared-norm error bound.

To derive the norm error bound, we derive

$$
\begin{aligned}
&\mathbb{E}[\|u_{i,t+1} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t+1})\|]^2 \\
&\leq \mathbb{E}[\|u_{i,t+1} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t+1})\|^2] \\
&\leq (1 - \frac{B_1 \tau_2}{2n_1})^{2(t+1)}\|u_{i,0} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,0})\|^2 + \frac{4\tau_2 \sigma^2}{B_2} + \frac{128 n_1 B_2 \tau_1^2 M^2 C_g^2}{B_1 n_2 \tau_2} + \frac{128 n_1^3 n_2 C_h^2 \eta^2 M^2 C_g^2}{B_1^3 B_2 \tau_2} \\
&\leq \left[(1 - \frac{B_1 \tau_2}{2n_1})^{t+1}\|u_{i,0} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,0})\| + \frac{2\tau_2^{1/2}\sigma}{B_2^{1/2}} + \frac{8\sqrt{2} n_1^{1/2} B_2^{1/2} \tau_1 M C_g}{B_1^{1/2} n_2^{1/2} \tau_2^{1/2}} + \frac{8\sqrt{2} n_1^{3/2} n_2^{1/2} C_h \eta M C_g}{B_1^{3/2} B_2^{1/2} \tau_2^{1/2}}\right]^2
\end{aligned}
$$

Taking squared root on both sides and taking average over $i \in S_1$, we obtain the norm error bound

$$\mathbb{E}\left[\frac{1}{n_1}\sum_{i\in\mathcal{S}_1}\|u_{i,t+1}-\frac{1}{n_2}\sum_{j\in\mathcal{S}_2}g_i(v_{i,j,t+1})\|\right]$$

$$\leq (1-\frac{B_1\tau_2}{2n_1})^{t+1}\frac{1}{n_1}\sum_{i\in\mathcal{S}_1}\|u_{i,0}-\frac{1}{n_2}\sum_{j\in\mathcal{S}_2}g_i(v_{i,j,0})\| + \frac{2\tau_2^{1/2}\sigma}{B_2^{1/2}} + \frac{C_2 n_1^{1/2}B_2^{1/2}\tau_1}{B_1^{1/2}n_2^{1/2}\tau_2^{1/2}} + \frac{C_2 n_1^{3/2}n_2^{1/2}\eta}{B_1^{3/2}B_2^{1/2}\tau_2^{1/2}}$$

where $C_2 = \max\{8\sqrt{2}MC_g, 8\sqrt{2}C_hMC_g\}$. □

D.7 Proof of Lemma B.2

Proof of Lemma B.2. Define

$$\tilde{u}_{i,t} = (1-\tau)u_{i,t} + \tau g_i(\mathbf{w}_t;\mathcal{B}_{2,i}^t)$$

Then we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{B}_{2,i}^t}[\|\tilde{u}_{i,t}-g_i(\mathbf{w}_t)\|^2] \\
&= \mathbb{E}_{\mathcal{B}_{2,i}^t}[\|(1-\tau)(u_{i,t}-g_i(\mathbf{w}_t)) + \tau(g_i(\mathbf{w}_t;\mathcal{B}_{2,i}^t)-g_i(\mathbf{w}_t))\|^2] \\
&= \mathbb{E}_{\mathcal{B}_{2,i}^t}[(1-\tau)^2\|u_{i,t}-g_i(\mathbf{w}_t)\|^2 + \tau^2\|g_i(\mathbf{w}_t;\mathcal{B}_{2,i}^t)-g_i(\mathbf{w}_t)\|^2 \\
&\quad + 2(1-\tau)\tau\langle u_{i,t}-g_i(\mathbf{w}_t), g_i(\mathbf{w}_t;\mathcal{B}_{2,i}^t)-g_i(\mathbf{w}_t)\rangle] \\
&\leq (1-\tau)^2\|u_{i,t}-g_i(\mathbf{w}_t)\|^2 + \frac{\tau^2\sigma^2}{B_2}
\end{aligned}$$

