Supplementary: Optimal Distributed Learning with Multi-pass Stochastic Gradient Methods

In this appendix, we provide the proofs of our main theorems for distributed SGM. We begin with some basic notations.

A. Notations

We introduce the inclusion operator $\mathcal{S}_\rho : H \to L^2_{\rho_X}$, which is continuous under Assumption (8). Furthermore, we consider the adjoint operator $\mathcal{S}^*_\rho : L^2_{\rho_X} \to H$, the covariance operator $\mathcal{T} : H \to H$ given by $\mathcal{T} = \mathcal{S}^*_\rho \mathcal{S}_\rho$, and the operator $\mathcal{L} : L^2_{\rho_X} \to L^2_{\rho_X}$ given by $\mathcal{S}_\rho \mathcal{S}^*_\rho$. It can be easily proved that $\mathcal{S}^*_\rho f = \int_X K_x f(x) d\rho_X(x)$ and $\mathcal{T} = \int_X \langle \cdot, K_x \rangle_H K_x d\rho_X(x)$. The operators $\mathcal{T}$ and $\mathcal{L}$ can be proved to be positive trace class operators (and hence compact). In fact, by Assumption (8),

$$\|\mathcal{L}\| = \|\mathcal{T}\| \le \mathrm{tr}(\mathcal{T}) = \int_X \mathrm{tr}(K_x \otimes K_x) d\rho_X(x) = \int_X \|K_x\|^2_H d\rho_X(x) \le \kappa^2. \tag{24}$$

For any function $f \in H$, the H-norm can be related to the $L^2_{\rho_X}$-norm by $\sqrt{\mathcal{T}}$ (Bauer et al., 2007):

$$\|f\|_\rho = \|\mathcal{S}_\rho f\|_\rho = \left\| \sqrt{\mathcal{T}} f \right\|_H, \tag{25}$$

and furthermore

$$\|\mathcal{L}^{-\frac{1}{2}} \mathcal{S}_\rho f\|_\rho \le \|f\|_H. \tag{26}$$

We define the sampling operator (with respect to any given set $\mathbf{x} \subseteq X$ of cardinality n) $\mathcal{S}_\mathbf{x} : H \to \mathbb{R}^n$ by $(\mathcal{S}_\mathbf{x} f)_i = f(x_i) = \langle f, K_{x_i} \rangle_H$, $i \in [n]$, where the norm $\|\cdot\|_{\mathbb{R}^n}$ is the standard Euclidean norm times $1/\sqrt{n}$. Its adjoint operator $\mathcal{S}^*_\mathbf{x} : \mathbb{R}^n \to H$, defined by $\langle \mathcal{S}^*_\mathbf{x} \mathbf{y}, f \rangle_H = \langle \mathbf{y}, \mathcal{S}_\mathbf{x} f \rangle_{\mathbb{R}^n}$ for $\mathbf{y} \in \mathbb{R}^n$ is thus given by

$$\mathcal{S}^*_\mathbf{x} \mathbf{y} = \frac{1}{n} \sum_{i=1}^n y_i K_{x_i}. \tag{27}$$

Moreover, we can define the empirical covariance operator (with respect to $\mathbf{x}$) $\mathcal{T}_\mathbf{x} : H \to H$ such that $\mathcal{T}_\mathbf{x} = \mathcal{S}^*_\mathbf{x} \mathcal{S}_\mathbf{x}$. Obviously,

$$\mathcal{T}_\mathbf{x} = \frac{1}{n} \sum_{i=1}^n \langle \cdot, K_{x_i} \rangle_H K_{x_i}.$$

By Assumption (8), similar to (24), we have

$$\|\mathcal{T}_\mathbf{x}\| \le \mathrm{tr}(\mathcal{T}_\mathbf{x}) \le \kappa^2. \tag{28}$$

For any given inputs set $\mathbf{x} \subseteq X$, $\mathcal{L}_\mathbf{x} : L^2_{\rho_X} \to H$ is defined as that for any $f \in L^2_{\rho_X}$ such that $\|f\|_\infty < \infty$,

$$\mathcal{L}_\mathbf{x} f = \frac{1}{|\mathbf{x}|} \sum_{x \in \mathbf{x}} f(x) K_x. \tag{29}$$

For any $\tilde{\lambda} > 0$, for notational simplicity, we let $\mathcal{T}_{\tilde{\lambda}} = \mathcal{T} + \tilde{\lambda}$, $\mathcal{T}_{\mathbf{x}\tilde{\lambda}} = \mathcal{T}_\mathbf{x} + \tilde{\lambda}$, and

$$\mathcal{N}(\tilde{\lambda}) = \mathrm{tr}(\mathcal{L}(\mathcal{L} + \tilde{\lambda})^{-1}) = \mathrm{tr}(\mathcal{T}(\mathcal{T} + \tilde{\lambda})^{-1}).$$

For any $f \in H$ and $x \in X$, the following well known reproducing property holds:

$$\langle f, K_x \rangle_H = f(x). \tag{30}$$

and following from the above, Cauchy-Schwarz inequality and (8), one can prove that

$$|f(x)| = |\langle f, K_x \rangle_H| \le \|f\|_H \|K_x\|_H \le \kappa \|f\|_H \tag{31}$$

$\mathbb{E}[\xi]$ denotes the expectation of a random variable ξ. $\|\cdot\|_\infty$ denotes the supreme norm with respect to ρ_X. For a given bounded operator $L : H \to H'$, $\|L\|$ denotes the operator norm of L, i.e., $\|L\| = \sup_{f \in H, \|f\|_H = 1} \|Lf\|_{H'}$. Here H' could be another separable Hilbert space different from H.

For any $s \in [m]$, we denote the set of random variables $\{j_{s,i}\}_{b(t-1)+1 \le i \le bt}$ by $\mathbf{J}_{s,t}$, $\{j_{s,1}, j_{s,2}, \cdots, j_{s,bT}\}$ by $\mathbf{J}_s$, and $\{\mathbf{J}_1, \cdots, \mathbf{J}_m\}$ by $\mathbf{J}$. Note that $j_{s,1}, j_{s,2}, \cdots, j_{s,bT}$ are conditionally independent given $\mathbf{z}_s$.

B. Proof for Error Decomposition

Proof of Proposition 1. For any $s \in [m]$, using an inductive argument, one can prove that (Lin & Rosasco, 2017b)

$$\mathbb{E}_{\mathbf{J}_s|\mathbf{z}_s}[f_{s,t}] = g_{s,t}. \tag{32}$$

Here $\mathbb{E}_{\mathbf{J}_s|\mathbf{z}_s}$ (or abbreviated as $\mathbb{E}_{\mathbf{J}_s}$) denotes the conditional expectation with respect to $\mathbf{J}_s$ given $\mathbf{z}_s$. Indeed, taking the conditional expectation with respect to $\mathbf{J}_{s,t}$ (given $\mathbf{z}_s$) on both sides of (4), and noting that $f_{s,t}$ depends only on $\mathbf{J}_{s,1}, \cdots, \mathbf{J}_{s,t-1}$ (given $\mathbf{z}_s$), one has

$$\mathbb{E}_{\mathbf{J}_{s,t}}[f_{s,t+1}] = f_{s,t} - \eta_t \frac{1}{n}\sum_{i=1}^{n}(f_{s,t}(x_{s,i}) - y_{s,i})K_{x_{s,i}},$$

and thus,

$$\mathbb{E}_{\mathbf{J}_s}[f_{s,t+1}] = \mathbb{E}_{\mathbf{J}_s}[f_{s,t}] - \eta_t \frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}_{\mathbf{J}_s}[f_{s,t}](x_{s,i}) - y_{s,i})K_{x_{s,i}}, \quad t = 1, \ldots, T,$$

which satisfies the iterative relationship given in (17). Similarly, using the definition of the regression function (2) and an inductive argument, one can also prove that

$$\mathbb{E}_{\mathbf{y}_s}[g_{s,t}] = h_{s,t}. \tag{33}$$

Here, $\mathbb{E}_{\mathbf{y}_s}$ denotes the conditional expectation with respect to $\mathbf{y}_s$ given $\mathbf{x}_s$.

Following from (3), we have

$$\mathcal{E}(\bar{f}_t) - \mathcal{E}(f_\rho) = \|\mathcal{S}_\rho \bar{f}_t - f_\rho\|_\rho^2 = \|\mathcal{S}_\rho \bar{f}_t - \mathcal{S}_\rho \bar{g}_t\|_\rho^2 + \|\mathcal{S}_\rho \bar{g}_t - f_\rho\|_\rho^2 + 2\langle \mathcal{S}_\rho \bar{f}_t - \mathcal{S}_\rho \bar{g}_t, \mathcal{S}_\rho \bar{g}_t - f_\rho\rangle.$$

Taking the conditional expectation with respect to $\mathbf{J}$ (given $\mathbf{z}$) on both sides, using (32) which implies

$$\mathbb{E}_{\mathbf{J}}\mathcal{S}_\rho(\bar{f}_t - \bar{g}_t) = \frac{1}{m}\sum_{s=1}^{m}\mathcal{S}_\rho \mathbb{E}_{\mathbf{J}_s}[f_{s,t} - g_{s,t}] = 0,$$

we thus have

$$\mathbb{E}_{\mathbf{J}}\|\mathcal{S}_\rho \bar{f}_t - f_\rho\|_\rho^2 = \mathbb{E}_{\mathbf{J}}\|\mathcal{S}_\rho \bar{f}_t - \mathcal{S}_\rho \bar{g}_t\|_\rho^2 + \|\mathcal{S}_\rho \bar{g}_t - f_\rho\|_\rho^2.$$

Taking the conditional expectation with respect to $\bar{\mathbf{y}} = \{\mathbf{y}_1, \cdots, \mathbf{y}_m\}$ (given $\bar{\mathbf{x}} = \{\mathbf{x}_1, \cdots, \mathbf{x}_m\}$), noting that

$$\mathbb{E}_{\bar{\mathbf{y}}}\|\mathcal{S}_\rho \bar{g}_t - f_\rho\|_\rho^2 = \mathbb{E}_{\bar{\mathbf{y}}}[\|\mathcal{S}_\rho(\bar{g}_t - \bar{h}_t)\|_\rho^2] + \|\mathcal{S}_\rho \bar{h}_t - f_\rho\|_\rho^2 + 2\langle \mathcal{S}_\rho \mathbb{E}_{\bar{\mathbf{y}}}[\bar{g}_t - \bar{h}_t], \mathcal{S}_\rho \bar{h}_t - f_\rho\rangle_\rho$$

and that from (33),

$$\langle \mathcal{S}_\rho \mathbb{E}_{\bar{\mathbf{y}}}[\bar{g}_t - \bar{h}_t], \mathcal{S}_\rho \bar{h}_t - f_\rho\rangle_\rho = \frac{1}{m}\sum_{s=1}^{m}\langle \mathcal{S}_\rho \mathbb{E}_{\mathbf{y}_s}(g_{s,t} - h_{s,t}), \mathcal{S}_\rho \bar{h}_t - f_\rho\rangle_\rho = 0,$$

we know that

$$\mathbb{E}_{\bar{\mathbf{y}}}\mathbb{E}_{\mathbf{J}}\mathcal{E}(\bar{f}_t) - \mathcal{E}(f_\rho) = \mathbb{E}_{\bar{\mathbf{y}}}\mathbb{E}_{\mathbf{J}}\|\mathcal{S}_\rho \bar{f}_t - \mathcal{S}_\rho \bar{g}_t\|_\rho^2 + \mathbb{E}_{\bar{\mathbf{y}}}[\|\mathcal{S}_\rho(\bar{g}_t - \bar{h}_t)\|_\rho^2] + \|\mathcal{S}_\rho \bar{h}_t - f_\rho\|_\rho^2,$$

which leads to the desired result. □

C. Estimating Bias

In this section, we estimate bias, i.e., $\mathbb{E}\|\mathcal{S}_\rho \bar{h}_t - f_\rho\|_\rho^2$. We first give the following lemma, which asserts that the bias term can be estimated in terms of the bias of a local estimator.

Lemma 1. *For any $t \in [T]$, we have*

$$\mathbb{E}\|\mathcal{S}_\rho \bar{h}_t - f_\rho\|_\rho^2 \le \mathbb{E}\|\mathcal{S}_\rho h_{1,t} - f_\rho\|_\rho^2.$$

Proof. By Hölder's inequality, we know that

$$\mathbb{E}\|\mathcal{S}_\rho \bar{h}_t - f_\rho\|_\rho^2 = \frac{1}{m^2}\mathbb{E}\left\|\sum_{s=1}^{m}(\mathcal{S}_\rho h_{s,t} - f_\rho)\right\|_\rho^2 \le \frac{1}{m}\mathbb{E}\sum_{s=1}^{m}\|\mathcal{S}_\rho h_{s,t} - f_\rho\|_\rho^2 = \mathbb{E}\|\mathcal{S}_\rho h_{1,t} - f_\rho\|_\rho^2.$$

□

Given the above lemma, in what follows, we will estimate the bias of the local estimator, $\mathbb{E}\|\mathcal{S}_\rho h_{1,t} - f_\rho\|_\rho^2$. To do so, we need to introduce some preliminary notations and lemmas.

