Supplementary Material

A. Proof of Lemma 1

Proof. We begin by introducing a Lagrange multiplier η for the constraint $\mathbb{E}[q_i] = 1$, and form the Lagrangian

$$L(\eta, q) := \mathbb{E}[q_i f^i(\omega)] + \eta(1 - \mathbb{E}[q_i]) = \mathbb{E}[q_i(f^i(\omega) - \eta)] + \eta.$$

Thus, f_α is equivalent to

$$\sup_{q:q_i \in [0,1/a_i]} \inf_{\eta \in \mathbb{R}} L(\eta, q).$$

By switching $\inf$ and $\sup$, we obtain the following inequality

$$\sup_{q:q_i \in [0,1/a_i]} \inf_{\eta \in \mathbb{R}} L(\eta, q) \leq \inf_{\eta \in \mathbb{R}} \sup_{q:q_i \in [0,1/a_i]} L(\eta, q). \tag{A.1}$$

The inner maximization problem in the right hand side can be solved exactly by letting $q_i = 0$ if $f^i(\omega) - \eta < 0$ and $q_i = a_i^{-1}$ if $f^i(\omega) - \eta \geq 0$, leading to

$$\inf_{\eta \in \mathbb{R}} \sup_{q:q_i \in [0,1/a_i]} L(q, \eta) = \inf_{\eta \in \mathbb{R}} \left\{ \mathbb{E}\left[\frac{1}{a_i} (f^i(\omega) - \eta)_+ \right] + \eta \right\}.$$

Therefore, to prove the first part of the lemma, it remains to show that equation (A.1) holds with equality. Denote $L = \min_i f^i(\omega)$ and $U = \max_i f^i(\omega)$. Since $\eta \to g(\eta) := \mathbb{E}[\frac{1}{a_i}(f^i(\omega) - \eta)_+] + \eta$ is strictly increasing on $[U, \infty)$, we have $g(\eta) \geq g(U)$ for $\eta \in [U, \infty)$. For $\eta \leq L$, we have $g(\eta) = \mathbb{E}[\frac{1}{a_i}(f^i(\omega)] + \eta(1 - \mathbb{E}(\frac{1}{a_i}))$ which is non-increasing as $\mathbb{E}(\frac{1}{a_i}) \geq 1$. So $g(\eta) \geq g(L)$ for $\eta \leq L$. Therefore, we may restrict the domain of η on a compact convex domain $[L, U]$. Now since $\eta \to L(\eta, q)$ is linear and thus convex, $q \to L(\eta, q)$ is linear and thus concave, and the domain of q and η are both compact and convex, the von Neumann's minimax theorem (Neumann, 1928) implies that the equality holds, which completes the proof of the first part.

The second part can be proven as follows.

$$\begin{aligned}
&\inf_{\eta \in \mathbb{R}} \left\{ \mathbb{E}\left[\frac{1}{a_i} (f^i(\omega) - \eta)_+ \right] + \eta \right\} \\
&= \inf_{\eta \in \mathbb{R}} \left\{ \frac{1}{\tau_\alpha^{-1}} \mathbb{E}_\alpha \left[(f^i(\omega) - \eta)_+ \right] + \eta \right\}, \\
&= \mathbb{E}_\alpha[f^i(\omega) | f^i(\omega) > \mathbb{Q}_{1-1/\tau_\alpha}(f^i(\omega))] \\
&= \text{HCVaR}_{1-\alpha}(f^i(\omega))
\end{aligned}$$

where the second inequality follows from Theorem 1 of Rockafellar et al. (2000). □

B. Proof of Lemma 2

The proof of Lemma 2 relies on the following result.

Proposition 1 ((Bullins, 2020), Lemma 3)**.** *For $x \in \mathbb{R}$,*

$$(x)_+ \leq \phi_\mu(x) \leq (x)_+ + \mu.$$

Proof. By Proposition 1, we have $(\hat{f}^i(\omega) - \eta)_+ \leq \phi_\mu(\hat{f}^i(\omega) - \eta) \leq (\hat{f}^i(\omega) - \eta)_+ + \mu$. Multiplying by $\frac{1}{\alpha_i}$ on both sides and summing them up give us the desired result. □

C. Proof of Theorem 1

This section includes the full proof of Theorem 1. For ease of notation, we introduce the following shorthand notations. We denote $\theta := (\omega, \eta) \in \mathcal{W} \times \mathbb{R}$. Then the local and global losses can be rewritten as $\hat{f}^{\mu,i}(\theta) := \hat{f}^{\mu,i}(\omega, \eta)$, $\hat{F}^{\mu}_{\alpha}(\theta) := \hat{F}^{\mu}_{\alpha}(\omega, \eta)$, and $\hat{F}_{\alpha}(\theta) := \hat{F}_{\alpha}(\omega, \eta)$. Let the average model at iteration t be $\bar{\theta}_t = \sum_{i=1}^{n} p_i \theta_t^i$ and the minimizer $\theta^* := \arg\min \hat{F}^{\mu}_{\alpha}(\theta)$. We first establish the convexity and Lipschitz gradient property of $\hat{f}^{\mu,i}$ and $\hat{F}^{\mu}_{\alpha}$ with respect to the parameter θ.

Lemma C.1. *If the empirical loss $\hat{f}^i(\omega)$ satisfies Assumption 1 and 2, then $\hat{f}^{\mu,i}(\theta)$ and $\hat{F}^{\mu}_{\alpha}(\theta)$ are convex and have Lipschitz gradients as follows, for any θ, θ',*

$$\begin{array}{l} ||\nabla \hat{f}^{\mu,i}(\theta) - \nabla \hat{f}^{\mu,i}(\theta')|| \leq L||\theta - \theta'|| \\ ||\nabla \hat{F}^{\mu}_{\alpha}(\theta) - \nabla \hat{F}^{\mu}_{\alpha}(\theta')|| \leq L||\theta - \theta'|| \end{array},$$

where $L := (L_1 + \dfrac{L_2^2 + 1}{\mu}) \max_i \dfrac{1}{\alpha_i}$.

