A Supplementary Materials

All our proofs follow the framework of Dai et al (2014).

Lemma 5.2

Proof: Denote $V_i(x) = V_i(x;\mathcal{D}^i, r^i, \omega^i) := [v_t^i \zeta_{r_i,i}(x) - v_t^i \xi_{r_i,i}(x)]$. We have

$$|V_i(x)| \le |v_t^i|\phi M + |\hat{v}_t^i|\kappa M =$$

$$|a_t^i|[\frac{\phi M}{\psi_i} + \frac{\kappa M}{\hat{\psi}_i}] \le |a_t^i|(\kappa+\phi)M$$

The last inequality is because $\psi_i \ge 1, \hat{\psi}_i \ge 1$. The rest of the proofs can follow Dai et al (2014) Lemma 7. □

Lemma 5.3

Proof: Now define:

$$g_t := l'(F_t(x_t), y_t)k_{r_t}(x_t,\cdot) + \nu\hat{\psi}_t h_{r_t,t} \tag{32}$$

$$\hat{g}_t := \nabla R_t(h_{r_t}) = \hat{\xi}_{r_t,t} + \nu\hat{\psi}_t h_{r_t,t}$$

$$= l'(H_t(x_t), y_t)k_{r_t}(x_t,\cdot) + \nu\hat{\psi}_t h_{r_t,t} \tag{33}$$

$$\bar{g}_t := \mathbb{E}_{D_t}[\nabla R_t(h_{r_t})] = \mathbb{E}_{D_t}[\hat{g}_t]$$

$$= \mathbb{E}_{D_t}[l'(H_t(x_t), y_t)k_{r_t}(x_t,\cdot)] + \nu\hat{\psi}_t h_{r_t,t} \tag{34}$$

Denote $A_t = ||H_t - F_*||^2_{\mathcal{H}}$. We have $A_{t+1} =$

$$||H_t - F_* - \eta_t g_t||^2_{\mathcal{H}} = A_t + \eta_t^2 ||g_t||^2_{\mathcal{H}} - 2\eta_t\langle H_t - F_*, g_t\rangle_{\mathcal{H}}$$

$$= A_t + \eta_t^2||g_t||^2_{\mathcal{H}} - 2\eta_t\langle H_t - F_*, \bar{g}_t\rangle_{\mathcal{H}}$$

$$+2\eta_t\langle H_t - F_*, \bar{g}_t - \hat{g}_t\rangle_{\mathcal{H}} + 2\eta_t\langle H_t - F_*, \hat{g}_t - g_t\rangle_{\mathcal{H}} \tag{35}$$

Since (11) is $\frac{\nu(p-1)}{2p}$-strongly convex, the optimality condition gives:

$$\langle H_t - F_*, \bar{g}_t\rangle_{\mathcal{H}} \ge \frac{\nu(p-1)}{2p}||H_t - F_*||^2_{\mathcal{H}} \tag{36}$$

thus, $A_{t+1} \le$

$$(1 - \frac{\eta_t\nu(p-1)}{p})A_t + \eta_t^2||g_t||^2_{\mathcal{H}} - 2\eta_t\langle H_t - F_*, \bar{g}_t\rangle_{\mathcal{H}}$$

$$+2\eta_t\langle H_t - F_*, \bar{g}_t - \hat{g}_t\rangle_{\mathcal{H}} + 2\eta_t\langle H_t - F_*, \hat{g}_t - g_t\rangle_{\mathcal{H}} \tag{37}$$

denote

$$\mathcal{M}_t = ||g_t||^2_{\mathcal{H}} \tag{38}$$

$$\mathcal{N}_t = \langle H_t - F_*, \bar{g}_t - \hat{g}_t\rangle_{\mathcal{H}} \tag{39}$$

$$\mathcal{R}_t = \langle H_t - F_*, \hat{g}_t - g_t\rangle_{\mathcal{H}} \tag{40}$$

which will be separately bounded in the Lemma A.1. Denote $e_t = \mathbb{E}_{\mathcal{D}^{t-1}, r^{t-1}, \omega^{t-1}}[A_t]$, using the bounds derived in Lemma A.1, we have

$$e_{t+1} \le (1 - \frac{\eta_t\nu(p-1)}{p})e_t + \kappa M^2\gamma_t^2(1+\nu c_t)^2$$

$$+2\gamma_t\kappa^{1/2}LB_{1,t}\sqrt{e_t} \tag{41}$$

Applying $|a_t^i| \le \frac{\theta}{t}$ (Lemma 8(1) in Dai et al (2014)). We have $c_t =$

$$\sqrt{\sum_{i,j\in I^{t-1}(r_t)} |\hat{v}_{t-1}^i||\hat{v}_{t-1}^j|} = \sqrt{\sum_{i,j\in I^{t-1}(r_t)} \frac{|a_{t-1}^i||a_{t-1}^j|}{\hat{\psi}_i\hat{\psi}_j}}$$

$$\le \sqrt{(\frac{\theta}{t-1})^2 \sum_{i,j\in I^{t-1}(r_t)} \frac{1}{\hat{\psi}_i\hat{\psi}_j}}$$

$$\le \sqrt{(\frac{\theta}{t-1})^2 \cdot [C^{t-1}(r_t)]^2} = \theta\frac{C^{t-1}(r_t)}{t-1} \le \theta \tag{42}$$

where we have used the fact that $\hat{\psi}_i \ge 1$, and $\sum_{i\in I^{t-1}(r_t)} 1 = C^{t-1}(r_t)$, i.e., the number of iterations up to $t-1$ that kernel r_t gets selected.