It follows

$$\begin{aligned}
&\mathbb{E}_{\mathcal{B}_{2,i}^t}\mathbb{E}_{\mathcal{B}_1^t}[\|u_{i,t+1}-g_i(\mathbf{w}_t)\|^2] \\
&= \frac{B_1}{n_1}\mathbb{E}_{\mathcal{B}_{2,i}^t}[\|\tilde{u}_{i,t}-g_i(\mathbf{w}_t)\|^2] + (1-\frac{B_1}{n_1})\|u_{i,t}-g_i(\mathbf{w}_t)\|^2 \\
&\leq \frac{B_1}{n_1}(1-\tau)^2\|u_{i,t}-g_i(\mathbf{w}_t)\|^2 + \frac{B_1\tau^2\sigma^2}{n_1B_2} + (1-\frac{B_1}{n_1})\|u_{i,t}-g_i(\mathbf{w}_t)\|^2 \\
&\leq (1-\frac{B_1\tau}{2n_1})^2\|u_{i,t}-g_i(\mathbf{w}_t)\|^2 + \frac{B_1\tau^2\sigma^2}{n_1B_2}
\end{aligned}$$

where we use

$$\begin{aligned}
\frac{B_1}{n_1}(1-\tau)^2 + (1-\frac{B_1}{n_1}) &= \frac{B_1}{n_1}(1-2\tau+\tau^2) + 1 - \frac{B_1}{n_1} \\
&= 1 - 2\tau\frac{B_1}{n_1} + \tau^2\frac{B_1}{n_1} \\
&\leq 1 - \tau\frac{B_1}{n_1} \\
&\leq 1 - \tau\frac{B_1}{n_1} + (\frac{\tau B_1}{2n_1})^2 = (1-\frac{\tau B_1}{2n_1})^2
\end{aligned}$$

Then

$$\begin{aligned}
&\mathbb{E}_t[\|u_{i,t+1}-g_i(\mathbf{w}_{t+1})\|^2] \\
&\leq \mathbb{E}_t\left[(1+\frac{B_1\tau}{4n_1})\|u_{i,t+1}-g_i(\mathbf{w}_t)\|^2 + (1+\frac{4n_1}{B_1\tau})\|g_i(\mathbf{w}_t)-g_i(\mathbf{w}_{t+1})\|^2\right] \\
&\leq (1+\frac{B_1\tau}{4n_1})(1-\frac{B_1\tau}{2n_1})^2\|u_{i,t}-g_i(\mathbf{w}_t)\|^2 + (1+\frac{B_1\tau}{4n_1})\frac{B_1\tau^2\sigma^2}{n_1B_2} \\
&\quad + (1+\frac{4n_1}{B_1\tau})C_g^2\mathbb{E}_t\|\mathbf{w}_t-\mathbf{w}_{t+1}\|^2] \\
&\leq (1-\frac{B_1\tau}{4n_1})^2\|u_{i,t}-g_i(\mathbf{w}_t)\|^2 + \frac{2B_1\tau^2\sigma^2}{n_1B_2} + \frac{8n_1}{B_1\tau}C_g^2\mathbb{E}_t\|\mathbf{w}_t-\mathbf{w}_{t+1}\|^2] \\
&\leq (1-\frac{B_1\tau}{4n_1})^2\|u_{i,t}-g_i(\mathbf{w}_t)\|^2 + \frac{2B_1\tau^2\sigma^2}{n_1B_2} + \frac{8n_1C_g^2M^2\eta^2}{B_1\tau}
\end{aligned}$$

where we use $\frac{B_1\tau}{4n_1} \leq 1$. Applying this inequality recursively, we obtain

$$\begin{aligned}
&\mathbb{E}[\|u_{i,t+1} - g_i(\mathbf{w}_{t+1})\|^2] \\
&\leq (1 - \frac{B_1\tau}{4n_1})^2 \mathbb{E}[\|u_{i,t} - g_i(\mathbf{w}_t)\|^2] + \frac{2B_1\tau^2\sigma^2}{n_1B_2} + \frac{8n_1C_g^2M^2\eta^2}{B_1\tau} \\
&\leq (1 - \frac{B_1\tau}{4n_1})^{2(t+1)} \|u_{i,0} - g_i(\mathbf{w}_0)\|^2 + \sum_{j=0}^{t}(1 - \frac{B_1\tau}{4n_1})^{2(t-j)}\left[\frac{2B_1\tau^2\sigma^2}{n_1B_2} + \frac{8n_1C_g^2M^2\eta^2}{B_1\tau}\right] \\
&\leq (1 - \frac{B_1\tau}{4n_1})^{2(t+1)} \|u_{i,0} - g_i(\mathbf{w}_0)\|^2 + \frac{8\tau\sigma^2}{B_2} + \frac{32n_1^2C_g^2M^2\eta^2}{B_1^2\tau^2}
\end{aligned}$$

where we use $\sum_{j=0}^{t}(1 - \frac{B_1\tau}{4n_1})^{2(t-j)} \leq \frac{4n_1}{B_1\tau}$.