$\Pi_{t+1}^T(L) = \prod_{k=t+1}^T (I - \eta_k L)$ for $t \in [T-1]$ and $\Pi_{T+1}^T(L) = I$, for any operator $L : H \to H$, where H is a Hilbert space and I denotes the identity operator on H. Let $k, t \in \mathbb{N}$. We will use the following conventional notations: $1/0 = +\infty$, $\prod_k^t = 1$ and $\sum_k^t = 0$ whenever $k > t$. $\Sigma_k^t = \sum_{i=k}^t \eta_i$, $\lambda_{k:t} = (\Sigma_k^t)^{-1}$, and specially $\lambda_{1:t}$ is abbreviated as λ_t. Define the function $G_t : \mathbb{R} \to \mathbb{R}$ by

$$G_t(u) = \sum_{k=1}^t \eta_k \Pi_{k+1}^t(u). \tag{34}$$

Throughout this paper, we assume that the step-size sequence satisfies $\eta_t \in]0, \kappa^{-2}]$ for all $t \in \mathbb{N}$. Thus, $G_t(u)$ and $\Pi_k^t(u)$ are non-negative on $]0, \kappa^2]$. For notational simplicity, throughout the rest of this subsection, we will drop the index $s = 1$ for the first local estimator whenever it shows up, i.e, we abbreviate $h_{1,t}$ as h_t, $\mathbf{z}_1$ as $\mathbf{z}$, and $\mathcal{T}_{\mathbf{x}_1}$ as $\mathcal{T}_{\mathbf{x}}$, etc.

The key idea for our estimation on bias is that $\{h_t\}_t$ can be well approximated by the population sequence $\{r_t\}_t$. Recall that the population sequence is defined by $r_1 = 0$ and

$$r_{t+1} = (I - \mathcal{T})r_t + \mathcal{S}_\rho^* f_\rho. \tag{35}$$

It is easy to see that the population sequence is deterministic, and it depends on the regression function f_ρ.

We first have the following observations.

Lemma 2. *The sequence $\{r_t\}_t$ defined by (35) can be rewritten as*

$$r_{t+1} = G_t(\mathcal{T})\mathcal{S}_\rho^* f_\rho. \tag{36}$$

Similarly, for any $s \in [m]$, the sequences $\{g_{s,t}\}_t$ and $\{h_{s,t}\}_t$ defined by (17) and (19) can be rewritten as

$$g_{s,t+1} = G_t(\mathcal{T}_{\mathbf{x}_s})\mathcal{S}_{\mathbf{x}_s}^* \mathbf{y}_s,$$

and

$$h_{s,t+1} = G_t(\mathcal{T}_{\mathbf{x}_s})\mathcal{L}_{\mathbf{x}_s}^* f_\rho.$$

Proof. Using the relationship (35) iteratively, introducing with $r_1 = 0$, one can prove the first conclusion. □

According to the above lemma, we know that GM can be rewritten as a form of SRA with filter function $\widetilde{G}_\lambda(\cdot) = G_t(\cdot)$. In the next lemma, we will further develop some basic properties for this filter function.

Lemma 3. *For all $u \in [0, \kappa^2]$,*
1) $u^\alpha G_t(u) \le \lambda_t^{\alpha-1}$, $\forall \alpha \in [0, 1]$.
2) $(1 - uG_t(u))u^\alpha = \Pi_1^t(u)u^\alpha \le (\alpha/\mathrm{e})^\alpha \lambda_t^\alpha$, $\quad \forall \alpha \in [0, \infty[$.
3) $\Pi_k^t(u)u^\alpha \le (\alpha/\mathrm{e})^\alpha \lambda_{k:t}^\alpha$, $\quad \forall t, k \in \mathbb{N}$.

Proof. 1). For $\alpha = 0$ or 1, the proof is straightforward and can be found in (Yao et al., 2007). Indeed, for all $u \in [0, \kappa^2]$, $\Pi_{k+1}^t(u) \le 1$ and thus $G_t(u) \le \sum_{k=1}^t \eta_k = \lambda_t^{-1}$. Moreover, writing $\eta_k u = 1 - (1 - \eta_k u)$, we have

$$uG_t(u) = \sum_{k=1}^t (\eta_k u)\Pi_{k+1}^t(u) = \sum_{k=1}^t (\Pi_{k+1}^t(u) - \Pi_k^t(u)) = 1 - \Pi_1^t(u) \le 1. \tag{37}$$

Now we consider the case $0 < \alpha < 1$. We have

$$u^\alpha G_t(u) = |uG_t(u)|^\alpha |G_t(u)|^{1-\alpha} \le \lambda_t^{\alpha-1},$$

where we used $uG_t(u) \le 1$ and $G_t(u) \le \lambda_t^{-1}$ in the above.

2) By (37), we have $(1 - uG_t(u))u^\alpha = \Pi_1^t(u)u^\alpha$. Then the desired result is a direct consequence of Conclusion 3).

3) The proof can be also found, e.g., in (Lin & Rosasco, 2017b). Using the basic inequality

$$1 + x \le \mathrm{e}^x \qquad \text{for all } x \ge -1, \tag{38}$$

with $\eta_l \kappa^2 \le 1$, we get

$$\Pi_{k+1}^t(u)u^\alpha \le \exp\left\{-u\Sigma_{k+1}^t\right\} u^\alpha.$$

The maximum of the function $g(u) = \mathrm{e}^{-cu}u^\alpha$ (with $c > 0$) over $\mathbb{R}_+$ is achieved at $u_{\max} = \alpha/c$, and thus

$$\sup_{u\geq 0} \mathrm{e}^{-cu}u^\alpha = \Big(\frac{\alpha}{\mathrm{e}c}\Big)^\alpha. \tag{39}$$

Using this inequality with $c = \Sigma_{k+1}^t$, one can prove the desired result. □

According to Lemma 3, $G_t(\cdot)$ is a filter function indexed with regularization parameter $\lambda = \lambda_t$, and the qualification τ can be any positive number, and $E = 1$, $F_\tau = (\tau/\mathrm{e})^\tau$. Using Lemma 3 and the spectral theorem, one can get the following results.

Lemma 4. *Let L be a compact, positive operator on a separable Hilbert space H such that $\|L\| \leq \kappa^2$. Then for any $\tilde{\lambda} \geq 0$,*
1) $\|(L+\tilde{\lambda})^\alpha G_t(L)\| \leq \lambda_t^{\alpha-1}(1+(\tilde{\lambda}/\lambda_t)^\alpha)$, $\forall \alpha \in [0,1]$.
2) $\|(I - LG_t(L))(L+\tilde{\lambda})^\alpha\| = \|\Pi_1^t(L)(L+\tilde{\lambda})^\alpha\| \leq 2^{(\alpha-1)_+}((\alpha/\mathrm{e})^\alpha + (\tilde{\lambda}/\lambda_t)^\alpha)\lambda_t^\alpha$, $\forall \alpha \in [0,\infty[$.
3) $\|\Pi_{k+1}^t(L)L^\alpha\| \leq (\alpha/\mathrm{e})^\alpha \lambda_{k:t}^\alpha$, $\forall k,t \in \mathbb{N}$.

Proof. 1) Following from the spectral theorem, one has

$$\|(L+\tilde{\lambda})^\alpha G_t(L)\| \leq \sup_{u\in[0,\kappa^2]} (u+\tilde{\lambda})^\alpha G_t(u) \leq \sup_{u\in[0,\kappa^2]} (u^\alpha + \tilde{\lambda}^\alpha) G_t(u).$$

Using Part 1) of Lemma 3 to the above, one can prove the first conclusion.
2) Using the spectral theorem,

$$\|\Pi_1^t(L)(L+\tilde{\lambda})^\alpha\| \leq \sup_{u\in[0,\kappa^2]} (u+\tilde{\lambda})^\alpha \Pi_1^t(u).$$

When $\alpha \leq 1$,

$$\sup_{u\in[0,\kappa^2]} (u+\tilde{\lambda})^\alpha \Pi_1^t(u) \leq \sup_{u\in[0,\kappa^2]} (u^\alpha+\tilde{\lambda}^\alpha)\Pi_1^t(u) \leq (\alpha/\mathrm{e})^\alpha \lambda_t^\alpha + \tilde{\lambda}^\alpha,$$

where for the last inequality, we used Part 2) of Lemma 3. Similarly, when $\alpha > 1$, by Hölder's inequality, and Part 2) of Lemma 3,

$$\sup_{u\in[0,\kappa^2]} (u+\tilde{\lambda})^\alpha \Pi_1^t(u) \leq 2^{\alpha-1} \sup_{u\in[0,\kappa^2]} (u^\alpha+\tilde{\lambda}^\alpha)\Pi_1^t(u) \leq 2^{\alpha-1}((\alpha/\mathrm{e})^\alpha \lambda_t^\alpha + \tilde{\lambda}^\alpha).$$

From the above analysis, one can prove the second conclusion.
3) Simply applying the spectral theorem and 3) of Lemma 3, one can prove the third conclusion.

□

Using Lemma 4, one can prove the following results, which give some basic properties for the population sequence $\{r_t\}_t$.

Lemma 5. *Let $a \in \mathbb{R}$. Under Assumption 3, the following results hold.*
1) For any $a \leq \zeta$, we have

$$\|\mathcal{L}^{-a}\left(\mathcal{S}_\rho r_{t+1} - f_\rho\right)\|_\rho \leq ((\zeta-a)/\mathrm{e})^{\zeta-a} R\lambda_t^{\zeta-a}.$$

2) We have

$$\|\mathcal{T}^{a-1/2} r_{t+1}\|_H \leq R \cdot \begin{cases} \lambda_t^{\zeta+a-1}, & \text{if } -\zeta \leq a \leq 1-\zeta, \\ \kappa^{2(\zeta+a-1)}, & \text{if } a \geq 1-\zeta. \end{cases} \tag{40}$$

Proof. 1) Using (36) and noting that

$$\mathcal{S}_\rho G_t(\mathcal{T})\mathcal{S}_\rho^* = \mathcal{S}_\rho G_t(\mathcal{S}_\rho^*\mathcal{S}_\rho)\mathcal{S}_\rho^* = G_t(\mathcal{S}_\rho\mathcal{S}_\rho^*)\mathcal{S}_\rho\mathcal{S}_\rho^* = G_t(\mathcal{L})\mathcal{L}.$$

We thus have

$$\mathcal{L}^{-a}(\mathcal{S}_\rho r_{t+1} - f_\rho) = \mathcal{L}^{-a}\left(G_t(\mathcal{L})\mathcal{L} - I\right) f_\rho.$$

Taking the ρ-norm, applying Assumption 3 and (37), we have

$$\|\mathcal{L}^{-a}(\mathcal{S}_\rho r_{t+1} - f_\rho)\|_\rho \leq \|\mathcal{L}^{\zeta-a}(G_t(\mathcal{L})\mathcal{L} - I)\|R = \|\mathcal{L}^{\zeta-a}\Pi_1^t(\mathcal{L})\|R.$$

Note that the condition (8) implies (24). Applying Part 3) of Lemma 4, one can prove the first desired result.

2) By (36) and Assumption 3,

$$\|\mathcal{T}^{a-1/2} r_{t+1}\|_H = \|\mathcal{T}^{a-1/2} G_t(\mathcal{T})\mathcal{S}_\rho^* f_\rho\|_H \leq \|\mathcal{T}^{a-1/2} G_t(\mathcal{T})\mathcal{S}_\rho^*\mathcal{L}^\zeta\|R.$$

Noting that

$$\begin{aligned}\|\mathcal{T}^{a-1/2}G_t(\mathcal{T})\mathcal{S}_\rho^*\mathcal{L}^\zeta\| &= \|\mathcal{T}^{a-1/2}G_t(\mathcal{T})\mathcal{S}_\rho^*\mathcal{L}^{2\zeta}\mathcal{S}_\rho G_t(\mathcal{T})\mathcal{T}^{a-1/2}\|^{1/2}\\ &= \|G_t^2(\mathcal{T})\mathcal{T}^{2\zeta+2a}\|^{1/2} = \|G_t(\mathcal{T})\mathcal{T}^{\zeta+a}\|,\end{aligned}$$

we thus have

$$\|\mathcal{T}^{a-1/2}r_{t+1}\|_H \le \|G_t(\mathcal{T})\mathcal{T}^{\zeta+a}\|R.$$

If $0 \le \zeta + a \le 1$, i.e., $-\zeta \le a \le 1-\zeta$, then by using 1) of Lemma 4, we get

$$\|\mathcal{T}^{a-1/2}r_{t+1}\|_H \le \lambda_t^{\zeta+a-1}R.$$

Similarly, when $a \ge 1-\zeta$, we have

$$\|\mathcal{T}^{a-1/2}r_{t+1}\|_H \le \|G_t(\mathcal{T})\mathcal{T}\|\|\mathcal{T}\|^{\zeta+a-1}R \le \kappa^{2(\zeta+a-1)}R,$$

where for the last inequality we used 1) of Lemma 4 and (24). This thus proves the second desired result. □

We also need the following two lemmas on operator inequalities.

