Supplementary Materials: SGD Meets Distribution Regression

Contents.

1 First Results

We introduce some auxiliary operators, being useful in our proofs. These operators have been introduced in a variety of previous works, see e.g. Caponnetto and De Vito (2006); Dieuleveut and Bach (2016); Blanchard and Mücke (2018). Recall that $S_K : \mathcal{H}_K \hookrightarrow L^2(\mu(\mathcal{M}^+(\mathcal{X})), \rho_\mu)$ denotes the canonical injection map. The adjoint $S_K^* : L^2(\mu(\mathcal{M}^+(\mathcal{X})), \rho_\mu) \to \mathcal{H}_K$ is given by

$$S_K^* g = \int_{\mu(\mathcal{M}^+(\mathcal{X}))} g(\tilde{\mu}) K_{\tilde{\mu}} \, \rho_\mu(d\tilde{\mu}) \, ,$$

where we remind at the notation $K_{\tilde{\mu}} = K(\tilde{\mu}, \cdot)$. The covariance operator is $T_K := S_K^* S_K : \mathcal{H}_K \to \mathcal{H}_K$, with

$$T_K f = \int_{\mu(\mathcal{M}^+(\mathcal{X}))} \langle K_{\tilde{\mu}}, f \rangle_{\mathcal{H}_K} K_{\tilde{\mu}} \, \rho_\mu(d\tilde{\mu})$$

and the kernel integral operator $L_K : L^2(\mu(\mathcal{M}^+(\mathcal{X})), \rho_\mu) \to L^2(\mu(\mathcal{M}^+(\mathcal{X})), \rho_\mu)$ is

$$L_K g = \int_{\mu(\mathcal{M}^+(\mathcal{X}))} g(\tilde{\mu}) K(\tilde{\mu}, \cdot) \, \rho_\mu(d\tilde{\mu}) \, .$$

We further introduce the empirical counterparts:

$$T_{\mathbf{x}} := \frac{1}{n} \sum_{j=1}^n \Big\langle K_{\mu_{x_j}}, \cdot \Big\rangle_{\mathcal{H}_K} K_{\mu_{x_j}} \, , \qquad T_{\hat{\mathbf{x}}} := \frac{1}{n} \sum_{j=1}^n \Big\langle K_{\mu_{\hat{x}_j}}, \cdot \Big\rangle_{\mathcal{H}_K} K_{\mu_{\hat{x}_j}}$$

$$g_{\mathbf{z}} := \frac{1}{n} \sum_{j=1}^n y_j K_{\mu_{x_j}} \, , \qquad g_{\hat{\mathbf{z}}} := \frac{1}{n} \sum_{j=1}^n y_j K_{\mu_{\hat{x}_j}} \, ,$$

where $T_{\mathbf{x}}, T_{\hat{\mathbf{x}}} : \mathcal{H}_K \to \mathcal{H}_K$, $g_{\mathbf{z}}, g_{\hat{\mathbf{z}}} \in \mathcal{H}_K$.

We collect some preliminary results.

Lemma 1.1. *Suppose Assumptions 2.1, 2.2 and 3.1 are satisfied. Then*

$$\mathbb{E}_{\hat{D}|D}[||T_K(T_{\hat{\mathbf{x}}} + \lambda Id)^{-1}||] \leq \frac{1}{\lambda}||T_K - T_{\mathbf{x}}|| + \frac{c_\alpha \gamma^\alpha LM}{\sqrt{\lambda} N^{\frac{\alpha}{2}}} + 1 \, ,$$

for some $c_\alpha < \infty$. Moreover, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ w.r.t. the data D, one has

$$\mathbb{E}_{\hat{D}|D}[||T_K(T_{\hat{\mathbf{x}}} + \lambda Id)^{-1}||] \leq 6 \log(2/\delta) \frac{1}{\lambda \sqrt{n}} + \frac{c_\alpha \gamma^\alpha LM}{\sqrt{\lambda} N^{\frac{\alpha}{2}}} + 1 \, .$$

Proof of 1.1. Let us bound for any $\lambda > 0$

$$\begin{aligned} \mathbb{E}_{\hat{D}|D}[||T_K(T_{\hat{\mathbf{x}}} + \lambda Id)^{-1}||] &\leq ||T_K(T_K + \lambda Id)^{-1}|| \cdot \mathbb{E}_{\hat{D}|D}[||(T_K + \lambda Id)(T_{\hat{\mathbf{x}}} + \lambda Id)^{-1}||] \\ &\leq \mathbb{E}_{\hat{D}|D}[||(T_K + \lambda Id)(T_{\hat{\mathbf{x}}} + \lambda Id)^{-1}||] \, . \end{aligned}$$