To obtain the absolute bound, we derive

$$\begin{aligned}
\mathbb{E}[\|u_{i,t+1} - g_i(\mathbf{w}_{t+1})\|]^2 &\leq \mathbb{E}[\|u_{i,t+1} - g_i(\mathbf{w}_{t+1})\|^2] \\
&\leq (1 - \frac{B_1\tau}{4n_1})^{2(t+1)} \|u_{i,0} - g_i(\mathbf{w}_0)\|^2 + \frac{8\tau\sigma^2}{B_2} + \frac{32n_1^2C_g^2M^2\eta^2}{B_1^2\tau^2} \\
&\leq \left[(1 - \frac{B_1\tau}{4n_1})^{t+1} \|u_{i,0} - g_i(\mathbf{w}_0)\| + \frac{2\sqrt{2}\tau^{1/2}\sigma}{B_2^{1/2}} + \frac{4\sqrt{2}n_1C_gM\eta}{B_1\tau}\right]^2
\end{aligned}$$

The desired result follows by taking squared root on both sides. ☐

D.8 Proof of Lemma B.5

Proof of Lemma B.5. The proof of Lemma B.5 is the same as Lemma B.2. ☐

D.9 Proof of Lemma B.6

Proof of Lemma B.6. Define

$$\tilde{u}_{i,t} = (1 - \tau_2)u_{i,t} + \tau_2 \frac{1}{B_2} \sum_{j \in \mathcal{B}_{2,i}^t} g_i(v_{i,j,t})$$

Then we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{B}_2^t}[\|\tilde{u}_{i,t} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t})\|^2] \\
&= \mathbb{E}_{\mathcal{B}_2^t}[\|(1-\tau_2)(u_{i,t} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t})) + \tau_2(\frac{1}{B_2}\sum_{j\in\mathcal{B}_2^t} g_i(v_{i,j,t}) - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t}))\|^2] \\
&= \mathbb{E}_{\mathcal{B}_2^t}[(1-\tau_2)^2\|u_{i,t} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t})\|^2 + \tau_2^2\|\frac{1}{B_2}\sum_{j\in\mathcal{B}_2^t} g_i(v_{i,j,t}) - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t})\|^2 \\
&\quad + 2(1-\tau_2)\tau_2\langle u_{i,t} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t}), \frac{1}{B_2}\sum_{j\in\mathcal{B}_2^t} g_i(v_{i,j,t}) - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t})\rangle] \\
&\leq (1-\tau_2)^2\|u_{i,t} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t})\|^2 + \frac{\tau_2^2\sigma^2}{B_2}
\end{aligned}$$

It follows

$$\begin{aligned}
&\mathbb{E}_{\mathcal{B}_{2,i}^t}\mathbb{E}_{\mathcal{B}_1^t}[\|u_{i,t+1} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t})\|^2] \\
&= \frac{B_1}{n_1}\mathbb{E}_{\mathcal{B}_2^t}[\|\tilde{u}_{i,t} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t})\|^2] + (1-\frac{B_1}{n_1})\|u_{i,t} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t})\|^2 \\
&\leq \frac{B_1}{n_1}(1-\tau_2)^2\|u_{i,t} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t})\|^2 + \frac{B_1\tau_2^2\sigma^2}{n_1B_2} + (1-\frac{B_1}{n_1})\|u_{i,t} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t})\|^2 \\
&\leq (1-\frac{B_1\tau_2}{2n_1})^2\|u_{i,t} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t})\|^2 + \frac{B_1\tau^2\sigma^2}{n_1B_2}
\end{aligned}$$

where we use

$$\frac{B_1}{n_1}(1-\tau_2)^2 + (1-\frac{B_1}{n_1}) \leq (1-\frac{\tau_2B_1}{2n_1})^2$$