Lemma 6. *(Fujii et al., 1993) Let A and B be two positive bounded linear operators on a separable Hilbert space. Then*

$$\|A^sB^s\| \le \|AB\|^s, \quad \textit{when } 0 \le s \le 1.$$

Lemma 7. *Let A and B be two non-negative bounded linear operators on a separable Hilbert space with $\max(\|A\|, \|B\|) \le \kappa^2$ for some non-negative κ^2. Then for any $\zeta > 0$,*

$$\|A^\zeta - B^\zeta\| \le C_{\zeta,\kappa}\|A-B\|^{\zeta\wedge 1}, \tag{41}$$

where

$$C_{\zeta,\kappa} = \begin{cases} 1 & \textit{when } \zeta \le 1,\\ 2\zeta\kappa^{2\zeta-2} & \textit{when } \zeta > 1.\end{cases} \tag{42}$$

Proof. Following from (Mathé & Pereverzev, 2002), one can prove the desired result for $\zeta \le 1$. For $\zeta \ge 1$, the proof can be found in (Dicker et al., 2017), see also (Blanchard & Mücke, 2017). □

With the above lemmas, we can prove the the following analytic result, which enables us to estimate the bias term in terms of several random quantities.

Lemma 8. *Under Assumption 3, let $\tilde\lambda > 0$,*

$$\Delta_1^{\mathbf{z}} = \|\mathcal{T}_{\tilde\lambda}^{1/2}\mathcal{T}_{\mathbf{x}\tilde\lambda}^{-1/2}\|^2 \vee 1, \qquad \Delta_3^{\mathbf{z}} = \|\mathcal{T} - \mathcal{T}_{\mathbf{x}}\|$$

and

$$\Delta_2^{\mathbf{z}} = \|\mathcal{L}_{\mathbf{x}}f_\rho - \mathcal{S}_\rho^*f_\rho - \mathcal{T}_{\mathbf{x}}r_{t+1} + \mathcal{T}r_{t+1}\|_H.$$

Then the following results hold.
1) For $0 < \zeta \le 1$,

$$\|\mathcal{S}_\rho h_{t+1} - f_\rho\|_\rho \le \left(1 \vee \left(\frac{\tilde\lambda}{\lambda_t}\right)^{\zeta\vee\frac12}\right)(C_1(\Delta_1^{\mathbf{z}})^{\zeta\vee\frac12}\lambda_t^\zeta + 2\sqrt{\Delta_1^{\mathbf{z}}}\lambda_t^{-\frac12}\Delta_2^{\mathbf{z}}). \tag{43}$$

2) For $\zeta > 1$,

$$\|\mathcal{S}_\rho h_{t+1} - f_\rho\|_\rho \le \sqrt{\Delta_1^{\mathbf{z}}}\left(1 \vee \left(\frac{\tilde\lambda}{\lambda_t}\right)^{\zeta}\right)(C_2\lambda_t^\zeta + 2\lambda_t^{-\frac12}\Delta_2^{\mathbf{z}} + C_3\lambda_t^{\frac12}(\Delta_3^{\mathbf{z}})^{(\zeta-\frac12)\wedge 1}). \tag{44}$$

Here, C_1, C_2 and C_3 are positive constants depending only on ζ, κ, and R.

Proof. Using Lemma 2 with $s = 1$, we can estimate $\|\mathcal{S}_\rho h_{t+1} - f_\rho\|_\rho$ as

$$\begin{aligned}\|\mathcal{S}_\rho G_t(\mathcal{T}_{\mathbf{x}})\mathcal{L}_{\mathbf{x}} f_\rho - f_\rho\|_\rho \leq & \|\underbrace{\mathcal{S}_\rho G_t(\mathcal{T}_{\mathbf{x}})[\mathcal{L}_{\mathbf{x}} f_\rho - \mathcal{S}_\rho^* f_\rho - \mathcal{T}_{\mathbf{x}} r_{t+1} + \mathcal{T} r_{t+1}]}_{\textbf{Bias.1}}\|_\rho \\ & + \|\underbrace{\mathcal{S}_\rho G_t(\mathcal{T}_{\mathbf{x}})[\mathcal{S}_\rho^* f_\rho - \mathcal{T} r_{t+1}]}_{\textbf{Bias.2}}\|_\rho \\ & + \|\underbrace{\mathcal{S}_\rho[I - G_t(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}}] r_{t+1}}_{\textbf{Bias.3}}\|_\rho \\ & + \|\underbrace{\mathcal{S}_\rho r_{t+1} - f_\rho}_{\textbf{Bias.4}}\|_\rho. \end{aligned} \tag{45}$$

In the rest of the proof, we will estimate the four terms of the r.h.s separately.

Estimating Bias.4

Using 1) of Lemma 5 with $a = 0$, we get

$$\|\mathbf{Bias.4}\|_\rho \leq (\zeta/\mathrm{e})^\zeta \lambda_t^\zeta R. \tag{46}$$

Estimating Bias.1

By a simple calculation, we know that for any $f \in H$,

$$\|\mathcal{S}_\rho G_t(\mathcal{T}_{\mathbf{x}}) f\|_\rho \leq \|\mathcal{S}_\rho \mathcal{T}_{\tilde{\lambda}}^{-1/2}\| \|\mathcal{T}_{\tilde{\lambda}}^{1/2} \mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-1/2}\| \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} G_t(\mathcal{T}_{\mathbf{x}})\| \|f\|_H.$$

Note that

$$\|\mathcal{S}_\rho \mathcal{T}_{\tilde{\lambda}}^{-1/2}\| = \sqrt{\|\mathcal{S}_\rho \mathcal{T}_{\tilde{\lambda}}^{-1} \mathcal{S}_\rho^*\|} = \sqrt{\|\mathcal{L}\mathcal{L}_{\tilde{\lambda}}^{-1}\|} \leq 1, \tag{47}$$

and that applying 1) of Lemma 4, with (28), we have

$$\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} G_t(\mathcal{T}_{\mathbf{x}})\| \leq (1 + \sqrt{\tilde{\lambda}/\lambda_t})/\sqrt{\lambda_t}.$$

Thus for any $f \in H$, we have

$$\|\mathcal{S}_\rho G_t(\mathcal{T}_{\mathbf{x}}) f\|_\rho \leq (1 + \sqrt{\tilde{\lambda}/\lambda_t}) \lambda_t^{-\frac{1}{2}} \sqrt{\Delta_1^{\mathbf{z}}} \|f\|_H. \tag{48}$$

Therefore,

$$\|\mathbf{Bias.1}\|_\rho \leq (1 + \sqrt{\tilde{\lambda}/\lambda_t}) \lambda_t^{-\frac{1}{2}} \sqrt{\Delta_1^{\mathbf{z}}} \Delta_2^{\mathbf{z}}. \tag{49}$$

Estimating Bias.2

By (48), we have

$$\|\mathbf{Bias.2}\|_\rho \leq (1 + \sqrt{\tilde{\lambda}/\lambda_t}) \lambda_t^{-\frac{1}{2}} \sqrt{\Delta_1^{\mathbf{z}}} \|\mathcal{T} r_{t+1} - \mathcal{S}_\rho^* f_\rho\|_H.$$

Using (with $\mathcal{T} = \mathcal{S}_\rho^* \mathcal{S}_\rho$ and $\mathcal{L} = \mathcal{S}_\rho \mathcal{S}_\rho^*$)

$$\|\mathcal{T} r_{t+1} - \mathcal{S}_\rho^* f_\rho\|_H = \|\mathcal{S}_\rho^*(\mathcal{S}_\rho r_{t+1} - f_\rho)\|_H = \|\mathcal{L}^{1/2}(\mathcal{S}_\rho r_{t+1} - f_\rho)\|_\rho,$$

and applying 1) of Lemma 5 with $a = -1/2$, we get

$$\|\mathbf{Bias.2}\|_\rho \leq ((\zeta + 1/2)/\mathrm{e})^{\zeta+1/2} (1 + \sqrt{\tilde{\lambda}/\lambda_t}) \sqrt{\Delta_1^{\mathbf{z}}} \lambda_t^\zeta R. \tag{50}$$

Estimating Bias.3

By 2) of Lemma 3,

$$\mathbf{Bias.3} = \mathcal{S}_\rho \Pi_1^t(\mathcal{T}_{\mathbf{x}}) r_{t+1}.$$

When $\zeta \leq 1/2$, by a simple calculation, we have

$$\begin{aligned}\|\mathbf{Bias.3}\|_\rho \leq & \|\mathcal{S}_\rho \mathcal{T}_{\tilde{\lambda}}^{-1/2}\| \|\mathcal{T}_{\tilde{\lambda}}^{1/2} \mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-1/2}\| \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} \Pi_1^t(\mathcal{T}_{\mathbf{x}})\| \|r_{t+1}\|_H \\ \leq & \sqrt{\Delta_1^{\mathbf{z}}} \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} \Pi_1^t(\mathcal{T}_{\mathbf{x}})\| \|r_{t+1}\|_H, \end{aligned}$$

where for the last inequality, we used (47). By 2) of Lemma 4, with (28),

$$\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} \Pi_1^t(\mathcal{T}_{\mathbf{x}})\| \leq \sqrt{\lambda_t}(1/\sqrt{2\mathrm{e}} + \sqrt{\tilde{\lambda}/\lambda_t}), \tag{51}$$

and by 2) of Lemma 5,

$$\|r_{t+1}\|_H \leq R\lambda_t^{\zeta-1/2}.$$

It thus follows that

$$\|\mathbf{Bias.3}\|_\rho \leq \sqrt{\Delta_1^{\mathbf{z}}}(\sqrt{\tilde{\lambda}/\lambda_t}+1/\sqrt{2\mathrm{e}})R\lambda_t^{\zeta}.$$

When $1/2<\zeta\leq 1$, by a simple computation, we have

$$\|\mathbf{Bias.3}\|_\rho \leq \|\mathcal{S}_\rho\mathcal{T}_{\tilde{\lambda}}^{-1/2}\|\|\mathcal{T}_{\tilde{\lambda}}^{1/2}\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-1/2}\|\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2}\Pi_1^t(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\zeta-1/2}\|\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2-\zeta}\mathcal{T}_{\tilde{\lambda}}^{\zeta-1/2}\|\|\mathcal{T}_{\tilde{\lambda}}^{1/2-\zeta}r_{t+1}\|_H.$$

Applying (47) and 2) of Lemma 5, we have

$$\|\mathbf{Bias.3}\|_\rho \leq \sqrt{\Delta_1^{\mathbf{z}}}\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2}\Pi_1^t(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\zeta-1/2}\|\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2-\zeta}\mathcal{T}_{\tilde{\lambda}}^{\zeta-1/2}\|R.$$

By 2) of Lemma 4,

$$\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2}\Pi_1^t(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\zeta-1/2}\| = \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\zeta}\Pi_1^t(\mathcal{T}_{\mathbf{x}})\| \leq ((\zeta/\mathrm{e})^\zeta+(\tilde{\lambda}/\lambda_t)^\zeta)\lambda_t^\zeta.$$

Besides, by $\zeta\leq 1$ and Lemma 6,

$$\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2-\zeta}\mathcal{T}_{\tilde{\lambda}}^{\zeta-1/2}\| = \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-\frac{1}{2}(2\zeta-1)}\mathcal{T}_{\tilde{\lambda}}^{\frac{1}{2}(2\zeta-1)}\| \leq \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-\frac{1}{2}}\mathcal{T}_{\tilde{\lambda}}^{\frac{1}{2}}\|^{2\zeta-1} \leq (\Delta_1^{\mathbf{z}})^{\zeta-\frac{1}{2}}.$$

It thus follows that

$$\|\mathbf{Bias.3}\|_\rho \leq (\Delta_1^{\mathbf{z}})^\zeta((\tilde{\lambda}/\lambda_t)^\zeta+(\zeta/\mathrm{e})^\zeta)R\lambda_t^\zeta.$$