We proceed by writing

$$\begin{aligned} (T_K + \lambda Id)(T_{\hat{\mathbf{x}}} + \lambda Id)^{-1} &= (T_K + \lambda Id)((T_{\hat{\mathbf{x}}} + \lambda Id)^{-1} - (T_K + \lambda Id)^{-1}) + (T_K + \lambda Id)(T_K + \lambda Id)^{-1} \\ &= ((T_K - T_{\mathbf{x}}) + (T_{\mathbf{x}} - T_{\hat{\mathbf{x}}}))(T_{\hat{\mathbf{x}}} + \lambda Id)^{-1} + Id \, . \end{aligned}$$

Since $||(T_{\hat{\mathbf{x}}} + \lambda Id)^{-1}|| \leq 1/\lambda$, this leads to

$$\mathbb{E}_{\hat{D}|D}[||T_K(T_{\hat{\mathbf{x}}} + \lambda Id)^{-1}||] \leq \frac{1}{\lambda}||T_K - T_{\mathbf{x}}|| + \frac{1}{\lambda}\mathbb{E}_{\hat{D}|D}[||T_{\mathbf{x}} - T_{\hat{\mathbf{x}}}||] + 1 \, .$$

The first result follows then from Lemma 1.3 and Jensen's inequality. The second result follows from the first one by applying Proposition 5.5. in Blanchard and Mücke (2018). □

Lemma 1.2. *(Fang et al., 2020, Eq. (37)) Suppose Assumptions 2.1, 2.2 and 3.1 are satisfied. Then*

$$\mathbb{E}_{\hat{D}|D}[||g_{\hat{\mathbf{z}}} - g_{\mathbf{z}}||^2_{\mathcal{H}_K}] \leq c_\alpha L^2 M^2 \frac{\gamma^{2\alpha}}{N^\alpha} ,$$

for some $c_\alpha < \infty$.

Lemma 1.3. *(Fang et al., 2020, Eq. (38)) Suppose Assumptions 2.1, 2.2 and 3.1 are satisfied. Then*

$$\mathbb{E}_{\hat{D}|D}[||T_{\hat{\mathbf{x}}} - T_{\mathbf{x}}||^2] \leq c_\alpha \kappa^2 L^2 \frac{\gamma^{2\alpha}}{N^\alpha} ,$$

for some $c_\alpha < \infty$.

Lemma 1.4. *(Fang et al., 2020, Lemma 8) Suppose Assumptions 2.1, 2.2 and 3.1 are satisfied and let $\lambda > 0$. Define*

$$\mathcal{C}_{\mathbf{x}}(\lambda) := \left(\frac{\mathcal{A}_{\mathbf{x}}(\lambda)}{\lambda} + \frac{1}{\lambda^{\frac{3}{2}} N^{\frac{\alpha}{2}}} + \frac{1}{\sqrt{\lambda}} \right) , \tag{1}$$

where

$$\mathcal{A}_{\mathbf{x}}(\lambda) := ||(T_K + \lambda Id)^{-\frac{1}{2}}(T_K - T_{\mathbf{x}})|| .$$

Then

$$\mathbb{E}_{\hat{D}|D}[||T_K^{\frac{1}{2}}(T_{\hat{\mathbf{x}}} + \lambda Id)^{-1}||^2] \leq C^2_{\kappa,\gamma,L,\alpha} \mathcal{C}_{\mathbf{x}}(\lambda)^2 ,$$

for some $C_{\kappa,\gamma,L,\alpha} < \infty$.

2 Solving Distribution Regression with Tail-Averaged Gradient Descent

In this section we derive the learning properties of tail-averaged two-stage Gradient Descent. This is a necessary step for deriving our learning bounds for SGD on distribution regression problems and is of independent interest.