Proof. Denote $g(\theta) = \hat{f}^i(\omega) - \eta$. By Assumption 1 (smoothness) and 2, we have $||\nabla g(\theta) - \nabla g(\theta')|| \leq L_1 ||\theta - \theta'||$ and $||\nabla g(\theta)|| \leq \sqrt{L_2^2 + 1}$. Then, the Lipschitz gradient parameter for $\phi_\mu(g(\theta))$ can be calculated as follows.

$$\begin{array}{l} ||\nabla \phi_\mu(g(\theta)) - \nabla \phi_\mu(g(\theta'))|| \\ = ||\phi'_\mu(g(\theta))\nabla g(\theta) - \phi'_\mu(g(\theta'))\nabla g(\theta')|| \\ = ||\phi'_\mu(g(\theta))\nabla g(\theta) - \phi'_\mu(g(\theta))\nabla g(\theta') + \phi'_\mu(g(\theta))\nabla g(\theta') - \phi'_\mu(g(\theta'))\nabla g(\theta')|| \\ \leq \phi'_\mu(g(\theta))||\nabla g(\theta) - \nabla g(\theta')|| + |\phi'_\mu(g(\theta)) - \phi'_\mu(g(\theta'))|||\nabla g(\theta')|| \\ \leq L_1||\theta - \theta'|| + \dfrac{1}{\mu}|g(\theta) - g(\theta')|||\nabla g(\theta')|| \\ \leq (L_1 + \dfrac{L_2^2 + 1}{\mu})||\theta - \theta'|| \end{array},$$

where the second inequality uses $\phi'_\mu(\cdot) \leq 1$ and $\frac{1}{\mu}$-smoothness of ϕ_μ. Since $\hat{f}^{\mu,i}(\theta) = \dfrac{1}{\alpha_i}\phi_\mu(\hat{f}^i(\omega) - \eta) + \eta$ and $\hat{F}^{\mu}_{\alpha}(\theta) = \sum_{i=1}^{n} p_i \hat{f}^{\mu,i}(\theta)$, we obtain

$$||\nabla \hat{f}^{\mu,i}(\theta) - \nabla \hat{f}^{\mu,i}(\theta')|| \leq (L_1 + \frac{L_2^2 + 1}{\mu})\frac{1}{\alpha_i}||\theta - \theta'||.$$

And,

$$||\nabla \hat{F}^{\mu}_{\alpha}(\theta) - \nabla \hat{F}^{\mu}_{\alpha}(\theta')|| \leq (L_1 + \frac{L_2^2 + 1}{\mu})\sum_{i=1}^{n}\frac{p_i}{\alpha_i}||\theta - \theta'||.$$

To show the convexity of $\hat{f}^{\mu,i}(\theta)$ and $\hat{F}^{\mu}_{\alpha}(\theta)$, we first observe that $g(\theta)$ is convex with respect to θ as $g(\theta) = \hat{f}^i(\omega) - \eta \geq \hat{f}^i(\omega') - \eta' + < \nabla \hat{f}^i(\omega'), \omega - \omega' > + \eta' - \eta = g(\theta') + < \nabla g(\theta'), \theta - \theta' >$ where the first inequality uses Assumption 1 (convexity). Then since $\phi_\mu(\cdot)$ is convex and non-decreasing, we can conclude that $\phi_\mu(g(\theta))$ is also convex with respect to θ. Therefore, $\hat{f}^{\mu,i}(\theta)$ and $\hat{F}^{\mu}_{\alpha}(\theta)$ are convex. □

We next bound the average deviation of local models from their average over T iterations. For this purpose, we first prove a technical lemma given as follows.

Lemma C.2. *Suppose that two non-negative sequences $\{I_t\}_{t\geq 0} (I_0, I_{\mathbb{Z}^+ \cdot \kappa + 1} = 0)$ and $\{H_t\}_{t \geq 0}$ satisfy the following inequality for each iteration $t \geq 0$ and some constants $C_1 \geq 0, C_2 \geq 0$ and $C_3 \geq 0$:*

$$I_t \leq C_1 \sum_{l=t_\kappa+1}^{t-1} I_l + C_2 \sum_{l=t_\kappa+1}^{t-1} H_l + C_3, \tag{C.1}$$

where $t_\kappa := \lfloor \frac{t-1}{\kappa} \rfloor \kappa$. If further assuming that $C_1(\kappa - 1) \leq \frac{1}{2}$, then we have

$$\sum_{t=0}^{T-1} I_t \leq 2C_2(\kappa - 1)\sum_{t=0}^{T-1} H_t + 2C_3 T.$$

Proof. We apply the inequality (C.1) to each iteration $t = 0, \cdots, T-1$ and obtain

$$
\begin{cases}
I_0 = 0 \\
I_1 = 0 \\
I_2 \leq C_1 I_1 + C_2 H_1 + C_3 \\
\vdots \\
I_\kappa \leq C_1(I_1 + \cdots + I_{\kappa-1}) + C_2(H_1 + \cdots + H_{\kappa-1}) + C_3
\end{cases}
\begin{cases}
I_{\kappa+1} = 0 \\
I_{\kappa+2} \leq C_1 I_{\kappa+1} + C_2 H_{\kappa+1} + C_3 \\
\vdots \\
I_{2\kappa} \leq C_1(I_{\kappa+1} \cdots + I_{2\kappa-1}) + C_2(H_{\kappa+1} \cdots + H_{2\kappa-1}) + C_3
\end{cases}
\vdots
\begin{cases}
I_{(T-1)_\kappa+1} = 0 \\
I_{(T-1)_\kappa+2} \leq C_1 I_{(T-1)_\kappa+1} + C_2 H_{(T-1)_\kappa+1} + C_3 \\
\vdots \\
I_{T-1} \leq C_1(I_{(T-1)_\kappa+1} \cdots + I_{T-2}) + C_2(H_{(T-1)_\kappa+1} \cdots + H_{T-2}) + C_3
\end{cases}
$$

Summing the above inequalities yields that

$$
\sum_{t=0}^{T-1} I_t \leq C_1(\kappa - 1) \sum_{t=0}^{T-1} I_t + C_2(\kappa - 1) \sum_{t=0}^{T-1} H_t + C_3 T.
$$

As $C_1(\kappa - 1) \leq \frac{1}{2}$ by assumption, rearranging the terms gives

$$
\sum_{t=0}^{T-1} I_t \leq 2C_2(\kappa - 1) \sum_{t=0}^{T-1} H_t + 2C_3 T.
$$

□

The following lemma bounds the sum of model variance from iteration $t = 0$ to $T - 1$.

Lemma C.3. *If the Assumption 1 and 2 hold and the learning rate β satisfies $6L^2\beta^2(\kappa-1)^2 \leq 1$, then*

$$
\textstyle\sum_{t=0}^{T-1} \sum_{i=1}^n p_i ||\theta_t^i - \bar{\theta}_t||^2 \leq 8L\beta^2(\kappa-1)^2 \sum_{i=0}^{T-1} (\hat{F}_\alpha^\mu(\bar{\theta}_i) - \hat{F}_\alpha^\mu(\theta^*)) + 12T\beta^2(\kappa-1)^2\rho^2 \ ,
$$

where $\rho^2 := \sum_{i=1}^n p_i ||\nabla \hat{f}^{\mu,i}(\theta^*)||^2$.