When $\zeta>1$, we rewrite **Bias.3** as

$$\mathcal{S}_\rho\mathcal{T}_{\tilde{\lambda}}^{-1/2}\cdot\mathcal{T}_{\tilde{\lambda}}^{1/2}\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-1/2}\cdot\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2}\Pi_1^t(\mathcal{T}_{\mathbf{x}})(\mathcal{T}_{\mathbf{x}}^{\zeta-1/2}+\mathcal{T}^{\zeta-1/2}-\mathcal{T}_{\mathbf{x}}^{\zeta-1/2})\mathcal{T}^{1/2-\zeta}r_{t+1}.$$

By a simple calculation, we can upper bound $\|\mathbf{Bias.3}\|_\rho$ by

$$\leq \|\mathcal{S}_\rho\mathcal{T}_{\tilde{\lambda}}^{-1/2}\|\|\mathcal{T}_{\tilde{\lambda}}^{1/2}\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-1/2}\|(\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2}\Pi_1^t(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}}^{\zeta-1/2}\|+\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2}\Pi_1^t(\mathcal{T}_{\mathbf{x}})\|\|\mathcal{T}^{\zeta-1/2}-\mathcal{T}_{\mathbf{x}}^{\zeta-1/2}\|)\|\mathcal{T}^{1/2-\zeta}r_{t+1}\|.$$

Introducing with (47) and (51), and applying 2) of Lemma 5,

$$\|\mathbf{Bias.3}\|_\rho \leq \sqrt{\Delta_1^{\mathbf{z}}}(\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2}\Pi_1^t(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}}^{\zeta-1/2}\|+(1/\sqrt{2\mathrm{e}}+\sqrt{\tilde{\lambda}/\lambda_t})\sqrt{\lambda_t}\|\mathcal{T}^{\zeta-1/2}-\mathcal{T}_{\mathbf{x}}^{\zeta-1/2}\|)R.$$

By 2) of Lemma 4,

$$\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2}\Pi_1^t(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}}^{\zeta-1/2}\| \leq \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\zeta}\Pi_1^t(\mathcal{T}_{\mathbf{x}})\| \leq 2^{\zeta-1}((\zeta/\mathrm{e})^\zeta+(\tilde{\lambda}/\lambda_t)^\zeta)\lambda_t^\zeta.$$

Moreover, by Lemma 7 and $\max(\|\mathcal{T}\|,\|\mathcal{T}_{\mathbf{x}}\|)\leq\kappa^2$,

$$\|\mathcal{T}^{\zeta-1/2}-\mathcal{T}_{\mathbf{x}}^{\zeta-1/2}\| \leq (2\zeta\kappa^{2\zeta-3})^{\mathbf{1}_{\{2\zeta\geq 3\}}}\|\mathcal{T}-\mathcal{T}_{\mathbf{x}}\|^{(\zeta-1/2)\wedge 1}.$$

Therefore, when $\zeta>1$, **Bias.3** can be estimated as

$$\begin{aligned}&\|\mathbf{Bias.3}\|_\rho\\ \leq&\sqrt{\Delta_1^{\mathbf{z}}}\left(2^{\zeta-1}((\zeta/\mathrm{e})^\zeta+(\tilde{\lambda}/\lambda_t)^\zeta)\lambda_t^\zeta+(2\zeta\kappa^{2\zeta-3})^{\mathbf{1}_{\{2\zeta\geq 3\}}}(1/\sqrt{2\mathrm{e}}+\sqrt{\tilde{\lambda}/\lambda_t})\sqrt{\lambda_t}(\Delta_3^{\mathbf{z}})^{(\zeta-1/2)\wedge 1}\right)R.\end{aligned}$$

From the above analysis, we know that $\|\mathbf{Bias.3}\|_\rho$ can be upper bounded by

$$\begin{cases}\sqrt{\Delta_1^{\mathbf{z}}}(\sqrt{\tilde{\lambda}/\lambda_t}+1/\sqrt{2\mathrm{e}})R\lambda_t^\zeta, & \text{if } \zeta\in]0,1/2],\\ (\Delta_1^{\mathbf{z}})^\zeta(\left(\tilde{\lambda}/\lambda_t\right)^\zeta+(\zeta/\mathrm{e})^\zeta)R\lambda_t^\zeta, & \text{if } \zeta\in]1/2,1],\\ \sqrt{\Delta_1^{\mathbf{z}}}\left(2^{\zeta-1}((\frac{\zeta}{\mathrm{e}})^\zeta+(\frac{\tilde{\lambda}}{\lambda_t})^\zeta)\lambda_t^\zeta+(2\zeta\kappa^{2\zeta-3})^{\mathbf{1}_{\{2\zeta\geq 3\}}}(\frac{1}{\sqrt{2\mathrm{e}}}+\sqrt{\frac{\tilde{\lambda}}{\lambda_t}})\sqrt{\lambda_t}(\Delta_3^{\mathbf{z}})^{(\zeta-\frac{1}{2})\wedge 1}\right)R, & \text{if } \zeta\in]1,\infty[.\end{cases}\tag{52}$$

Introducing (46), (49), (50) and (52) into (45), and by a simple calculation, one can prove the desired results with

$$C_1 = R\left((\zeta/\mathrm{e})^\zeta+2((\zeta+\frac{1}{2})/\mathrm{e})^{\zeta+\frac{1}{2}}+((\zeta\vee\frac{1}{2})/\mathrm{e})^{\zeta\vee\frac{1}{2}}+1\right),$$

$$C_2 = R\left((2^{\zeta-1}+1)(\zeta/\mathrm{e})^\zeta+2((\zeta+\frac{1}{2})/\mathrm{e})^{\zeta+\frac{1}{2}}+2^{\zeta-1}\right),$$

$$\text{and}\quad C_3 = (2\zeta\kappa^{2\zeta-3})^{\mathbf{1}_{\{2\zeta\geq 3\}}}(1/\sqrt{2\mathrm{e}}+1).$$

□

The upper bounds in (43) and (44) depend on three random quantities, $\Delta_1^{\mathbf{z}}$, $\Delta_3^{\mathbf{z}}$ and $\Delta_2^{\mathbf{z}}$. To derive error bounds for the bias term from Lemma 8, it is necessary to estimate these three random quantities.

We first introduce the following concentration result for Hilbert space valued random variable used in (Caponnetto & De Vito, 2007) and based on the results in (Pinelis & Sakhanenko, 1986).

Lemma 9. *Let $w_1, \cdots, w_m$ be i.i.d random variables in a separable Hilbert space with norm $\|\cdot\|$. Suppose that there are two positive constants B and σ^2 such that*

$$\mathbb{E}[\|w_1 - \mathbb{E}[w_1]\|^l] \leq \frac{1}{2} l! B^{l-2} \sigma^2, \quad \forall l \geq 2. \tag{53}$$

Then for any $0 < \delta < 1/2$, the following holds with probability at least $1 - \delta$,

$$\left\| \frac{1}{m} \sum_{k=1}^{m} w_m - \mathbb{E}[w_1] \right\| \leq 2 \left(\frac{B}{m} + \frac{\sigma}{\sqrt{m}} \right) \log \frac{2}{\delta}.$$

In particular, (53) holds if

$$\|w_1\| \leq B/2 \ \text{ a.s.}, \quad \text{and} \quad \mathbb{E}[\|w_1\|^2] \leq \sigma^2. \tag{54}$$

Using the above lemma, we can prove the following two results.

Lemma 10. *Let $f : X \to Y$ be a measurable function such that $\|f\|_\infty < \infty$, then with probability at least $1 - \delta$ $(0 < \delta < 1/2)$,*

$$\|\mathcal{L}_{\mathbf{x}} f - \mathcal{L} f\|_H \leq 2\kappa \left(\frac{2\|f\|_\infty}{|\mathbf{x}|} + \frac{\|f\|_\rho}{\sqrt{|\mathbf{x}|}} \right) \log \frac{2}{\delta}.$$

Proof. Let $\xi_i = f(x_i) K_{x_i}$ for $i = 1, \cdots, |\mathbf{x}|$. Obviously,

$$\mathcal{L}_{\mathbf{x}} f - \mathcal{L} f = \frac{1}{|\mathbf{x}|} \sum_{i=1}^{|\mathbf{x}|} (\xi_i - \mathbb{E}[\xi_i]),$$

and by Assumption (8), we have

$$\|\xi\|_H \leq \|f\|_\infty \|K_x\|_H \leq \kappa \|f\|_\infty$$

and

$$\mathbb{E}\|\xi\|_H^2 \leq \kappa^2 \|f\|_\rho^2.$$

Applying Lemma 9 with $B' = 2\kappa\|f\|_\infty$ and $\sigma = \kappa\|f\|_\rho$, one can prove the desired result. □

Lemma 11. *Let $0 < \delta < 1/2$. It holds with probability at least $1 - \delta$:*

$$\|\mathcal{T} - \mathcal{T}_{\mathbf{x}}\|_{HS} \leq \frac{6\kappa^2}{\sqrt{|\mathbf{x}|}} \log \frac{2}{\delta}.$$

Here, $\|\cdot\|_{HS}$ denotes the Hilbert-Schmidt norm.

Proof. Let $\xi_i = K_{x_i} \otimes K_{x_i}$, for all $i \in [|\mathbf{x}|]$. Obviously,

$$\mathcal{T} - \mathcal{T}_{\mathbf{x}} = \frac{1}{|\mathbf{x}|} \sum_{i=1}^{|\mathbf{x}|} (\mathbb{E}[\xi_i] - \xi_i),$$

and by Assumption (8), $\|\xi_i\|_{HS} = \|K_{x_i}\|_H^2 \leq \kappa^2$. Applying Lemma 9 with $B' = 2\kappa^2$ and $\sigma' = \kappa^2$, one can prove the desire result. □

We next introduce the following concentration inequality for norms of self-adjoint operators on a Hilbert space.

Lemma 12. *Let $\mathcal{X}_1, \cdots, \mathcal{X}_m$ be a sequence of independently and identically distributed self-adjoint Hilbert-Schmidt operators on a separable Hilbert space. Assume that $\mathbb{E}[\mathcal{X}_1] = 0$, and $\|\mathcal{X}_1\| \leq B$ almost surely for some $B > 0$. Let $\mathcal{V}$ be a positive trace-class operator such that $\mathbb{E}[\mathcal{X}_1^2] \preccurlyeq \mathcal{V}$. Then with probability at least $1 - \delta$, $(\delta \in]0, 1[)$, there holds*

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \mathcal{X}_i \right\| \leq \frac{2B\beta}{3m} + \sqrt{\frac{2\|\mathcal{V}\|\beta}{m}}, \qquad \beta = \log \frac{4 \operatorname{tr} \mathcal{V}}{\|\mathcal{V}\|\delta}.$$

Proof. The proof can be found in, e.g., (Rudi et al., 2015; Dicker et al., 2017). Following from the argument in (Minsker, 2011), we can generalize (Tropp, 2012) from a sequence of self-adjoint matrices to a sequence of self-adjoint Hilbert-Schmidt operators on a separable Hilbert space, and get that for any $t \geq \sqrt{\frac{\|\mathcal{V}\|}{m}} + \frac{B}{3m}$,

$$\Pr\left(\left\|\frac{1}{m}\sum_{i=1}^{m}\mathcal{X}_i\right\| \geq t\right) \leq \frac{4\operatorname{tr}\mathcal{V}}{\|\mathcal{V}\|}\exp\left(\frac{-mt^2}{2\|\mathcal{V}\| + 2Bt/3}\right). \tag{55}$$

Rewriting

$$\frac{4\operatorname{tr}\mathcal{V}}{\|\mathcal{V}\|}\exp\left(\frac{-mt^2}{2\|\mathcal{V}\| + 2Bt/3}\right) = \delta,$$

as a quadratic equation with respect to the variable t, and then solving the quadratic equation, we get

$$t_0 = \frac{B\beta}{3m} + \sqrt{\left(\frac{B\beta}{3m}\right)^2 + \frac{2\beta\|\mathcal{V}\|}{m}} \leq \frac{2B\beta}{3m} + \sqrt{\frac{2\beta\|\mathcal{V}\|}{m}} := t^*,$$

where we used $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}, \forall a, b > 0$. Note that $\beta > 1$, and thus $t_0 \geq \sqrt{\frac{\|\mathcal{V}\|}{m}} + \frac{B}{3m}$. By

$$\Pr\left(\left\|\frac{1}{m}\sum_{i=1}^{m}\mathcal{X}_i\right\| \geq t_*\right) \leq \Pr\left(\left\|\frac{1}{m}\sum_{i=1}^{m}\mathcal{X}_i\right\| \geq t_0\right),$$

and applying (55) to bound the left-hand side, one can get the desire result. □

Lemma 13. *Let $0 < \delta < 1$ and $\lambda > 0$. With probability at least $1 - \delta$, the following holds:*

$$\left\|(\mathcal{T}+\lambda)^{-1/2}(\mathcal{T}-\mathcal{T}_{\mathbf{x}})(\mathcal{T}+\lambda)^{-1/2}\right\| \leq \frac{4\kappa^2\beta}{3|\mathbf{x}|\lambda} + \sqrt{\frac{2\kappa^2\beta}{|\mathbf{x}|\lambda}}, \quad \beta = \log\frac{4\kappa^2(\mathcal{N}(\lambda)+1)}{\delta\|\mathcal{T}\|}.$$