Let us begin with introducing the gradient updates using the second-stage data $D = \{(\mu_{\hat{x}_j}, y_j)\}_{j=1}^n \subset \mu(\mathcal{M}^+(\mathcal{X})) \times \mathcal{Y}$ as $\hat{f}_0 = 0$ and for $t \geq 1$

$$\hat{f}_{t+1} = \hat{f}_t - \eta \frac{1}{n} \sum_{j=1}^n (\hat{f}_t(\mu_{\hat{x}_j}) - y_j) K_{\mu_{\hat{x}_j}} .$$

Here, $\eta > 0$ is the constant step-size. We furthermore set

$$\bar{\hat{f}}_T := \frac{2}{T} \sum_{t=\lfloor T/2 \rfloor + 1}^T \hat{f}_t . \tag{2}$$

Similarly, we introduce the Gradient Descent updates using the first-stage data $D = \{(\mu_{x_j}, y_j)\}_{j=1}^n \subset \mu(\mathcal{M}^+(\mathcal{X})) \times \mathcal{Y}$ with initialization $f_0 = 0$ as

$$f_{t+1} = f_t - \eta \frac{1}{n} \sum_{j=1}^n (f_t(\mu_{x_j}) - y_j) K_{\mu_{x_j}}$$

and

$$\bar{f}_T := \frac{2}{T} \sum_{t=\lfloor T/2 \rfloor + 1}^T f_t .$$

We analyze the learning properties of (2) based on the decomposition

$$\bar{\hat{f}}_T - f_\rho = \underbrace{(\bar{\hat{f}}_T - \bar{f}_T)}_{GD\ Variance\ 2.\ stage} + \underbrace{(\bar{f}_T - f_\rho)}_{Error\ GD\ first\ stage} .$$

The bound for the second stage GD variance is derived in Section 2.3. The error estimate for first stage Gradient Descent is known from previous results. For completeness sake we review the main results in our setting in Section 2.2

2.1 A General Result

In this section we state a general result for spectral regularization algorithms. Those bounds are known for some time in learning theory, see e.g. Bauer et al. (2007). We collect some results from Caponnetto and De Vito (2006); Fischer and Steinwart (2017); Blanchard and Mücke (2018); Lin et al. (2020); Mücke et al. (2019).

We let $\{g_\lambda : [0, ||T_K||] \to [0, \infty) \ : \ \lambda \in (0, ||T_K||]\}$ be a family of filter functions (for the definition we refer to one of the above mentioned papers). We define

$$\hat{u}_\lambda := g_\lambda(T_{\mathbf{x}}) S^*_{\mathbf{x}} \mathbf{y} \ , \quad u_\lambda := g_\lambda(T_K) S^*_K f_\rho \ .$$

Our aim is to provide a bound of the estimation error in different norms $||T_K^a(\hat{u}_\lambda - u_\lambda)||_{\mathcal{H}_K}$, for $a \in [0, 1/2]$. To this end, we require a condition for the observation noise.

Assumption 2.1 (Bernstein Observation Noise)**.** *For some $\sigma > 0$, $B > 0$ and for all $m \geq 2$ we have almost surely*

$$\int_{\mathcal{Y}} |y - f_\rho(x)|^m \ \rho(dy|x) \leq \frac{1}{2} m! \sigma^2 B^{m-2} \ .$$

Finally, we need

Assumption 2.2. *Let $\lambda > 0$. Suppose that*

$$n \geq \frac{32 \kappa^2 \log(4/\delta)}{\lambda} \log\left(e \mathcal{N}(\lambda) \left(1 + \frac{\lambda}{||T_K||} \right) \right) .$$

Proposition 2.3 (Estimation Error)**.** *Let $a \in [0, 1/2]$. Let further $\delta \in (0, 1]$ and suppose Assumptions 2.1, 2.2 are satisfied. Denote*

$$B_\lambda = \max\{B, ||S_K u_\lambda - f_\rho||_\infty\} \ .$$

1. *Assume $f_\rho \in Range(L_K^\zeta)$, for some $\zeta \in (0, 1]$. With probability not less than $1 - \delta$,*

$$||T_K^a(\hat{u}_\lambda - u_\lambda)||_{\mathcal{H}_K} \leq C_1 \log(12/\delta) \frac{\lambda^{a-1/2}}{\sqrt{n}} \left(\sigma \sqrt{\mathcal{N}(\lambda)} + \frac{||S_K u_\lambda - f_\rho||_{L^2}}{\sqrt{\lambda}} + \frac{B_\lambda}{\sqrt{n\lambda}} \right) + C_2 \lambda^{a+1/2} \ ||T_\lambda^{-1/2} u_\lambda||_{\mathcal{H}_K} + C_3 \lambda^{a-1/2} ||S_K u_\lambda - f_\rho||_{L^2} \ ,$$

   for some $C_1 > 0$, $C_2 > 0$ and $C_3 > 0$.