Proof. Consider an iteration t and denote by t_κ the step of the most recent communication between the clients and the server, i.e., $t_\kappa = \lfloor \frac{t-1}{\kappa} \rfloor \kappa$. Then by the update rule of Algorithm 1, all the clients have the same local model at iteration $t_\kappa + 1$, i.e., $\theta_{t_\kappa+1}^1 = \cdots = \theta_{t_\kappa+1}^n = \bar{\theta}_{t_\kappa+1}$, and for each client we can write $\theta_t^i = \theta_{t_\kappa+1}^i - \beta \sum_{l=t_\kappa+1}^{t-1} \nabla \hat{f}^{\mu,i}(\theta_l^i)$. Therefore, we can upper bound $\sum_{i=1}^n p_i ||\theta_t^i - \bar{\theta}_t||^2$ as follows.

$$
\begin{aligned}
&\textstyle\sum_{i=1}^n p_i ||\theta_t^i - \bar{\theta}_t||^2 \\
&= \beta^2 \textstyle\sum_{i=1}^n p_i || \sum_{l=t_\kappa+1}^{t-1} \nabla \hat{f}^{\mu,i}(\theta_l^i) - \sum_{l=t_\kappa+1}^{t-1} \mathbb{E}[\nabla \hat{f}^{\mu,i}(\theta_l^i)]||^2 \\
&\leq \beta^2 (t - 1 - t_\kappa) \textstyle\sum_{i=1}^n p_i \sum_{l=t_\kappa+1}^{t-1} ||\nabla \hat{f}^{\mu,i}(\theta_l^i) - \mathbb{E}[\nabla \hat{f}^{\mu,i}(\theta_l^i)]||^2 \ , \\
&\leq \beta^2 (\kappa - 1) \textstyle\sum_{l=t_\kappa+1}^{t-1} \sum_{i=1}^n p_i ||\nabla \hat{f}^{\mu,i}(\theta_l^i) - \mathbb{E}[\nabla \hat{f}^{\mu,i}(\theta_l^i)]||^2 \\
&\leq \beta^2 (\kappa - 1) \textstyle\sum_{l=t_\kappa+1}^{t-1} \sum_{i=1}^n p_i ||\nabla \hat{f}^{\mu,i}(\theta_l^i)||^2
\end{aligned}
\tag{C.2}
$$

where the first inequality follows from the Jensen's inequality, the second inequality is due to the fact that $t - t_\kappa \leq \kappa$ by definition, and the last inequality uses $\text{Var}(Z) \leq \mathbb{E}(Z^2)$.

Now we proceed to bound $\sum_{i=1}^n p_i||\nabla \hat{f}^{\mu,i}(\theta_l^i)||^2$.

$$\begin{aligned}
&\sum_{i=1}^n p_i||\nabla \hat{f}^{\mu,i}(\theta_l^i)||^2 \\
&= \sum_{i=1}^n p_i||\nabla \hat{f}^{\mu,i}(\theta_l^i) - \nabla \hat{f}^{\mu,i}(\bar{\theta}_l) + \nabla \hat{f}^{\mu,i}(\bar{\theta}_l) - \nabla \hat{f}^{\mu,i}(\theta^*) + \nabla \hat{f}^{\mu,i}(\theta^*)||^2 \\
&\leq \sum_{i=1}^n p_i(3||\nabla \hat{f}^{\mu,i}(\theta_l^i) - \nabla \hat{f}^{\mu,i}(\bar{\theta}_l)||^2 + 2||\nabla \hat{f}^{\mu,i}(\bar{\theta}_l) - \nabla \hat{f}^{\mu,i}(\theta^*)||^2 + 6||\nabla \hat{f}^{\mu,i}(\theta^*)||^2) \\
&\leq \sum_{i=1}^n p_i(3L^2||\theta_l^i - \bar{\theta}_l||^2 + 2||\nabla \hat{f}^{\mu,i}(\bar{\theta}_l) - \nabla \hat{f}^{\mu,i}(\theta^*)||^2 + 6||\nabla \hat{f}^{\mu,i}(\theta^*)||^2) \\
&\leq \sum_{i=1}^n p_i(3L^2||\theta_l^i - \bar{\theta}_l||^2 + 4L(\hat{f}^{\mu,i}(\bar{\theta}_l) - \hat{f}^{\mu,i}(\theta^*) - < \nabla \hat{f}^{\mu,i}(\theta^*), \bar{\theta}_l - \theta^* >) + 6||\nabla \hat{f}^{\mu,i}(\theta^*)||^2) \\
&= 3L^2 \sum_{i=1}^n p_i||\theta_l^i - \bar{\theta}_l||^2 + 4L(\hat{F}_\alpha^\mu(\bar{\theta}_l) - \hat{F}_\alpha^\mu(\theta^*) - < \nabla \hat{F}_\alpha^\mu(\theta^*), \bar{\theta}_l - \theta^* >) + 6\sum_{i=1}^n p_i||\nabla \hat{f}^{\mu,i}(\theta^*)||^2 \\
&= 3L^2 \sum_{i=1}^n p_i||\theta_l^i - \bar{\theta}_l||^2 + 4L(\hat{F}_\alpha^\mu(\bar{\theta}_l) - \hat{F}_\alpha^\mu(\theta^*)) + 6\rho^2
\end{aligned} \quad , \tag{C.3}$$

where the first inequality uses AM-GM inequality, the second inequality follows from the Lipschitz gradient, the third inequality uses the co-coercivity of convex and smooth function, and the last equality holds as $\nabla \hat{F}_\alpha^\mu(\theta^*) = 0$.

Plugging (C.3) back in (C.2) yields

$$\sum_{i=1}^n p_i||\theta_t^i - \bar{\theta}_t||^2 \leq 3L^2\beta^2(\kappa-1)\sum_{l=t_\kappa+1}^{t-1}\sum_{i=1}^n p_i||\theta_l^i - \bar{\theta}_l||^2 + 4L\beta^2(\kappa-1)\sum_{l=t_\kappa+1}^{t-1}(\hat{F}_\alpha^\mu(\bar{\theta}_l) - \hat{F}_\alpha^\mu(\theta^*)) + 6\beta^2(\kappa-1)^2\rho^2 \quad . \tag{C.4}$$

Since $3L^2\beta^2(\kappa-1)^2 \leq \frac{1}{2}$ by assumption, we apply Lemma C.2 to derive the desired result. □

We now return to the proof of Theorem 1.