Proof. The proof can be also found in (Rudi et al., 2015; Dicker et al., 2017). We will use Lemma 12 to prove the result. Let $|\mathbf{x}| = m$ and $\mathcal{X}_i = \mathcal{T}_{\tilde{\lambda}}^{-1/2}(\mathcal{T} - \mathcal{T}_{x_i})\mathcal{T}_{\tilde{\lambda}}^{-1/2}$, for all $i \in [m]$. Then $\mathcal{T}_{\tilde{\lambda}}^{-1/2}(\mathcal{T} - \mathcal{T}_{\mathbf{x}})\mathcal{T}_{\tilde{\lambda}}^{-1/2} = \frac{1}{m}\sum_{i=1}^{m}\mathcal{X}_i$. Obviously, for any $\mathcal{X} = \mathcal{X}_i$, $\mathbb{E}[\mathcal{X}] = 0$, and

$$\|\mathcal{X}\| \leq \mathbb{E}\left[\|\mathcal{T}_{\tilde{\lambda}}^{-1/2}\mathcal{T}_x\mathcal{T}_{\tilde{\lambda}}^{-1/2}\|\right] + \|\mathcal{T}_{\tilde{\lambda}}^{-1/2}\mathcal{T}_x\mathcal{T}_{\tilde{\lambda}}^{-1/2}\| \leq 2\kappa^2/\tilde{\lambda},$$

where for the last inequality, we used Assumption (8) which implies

$$\|\mathcal{T}_{\tilde{\lambda}}^{-1/2}\mathcal{T}_x\mathcal{T}_{\tilde{\lambda}}^{-1/2}\| \leq \operatorname{tr}(\mathcal{T}_{\tilde{\lambda}}^{-1/2}\mathcal{T}_x\mathcal{T}_{\tilde{\lambda}}^{-1/2}) = \operatorname{tr}(\mathcal{T}_{\tilde{\lambda}}^{-1}\mathcal{T}_x) = \langle\mathcal{T}_{\tilde{\lambda}}^{-1}K_x, K_x\rangle_H \leq \kappa^2/\tilde{\lambda}.$$

Also, by $\mathbb{E}(A - \mathbb{E}A)^2 \preccurlyeq \mathbb{E}A^2$,

$$\begin{aligned}\mathbb{E}\mathcal{X}^2 &\preccurlyeq \mathbb{E}(\mathcal{T}_{\tilde{\lambda}}^{-1/2}\mathcal{T}_{\tilde{x}}\mathcal{T}_{\tilde{\lambda}}^{-1/2})^2 = \mathbb{E}[\langle\mathcal{T}_{\tilde{\lambda}}^{-1}K_x, K_x\rangle_H\mathcal{T}_{\tilde{\lambda}}^{-1/2}K_x \otimes K_x\mathcal{T}_{\tilde{\lambda}}^{-1/2}] \\ &\preccurlyeq \frac{\kappa^2}{\tilde{\lambda}}\mathbb{E}[\mathcal{T}_{\tilde{\lambda}}^{-1/2}K_x \otimes K_x\mathcal{T}_{\tilde{\lambda}}^{-1/2}] = \frac{\kappa^2}{\tilde{\lambda}}\mathcal{T}_{\tilde{\lambda}}^{-1}\mathcal{T} = \mathcal{V},\end{aligned}$$

Note that $\|\mathcal{T}_{\tilde{\lambda}}^{-1}\mathcal{T}\| = \frac{\|\mathcal{T}\|}{\|\mathcal{T}\|+\tilde{\lambda}} \leq 1$. Therefore, $\|\mathcal{V}\| \leq \frac{\kappa^2}{\tilde{\lambda}}$ and

$$\frac{\operatorname{tr}(\mathcal{V})}{\|\mathcal{V}\|} = \frac{\mathcal{N}(\tilde{\lambda})\|\mathcal{T}\| + \operatorname{tr}(\mathcal{T}_{\tilde{\lambda}}^{-1}\mathcal{T})\tilde{\lambda}}{\|\mathcal{T}\|} \leq \frac{\mathcal{N}(\tilde{\lambda})\|\mathcal{T}\| + \operatorname{tr}(\mathcal{T})}{\|\mathcal{T}\|} \leq \frac{\kappa^2(\mathcal{N}(\tilde{\lambda})+1)}{\|\mathcal{T}\|},$$

where for the last inequality we used (24). Now, the result can be proved by applying Lemma 12. □

We will use Lemmas 10 and 5 to estimate the quantity $\Delta_2^{\mathbf{z}}$. The quantity $\Delta_3^{\mathbf{z}}$ can be estimated by Lemma 11 directly, as $\|\mathcal{T} - \mathcal{T}_{\mathbf{x}}\| \leq \|\mathcal{T} - \mathcal{T}_{\mathbf{x}}\|_{HS}$. The quantity $\Delta_1^{\mathbf{z}}$ can be estimated by the following lemma, whose proof is based on Lemma 13.

Lemma 14. *Under Assumption 4, let $c, \delta \in (0,1)$, $\lambda = |\mathbf{x}|^{-\theta}$ for some $\theta \geq 0$, and*

$$a_{|\mathbf{x}|,\delta,\gamma}(c,\theta) = \frac{32\kappa^2}{(\sqrt{9+24c}-3)^2}\left(\log\frac{4\kappa^2(c_\gamma+1)}{\delta\|\mathcal{T}\|} + \theta\gamma\min\left(\frac{1}{\mathrm{e}(1-\theta)_+}, \log|\mathbf{x}|\right)\right). \tag{56}$$

Then with probability at least $1-\delta$,

$$\|(\mathcal{T}+\lambda)^{-1/2}(\mathcal{T}_{\mathbf{x}}+\lambda)^{1/2}\|^2 \leq (1+c)a_{|\mathbf{x}|,\delta,\gamma}(c,\theta)(1\vee|\mathbf{x}|^{\theta-1}), \text{ and}$$

$$\|(\mathcal{T}+\lambda)^{1/2}(\mathcal{T}_{\mathbf{x}}+\lambda)^{-1/2}\|^2 \leq (1-c)^{-1}a_{|\mathbf{x}|,\delta,\gamma}(c,\theta)(1\vee|\mathbf{x}|^{\theta-1}).$$

Remark 1. *Typically, we will choose $c = 2/3$. In this case,*

$$a_{|\mathbf{x}|,\delta,\gamma}(2/3,\theta) = 8\kappa^2\left(\log\frac{4\kappa^2(c_\gamma+1)}{\delta\|\mathcal{T}\|} + \theta\gamma\min\left(\frac{1}{\mathrm{e}(1-\theta)_+}, \log|\mathbf{x}|\right)\right). \tag{57}$$

We have with probability at least $1-\delta$,

$$\|(\mathcal{T}+\lambda)^{1/2}(\mathcal{T}_{\mathbf{x}}+\lambda)^{-1/2}\|^2 \leq 3a_{|\mathbf{x}|,\delta,\gamma}(2/3,\theta)(1\vee|\mathbf{x}|^{\theta-1}).$$

Proof. We use Lemma 13 to prove the result. Let $c \in (0,1]$. By a simple calculation, we have that if $0 \leq u \leq \frac{\sqrt{9+24c}-3}{4}$, then $2u^2/3 + u \leq c$. Letting $\sqrt{\frac{2\kappa^2\beta}{|\mathbf{x}|\lambda'}} = u$, and combining with Lemma 13, we know that if

$$\sqrt{\frac{2\kappa^2\beta}{|\mathbf{x}|\lambda'}} \leq \frac{\sqrt{9+24c}-3}{4},$$

which is equivalent to

$$|\mathbf{x}| \geq \frac{32\kappa^2\beta}{(\sqrt{9+24c}-3)^2\lambda'}, \quad \beta = \log\frac{4\kappa^2(1+\mathcal{N}(\lambda'))}{\delta\|\mathcal{T}\|}, \tag{58}$$

then with probability at least $1-\delta$,

$$\left\|\mathcal{T}_{\lambda'}^{-1/2}(\mathcal{T}-\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\lambda'}^{-1/2}\right\| \leq c. \tag{59}$$

Note that from (59), we can prove

$$\|\mathcal{T}_{\lambda'}^{-1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{1/2}\|^2 \leq c+1, \quad \|\mathcal{T}_{\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{-1/2}\|^2 \leq (1-c)^{-1}. \tag{60}$$

Indeed, by simple calculations,

$$\begin{aligned}&\|\mathcal{T}_{\lambda'}^{-1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{1/2}\|^2 = \|\mathcal{T}_{\lambda'}^{-1/2}\mathcal{T}_{\mathbf{x}\lambda'}\mathcal{T}_{\lambda'}^{-1/2}\| = \|\mathcal{T}_{\lambda'}^{-1/2}(\mathcal{T}-\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\lambda'}^{-1/2}+I\| \\ &\leq \|\mathcal{T}_{\lambda'}^{-1/2}(\mathcal{T}-\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\lambda'}^{-1/2}\| + \|I\| \leq c+1,\end{aligned}$$

and (Caponnetto & De Vito, 2007)

$$\|\mathcal{T}_{\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{-1/2}\|^2 = \|\mathcal{T}_{\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{-1}\mathcal{T}_{\lambda'}^{1/2}\| = \|(I-\mathcal{T}_{\lambda'}^{-1/2}(\mathcal{T}-\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\lambda'}^{-1/2})^{-1}\| \leq (1-c)^{-1}.$$

From the above analysis, we know that for any fixed $\lambda' > 0$ such that (58), then with probability at least $1-\delta$, (60) hold.

Now let $\lambda' = a\lambda$ when $\theta \in [0,1)$ and $\lambda' = a|\mathbf{x}|^{-1}$ when $\theta \geq 1$, where for notational simplicity, we denote $a_{|\mathbf{x}|,\delta,\gamma}(c,\theta)$ by a. We will prove that the choice on λ' ensures the condition (58) is satisfied, as thus with probability at least $1-\delta$, (60) holds. Obviously, one can easily prove that $a \geq 1$, using $\kappa^2 \geq 1$ and (24). Therefore, $\lambda' \geq \lambda$, and

$$\|\mathcal{T}_{\lambda}^{1/2}\mathcal{T}_{\mathbf{x}\lambda}^{-1/2}\| \leq \|\mathcal{T}_{\lambda}^{1/2}\mathcal{T}_{\lambda'}^{-1/2}\|\|\mathcal{T}_{\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{-1/2}\|\|\mathcal{T}_{\mathbf{x}\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda}^{-1/2}\| \leq \|\mathcal{T}_{\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{-1/2}\|\sqrt{\lambda'/\lambda},$$

where for the last inequality, we used $\|\mathcal{T}_{\lambda}^{1/2}\mathcal{T}_{\lambda'}^{-1/2}\|^2 \leq \sup_{u\geq 0}\frac{u+\lambda}{u+\lambda'} \leq 1$ and $\|\mathcal{T}_{\mathbf{x}\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda}^{-1/2}\|^2 \leq \sup_{u\geq 0}\frac{u+\lambda'}{u+\lambda} \leq \lambda'/\lambda$. Similarly,

$$\|\mathcal{T}_{\lambda}^{-1/2}\mathcal{T}_{\mathbf{x}\lambda}^{1/2}\| \leq \|\mathcal{T}_{\lambda'}^{-1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{1/2}\|\sqrt{\lambda'/\lambda}.$$

Combining with (60), and by a simple calculation, one can prove the desired bounds. What remains is to prove that the condition (58) is satisfied. By Assumption 4 and $a \geq 1$,

$$\beta \leq \log \frac{4\kappa^2(1+c_\gamma a^{-\gamma}|\mathbf{x}|^{(\theta\wedge 1)\gamma})}{\delta\|\mathcal{T}\|} \leq \log \frac{4\kappa^2(1+c_\gamma)|\mathbf{x}|^{\theta\gamma}}{\delta\|\mathcal{T}\|} = \log \frac{4\kappa^2(1+c_\gamma)}{\delta\|\mathcal{T}\|} + \theta\gamma \log|\mathbf{x}|.$$

If $\theta \geq 1$, or $\theta\gamma = 0$, or $\log|\mathbf{x}| \leq \frac{1}{(1-\theta)_+ \mathrm{e}}$, then the condition (58) follows trivially. Now consider the case $\theta \in (0,1)$, $\theta\gamma \neq 0$ and $\log|\mathbf{x}| \geq \frac{1}{(1-\theta)_+\mathrm{e}}$. In this case, we apply (38) to get $\frac{\theta\gamma}{1-\theta}\log|\mathbf{x}|^{1-\theta} \leq \frac{\theta\gamma}{1-\theta}\frac{|\mathbf{x}|^{1-\theta}}{\mathrm{e}}$, and thus

$$\beta \leq \log \frac{4\kappa^2(1+c_\gamma)}{\delta\|\mathcal{T}\|} + \frac{\theta\gamma}{1-\theta}\frac{|\mathbf{x}|^{1-\theta}}{\mathrm{e}}.$$

Therefore, a sufficient condition for (58) is

$$\frac{|\mathbf{x}|^{1-\theta}a}{g(c)} \geq \log \frac{4\kappa^2(1+c_\gamma)}{\delta\|\mathcal{T}\|} + \frac{\theta\gamma}{\mathrm{e}(1-\theta)}|\mathbf{x}|^{1-\theta}, \quad g(c) = \frac{32\kappa^2}{(\sqrt{9+24c}-3)^2}.$$

From the definition of a in (56),

$$a = g(c)\left(\log \frac{4\kappa^2(c_\gamma+1)}{\delta\|\mathcal{T}\|} + \frac{\theta\gamma}{\mathrm{e}(1-\theta)_+}\right),$$

and by a direct calculation, one can prove that the condition (58) is satisfied. The proof is complete. □

We also need the following lemma, which enables one to derive convergence results in expectation from convergence results in high probability.