Then

$$\begin{aligned}
&\mathbb{E}_t[\|u_{i,t+1} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t+1})\|^2] \\
&\leq \mathbb{E}_t\left[(1+\frac{B_1\tau_2}{4n_1})\|u_{i,t+1} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t})\|^2 + (1+\frac{4n_1}{B_1\tau_2})\|\frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t}) - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t+1})\|^2\right] \\
&\leq (1+\frac{B_1\tau_2}{4n_1})(1-\frac{B_1\tau_2}{2n_1})^2\|u_{i,t} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t})\|^2 + (1+\frac{B_1\tau_2}{4n_1})\frac{B_1\tau_2^2\sigma^2}{n_1B_2} \\
&\quad + (1+\frac{4n_1}{B_1\tau_2})C_g^2\mathbb{E}_t[\frac{1}{n_2}\sum_{j\in\mathcal{S}_2}\|v_{i,j,t} - v_{i,j,t+1}\|^2] \\
&\leq (1-\frac{B_1\tau_2}{4n_1})^2\|u_{i,t} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t})\|^2 + \frac{2B_1\tau_2^2\sigma^2}{n_1B_2} + \frac{8C_g^2M^2B_2\tau_1^2}{n_2\tau_2}
\end{aligned}$$

where we use $\frac{B_1\tau_2}{4n_1} \leq 1$, and

$$\mathbb{E}_t[\|v_{i,j,t} - v_{i,j,t+1}\|^2] = \frac{B_1B_2}{n_1n_2}\mathbb{E}_{\mathcal{B}_{3,i,j}^t}\|\tau_1v_{i,j,t} - \tau_1h_{i,j}(\mathbf{w}_t;\mathcal{B}_{3,i,j}^t)\|^2 \leq \frac{B_1B_2\tau_1^2M^2}{n_1n_2}.$$

Applying this inequality recursively, we obtain

$$\begin{aligned}
&\mathbb{E}[\|u_{i,t+1} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t+1})\|^2] \\
&\leq (1-\frac{B_1\tau_2}{4n_1})^2\mathbb{E}[\|u_{i,t} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t})\|^2] + \frac{2B_1\tau_2^2\sigma^2}{n_1B_2} + \frac{8C_g^2M^2B_2\tau_1^2}{n_2\tau_2} \\
&\leq (1-\frac{B_1\tau_2}{4n_1})^{2(t+1)}\|u_{i,0} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,0})\|^2 + \sum_{j=0}^{t}(1-\frac{B_1\tau_2}{4n_1})^{2(t-j)}\left[\frac{2B_1\tau_2^2\sigma^2}{n_1B_2} + \frac{8C_g^2M^2B_2\tau_1^2}{n_2\tau_2}\right] \\
&\leq (1-\frac{B_1\tau_2}{4n_1})^{2(t+1)}\|u_{i,0} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,0})\|^2 + \frac{8\tau_2\sigma^2}{B_2} + \frac{32C_g^2M^2n_1B_2\tau_1^2}{B_1n_2\tau_2^2}
\end{aligned}$$

where we use $\sum_{j=0}^{t}(1-\frac{B_1\tau_2}{4n_1})^{2(t-j)} \leq \frac{4n_1}{B_1\tau_2}$.

To obtain the absolute bound, we derive

$$
\begin{aligned}
&\mathbb{E}[\|u_{i,t+1} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t+1})\|]^2 \\
&\leq \mathbb{E}[\|u_{i,t+1} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,t+1})\|^2] \\
&\leq (1-\frac{B_1\tau_2}{4n_1})^{2(t+1)}\|u_{i,0} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,0})\|^2 + \frac{8\tau_2\sigma^2}{B_2} + \frac{32C_g^2M^2n_1B_2\tau_1^2}{B_1n_2\tau_2^2} \\
&\leq \left[(1-\frac{B_1\tau_2}{4n_1})^{t+1}\|u_{i,0} - \frac{1}{n_2}\sum_{j\in\mathcal{S}_2} g_i(v_{i,j,0})\| + \frac{2\sqrt{2}\tau_2^{1/2}\sigma}{B_2^{1/2}} + \frac{4\sqrt{2}C_gMn_1^{1/2}B_2^{1/2}\tau_1}{B_1^{1/2}n_2^{1/2}\tau_2}\right]^2
\end{aligned}
$$