Proof. We begin by noting that $\bar{\theta}_{t+1} = \sum_{i=1}^n p_i(\theta_t^i - \beta \nabla \hat{f}^{\mu,i}(\theta_t^i))$ always holds by the update rule of `rFedFair`. Then we can write

$$\begin{aligned}
&||\bar{\theta}_{t+1} - \theta^*||^2 \\
&= ||\sum_{i=1}^n p_i(\theta_t^i - \beta \nabla \hat{f}^{\mu,i}(\theta_t^i)) - \theta^*||^2 \\
&= ||\bar{\theta}_t - \beta \sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\theta_t^i) - \theta^*||^2 \\
&= ||\bar{\theta}_t - \theta^*||^2 + \beta^2||\sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\theta_t^i)||^2 - 2\beta < \bar{\theta}_t - \theta^*, \sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\theta_t^i) >
\end{aligned} \quad . \tag{C.5}$$

For the term $||\sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\theta_t^i)||^2$, it can be further decomposed as

$$\begin{aligned}
&||\sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\theta_t^i)||^2 \\
&= ||\sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\theta_t^i) - \sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\bar{\theta}_t) + \sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\bar{\theta}_t)||^2 \\
&\leq 2||\sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\theta_t^i) - \sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\bar{\theta}_t)||^2 + 2||\sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\bar{\theta}_t)||^2 \\
&\leq 2\sum_{i=1}^n L^2 p_i||\theta_t^i - \bar{\theta}_t||^2 + 2||\sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\bar{\theta}_t)||^2 \\
&= 2\sum_{i=1}^n L^2 p_i||\theta_t^i - \bar{\theta}_t||^2 + 2||\nabla \hat{F}_\alpha^\mu(\bar{\theta}_t)||^2 \\
&\leq 2\sum_{i=1}^n L^2 p_i||\theta_t^i - \bar{\theta}_t||^2 + 4L(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*))
\end{aligned} \quad , \tag{C.6}$$

where the first inequality uses $||a+b||^2 \leq 2||a||^2 + 2||b||^2$, the second and last inequalities follow from the Lipschitz gradient of $\hat{f}^{\mu,i}(\theta)$ and $\hat{F}_\alpha^\mu(\theta)$ by Lemma C.1.

We also upper bound the last term as follows.

$$\begin{aligned}
&-2\beta < \bar{\theta}_t - \theta^*, \sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\theta_t^i) > \\
&= \beta \sum_{i=1}^n -2p_i < \bar{\theta}_t - \theta^*, \nabla \hat{f}^{\mu,i}(\theta_t^i) > \\
&= \beta \sum_{i=1}^n p_i[-2 < \theta_t^i - \theta^*, \nabla \hat{f}^{\mu,i}(\theta_t^i) > -2 < \bar{\theta}_t - \theta_t^i, \nabla \hat{f}^{\mu,i}(\theta_t^i) >] \\
&\leq \beta \sum_{i=1}^n p_i[2(\hat{f}^{\mu,i}(\theta^*) - \hat{f}^{\mu,i}(\theta_t^i)) - 2 < \bar{\theta}_t - \theta_t^i, \nabla \hat{f}^{\mu,i}(\theta_t^i) >] \\
&\leq \beta \sum_{i=1}^n p_i[2(\hat{f}^{\mu,i}(\theta^*) - \hat{f}^{\mu,i}(\bar{\theta}_t)) + L||\bar{\theta}_t - \theta_t^i||^2] \\
&= \beta[2(\hat{F}_\alpha^\mu(\theta^*) - \hat{F}_\alpha^\mu(\bar{\theta}_t)) + \sum_{i=1}^n p_i L||\bar{\theta}_t - \theta_t^i||^2]
\end{aligned} \quad , \tag{C.7}$$

where the first inequality uses the convexity of $\hat{f}^{\mu,i}(\theta)$, and the second inequality uses the Lipschitz gradient of $\hat{f}^{\mu,i}(\theta)$.

Plugging (C.6) and (C.7) back in (C.5) implies that

$$\begin{aligned}
&||\bar{\theta}_{t+1} - \theta^*||^2 \\
&\leq ||\bar{\theta}_t - \theta^*||^2 + (2L^2\beta^2 + \beta L)\sum_{i=1}^n p_i||\theta_t^i - \bar{\theta}_t||^2 + (4L\beta^2 - 2\beta)(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*)) \\
&\leq ||\bar{\theta}_t - \theta^*||^2 + \frac{21}{20}\beta L \sum_{i=1}^n p_i||\theta_t^i - \bar{\theta}_t||^2 - \frac{19}{10}\beta(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*))
\end{aligned} \quad , \tag{C.8}$$

where the second inequality uses the assumption that $\beta \leq \frac{1}{40L}$. Summing up all the T inequalities in (C.8) from $t = 0, 1, \cdots, T-1$ gives

$$\begin{aligned}
&||\bar{\theta}_T - \theta^*||^2 - ||\bar{\theta}_0 - \theta^*||^2 \\
&\leq \frac{21}{20}\beta L \textstyle\sum_{t=0}^{T-1}\sum_{i=1}^n p_i||\theta_t^i - \bar{\theta}_t||^2 - \frac{19}{10}\beta\sum_{t=0}^{T-1}(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*)) \\
&\leq (\frac{21}{20}\beta 8L^2\beta^2(\kappa-1)^2 - \frac{19}{10}\beta)\textstyle\sum_{t=0}^{T-1}(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*)) + \frac{21}{20}\beta L 12T\beta^2(\kappa-1)^2\rho^2 \quad , \\
&\leq -\frac{1}{2}\beta\textstyle\sum_{t=0}^{T-1}(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*)) + 13\beta TL\beta^2(\kappa-1)^2\rho^2
\end{aligned} \tag{C.9}$$

where the second inequality uses Lemma C.3, and the last inequality is due to the assumption that $6L^2\beta^2(\kappa-1)^2 \leq 1$. Rearranging the terms and dividing both sides by $\frac{1}{2}\beta T$ yield that

$$\frac{1}{T}\textstyle\sum_{t=0}^{T-1}\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*) \leq \frac{2||\bar{\theta}_0 - \theta^*||^2}{\beta T} + 26L\beta^2(\kappa-1)^2\rho^2, \ . \tag{C.10}$$

Finally, we lower bound LHS of (C.10).