Lemma 15. *Let $F :]0,1] \to \mathbb{R}_+$ be a monotone non-increasing, continuous function, and ξ a nonnegative real random variable such that*

$$\Pr[\xi > F(t)] \leq t, \quad \forall t \in (0,1].$$

Then

$$\mathbb{E}[\xi] \leq \int_0^1 F(t)dt.$$

The proof of the above lemma can be found in, e.g., (Blanchard & Mücke, 2017). Now we are ready to state and prove the following result for the local bias.

Proposition 2. *Under Assumptions 3 and 4, we let $\tilde{\lambda} = n^{-1+\theta}$ for some $\theta \in [0,1]$. Then for any $t \in [T]$, the following results hold.*
1) For $0 < \zeta \leq 1$,

$$\mathbb{E}\|\mathcal{S}_\rho h_{t+1} - f_\rho\|_\rho^2 \leq C_5\left(1 \vee \frac{\tilde{\lambda}^2}{\lambda_t^2} \vee [\gamma(\theta^{-1}\wedge\log n)]^{2\zeta\vee 1}\right)\lambda_t^{2\zeta}, \quad \lambda_t = \frac{1}{\sum_{k=1}^t \eta_k}.$$

2)For $\zeta > 1$,

$$\mathbb{E}\|\mathcal{S}_\rho h_{t+1} - f_\rho\|_\rho^2 \leq C_6\left(1 \vee \frac{\tilde{\lambda}^{2\zeta}}{\lambda_t^{2\zeta}} \vee \lambda_t^{1-2\zeta}\left(\frac{1}{n}\right)^{(\zeta-\frac{1}{2})\wedge 1} \vee [\gamma(\theta^{-1}\wedge\log n)]\right)\lambda_t^{2\zeta}.$$

Here, C_5 and C_6 are positive constants depending only on κ, ζ, R, M and can be given explicitly in the proof.

Remark 2. *It should be noted that the constants C_5 and C_6 can be further optimized if one considers a delicate but fundamental calculation in the proof, or one considers the special case, e.g., $\gamma = 0$.*

Proof. We will use Lemma 8 to prove the results. To do so, we need to estimate $\Delta_1^{\mathbf{z}}$, $\Delta_2^{\mathbf{z}}$ and $\Delta_3^{\mathbf{z}}$.

By Lemma 14, we have that with probability at least $1-\delta$,

$$\Delta_1^{\mathbf{z}} \leq 3a_{n,\delta,\gamma}(1-\theta) \leq (1 \vee \gamma[\theta^{-1}\wedge\log n])24\kappa^2 \log\frac{4\kappa^2\mathrm{e}(c_\gamma+1)}{\delta\|\mathcal{T}\|}, \tag{61}$$

where $a_{n,\delta,\gamma}(1-\theta) = a_{n,\delta,\gamma}(2/3, 1-\theta)$, given by (57). By Lemma 10, we have that with probability at least $1-\delta$,

$$\Delta_2^{\mathbf{z}} \leq 2\kappa\left(\frac{2\|r_{t+1}-f_\rho\|_\infty}{n} + \frac{\|\mathcal{S}_\rho r_{t+1} - f_\rho\|_\rho}{\sqrt{n}}\right)\log\frac{2}{\delta}.$$

Applying Part 1) of Lemma 5 with $a = 0$ to estimate $\|\mathcal{S}_\rho r_{t+1} - f_\rho\|_\rho$, we get that with probability at least $1-\delta$,

$$\Delta_2^{\mathbf{z}} \leq 2\kappa\left(2\|r_{t+1}-f_\rho\|_\infty/n + (\zeta/\mathrm{e})^\zeta R\lambda_t^\zeta/\sqrt{n}\right).$$

When $\zeta \geq 1/2$, we know that there exists a $f_H \in H$ such that $\mathcal{S}_\rho f_H = f_\rho$ (Steinwart & Christmann, 2008) and thus

$$\begin{aligned}\|r_{t+1} - f_\rho\|_\infty =& \|r_{t+1} - f_H\|_\infty \\ \leq& \kappa \|r_{t+1} - f_H\|_H \\ \leq& \kappa \|\mathcal{L}^{-1/2}(\mathcal{S}_\rho r_{t+1} - \mathcal{S}_\rho f_H)\|_\rho \\ \leq& \kappa \|\mathcal{L}^{-1/2}(\mathcal{S}_\rho r_{t+1} - f_\rho)\|_\rho \\ \leq& \kappa((\zeta - 1/2)/\mathrm{e})^{\zeta - 1/2} R \lambda_t^{\zeta - 1/2}.\end{aligned}$$

In the above, we used (31) for the second inequality, (26) for the third inequality, and Lemma 5 for the last inequality. When $\zeta < 1/2$, by Part 2) of Lemma 5, $\|r_{t+1}\|_H \leq R\lambda_t^{\zeta - 1/2}$. Combining with (31) and (10), we have

$$\|r_{t+1} - f_\rho\|_\infty \leq \kappa \|r_{t+1}\|_H + \|f_\rho\|_\infty \leq \kappa \lambda_t^{\zeta - 1/2} R + M.$$

From the above analysis, we get that with probability at least $1 - \delta$,

$$\Delta_2^{\mathbf{z}} \leq \log \frac{2}{\delta} \begin{cases} 2\kappa R\big(2\kappa((\zeta - 1/2)/\mathrm{e})^{\zeta - 1/2}/(\lambda_t n) + (\zeta/\mathrm{e})^\zeta/\sqrt{\lambda_t n}\big)\lambda_t^{\zeta + 1/2}, & \text{if } \zeta \geq 1/2, \\ 2\kappa\big(2\kappa R/(\lambda_t n) + 2M(n\lambda_t)^{-\zeta - 1/2} + (\zeta/\mathrm{e})^\zeta R/\sqrt{n\lambda_t}\big)\lambda_t^{\zeta + 1/2}, & \text{if } \zeta < 1/2, \end{cases}$$

which can be further relaxed as

$$\Delta_2^{\mathbf{z}} \leq C_4 (1 \vee (\lambda_t n)^{-1}) \lambda_t^{\zeta + 1/2} \log \frac{2}{\delta}, \tag{62}$$

where

$$C_4 \leq \begin{cases} 2\kappa R\big(2\kappa((\zeta - 1/2)/\mathrm{e})^{\zeta - 1/2} + (\zeta/\mathrm{e})^\zeta\big), & \text{if } \zeta \geq 1/2, \\ 2\kappa\big(2\kappa R + 2M + (\zeta/\mathrm{e})^\zeta R\big), & \text{if } \zeta < 1/2. \end{cases}$$

Applying Lemma 11, and combining with the fact that $\|\mathcal{T} - \mathcal{T}_{\mathbf{x}}\| \leq \|\mathcal{T} - \mathcal{T}_{\mathbf{x}}\|_{HS}$, we have that with probability at least $1 - \delta$,

$$\Delta_3^{\mathbf{z}} \leq \frac{6\kappa^2}{\sqrt{n}} \log \frac{2}{\delta}. \tag{63}$$

For $0 < \zeta \leq 1$, by Pat 1) of Lemma 8, (61) and (62), we have that with probability at least $1 - 2\delta$,

$$\|\mathcal{S}_\rho h_{t+1} - f_\rho\|_\rho \leq \left(3^{\zeta \vee \frac{1}{2}} C_1 a_{n,\delta,\gamma}^{\zeta \vee \frac{1}{2}}(1 - \theta) + 2\sqrt{3} C_4 a_{n,\delta,\gamma}^{\frac{1}{2}}(1 - \theta) \log \frac{2}{\delta}\right) \left(1 \vee \left(\frac{\tilde{\lambda}}{\lambda_t}\right)^{\zeta \vee \frac{1}{2}} \vee \frac{1}{n\lambda_t}\right) \lambda_t^\zeta.$$

Rescaling δ, and then combining with Lemma 15, we get

$$\begin{aligned}&\mathbb{E}\|\mathcal{S}_\rho h_{t+1} - f_\rho\|_\rho^2 \\ \leq& \int_0^1 \left(3^{\zeta \vee \frac{1}{2}} C_1 a_{n,\delta/2,\gamma}^{\zeta \vee \frac{1}{2}}(1 - \theta) + 2\sqrt{3} C_4 a_{n,\delta/2,\gamma}^{\frac{1}{2}}(1 - \theta) \log \frac{4}{\delta}\right)^2 d\delta \left(1 \vee \left(\frac{\tilde{\lambda}}{\lambda_t}\right)^{2\zeta \vee 1} \vee \frac{1}{n^2 \lambda_t^2}\right) \lambda_t^{2\zeta}.\end{aligned}$$

By a direct computation, noting that since $\tilde{\lambda} \geq n^{-1}$ and $2\zeta \leq 2$,

$$1 \vee \left(\frac{\tilde{\lambda}}{\lambda_t}\right)^{2\zeta \vee 1} \vee \frac{1}{n^2 \lambda_t^2} \leq 1 \vee \left(\frac{\tilde{\lambda}}{\lambda_t}\right)^2,$$

and that for all $b \in \mathbb{R}_+$,

$$\int_0^1 \log^b \frac{1}{t} dt = \Gamma(b + 1), \tag{64}$$

one can prove the first desired result with

$$C_5 = 2[C_1^2 (48\kappa^2)^{2\zeta \vee 1}(A^{2\zeta \vee 1} + 2) + 192\kappa^2 C_4^2 (A(\log^2 4 + 2 + 2\log 4) + \log^2 4 + 4\log 4 + 6)],\ A = \log \frac{8\kappa^2 (c_\gamma + 1)\mathrm{e}}{\|\mathcal{T}\|}.$$

For $\zeta > 1$, by Part 2) of Lemma 8, (61), (62) and (63), we know that with probability at least $1 - 3\delta$,

$$\begin{aligned}&\|\mathcal{S}_\rho h_{t+1} - f_\rho\|_\rho \\ &\leq \sqrt{3}(C_2 + 2C_4 + 6\kappa^2 C_3) a_{n,\delta,\gamma}^{\frac{1}{2}}(1-\theta)\log\frac{2}{\delta}\left(1 \vee \frac{\tilde{\lambda}^\zeta}{\lambda_t^\zeta} \vee \frac{1}{n\lambda_t} \vee \lambda_t^{\frac{1}{2}-\zeta}\left(\frac{1}{n}\right)^{\frac{(\zeta-\frac{1}{2})\wedge 1}{2}}\right)\lambda_t^\zeta.\end{aligned}$$

Rescaling δ, and applying Lemma 15, we get

$$\begin{aligned}&\mathbb{E}\|\mathcal{S}_\rho h_{t+1} - f_\rho\|_\rho^2 \\ &\leq 3(C_2 + 2C_4 + 6\kappa^2 C_3)^2 \int_0^1 a_{n,\delta/3,\gamma}(1-\theta)\log^2\frac{6}{\delta} d\delta \left(1 \vee \frac{\tilde{\lambda}^{2\zeta}}{\lambda_t^{2\zeta}} \vee \frac{1}{n^2\lambda_t^2} \vee \lambda_t^{1-2\zeta}\left(\frac{1}{n}\right)^{(\zeta-\frac{1}{2})\wedge 1}\right)\lambda_t^{2\zeta}.\end{aligned}$$

This leads to the second desired result with

$$C_6 = 24\kappa^2(C_2 + 2C_4 + 6\kappa^2 C_3)^2((A+1)\log^2 6 + 2(A+2)\log 6 + 2A + 6),\ A = \log\frac{12\kappa^2(c_\gamma+1)\mathrm{e}}{\|\mathcal{T}\|},$$

by noting that $n^{-1} \leq \tilde{\lambda}$. The proof is complete. □

Combining Proposition 2 with Lemma 1, we get the following results for the bias of the fully averaged estimator.