From Lemma 5.2 and $\sum_{i=1}^t |a_t^i|^2 \le \frac{\theta^2}{t}$ (Lemma 8(1) in Dai et al (2014)), we have

$$B_{1,t}^2 \le 4M^2(\kappa+\phi)^2\frac{\theta^2}{t-1} \tag{43}$$

Plug it into (37) together with

$$\frac{\gamma_t}{C_q} \le \eta_t \le \gamma_t = \frac{\theta}{t} \tag{44}$$

leads to

$$e_{t+1} \le (1 - \frac{2\nu'\theta}{t})e_t + \kappa M^2(\frac{\theta}{t})^2(1+\nu\theta)^2$$

$$+2\kappa^{1/2}L\frac{\theta}{t}\sqrt{4M^2(\kappa+\phi)^2\frac{\theta^2}{t-1}}\sqrt{e_t} \tag{45}$$

where $\nu' = \frac{\nu(p-1)}{2pC_q}$ further rewritten as

$$e_{t+1} \le (1 - 2\frac{\nu'\theta}{t})e_t + \frac{\beta_2}{t^2} + \frac{\beta_1}{t}\sqrt{\frac{e_t}{t}} \tag{46}$$

with $\beta_1 = 4\sqrt{2}\kappa^{1/2}LM(\kappa+\phi)\theta^2$, $\beta_2 = \kappa M^2(1+\nu\theta)^2\theta^2$.

Using Lemma 14 of Dai et al (2014), we have

$$e_t \le \frac{R}{t} \tag{47}$$

where $R = \max\{||F_*||_{\mathcal{H}}, R_0^2\}$,

$$R_0 = \frac{Q_0 + \sqrt{Q_0^2 + (2\nu'\theta - 1)\kappa M^2 (1+\nu\theta)^2(\theta)^2}}{(2\nu'\theta - 1)}$$

where $Q_0 = 2\sqrt{2}\kappa^{1/2} LM(\kappa+\phi)\theta^2$. □

The following is an auxiliary lemma used in the above proof, which follows the logic flow of Lemma 10 in Dai et al (2014).

Lemma A.1 *(1)* $\eta_t^2 \mathcal{M}_t \leq \kappa M^2 (1+\nu c_t)^2$ *where* $c_t = \sqrt{\sum_{i,j\in I^{t-1}(r_t)} |\hat{v}_{t-1}^i||\hat{v}_{t-1}^j|}$ *for* $t \geq 2$ *and* $\forall r, c_1(r) = 0$*;*
(2) $\mathbb{E}_{\mathcal{D}^t, r^t, \omega^t}[\eta_t \mathcal{N}_t] \leq 0$*;*
(3) $\mathbb{E}_{\mathcal{D}^t, r^t, \omega^t}[\eta_t \mathcal{R}_t] \leq \kappa^{1/2} L B_{1,t} \sqrt{\mathbb{E}_{\mathcal{D}^{t-1}, \omega^{t-1}}[A_t]}$*, where* $B_{1,t}^2 := 4M^2(\kappa+\phi)^2 \sum_{i=1}^{t-1} |\hat{a}_t^i|^2$ *for* $t \geq 2$ *and* $B_{1,1} = 0$*;*

Proof: (1) $\mathcal{M}_t = ||g_t||_{\mathcal{H}}^2 = ||\xi_{r_t,t} + \nu\hat{\psi}_t h_{r_t,t}||_{\mathcal{H}}^2$

$$\leq (||\xi_{r_t,t}||_{\mathcal{H}} + \nu\hat{\psi}_t ||h_{r_t,t}||_{\mathcal{H}})^2$$

where

$$||\xi_{r_t,t}||_{\mathcal{H}} = ||l'(F_t(x_t), y_t) k_{r_t}(x_t, \cdot)||_{\mathcal{H}} \leq \kappa^{1/2} M$$

and $||h_{r_t,t}||_{\mathcal{H}}^2$ equals

$$\langle \sum_{i\in I^{t-1}(r_t)} \hat{u}_{t-1}^i(r_t)\xi_{r_t,i}(\cdot), \sum_{j\in I^{t-1}(r_t)} \hat{u}_{t-1}^j(r_t)\xi_{r_t,i}(\cdot)\rangle_{\mathcal{H}}$$

$$= \sum_{i\in I^{t-1}(r_t)} \sum_{j\in I^{t-1}(r_t)} \Big[\hat{u}_{t-1}^i(r_t)\hat{u}_{t-1}^j(r_t)\cdot$$