$$\begin{aligned}
&\frac{1}{T}\textstyle\sum_{t=0}^{T-1}\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*) \\
&\geq \hat{F}_\alpha^\mu(\tilde{\theta}_T) - \hat{F}_\alpha^\mu(\theta^*) \\
&\geq \hat{F}_\alpha(\tilde{\theta}_T) - \hat{f}_\alpha^* - \mu\textstyle\sum_{i=1}^n \frac{p_i}{\alpha_i} \quad , \\
&\geq \hat{f}_\alpha(\tilde{\omega}_T) - \hat{f}_\alpha^* - \mu\textstyle\sum_{i=1}^n \frac{p_i}{\alpha_i}
\end{aligned} \tag{C.11}$$

where $\tilde{\theta}_T := \frac{1}{T}\sum_{t=0}^{T-1}\bar{\theta}_t$ and the first inequality uses Jensen's inequality, the second inequality follows from Lemma 2 which shows that $\hat{F}_\alpha(\tilde{\theta}_T) \leq \hat{F}_\alpha^\mu(\tilde{\theta}_T)$ and $\hat{F}_\alpha^\mu(\theta^*) \leq \hat{f}_\alpha^* + \mu\sum_{i=1}^n \frac{p_i}{\alpha_i}$, and the last inequality is by the definition of $\hat{f}_\alpha$.

Combining (C.10) and (C.11) gives us

$$\hat{f}_\alpha(\tilde{\omega}_T) - \hat{f}_\alpha^* \leq \frac{2||\bar{\theta}_0 - \theta^*||^2}{\beta T} + 26L\beta^2(\kappa-1)^2\rho^2 + \mu\sum_{i=1}^n \frac{p_i}{\alpha_i}.$$

□

D. Proof of Theorem 2

In this section, we prove the convergence of `rFedFair` with partial participation. Note that the only difference here is that at each communication round the server performs averaging step over a random selection of clients sampled with probability $p_1, p_2, \cdots, p_n$ instead of all clients. If that average does not deviate much from the average model across all clients, one may expect to use similar technique to prove the result for partial participation. Towards this end, we introduce a virtual sequence and rewrite the update rule of Algorithm 1 equivalently as:

$$\begin{aligned}
\vartheta_{t+1}^i &= \theta_t^i - \beta\nabla\hat{f}^{\mu,i}(\theta_t^i) \\
\theta_{t+1}^i &= \begin{cases} \vartheta_{t+1}^i \text{ if } t \text{ does not divide } \kappa \\ \frac{1}{K}\sum_{i\in Z_{t+1}}\vartheta_{t+1}^i \text{ otherwise} \end{cases} \quad ,
\end{aligned} \tag{D.1}$$

where $\{\vartheta_t^i\}_{t\geq 0}$ is a virtual sequence used just for the analysis. We denote by $\bar{\vartheta}_t = \sum_{i=1}^n p_i\vartheta_t^i$ the average virtual model at iteration t. The following lemma bounds how far the true average model $\bar{\theta}_t$ can derivate from the virtual average over T iterations. To simplify the notation, in what follows we simply use $\mathbb{E}[\cdot]$ to denote expectation with respect to sampling of clients at each communication round.

The proof of lemma relies on the following result.

Proposition 2. *Let $\{e_i\}_{i=1}^n$ denote any fixed deterministic sequence. We uniformly sample a subset with size K where e_i is sampled with probability p_i for $1 \leq i \leq n$ with replacement. Let $Z = \{i_1, \cdots, i_K\} \subset [n]$. Then,*

$$\mathbb{E}_Z\left[\sum_{i\in Z} e_i\right] = \mathbb{E}_Z\left[\sum_{j=1}^K e_{i_j}\right] = K\left[\sum_{i=1}^n p_i e_i\right].$$

Lemma D.1. *If the Assumption 1 and 2 hold, then*

$$\mathbb{E}\sum_{t=0}^{T-1}||\bar{\theta}_{t+1}-\bar{\vartheta}_{t+1}||^2 \leq \tfrac{3}{K}L^2\beta^2\kappa\mathbb{E}\sum_{t=0}^{T-1}\sum_{i=1}^{n}p_i||\theta_t^i-\bar{\theta}_t||^2+\tfrac{4}{K}L\beta^2\kappa\mathbb{E}\sum_{t=0}^{T-1}(\hat{F}_\alpha^\mu(\bar{\theta}_t)-\hat{F}_\alpha^\mu(\theta^*))+\tfrac{6}{K}\beta^2\kappa^2\rho^2T_N \ , \tag{D.2}$$

where $T_N := \lfloor \frac{T-1}{\kappa} \rfloor$.

Proof. First note that if t does not divide κ, we have $\bar{\theta}_{t+1}=\bar{\vartheta}_{t+1}$ by (D.1) and $||\bar{\theta}_{t+1}-\bar{\vartheta}_{t+1}||^2=0$. We can write

$$\begin{aligned}
&\mathbb{E}\sum_{t=0}^{T-1}||\bar{\theta}_{t+1}-\bar{\vartheta}_{t+1}||^2\\
&=\mathbb{E}\sum_{r=0}^{T_N}||\bar{\theta}_{r\kappa+1}-\bar{\vartheta}_{r\kappa+1}||^2\\
&=\mathbb{E}\sum_{r=0}^{T_N}||\tfrac{1}{K}\sum_{i\in Z_r}\vartheta_{r\kappa+1}^i-\bar{\vartheta}_{r\kappa+1}||^2 \ ,\\
&=\mathbb{E}\sum_{r=0}^{T_N}\tfrac{1}{K^2}\sum_{i\in Z_r}||\vartheta_{r\kappa+1}^i-\bar{\vartheta}_{r\kappa+1}||^2\\
&=\mathbb{E}\tfrac{1}{K}\sum_{r=0}^{T_N}\sum_{i=1}^{n}p_i||\vartheta_{r\kappa+1}^i-\bar{\vartheta}_{r\kappa+1}||^2
\end{aligned} \tag{D.3}$$

where the third equality is due to the independent and unbiased sampling of clients, and the last equality follows from Proposition 2.