2. *Assume $f_\rho \in Range(L_K^\zeta)$, for some $\zeta > 1$. With probability not less than $1 - \delta$,*

$$||T_K^a(\hat{u}_\lambda - u_\lambda)||_{\mathcal{H}_K} \leq C_1' \log(12/\delta) \frac{\lambda^{a-1/2}}{\sqrt{n}} \left(\sigma \sqrt{\mathcal{N}(\lambda)} + \frac{||S_K u_\lambda - f_\rho||_{L^2}}{\sqrt{\lambda}} + \frac{B_\lambda}{\sqrt{n\lambda}} \right) + C_2' \lambda^a \ ||T_K^{-\zeta} u_\lambda||_{\mathcal{H}_K} \left(\frac{\log(4/\delta)}{\sqrt{n}} + \lambda^\zeta \right) + C_3' \lambda^{a-1/2} ||S_K u_\lambda - f_\rho||_{L^2} \ ,$$

   for some $C_1' > 0$, $C_2' > 0$ and $C_3' > 0$.

Proof of Proposition 2.3. Let $a \in [0, 1/2]$. We write

$$\begin{aligned} &T_K^a(\hat{u}_\lambda - u_\lambda) \\ &= \underbrace{T_K^a g_\lambda(T_{\mathbf{x}})((S_{\mathbf{x}}^*\mathbf{y} - S_K^* f_\rho) - (T_{\mathbf{x}} u_\lambda - T_K u_\lambda))}_{\mathcal{T}_1} + \underbrace{T_K^a (g_\lambda(T_{\mathbf{x}}) T_{\mathbf{x}} - Id) u_\lambda}_{\mathcal{T}_2} + \underbrace{T_K^a g_\lambda(T_{\mathbf{x}})(S_K^* f_\rho - T_K u_\lambda)}_{\mathcal{T}_3} \, . \end{aligned}$$

We further set $T_{\mathbf{x},\lambda} := T_{\mathbf{x}} + \lambda$ and $T_\lambda := T_K + \lambda$.

Bounding $\mathcal{T}_1$. We further decompose

$$||\mathcal{T}_1||_{\mathcal{H}_K} \leq ||T_K^a T_\lambda^{-1/2}|| \cdot ||T_\lambda^{1/2} T_{\mathbf{x},\lambda}^{-1/2}|| \cdot ||T_{\mathbf{x},\lambda}^{1/2} g_\lambda(T_{\mathbf{x}}) T_{\mathbf{x},\lambda}^{1/2}|| \cdot ||T_{\mathbf{x},\lambda}^{-1/2}(S_{\mathbf{x}}^*\mathbf{y} - S_K^* f_\rho) - (T_{\mathbf{x}} u_\lambda - T_K u_\lambda))||_{\mathcal{H}_K} \, .$$

A short calculation shows that

$$||T_K^a T_\lambda^{-1/2}|| \leq \lambda^{a-1/2} \, .$$

By Caponnetto and De Vito (2006), Proof of Theorem 4, with probability at least $1 - \delta/6$

$$||T_\lambda^{1/2} T_{\mathbf{x},\lambda}^{-1/2}|| \leq \sqrt{2} \, ,$$

provided Assumption 2.2 is satisfied. Moreover, according to Bauer et al. (2007), Definition 1 we have almost surely

$$||T_{\mathbf{x},\lambda}^{1/2} g_\lambda(T_{\mathbf{x}}) T_{\mathbf{x},\lambda}^{1/2}|| \leq E \, ,$$

for some $E > 0$. Finally, Lemma 6.10 in Fischer and Steinwart (2017) shows that with probability at least $1 - \delta/6$

$$||T_{\mathbf{x},\lambda}^{-1/2}((S_{\mathbf{x}}^*\mathbf{y} - S_K^* f_\rho) - (T_{\mathbf{x}} u_\lambda - T_K u_\lambda))||_{\mathcal{H}_K}^2 \leq C_\kappa \log^2(12/\delta) \frac{1}{n} \left(\sigma^2 \mathcal{N}(T_K) + \frac{||S_K u_\lambda - f_\rho||_{L^2}^2}{\lambda} + \frac{B_\lambda^2}{n\lambda} \right) ,$$

for some $C_\kappa > 0$. Thus, combining the above gives us with probability at least $1 - \delta/3$

$$||\mathcal{T}_1||_{\mathcal{H}_K} \leq C'_\kappa \log(12/\delta) \frac{\lambda^{a-1/2}}{\sqrt{n}} \left(\sigma \sqrt{\mathcal{N}(T_K)} + \frac{||S_K u_\lambda - f_\rho||_{L^2}}{\sqrt{\lambda}} + \frac{B_\lambda}{\sqrt{n\lambda}} \right) , \tag{3}$$

for some $C'_\kappa > 0$.