$$l'(\sum_{s=1}^m f_s(x_i), y_i) l'(\sum_{s=1}^m f_s(x_j), y_j) k_{r_t}(x_i, x_j)\Big]$$

$$\leq \kappa M^2 \sum_{i\in I^{t-1}(r_t)} \sum_{j\in I^{t-1}(r_t)} |\hat{u}_{t-1}^i(r_t)| \cdot |\hat{u}_{t-1}^j(r_t)|$$

$$= \kappa M^2 \sum_{i\in I^{t-1}(r_t)} \sum_{j\in I^{t-1}(r_t)} |\hat{v}_{t-1}^i||\hat{v}_{t-1}^j|$$

thus

$$\eta_t^2 \mathcal{M}_t \leq \kappa M^2 \eta_t^2 (1+\nu\hat{\psi}_t c_t)^2$$
$$= \kappa M^2 (\eta_t + \nu\gamma_t c_t)^2 \leq \gamma_t^2 \kappa M^2 (1+\nu c_t)$$

(2) because $\mathcal{N}_t = \langle H_t - F_*, \bar{g}_t - \hat{g}_t\rangle_{\mathcal{H}}$, thus

$$\mathbb{E}_{\mathcal{D}^t, r^t, \omega^t}[\eta_t \mathcal{N}_t] \leq \mathbb{E}_{\mathcal{D}^t, r^t, \omega^t}[\gamma_t \mathcal{N}_t]$$

$$\leq \gamma_t \cdot \mathbb{E}_{\mathcal{D}^{t-1}, r^t, \omega^t}\left[\mathbb{E}_{\mathcal{D}^t}[\langle H_t - F_*, \bar{g}_t - \hat{g}_t\rangle_{\mathcal{H}} | \mathcal{D}^{t-1}, r^t, \omega^t]\right]$$
$$= \gamma_t \cdot \mathbb{E}_{\mathcal{D}^{t-1}, r^t, \omega^t}\left[\langle H_t - F_*, \mathbb{E}_{\mathcal{D}^t}[\bar{g}_t - \hat{g}_t]\rangle_{\mathcal{H}} | \mathcal{D}^{t-1}, r^t, \omega^t\right]$$
$$= 0$$

(3) because $\mathcal{R}_t = \langle H_t - F_*, \hat{g}_t - g_t\rangle_{\mathcal{H}}$, then

$$\mathbb{E}_{\mathcal{D}^t, r^t, \omega^t}\left[\mathcal{R}_t\right] = \mathbb{E}_{\mathcal{D}^t, r^t, \omega^t}\left[\langle H_t - F_*, \hat{g}_t - g_t\rangle_{\mathcal{H}}\right]$$
$$= \mathbb{E}_{\mathcal{D}^t, r^t, \omega^t}\big[\langle H_t - F_*, [l'(F_t(x_t), y_t)$$
$$- l'(H_t(x_t), y_t)] k_{r_t}(x_t, \cdot)\rangle_{\mathcal{H}}\big]$$
$$\leq \mathbb{E}_{\mathcal{D}^t, r^t, \omega^t}\big[|l'(F_t(x_t), y_t) - l'(H_t(x_t), y_t)|$$
$$\cdot ||k_{r_t}(x_t, \cdot)||_{\mathcal{H}} \cdot ||H_t - F_*||_{\mathcal{H}}\big]$$
$$\leq \kappa^{1/2} L \cdot \mathbb{E}_{\mathcal{D}^t, r^t, \omega^t}\left[|F_t(x_t) - H_t(x_t)| \cdot ||H_t - F_*||_{\mathcal{H}}\right]$$
$$\leq \kappa^{1/2} L \sqrt{\mathbb{E}_{\mathcal{D}^t, r^t, \omega^t}|F_t(x_t) - H_t(x_t)|^2}$$
$$\cdot \sqrt{\mathbb{E}_{\mathcal{D}^t, r^t, \omega^t}||H_t - F_*||_{\mathcal{H}}^2}$$
$$= \kappa^{1/2} L B_{1,t} \sqrt{\mathbb{E}_{\mathcal{D}^{t-1}, r^{t-1}, \omega^{t-1}}||A_t||_{\mathcal{H}}^2}$$

1st and 3rd inequalities: Cauchy-Schwarz;
2nd inequality: L-Lipschitz of $l'(\cdot,\cdot)$ and $||k_{r_t}(x_t,\cdot)||_{\mathcal{H}} = ||k_{r_t}(x_t,\cdot)||_{\mathcal{H}_r} \leq \kappa$
last step: due to Lemma 5.2 and definition of A_t. Thus

$$\mathbb{E}_{\mathcal{D}^t, r^t, \omega^t}[\eta_t \mathcal{R}_t] \leq \gamma_t \mathbb{E}_{\mathcal{D}^t, r^t, \omega^t}[\mathcal{R}_t]$$

$$\leq \gamma_t \kappa^{1/2} L B_{1,t} \sqrt{\mathbb{E}_{\mathcal{D}^{t-1}, r^{t-1}, \omega^{t-1}}||A_t||_{\mathcal{H}}^2} \qquad (48)$$

□