By the update rule (D.1), for each client i, we have $\vartheta_{r\kappa+1}^i=\theta_{(r-1)\kappa+1}^i-\beta\sum_{l=(r-1)\kappa+1}^{r\kappa}\nabla\hat{f}^{\mu,i}(\theta_l^i)$. Thus, the inner summation can be further upper bounded as follows.

$$\begin{aligned}
&\sum_{i=1}^{n}p_i||\vartheta_{r\kappa+1}^i-\bar{\vartheta}_{r\kappa+1}||^2\\
&=\beta^2\sum_{i=1}^{n}p_i||\sum_{l=(r-1)\kappa+1}^{r\kappa}\nabla\hat{f}^{\mu,i}(\theta_l^i)-\sum_{l=(r-1)\kappa+1}^{r\kappa}\mathbb{E}[\nabla\hat{f}^{\mu,i}(\theta_l^i)]||^2 \ ,\\
&\leq 3L^2\beta^2\kappa\sum_{l=(r-1)\kappa+1}^{r\kappa}\sum_{i=1}^{n}p_i||\theta_l^i-\bar{\theta}_l||^2+4L\beta^2\kappa\sum_{l=(r-1)\kappa+1}^{r\kappa}(\hat{F}_\alpha^\mu(\bar{\theta}_l)-\hat{F}_\alpha^\mu(\theta^*))+6\beta^2\kappa^2\rho^2
\end{aligned} \tag{D.4}$$

where the last inequality follows from (C.2) and (C.4).

Plugging (D.4) back in (D.3) yields that

$$\begin{aligned}
&\mathbb{E}\sum_{t=0}^{T-1}||\bar{\theta}_{t+1}-\bar{\vartheta}_{t+1}||^2\\
&\leq \tfrac{3}{K}L^2\beta^2\kappa\mathbb{E}\sum_{r=0}^{T_N}\sum_{l=(r-1)\kappa+1}^{r\kappa}\sum_{i=1}^{n}p_i||\theta_l^i-\bar{\theta}_l||^2+\tfrac{4}{K}L\beta^2\kappa\mathbb{E}\sum_{r=0}^{T_N}\sum_{l=(r-1)\kappa+1}^{r\kappa}(\hat{F}_\alpha^\mu(\bar{\theta}_l)-\hat{F}_\alpha^\mu(\theta^*))+\tfrac{6}{K}\beta^2\kappa^2\rho^2T_N \ ,\\
&\leq \tfrac{3}{K}L^2\beta^2\kappa\mathbb{E}\sum_{t=0}^{T-1}\sum_{i=1}^{n}p_i||\theta_t^i-\bar{\theta}_t||^2+\tfrac{4}{K}L\beta^2\kappa\mathbb{E}\sum_{t=0}^{T-1}(\hat{F}_\alpha^\mu(\bar{\theta}_t)-\hat{F}_\alpha^\mu(\theta^*))+\tfrac{6}{K}\beta^2\kappa^2\rho^2T_N
\end{aligned} \tag{D.5}$$

which completes the proof. □

Now we proceed to prove Theorem 2, which follows similar argument to that of Theorem 1.

Proof. We begin by decomposing the optimality gap as

$$\begin{aligned}
&\mathbb{E}||\bar{\theta}_{t+1}-\theta^*||^2\\
&=\mathbb{E}||\bar{\theta}_{t+1}-\bar{\vartheta}_{t+1}+\bar{\vartheta}_{t+1}-\theta^*||^2\\
&=\mathbb{E}||\bar{\theta}_{t+1}-\bar{\vartheta}_{t+1}||^2+\mathbb{E}||\bar{\vartheta}_{t+1}-\theta^*||^2+2\mathbb{E}<\bar{\theta}_{t+1}-\bar{\vartheta}_{t+1},\bar{\vartheta}_{t+1}-\theta^*> \ ,\\
&=\mathbb{E}||\bar{\theta}_{t+1}-\bar{\vartheta}_{t+1}||^2+\mathbb{E}||\bar{\vartheta}_{t+1}-\theta^*||^2
\end{aligned} \tag{D.6}$$

where the last equality holds since $\mathbb{E}_{Z_{t+1}}\bar{\theta}_{t+1}=\bar{\vartheta}_{t+1}$.

The second term in RHS of (D.6) can be bounded as follows.

$$\begin{aligned}
&\mathbb{E}||\bar{\vartheta}_{t+1}-\theta^*||^2\\
&=\mathbb{E}||\sum_{i=1}^{n}p_i(\theta_t^i-\beta\nabla\hat{f}^{\mu,i}(\theta_t^i))-\theta^*||^2\\
&=\mathbb{E}||\bar{\theta}_t-\beta\sum_{i=1}^{n}p_i\nabla\hat{f}^{\mu,i}(\theta_t^i)-\theta^*||^2 \ ,\\
&=\mathbb{E}[||\bar{\theta}_t-\theta^*||^2+\beta^2||\sum_{i=1}^{n}p_i\nabla\hat{f}^{\mu,i}(\theta_t^i)||^2-2\beta<\bar{\theta}_t-\theta^*,\sum_{i=1}^{n}p_i\nabla\hat{f}^{\mu,i}(\theta_t^i)>]\\
&\leq \mathbb{E}||\bar{\theta}_t-\theta^*||^2+(2L^2\beta^2+\beta L)\mathbb{E}\sum_{i=1}^{n}p_i||\theta_t^i-\bar{\theta}_t||^2+(4L\beta^2-2\beta)\mathbb{E}(\hat{F}_\alpha^\mu(\bar{\theta}_t)-\hat{F}_\alpha^\mu(\theta^*))
\end{aligned} \tag{D.7}$$

where the first equality uses (D.1), and the last inequality follows from (C.6) and (C.7).

Plugging (D.7) back in (D.6) and summing up from $t = 0, 1, \cdots, T-1$ yield that

$$\begin{aligned}
&\mathbb{E}||\bar{\theta}_T - \theta^*||^2 - ||\bar{\theta}_0 - \theta^*||^2 \\
&\leq \mathbb{E}\sum_{t=0}^{T-1} ||\bar{\theta}_{t+1} - \bar{\vartheta}_{t+1}||^2 + (2L^2\beta^2 + \beta L)\mathbb{E}[\sum_{t=0}^{T-1}\sum_{i=1}^n p_i||\theta_t^i - \bar{\theta}_t||^2] + (4L\beta^2 - 2\beta)\mathbb{E}\sum_{t=0}^{T-1}(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*)) \\
&\leq (2L^2\beta^2 + \beta L + \tfrac{3}{K}L^2\beta^2\kappa)\mathbb{E}[\sum_{t=0}^{T-1}\sum_{i=1}^n p_i||\theta_t^i - \bar{\theta}_t||^2] + (4L\beta^2 + \tfrac{4}{K}L\beta^2\kappa - 2\beta)\mathbb{E}\sum_{t=0}^{T-1}(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*)) + \tfrac{6}{K}\beta^2\kappa^2\rho^2 T_N \\
&\leq \frac{21}{20}\beta L\mathbb{E}\sum_{t=0}^{T-1}\sum_{i=1}^n p_i||\theta_t^i - \bar{\theta}_t||^2 - \frac{19}{10}\beta\mathbb{E}\sum_{t=0}^{T-1}(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*)) + \tfrac{6}{K}\beta^2\kappa^2\rho^2 T_N \\
&\leq -\frac{1}{2}\beta\mathbb{E}\sum_{t=0}^{T-1}(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*)) + 13\beta T L\beta^2(\kappa-1)^2\rho^2 + \tfrac{6}{K}\beta^2\kappa^2\rho^2 T_N
\end{aligned} \tag{D.8}$$