Bounding $\mathcal{T}_2$. Setting $r_\lambda(T_{\mathbf{x}}) = g_\lambda(T_{\mathbf{x}}) T_{\mathbf{x}} - Id$, we split once more and obtain with probability at least $1 - \delta$

$$\begin{aligned} ||\mathcal{T}_2||_{\mathcal{H}_K} &\leq ||T_K^a T_\lambda^{-1/2}|| \cdot ||T_\lambda^{1/2} T_{\mathbf{x},\lambda}^{-1/2}|| \cdot ||T_{\mathbf{x},\lambda}^{1/2} r_\lambda(T_{\mathbf{x}}) u_\lambda|| \\ &\leq \sqrt{2} \lambda^{a-1/2} ||T_{\mathbf{x},\lambda}^{1/2} r_\lambda(T_{\mathbf{x}}) u_\lambda||_{\mathcal{H}_K} \, . \end{aligned}$$

Now we follow the proof of Mücke et al. (2019), Proposition 2, slightly adapted to our purposes, and consider two different cases:

(a) Assume $f_\rho \in Range(L_K^\zeta)$, for some $\zeta \in (0, 1]$. Here, we write with probability at least $1 - \delta/3$

$$\begin{aligned} ||T_{\mathbf{x},\lambda}^{1/2} r_\lambda(T_{\mathbf{x}}) u_\lambda||_{\mathcal{H}_K} &\leq ||T_{\mathbf{x},\lambda} r_\lambda(T_{\mathbf{x}})|| \cdot ||T_{\mathbf{x},\lambda}^{-1/2} T_\lambda^{1/2}|| \cdot ||T_\lambda^{-1/2} u_\lambda||_{\mathcal{H}_K} \\ &\leq C_1 \lambda \, ||T_\lambda^{-1/2} u_\lambda||_{\mathcal{H}_K} \, , \end{aligned}$$

for some $C_1 > 0$. Thus,

$$||\mathcal{T}_2||_{\mathcal{H}_K} \leq C_1 \lambda^{a+1/2} \, ||T_\lambda^{-1/2} u_\lambda||_{\mathcal{H}_K} \, .$$

(b) Assume $f_\rho \in Range(L_K^\zeta)$, for some $\zeta > 1$. In this case we let $\zeta \geq 1$ and have for some $C_2 > 0$

$$\begin{aligned} ||T_{\mathbf{x},\lambda}^{1/2} r_\lambda(T_{\mathbf{x}}) u_\lambda||_{\mathcal{H}_K} &\leq ||T_{\mathbf{x},\lambda}^{1/2} r_\lambda(T_{\mathbf{x}})|| \, ||T_K^\zeta - T_{\mathbf{x}}^\zeta|| \, ||T_K^{-\zeta} u_\lambda||_{\mathcal{H}_K} + ||T_{\mathbf{x},\lambda}^{1/2} r_\lambda(T_{\mathbf{x}}) T_{\mathbf{x}}^\zeta|| \, ||T_K^{-\zeta} u_\lambda||_{\mathcal{H}_K} \\ &\leq \sqrt{\lambda} \, ||T_K^{-\zeta} u_\lambda||_{\mathcal{H}_K} \left(C_2 \, \frac{\log(4/\delta)}{\sqrt{n}} + \lambda^\zeta \right) , \end{aligned}$$

holding with probability at least $1-\delta/3$. Thus, for some $\tilde{C}_2 > 0$.

$$||\mathcal{T}_2||_{\mathcal{H}_K} \leq \tilde{C}_2 \lambda^a \, ||T_K^{-\zeta} u_\lambda||_{\mathcal{H}_K} \left(\frac{\log(4/\delta)}{\sqrt{n}} + \lambda^\zeta \right) .$$

Bounding $\mathcal{T}_3$. Applying standard arguments gives with probability at least $1-\delta/3$

$$\begin{aligned} ||\mathcal{T}_3||_{\mathcal{H}_K} &\leq ||T_K^a g_\lambda(T_{\mathbf{x}})(S_K^* f_\rho - T_K u_\lambda)||_{\mathcal{H}_K} \\ &\leq C_3 \lambda^{a-1/2} ||S_K u_\lambda - f_\rho||_{L^2} . \end{aligned}$$

Combining all of our findings leads to result. □

We summarize some results under refined assumptions on f_ρ, see e.g. Fischer and Steinwart (2017), Lemma 6.6. and Corollary 6.7 .