The desired result follows by taking squared root on both sides. □

E Group Distributionally Robust Optimization

NSWC FCCO finds an important application in group distributionally robust optimization (group DRO), particularly valuable in addressing distributional shift [25]. Consider N groups with different distributions. Each group k has an averaged loss $L_k(w) = \frac{1}{n_k}\sum_{i=1}^{n_k}\ell(f_w(x_i^k), y_i^k)$, where w is the the model parameter and (x_i^k, y_i^k) is a data point. For robust optimization, we assign different weights to different groups and form the following robust loss minimization problem:

$$\min_{w}\max_{p\in\Omega}\sum_{k=1}^{N} p_k L_k(w),$$

where $\Omega \subset \Delta$ and Δ denotes a simplex. A common choice for Ω is $\Omega = \{\mathbf{p} \in \Delta, p_i \leq 1/K\}$ where K is an integer, resulting in the so-called CVaR losses, i.e., average of top-K group losses. Consequently, the above problem can be equivalently reformulated as [23]:

$$\min_{w}\min_{s} F(w,s) = \frac{1}{K}\sum_{k=1}^{N}[L_k(w) - s]_+ + s.$$

This formulation can be mapped into non-smooth weakly-convex FCCO when the loss function $\ell(\cdot,\cdot)$ is weakly convex in terms of w. In comparison to directly solving the min-max problem, solving the above FCCO problem avoids the need of dealing with the projection onto the constraint Ω and expensive sampling as in existing works [4].

F More Information for Experiments

F.1 Dataset Statistics

Table 3: Datasets Statistics. The percentage in parenthesis represents the proportion of positive samples.

Dataset	Train	Validation	Test
moltox21(t0)	5834 (4.25%)	722 (4.01%)	709 (4.51%)
molmuv(t1)	11466 (0.18%)	1559 (0.13%)	1709 (0.35%)
molpcba(t0)	120762 (9.32%)	19865 (11.74%)	20397 (11.61%)

Table 4: Data statistics for the MIL datasets. D_+/D_- is the positive/negative bag number.

Data Format	Dataset	D_+	D_-	average bag size	#features
Tabular	MUSK2	39	63	64.69	166
	Fox	100	100	6.6	230
Histopathological	Lung	100	1000	256	32x32x3
Image	Lung	100	1000	256	32x32x3

F.2 Illustration for Histopathology Dataset on MIL Task

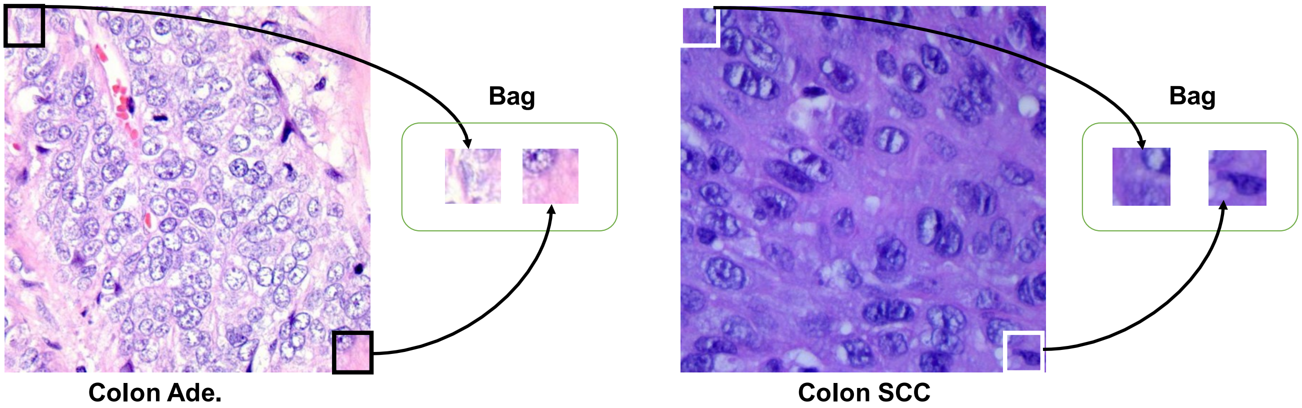

Figure 2: Illustration for Histopathology Dataset on MIL Task. Ade. is abbreviated for adenocarcinoma and SCC is short for squamous cell carcinoma. In this work, each RGB image is separated by 32×32 non-overlapped patches, which constitute the bag.