where the second inequality uses Lemma D.1, the third inequality holds since $L\beta(3\kappa/K + 2) \leq \frac{1}{20}$ by assumption, and the last inequality follows from Lemma C.3 and the assumption that $6L^2\beta^2(\kappa-1)^2 \leq 1$.

Rearranging the terms and dividing both sides by $\frac{1}{2}\beta T$ give

$$\mathbb{E}\frac{1}{T}\sum_{t=0}^{T-1}(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*)) \leq 2\frac{||\bar{\theta}_0 - \theta^*||^2}{\beta T} + 26L\beta^2(\kappa-1)^2\rho^2 + \frac{12}{K}\beta\kappa\rho^2, \tag{D.9}$$

where we use $\frac{T_N}{T} \leq \frac{1}{\kappa}$.

Again, we apply Lemma 2 to lower bound the LHS of (D.9) and obtain

$$\mathbb{E}\frac{1}{T}\sum_{t=0}^{T-1}\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*) \geq \mathbb{E}(\hat{f}_\alpha(\tilde{\omega}_T) - \hat{f}_\alpha^*) - \mu\tau_\alpha \ . \tag{D.10}$$

Finally, combining (D.9) and (D.10) concludes the proof. □

E. Proof of Theorem 3

Proof. We begin by rewriting $f_\alpha(\omega)$ using its dual representation

$$\begin{aligned}
&f_\alpha(\omega) \\
&= \inf_{\eta\in\mathbb{R}}\left\{\eta + \mathbb{E}\left[\frac{1}{\alpha_i}(f^i(\omega) - \eta)_+\right]\right\} \\
&= \inf_{\eta\in\mathbb{R}}\left\{\eta + \frac{1}{{\tau_\alpha}^{-1}}\mathbb{E}_\alpha\left[(f^i(\omega) - \eta)_+\right]\right\}
\end{aligned} \ . \tag{E.1}$$

By choosing $\eta = 0$ in (E.1), we obtain the following inequality which holds for any $\omega \in \mathcal{W}$

$$f_\alpha(\omega) \leq \frac{1}{{\tau_\alpha}^{-1}}\mathbb{E}_\alpha\left[(f^i(\omega))\right] \leq \frac{1}{{\tau_\alpha}^{-1}}\mathbb{E}_\alpha\left[(\hat{f}^i(\omega))\right] + \Psi(S) \ , \tag{E.2}$$

where $\Psi(S) := \sup_{\omega\in\mathcal{W}}\left\{\frac{1}{{\tau_\alpha}^{-1}}\mathbb{E}_\alpha\left[(f^i(\omega))\right] - \frac{1}{{\tau_\alpha}^{-1}}\mathbb{E}_\alpha\left[(\hat{f}^i(\omega))\right]\right\}$. The first term $\mathbb{E}_\alpha[(\hat{f}^i(\omega))]$ in the RHS of (E.2) can be bounded as follows.

$$\begin{aligned}
&\mathbb{E}_\alpha\left[(\hat{f}^i(\omega))\right] = \inf_{\eta\in\mathbb{R}}\left\{\eta + \mathbb{E}_\alpha\left[(\hat{f}^i(\omega) - \eta)\right]\right\} \\
&\leq \inf_{\eta\in\mathbb{R}}\left\{\eta + \frac{1}{{\tau_\alpha}^{-1}}\mathbb{E}_\alpha\left[(\hat{f}^i(\omega) - \eta)_+\right]\right\} \\
&= \inf_{\eta\in\mathbb{R}}\left\{\eta + \mathbb{E}\left[\frac{1}{\alpha_i}(\hat{f}^i(\omega) - \eta)_+\right]\right\} \\
&= \hat{f}_\alpha(\omega)
\end{aligned} \ , \tag{E.3}$$

where the first inequality uses $(\cdot) \leq \frac{1}{{\tau_\alpha}^{-1}}(\cdot)_+$ as $\tau_\alpha \geq 1$.

To bound the second term, we make use of McDiarmid's inequality. Let S' be a sample differing from S by exactly one

point, say (x_j^i, y_j^i) in S and (x'^i_j, y'^i_j) in S'. By definition of $\Psi(S)$, the following inequality holds:

$$\begin{aligned}
&\Psi(S) - \Psi(S') \\
&= \sup_{\omega\in\mathcal{W}} \left\{ \frac{1}{{\tau_\alpha}^{-1}} \mathbb{E}_\alpha \left[(f^i(\omega)) \right] - \frac{1}{{\tau_\alpha}^{-1}} \mathbb{E}_\alpha \left[(\hat{f}^i(\omega)) \right] \right\} - \sup_{\omega\in\mathcal{W}} \left\{ \frac{1}{{\tau_\alpha}^{-1}} \mathbb{E}_\alpha \left[(f^i(\omega)) \right] - \frac{1}{{\tau_\alpha}^{-1}} \mathbb{E}_\alpha \left[(\hat{f'}^i(\omega)) \right] \right\} \\
&= \sup_{\omega\in\mathcal{W}} \left\{ \mathbb{E}\left[\frac{1}{\alpha_i} f^i(\omega) \right] - \mathbb{E}\left[\frac{1}{\alpha_i} \hat{f}^i(\omega) \right] \right\} - \sup_{\omega\in\mathcal{W}} \left\{ \mathbb{E}\left[\frac{1}{\alpha_i} f^i(\omega) \right] - \mathbb{E}\left[\frac{1}{\alpha_i} \hat{f'}^i(\omega) \right] \right\} \\
&\le \sup_{\omega\in\mathcal{W}} \left\{ \mathbb{E}\left[\frac{1}{\alpha_i} \hat{f'}^i(\omega) \right] - \mathbb{E}\left[\frac{1}{\alpha_i} \hat{f}^i(\omega) \right] \right\} \\
&= \sup_{\omega\in\mathcal{W}} \left\{ \left[\frac{p_i}{\alpha_i} (\hat{f'}^i(\omega) - \hat{f}^i(\omega)) \right] \right\} \\
&= \sup_{\omega\in\mathcal{W}} \left\{ \left[\frac{p_i}{\alpha_i m_i} (l(f_\omega(x'^i_j), y'^i_j) - l(f_\omega(x^i_j), y^i_j)) \right] \right\} \\
&\le \frac{p_i B}{\alpha_i m_i}
\end{aligned} \tag{E.4}$$

where the first inequality uses the sub-additivity of sup, and the last inequality is due to the boundness assumption on the loss function.