Lemma 2.4. *Suppose Assumptions 2.2 and 3.2 are satisfied. Then*

1. $||u_\lambda||_{\mathcal{H}_K} \leq R\lambda^{r-\frac{1}{2}}$,
2. $||S_K u_\lambda - f_\rho||_{L^2} \leq R\lambda^r$,
3. $||S_K u_\lambda||_\infty \leq \kappa^2 R \lambda^{-|1/2-r|_+}$,
4. $||(T_K+\lambda)^{-1/2} u_\lambda||_{\mathcal{H}_K} \leq CR\lambda^{r-1}$,
5. *Assume* $||f_\rho||_\infty < \infty$. *Then* $||S_K u_\lambda - f_\rho||_\infty \leq (||f_\rho||_\infty + \kappa^2 R)\lambda^{-|1/2-r|_+}$.

2.2 Bounding First Stage Tail-Averaged Gradient Descent Error

Our aim is now to state an error bound for the first-stage tail-average GD algorithm, defined in (2). According to the results in Mücke et al. (2019), (2) can be rewritten as

$$\bar{f}_T = G_T(T_{\mathbf{x}}) S_{\mathbf{x}}^* \mathbf{y} ,$$

where G_T is defined in (2), Appendix B and constitutes a family of filter functions $\{G_\lambda\}_\lambda$, where we set $\lambda = 1/(\eta T)$. We then get

$$||S_K \bar{f}_T - f_\rho||_{L^2} \leq ||S_K(\bar{f}_T - \bar{u}_T)||_{L^2} + ||S_K \bar{u}_T - f_\rho||_{L^2} ,$$

where we set $\bar{u}_T = G_T(T_K) S_K^* f_\rho$. Thus, the first term corresponds the an estimation error which we bound by means of Proposition 2.3 with $a = 1/2$. Note that Assumption 2.1 is satisfied with $\sigma = B = 2M$, since $\mathcal{Y} \subseteq [-M, M]$ by assumption.

Proposition 2.5 (Excess Risk Tail-Ave GD First Stage)**.** *Suppose Assumptions 2.2, 2.2 are satisfied. Let* $T \in \mathbb{N}$ *and denote*

$$B_T = \max\{2M, ||S_K \bar{u}_T - f_\rho||_\infty\} .$$

Let further $\delta \in (0,1]$, $\lambda = (\eta T)^{-1}$.

1. *Assume* $f_\rho \in Range(L_K^\zeta)$, *for some* $\zeta \in (0,1]$. *With probability not less than* $1-\delta$,

$$\begin{aligned} ||S_K \bar{f}_T - f_\rho||_{L^2} \leq C_1 \log(12/\delta) \frac{1}{\sqrt{n}} \left(M\sqrt{\mathcal{N}(\lambda)} + \frac{||S_K \bar{u}_T - f_\rho||_{L^2}}{\sqrt{\lambda}} + \frac{B_T}{\sqrt{n\lambda}} \right) \\ + C_2 \lambda \, ||(T_K+\lambda)^{-1/2} \bar{u}_T||_{\mathcal{H}_K} + C_3 ||S_K \bar{u}_T - f_\rho||_{L^2} , \end{aligned}$$

for some $C_1 > 0$, $C_2 > 0$ *and* $C_3 > 0$.

2. Assume $f_\rho \in Range(L_K^\zeta)$, for some $\zeta > 1$. With probability not less than $1-\delta$,

$$||S_K \bar{f}_T - f_\rho||_{L^2} \leq C_1' \log(12/\delta) \frac{1}{\sqrt{n}} \left(M\sqrt{\mathcal{N}(\lambda)} + \frac{||S_K \bar{u}_T - f_\rho||_{L^2}}{\sqrt{\lambda}} + \frac{B_T}{\sqrt{n\lambda}} \right) + C_2' \lambda^{1/2} \, ||T_K^{-\zeta} \bar{u}_T||_{\mathcal{H}_K} \left(\frac{\log(4/\delta)}{\sqrt{n}} + \lambda^\zeta \right) + C_3' ||S_K \bar{u}_T - f_\rho||_{L^2} \, ,$$

for some $C_1' > 0$, $C_2' > 0$ and $C_3' > 0$.