Proposition 3. *Under Assumptions 3 and 4, let $0 < \zeta \leq 1$. For any $\tilde{\lambda} = n^{-1+\theta}$ with $\theta \in [0,1]$ and any $t \in [T]$, there holds*

$$\mathbb{E}\|\mathcal{S}_\rho \bar{h}_{t+1} - f_\rho\|_\rho^2 \leq C_5\left(1 \vee \frac{\tilde{\lambda}^2}{\lambda_t^2} \vee [\gamma(\theta^{-1} \wedge \log n)]^{2\zeta\vee 1}\right)\lambda_t^{2\zeta}, \quad \lambda_t = \frac{1}{\sum_{k=1}^t \eta_k}. \tag{65}$$

Here, C_5 is given by Proposition 2.

D. Estimating Sample Variance

In this section, we estimate sample variance $\|\mathcal{S}_\rho(\bar{g}_t - \bar{h}_t)\|_\rho$. We first introduce the following lemma.

Lemma 16. *For any $t \in [T]$, we have*

$$\mathbb{E}\|\mathcal{S}_\rho(\bar{g}_t - \bar{h}_t)\|_\rho = \frac{1}{m}\mathbb{E}\|\mathcal{S}_\rho(g_{1,t} - h_{1,t})\|_\rho^2. \tag{66}$$

Proof. Note that from the independence of $\mathbf{z}_1, \cdots, \mathbf{z}_m$ and (33), we have

$$\mathbb{E}_{\bar{\mathbf{y}}}\|\mathcal{S}_\rho(\bar{g}_t - \bar{h}_t)\|_\rho = \frac{1}{m^2}\sum_{s,l=1}^m \mathbb{E}_{\bar{\mathbf{y}}}\langle\mathcal{S}_\rho(g_{s,t} - h_{s,t}), \mathcal{S}_\rho(g_{l,t} - h_{l,t})\rangle_\rho = \frac{1}{m^2}\sum_{s=1}^m \mathbb{E}_{\mathbf{y}_s}\|\mathcal{S}_\rho(g_{s,t} - h_{s,t})\|_\rho^2.$$

Taking the expectation with respect to $\bar{\mathbf{x}}$, we get

$$\mathbb{E}\|\mathcal{S}_\rho(\bar{g}_t - \bar{h}_t)\|_\rho = \frac{1}{m^2}\sum_{s=1}^m \mathbb{E}\|\mathcal{S}_\rho(g_{s,t} - h_{s,t})\|_\rho^2 = \frac{1}{m}\mathbb{E}\|\mathcal{S}_\rho(g_{1,t} - h_{1,t})\|_\rho^2.$$

The proof is complete. □

According to Lemma 16, we know that the sample variance of the averaging over m local estimators can be well controlled in terms of the sample variance of a local estimator. In what follows, we will estimate the local sample variance, $\mathbb{E}\|\mathcal{S}_\rho(g_{1,t} - h_{1,t})\|_\rho^2$. Throughout the rest of this subsection, we shall drop the index $s = 1$ for the first local estimator whenever it shows up, i.e., we rewrite $g_{1,t}$ as g_t, $\mathbf{z}_1$ as $\mathbf{z}$, etc.

Proposition 4. *Under Assumption 4, let $\tilde{\lambda} = n^{\theta-1}$ for some $\theta \in [0,1]$. Then for any $t \in [T]$,*

$$\mathbb{E}\|\mathcal{S}_\rho(g_{t+1} - h_{t+1})\|_\rho^2 \leq C_8\frac{1}{n\tilde{\lambda}^\gamma}\left(1 \vee \frac{\tilde{\lambda}}{\lambda_t} \vee [\gamma(\theta^{-1} \wedge \log n])\right).$$

Here, C_8 is a positive constant depending only on $\sigma, \kappa, \gamma, c_\gamma, \|\mathcal{T}\|$ and will be given explicitly in the proof.

Proof. Following from Lemma 2,

$$g_{t+1} - h_{t+1} = G_t(\mathcal{T}_{\mathbf{x}})(\mathcal{S}_{\mathbf{x}}^*\mathbf{y} - \mathcal{L}_{\mathbf{x}} f_\rho).$$

For notational simplicity, we let $\epsilon_i = y_i - f_\rho(x_i)$ for all $i \in [n]$ and $\boldsymbol{\epsilon} = (\epsilon_i)_{1\le i\le n}$. Then the above can be written as

$$g_{t+1} - h_{t+1} = G_t(\mathcal{T}_{\mathbf{x}})\mathcal{S}_{\mathbf{x}}^*\boldsymbol{\epsilon}.$$

Using the above relationship and the isometric property (25), we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}}\|\mathcal{S}_\rho(g_{t+1} - h_{t+1})\|_\rho^2 &= \mathbb{E}_{\mathbf{y}}\|\mathcal{S}_\rho G_t(\mathcal{T}_{\mathbf{x}})\mathcal{S}_{\mathbf{x}}^*\boldsymbol{\epsilon}\|_\rho^2 \\
&= \mathbb{E}_{\mathbf{y}}\|\mathcal{T}^{1/2} G_t(\mathcal{T}_{\mathbf{x}})\mathcal{S}_{\mathbf{x}}^*\boldsymbol{\epsilon}\|_H^2 \\
&= \frac{1}{n^2}\sum_{l,k=1}^n \mathbb{E}_{\mathbf{y}}[\epsilon_l\epsilon_k]\,\mathrm{tr}\left(G_t(\mathcal{T}_{\mathbf{x}})\mathcal{T}G_t(\mathcal{T}_{\mathbf{x}})K_{x_l}\otimes K_{x_k}\right).
\end{aligned}$$

From the definition of f_ρ and the independence of z_l and z_k when $l \neq k$, we know that $\mathbb{E}_{\mathbf{y}}[\epsilon_l\epsilon_k] = 0$ whenever $l \neq k$. Therefore,

$$\mathbb{E}_{\mathbf{y}}\|\mathcal{S}_\rho(g_{t+1} - h_{t+1})\|_\rho^2 = \frac{1}{n^2}\sum_{k=1}^n \mathbb{E}_{\mathbf{y}}[\epsilon_k^2]\,\mathrm{tr}\left(G_t(\mathcal{T}_{\mathbf{x}})\mathcal{T}G_t(\mathcal{T}_{\mathbf{x}})K_{x_k}\otimes K_{x_k}\right).$$

Using Assumption 2,

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}}\|\mathcal{S}_\rho(g_{t+1} - h_{t+1})\|_\rho^2 \le& \frac{\sigma^2}{n^2}\sum_{k=1}^n \mathrm{tr}\left(G_t(\mathcal{T}_{\mathbf{x}})\mathcal{T}G_t(\mathcal{T}_{\mathbf{x}})K_{x_k}\otimes K_{x_k}\right) \\
=& \frac{\sigma^2}{n}\,\mathrm{tr}\left(\mathcal{T}(G_t(\mathcal{T}_{\mathbf{x}}))^2\mathcal{T}_{\mathbf{x}}\right) \\
\le& \frac{\sigma^2}{n}\,\mathrm{tr}(\mathcal{T}_{\tilde\lambda}^{-1/2}\mathcal{T}\mathcal{T}_{\tilde\lambda}^{-1/2})\|\mathcal{T}_{\tilde\lambda}^{1/2}G_t(\mathcal{T}_{\mathbf{x}})^2\mathcal{T}_{\mathbf{x}}\mathcal{T}_{\tilde\lambda}^{1/2}\| \\
\le& \frac{\sigma^2\mathcal{N}(\tilde\lambda)}{n}\|\mathcal{T}_{\tilde\lambda}^{1/2}\mathcal{T}_{\mathbf{x}\tilde\lambda}^{-1/2}\|\|\mathcal{T}_{\mathbf{x}\tilde\lambda}^{1/2}G_t(\mathcal{T}_{\mathbf{x}})^2\mathcal{T}_{\mathbf{x}}\mathcal{T}_{\mathbf{x}\tilde\lambda}^{1/2}\|\|\mathcal{T}_{\mathbf{x}\tilde\lambda}^{-1/2}\mathcal{T}_{\tilde\lambda}^{1/2}\| \\
\le& \frac{\sigma^2\mathcal{N}(\tilde\lambda)}{n}\Delta_1^{\mathbf{z}}\|G_t(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}}\|\|G_t(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}\tilde\lambda}\| \\
\le& \frac{\sigma^2\mathcal{N}(\tilde\lambda)}{n}\Delta_1^{\mathbf{z}}(1+\tilde\lambda/\lambda_t),
\end{aligned}$$

where $\Delta_1^{\mathbf{z}}$ is given by Lemme 8 and we used 1) of Lemma 4 for the last inequality. Taking the expectation with respect to $\mathbf{x}$, this leads to

$$\mathbb{E}\|\mathcal{S}_\rho(g_{t+1} - h_{t+1})\|_\rho^2 \le \frac{\sigma^2\mathcal{N}(\tilde\lambda)}{n}(1+\tilde\lambda/\lambda_t)\mathbb{E}[\Delta_1^{\mathbf{z}}].$$

Applying Lemmas 14 and 15, we get

$$\begin{aligned}
\mathbb{E}\|\mathcal{S}_\rho(g_{t+1} - h_{t+1})\|_\rho^2 \le& 6\frac{\sigma^2\mathcal{N}(\tilde\lambda)}{n}(1\vee(\tilde\lambda/\lambda_t))\int_0^1 a_{n,\delta,\gamma}(2/3, 1-\theta)d\delta \\
\le& C_7\frac{\sigma^2\mathcal{N}(\tilde\lambda)}{n}(1\vee(\tilde\lambda/\lambda_t)\vee[\gamma(\theta^{-1}\wedge\log n])),
\end{aligned}$$

where $C_7 = 48\kappa^2\log\frac{4\kappa^2(c_\gamma+1)\mathrm{e}}{\|\mathcal{T}\|}$. Using Assumption 4, we get the desired result with $C_8 = c_\gamma C_7\sigma^2$. □

Using the above proposition and Lemma 16, we derive the following results for sample variance.

Proposition 5. *Under Assumption 4, let $\tilde\lambda = n^{\theta-1}$ for some $\theta \in [0,1]$. Then for any $t \in [T]$,*

$$\mathbb{E}\|\mathcal{S}_\rho(\bar g_{t+1} - \bar h_{t+1})\|_\rho^2 \le C_8\frac{1}{N\tilde\lambda^\gamma}\left(1\vee\left(\frac{\tilde\lambda}{\lambda_t}\right)\vee[\gamma(\theta^{-1}\wedge\log n)]\right), \quad \lambda_t = \frac{1}{\sum_{k=1}^t\eta_k}. \tag{67}$$

Here, C_8 is a positive constant depending only on $\kappa^2, c_\gamma, \|\mathcal{T}\|$ and σ^2.

E. Estimating Computational Variance

In this section, we estimate computational variance, $\mathbb{E}[\|\mathcal{S}_\rho(\bar{f}_t - \bar{h}_t)\|_\rho^2]$. We begin with the following lemma, from which we can see that the global computational variance can be estimated in terms of local computational variances.