By McDiarmid's inequality, with probability at least $1-\delta$, the following inequality holds

$$\Psi(S) \le \mathbb{E}_S \Psi(S) + B \sqrt{\sum_{i=1}^n \frac{p_i^2 \log(\frac{1}{\delta})}{2\alpha_i^2 m_i}}. \tag{E.5}$$

The expectation on RHS of (E.5) can be further bounded as follows.

$$\begin{aligned}
&\mathbb{E}_S \Psi(S) \\
&= \mathbb{E}_S \sup_{\omega\in\mathcal{W}} \left\{ \mathbb{E}\left[\frac{1}{\alpha_i} f^i(\omega) \right] - \mathbb{E}\left[\frac{1}{\alpha_i} \hat{f}^i(\omega) \right] \right\} \\
&= \mathbb{E}_S \sup_{\omega\in\mathcal{W}} \left\{ \sum_{i=1}^n \left[\frac{p_i}{\alpha_i} f^i(\omega) \right] - \sum_{n=1}^n \left[\frac{p_i}{\alpha_i} \hat{f}^i(\omega) \right] \right\} \\
&= \mathbb{E}_S \sup_{\omega\in\mathcal{W}} \left\{ \mathbb{E}_{S'} \sum_{i=1}^n \left[\frac{p_i}{\alpha_i} \hat{f'}^i(\omega) \right] - \sum_{n=1}^n \left[\frac{p_i}{\alpha_i} \hat{f}^i(\omega) \right] \right\} \\
&\le \mathbb{E}_{S,S'} \sup_{\omega\in\mathcal{W}} \left\{ \sum_{i=1}^n \left[\frac{p_i}{\alpha_i} (\hat{f'}^i(\omega) - \hat{f}^i(\omega)) \right] \right\} \\
&= \mathbb{E}_{S,S',\sigma} \sup_{\omega\in\mathcal{W}} \left\{ \sum_{i=1}^n \sum_{j=1}^{m_i} \left[\frac{p_i}{\alpha_i m_i} \sigma_{ij} (l(f_\omega(x'^i_j), y'^i_j) - l(f_\omega(x^i_j), y^i_j)) \right] \right\} \\
&\le 2\mathbb{E}_{S,\sigma} \sup_{\omega\in\mathcal{W}} \left\{ \sum_{i=1}^n \sum_{j=1}^{m_i} \left[\frac{p_i}{\alpha_i m_i} \sigma_{ij} l(f_\omega(x^i_j), y^i_j) \right] \right\}
\end{aligned} \tag{E.6}$$

where the first inequality uses Jensen's inequality and the convexity of the supremum function, the last equality follows from the fact that the introduction of Rademacher variables does not change the expectation over all possible S and S', and the last inequality holds by the sub-additivity of sup and the fact that σ_{ij} and $-\sigma_{ij}$ have the same distribution.

Plugging (E.6), (E.5) and (E.3) into (E.2) concludes the proof of the theorem. □

F. Proof of Theorem 4

Proof. For any $\omega \in \mathcal{W}$, we have

$$\begin{aligned}
&f_\alpha(\omega) - \hat{f}_\alpha(\omega) \\
&= \inf_{\eta\in\mathbb{R}} \left\{ \eta + \mathbb{E}\left[\frac{1}{\alpha_i} (f^i(\omega) - \eta)_+ \right] \right\} - \inf_{\eta\in\mathbb{R}} \left\{ \eta + \mathbb{E}\left[\frac{1}{\alpha_i} (\hat{f}^i(\omega) - \eta)_+ \right] \right\} \\
&\le \mathbb{E}\left[\frac{1}{\alpha_i} (f^i(\omega) - \eta)_+ - \frac{1}{\alpha_i} (\hat{f}^i(\omega) - \eta)_+ \right] \\
&\le \mathbb{E}\left[\frac{1}{\alpha_i} |f^i(\omega) - \hat{f}^i(\omega)| \right]
\end{aligned} \tag{F.1}$$

where the first equality uses the variational representation of Q_α-weighted loss given in Lemma 1, the first inequality holds by selecting the first η to be identical to the second η. Taking supremum over $\mathcal{W}$, we get

$$\begin{aligned} &\sup_{\omega\in\mathcal{W}}\{f_\alpha(\omega) - \hat{f}_\alpha(\omega)\} \\ &\leq \sup_{\omega\in\mathcal{W}}\{\textstyle\sum_{i=1}^n \frac{p_i}{\alpha_i}|f^i(\omega) - \hat{f}^i(\omega)|\} \ , \\ &\leq \textstyle\sum_{i=1}^n \frac{p_i}{\alpha_i} \sup_{\omega\in\mathcal{W}} |f^i(\omega) - \hat{f}^i(\omega)| \end{aligned} \tag{F.2}$$

where the last inequality uses the sub-additivity of sup.

For a fixed m_i, by a standard Rademacher complexity bound (Mohri et al., 2018), for any $\delta > 0$, with probability at least $1 - \frac{\delta}{n}$, the following inequality holds

$$\sup_{\omega\in\mathcal{W}} |f^i(\omega) - \hat{f}^i(\omega)| \leq 2\mathfrak{R}^i_{m_i}(\mathcal{H}) + B\sqrt{\frac{1}{2m_i}\log\frac{2n}{\delta}}.$$

Plugging the above inequality back in (F.2) for each i and using a union bound yields that for every $\omega \in \mathcal{W}$,

$$f_\alpha(\omega) \leq \hat{f}_\alpha(\omega) + 2\textstyle\sum_{i=1}^n \frac{p_i}{\alpha_i}\mathfrak{R}^i_{m_i}(\mathcal{H}) + \sum_{i=1}^n \frac{p_i}{\alpha_i} B\sqrt{\frac{1}{2m_i}\log\frac{2n}{\delta}} \tag{F.3}$$

with probability at least $1 - \delta$. This completes the proof. □