Corollary 2.6 (Rate of Convergence First Stage Tail-Averaged Gradient Descent)**.** *Suppose all Assumptions of Proposition 2.5 are satisfied. Let additionally Assumptions 3.2 and 3.3 hold. Then with probability not less than $1-\delta$, the excess risk for the first stage tail-averaged Gradient Descent satisfies with probability not less than $1-\delta$:*

1. *If $2r + \nu > 1$: Let $\eta_n T_n = \left(\frac{R^2}{M^2} n\right)^{\frac{1}{2r+\nu}}$, then*

$$||S_K \bar{f}_T - f_\rho||_{L^2} \leq C \, \log(12/\delta) R \left(\frac{M^2}{R^2 n} \right)^{\frac{r}{2r+\nu}} .$$

for some constant $C < \infty$.

2. *If $2r + \nu \leq 1$: Let $\eta_n T_n = \frac{R^2 n}{M^2 \log^K(n)}$ for some $K > 1$, then*

$$||S_K \bar{f}_T - f_\rho||_{L^2} \leq C' \, \log(12/\delta) \left(\frac{M^2 \log^K(n)}{R^2 n} \right)^r .$$

for some constant $C' < \infty$.

Proof of Corollary 2.6. The proof follows from standard calculations by applying Lemma 2.4 and Proposition 2.5 with $\zeta = r$. □

2.3 Bounding Second Stage Tail-Averaged Gradient Descent Variance

Based on the notation introduced in Section 1, the GD updates can be rewritten as

$$\hat{f}_{t+1} = \hat{f}_t - \eta(T_{\hat{\mathbf{x}}} \hat{f}_t - g_{\hat{\mathbf{z}}}) \tag{4}$$

and

$$f_{t+1} = f_t - \eta(T_{\mathbf{x}} f_t - g_{\mathbf{z}}) \, .$$

We thus find for any $t \geq 1$

$$\hat{f}_{t+1} - f_{t+1} = (Id - \eta T_{\hat{\mathbf{x}}})(\hat{f}_t - f_t) + \eta \hat{\xi}_t \, ,$$

where we define the noise variables as

$$\hat{\xi}_t := \hat{\xi}_t^{(1)} + \hat{\xi}^{(2)} \quad \hat{\xi}_t^{(1)} := (T_{\mathbf{x}} - T_{\hat{\mathbf{x}}}) f_t \, , \quad \hat{\xi}^{(2)} := g_{\hat{\mathbf{z}}} - g_{\mathbf{z}} \, . \tag{5}$$

By induction we easily find

$$\hat{f}_{t+1} - f_{t+1} = \eta \sum_{s=0}^{t} (Id - \eta T_{\hat{\mathbf{x}}})^{t-s} \hat{\xi}_s \tag{6}$$

and

$$\bar{\hat{f}}_T - \bar{f}_T = \frac{2\eta}{T} \sum_{t=\lfloor T/2 \rfloor + 1}^{T} \sum_{s=0}^{t-1} (Id - \eta T_{\hat{\mathbf{x}}})^{t-1-s} \hat{\xi}_s \, .$$

As a first step we need to bound the norm of the noise variables (5). To this end, let us introduce the GD updates $u_t = g_t(T_K)S_K^* f_\rho$, where $\{g_t\}_t$ is a filter function, given in Eq. (23) in Mücke et al. (2019).

Proposition 2.7. *Let further $\delta \in (0,1]$ and suppose Assumptions 2.1, 2.2 are satisfied. Denote*

$$B_t = \max\{M, ||S_K u_t - f_\rho||_\infty\} .$$

1. *Assume $f_\rho \in Range(L_K^\zeta)$, for some $\zeta \in (0,1]$. With probability not less than $1-\delta$,*

$$||f_t||_{\mathcal{H}_K} \le C_1 \log(12/\delta)\sqrt{\frac{\eta t}{n}}\left(M\sqrt{\mathcal{N}(1/(\eta t))} + \sqrt{\eta t}||S_K u_t - f_\rho||_{L^2} + B_t\sqrt{\frac{\eta t}{n}}\right)$$
$$+ C_2(\eta t)^{-1/2}\,||(T_K + 1/(\eta t))^{-1/2}u_t||_{\mathcal{H}_K} + C_3\sqrt{\eta t}||S_K u_t - f_\rho||_{L^2} + ||u_t||_{\mathcal{H}_K} ,$$

   for some $C_1 > 0$, $C_2 > 0$ and $C_3 > 0$.