Lemma 17. *For any $t \in [T]$, we have*

$$\mathbb{E}\|\mathcal{S}_\rho(\bar{f}_t - \bar{g}_t)\|_\rho = \frac{1}{m^2}\sum_{s=1}^{m}\mathbb{E}\|\mathcal{S}_\rho(f_{s,t} - g_{s,t})\|_\rho^2. \tag{68}$$

Proof. Note that by (32) and from the conditional independence of $\mathbf{J}_s, \cdots \mathbf{J}_m$ (given $\bar{\mathbf{z}}$), we have

$$\mathbb{E}_{\mathbf{J}}\|\mathcal{S}_\rho(\bar{f}_t - \bar{g}_t)\|_\rho = \frac{1}{m^2}\sum_{s,l=1}^{m}\mathbb{E}_{\mathbf{J}}\langle\mathcal{S}_\rho(f_{s,t} - g_{s,t}), \mathcal{S}_\rho(f_{l,t} - g_{l,t})\rangle_\rho = \frac{1}{m^2}\sum_{s=1}^{m}\mathbb{E}_{\mathbf{J}_s}\|\mathcal{S}_\rho(f_{s,t} - g_{s,t})\|_\rho^2.$$

Taking the expectation with respect to $\bar{\mathbf{z}}$, we thus prove the desired result. The proof is complete. □

In what follows, we will estimate the local computational variance, i.e., $\mathbb{E}\|\mathcal{S}_\rho(f_{s,t} - g_{s,t})\|_\rho^2$. As in Subsections C and D, we will drop the index s for the s-th local estimator whenever it shows up. We first introduce the following two lemmas, whose proof can be found in (Lin & Rosasco, 2017b). The empirical risk $\mathcal{E}_{\mathbf{z}}(f)$ of a function f with respect to the samples $\mathbf{z}$ is defined as

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{n}\sum_{(x,y)\in\mathbf{z}}(f(x) - y)^2.$$

Lemma 18. *Assume that for all $t \in [T]$ with $t \geq 2$,*

$$\frac{1}{\eta_t}\sum_{k=1}^{t-1}\frac{1}{k(k+1)}\sum_{i=t-k}^{t-1}\eta_i^2 \leq \frac{1}{4\kappa^2}. \tag{69}$$

Then for all $t \in [T]$,

$$\sup_{k\in[t]}\mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(f_k)] \leq \frac{8\mathcal{E}_{\mathbf{z}}(0)\Sigma_1^t}{\eta_t t}. \tag{70}$$

Lemma 19. *For any $t \in [T]$, we have*

$$\mathbb{E}_{\mathbf{J}}\|\mathcal{S}_\rho f_{t+1} - \mathcal{S}_\rho g_{t+1}\|_\rho^2 \leq \frac{\kappa^2}{b}\sum_{k=1}^{t}\eta_k^2\left\|\mathcal{T}^{\frac{1}{2}}\Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\right\|^2\mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(f_k)]. \tag{71}$$

Now, we are ready to state and prove the result for local computational variance as follows.

Proposition 6. *Assume that (70) holds for any $t \in [T]$ with $t \geq 2$. Let $\tilde{\lambda} = n^{-\theta+1}$ for some $\theta \in [0,1]$. For any $t \in [T]$,*

$$\mathbb{E}\|\mathcal{S}_\rho f_{t+1} - \mathcal{S}_\rho g_{t+1}\|_\rho^2 \leq C_9(1 \vee [\gamma(\theta^{-1} \wedge \log n)])b^{-1}\sup_{k\in[t]}\left\{\frac{\Sigma_1^k}{\eta_k k}\right\}\left(\sum_{k=1}^{t-1}\eta_k^2(\tilde{\lambda} + \lambda_{k+1:t}\mathrm{e}^{-1}) + \eta_t^2\right).$$

Here, C_9 is a positive constant depending only on $\kappa, M, c_\gamma, \|\mathcal{T}\|$ and can be given explicitly in the proof.

Proof. Following from Lemmas 19 and 18, we have that,

$$\mathbb{E}_{\mathbf{J}}\|\mathcal{S}_\rho f_{t+1} - \mathcal{S}_\rho g_{t+1}\|_\rho^2 \leq \frac{8\kappa^2\mathcal{E}_{\mathbf{z}}(0)}{b}\sum_{k=1}^{t}\eta_k^2\left\|\mathcal{T}^{\frac{1}{2}}\Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\right\|^2\sup_{k\in[t]}\left\{\frac{\Sigma_1^k}{\eta_k k}\right\}.$$

Taking the expectation with respect to $\mathbf{y}|\mathbf{x}$ and then with respect to $\mathbf{x}$, noting that $\int_Y y^2 d\rho(y|x) \leq M$, we get

$$\mathbb{E}\|\mathcal{S}_\rho f_{t+1} - \mathcal{S}_\rho g_{t+1}\|_\rho^2 \leq \frac{8\kappa^2 M^2}{b}\sup_{k\in[t]}\left\{\frac{\Sigma_1^k}{\eta_k k}\right\}\sum_{k=1}^{t}\eta_k^2\mathbb{E}\left\|\mathcal{T}^{\frac{1}{2}}\Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\right\|^2.$$

Note that

$$\left\|\mathcal{T}^{\frac{1}{2}}\Pi_k^t(\mathcal{T}_{\mathbf{x}})\right\|^2 \leq \|\mathcal{T}^{\frac{1}{2}}\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-1/2}\|^2\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2}\Pi_k^t(\mathcal{T}_{\mathbf{x}})\|^2 \leq \Delta_1^{\mathbf{z}}\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}(\Pi_k^t(\mathcal{T}_{\mathbf{x}}))^2\|$$
$$\leq \Delta_1^{\mathbf{z}}(\|\mathcal{T}_{\mathbf{x}}\Pi_k^t(\mathcal{T}_{\mathbf{x}})\| + \tilde{\lambda}\|\Pi_k^t(\mathcal{T}_{\mathbf{x}})\|)\|\Pi_k^t(\mathcal{T}_{\mathbf{x}})\| \leq \Delta_1^{\mathbf{z}}(\lambda_{k:t}\mathrm{e}^{-1} + \tilde{\lambda}),$$

where $\Delta_1^{\mathbf{z}}$ is given by Lemma 8 and for the last inequality we used Part 2) of Lemma 4. Therefore,

$$\mathbb{E}\|\mathcal{S}_\rho f_{t+1} - \mathcal{S}_\rho g_{t+1}\|_\rho^2 \leq \mathbb{E}[\Delta_1^{\mathbf{z}}]\frac{8\kappa^2M^2}{b}\sup_{k\in[t]}\left\{\frac{\Sigma_1^k}{\eta_k k}\right\}\left(\sum_{k=1}^{t-1}\eta_k^2(\tilde{\lambda} + \lambda_{k+1:t}\mathrm{e}^{-1}) + \eta_t^2\right).$$

Using Lemmas 14 and 15, and by a simple calculation, one can upper bound $\mathbb{E}[\Delta_1^{\mathbf{z}}]$ and consequently prove the desired result with C_9 given by

$$C_9 = 192\kappa^4M^2\log\frac{4\kappa^2(c_\gamma+1)\mathrm{e}}{\|\mathcal{T}\|}.$$

The proof is complete. □

Combining Lemma 17 with Proposition 6, we have the following error bounds for computational variance.

Proposition 7. *Assume that (70) holds for any $t \in [T]$ with $t \geq 2$. Let $\tilde{\lambda} = n^{-\theta+1}$ for some $\theta \in [0,1]$. For any $t \in [T]$,*

$$\mathbb{E}\|\mathcal{S}_\rho(\bar{f}_{t+1} - \bar{g}_{t+1}\|_\rho^2 \leq C_9(1 \vee [\gamma(\theta^{-1} \wedge \log n)])\frac{1}{mb}\sup_{k\in[t]}\left\{\frac{\Sigma_1^k}{\eta_k k}\right\}\left(\sum_{k=1}^{t-1}\eta_k^2(\tilde{\lambda} + \lambda_{k+1:t}\mathrm{e}^{-1}) + \eta_t^2\right). \tag{72}$$

Here, C_9 is the positive constant from Proposition 6.

F. Deriving Total Errors

We are now ready to derive total error bounds for (distributed) SGM and to prove the main theorems for (distributed) SGM of this paper.

Proof of Theorem 1. We will use Propositions 1, 3, 5 and 7 to prove the result.

We first show that the condition (12) implies (69). Indeed, when $\eta_t = \eta$, for any $t \in [T]$

$$\frac{1}{\eta_t}\sum_{k=1}^{t-1}\frac{1}{k(k+1)}\sum_{i=t-k}^{t-1}\eta_i^2 = \eta\sum_{k=2}^{t}\frac{1}{k} \leq \eta\sum_{k=2}^{t}\int_{k-1}^{k}\frac{1}{x}dx = \eta\log t \leq \frac{1}{4\kappa^2}$$

where for the last inequality, we used the condition (12). Thus, by Proposition 7, (72) holds. Note also that $\lambda_{k+1:t} = \frac{1}{\eta(t-k)}$ and $\lambda_t = \frac{1}{\eta t}$ as $\eta_t = \eta$. It thus follows from (72) that

$$\mathbb{E}\|\mathcal{S}_\rho(\bar{f}_{t+1} - \bar{g}_{t+1}\|_\rho^2 \leq C_9(1 \vee [\gamma(\theta^{-1} \wedge \log n)])\frac{\eta}{mb}\left(\tilde{\lambda}\eta(t-1) + \sum_{k=1}^{t-1}\frac{1}{\mathrm{e}(t-k)} + \eta\right).$$

Applying

$$\sum_{k=1}^{t-1}\frac{1}{t-k} = \sum_{k=1}^{t-1}\frac{1}{k} \leq 1 + \sum_{k=2}^{t-1}\int_{k-1}^{k}\frac{1}{x}dx \leq 1 + \log t,$$

and (12), we get

$$\mathbb{E}\|\mathcal{S}_\rho(\bar{f}_{t+1} - \bar{g}_{t+1}\|_\rho^2 \leq C_9(1 \vee [\gamma(\theta^{-1} \wedge \log n)] \vee \tilde{\lambda}\eta t \vee \log t)\frac{\eta}{mb}\left(2 + \frac{1}{4\kappa^2}\right).$$

Introducing the above inequality, (65), and (67) into the error decomposition (23), by a direct calculation, one can prove the desired result. The proof is complete. □

Proof of Corollary 2. In Theorem 1, we let $\tilde{\lambda} = N^{-\frac{1}{2\zeta+\gamma}}$. In this case, with Condition (15), it is easy to show that

$$1 \geq \theta = \frac{\log\tilde{\lambda}}{\log n} + 1 = \frac{\log\tilde{\lambda}}{\log N - \log m} + 1 \geq -\frac{1}{2\zeta+\gamma}\frac{\log N}{\log N - \beta\log N} + 1 > 0.$$

The proof can be done by simply applying Theorem 1 and plugging with the specific choices of η_t, b, and T_*. □

Proof of Corollary 1. Since $f_\rho \in H$, we know from (26) that Assumption 3 holds with $\zeta = \frac{1}{2}$ and $R \leq \|f_\rho\|_H$. As noted in comments after Assumption 4, (11) trivially holds with $\gamma = 1$ and $c_\gamma = \kappa^2$. Applying Corollary 2, one can prove the desired results. □

Proof of Theorem 2. When $\zeta \leq 1$, we apply Theorem 1 with $m = 1$ and $n = N$ to get

$$\mathbb{E}\|\mathcal{S}_\rho \bar{f}_{t+1} - f_\rho\|_\rho^2 \lesssim ((\tilde{\lambda}\eta t)^2 \vee [\gamma(\theta^{-1} \wedge \log N)]^{2\zeta \vee 1} \vee 1 \vee \log t)[\frac{1}{(\eta t)^{2\zeta}} + \frac{1}{N\tilde{\lambda}^\gamma} + \frac{\eta}{b}]. \tag{73}$$

We let $\tilde{\lambda} = N^{\theta-1}$ with $\theta = 1 - \alpha$. Then it is easy to see that

$$\gamma(\theta^{-1} \wedge \log N) \leq \begin{cases} \frac{\gamma(2\zeta+\gamma)}{2\zeta+\gamma-1}, & \text{if } 2\zeta + \gamma > 1, \\ \gamma \log N, & \text{if } 2\zeta + \gamma \leq 1. \end{cases}$$

Following from the aboves and plugging with the specific choices on η_t, T_*, b, one can prove the desired error bounds for the case $\zeta \leq 1$.

The proof for the case $\zeta > 1$ is similar as that for the case $\zeta \leq 1$. Following the same lines as those for (73) (with Proposition 2.(1) replaced by Proposition 2.(2)), we get

$$\mathbb{E}\|\mathcal{S}_\rho \bar{f}_{t+1} - f_\rho\|_\rho^2 \lesssim ((\tilde{\lambda}\eta t)^{2\zeta} \vee [\gamma(\theta^{-1} \wedge \log N)] \vee (\frac{(\eta t)^{2\zeta-1}}{N^{(\zeta-1/2)\wedge 1}}) \vee 1 \vee \log t)[\frac{1}{(\eta t)^{2\zeta}} + \frac{1}{N\tilde{\lambda}^\gamma} + \frac{\eta}{b}].$$

Letting $\tilde{\lambda} = N^{-\alpha}$ and plugging with the specific choices on η_t, T_*, b and $\theta = 1 - \alpha$, one can prove the desired result for the case $\zeta \geq 1$. □