2. *Assume $f_\rho \in Range(L_K^\zeta)$, for some $\zeta \ge 1$. With probability not less than $1-\delta$,*

$$||f_t||_{\mathcal{H}_K} \le C_1' \log(12/\delta)\sqrt{\frac{\eta t}{n}}\left(M\sqrt{\mathcal{N}(1/(\eta t))} + \sqrt{\eta t}||S_K u_t - f_\rho||_{L^2} + B_t\sqrt{\frac{\eta t}{n}}\right)$$
$$+ C_2'\,||T_K^{-\zeta}u_t||_{\mathcal{H}_K}\left(\frac{\log(4/\delta)}{\sqrt{n}} + (\eta t)^{-\zeta}\right) + C_3'\sqrt{\eta t}||S_K u_t - f_\rho||_{L^2} + ||u_t||_{\mathcal{H}_K} ,$$

   for some $C_1' > 0$, $C_2' > 0$ and $C_3' > 0$.

Proof of Proposition 2.7. We decompose as

$$||f_t||_{\mathcal{H}_K} \le ||f_t - u_t||_{\mathcal{H}_K} + ||u_t||_{\mathcal{H}_K} .$$

The proof follows then from Proposition 2.5 with $a = 0$, $\sigma = B = M$. □

Corollary 2.8. *In addition to all assumptions of Proposition 2.7, suppose Assumptions 2.2, 3.2 and 3.3 are satisfied.*

1. *Let $0 < r \le 1/2$ and assume that*

$$n \ge 64\kappa^2 \log(12/\delta)(\eta t)\log((\eta t)^\nu) . \tag{7}$$

   Then Assumption 2.2 is satisfied and with probability not less than $1-\delta$

$$||f_t||_{\mathcal{H}_K} \le C'_{\kappa,M,R}(\eta t)^{\frac{1}{2}\max\{\nu,1-2r\}} ,$$

   for some $C'_{\kappa,M,R} > 0$.

2. *Let $1/2 \le r \le 1$ and assume that*

$$n \ge 64e\kappa^2 \log^2(12/\delta)(\eta t)^{1+\nu} . \tag{8}$$

   Then Assumption 2.2 is satisfied and

$$||f_t||_{\mathcal{H}_K} \le C_{\kappa,M,R} ,$$

   with probability not less than $1-\delta$ and for some $C_{\kappa,M,R} > 0$.

3. *Let $1 < r$ and assume that (8) is satisfied. Then*

$$||f_t||_{\mathcal{H}_K} \le C'_{\kappa,M,R} ,$$

   with probability not less than $1-\delta$ and for some $C'_{\kappa,M,R} > 0$.

Proof of Corollary 2.8. Recall that by Lemma 2.4 we have

$$||S_K u_t - f_\rho||_{L^2} \leq R(\eta t)^{-r}\,, \quad ||S_K u_t - f_\rho||_\infty \leq (M + \kappa^2 R)(\eta t)^{|1/2-r|_+}$$

and

$$||u_t||_{\mathcal{H}_K} \leq R(\eta t)^{\frac{1}{2}-r}\,, \quad ||(T_K + 1/(\eta t))^{-1/2} u_t|| \leq CR(\eta t)^{1-r}\,.$$

1. Now suppose that $0 < r \leq 1/2$. The first part of Proposition 2.7 yields with probability not less than $1-\delta$

$$||f_t||_{\mathcal{H}_K} \leq C_1 \log(12/\delta)\sqrt{\frac{\eta t}{n}}\left(M(\eta t)^{\nu/2} + R(\eta t)^{1/2-r} + (\eta t)^{|1/2-r|_+}\sqrt{\frac{\eta t}{n}}\right) + C_2\, R(\eta t)^{1/2-r}\,. \tag{9}$$

   Then (9) and (7) give with $\log(12/\delta) \geq 1$ the bound

$$\begin{aligned} ||f_t||_{\mathcal{H}_K} &\leq C_{\kappa,M,R} \log(12/\delta)\sqrt{\frac{\eta t}{n}}(\eta t)^{\frac{1}{2}\max\{\nu,1-2r\}} + C'_{\kappa,M,R}(\eta t)^{1/2-r} \\ &\leq C''_{\kappa,M,R}(\eta t)^{\frac{1}{2}\max\{\nu,1-2r\}}\,, \end{aligned}$$

   with probability not less than $1-\delta$.

2. Specifically, if $1/2 \leq r \leq 1$ and $1/(\eta t) \leq \kappa^2$, we have

$$\begin{aligned} ||f_t||_{\mathcal{H}_K} &\leq C'_{\kappa,r} \max\{M,R\} \log(12/\delta)\left(\frac{1}{\sqrt{n}}(\eta t)^{1/2+\nu/2} + 1 + \frac{\eta t}{n}\right) + C'_{\kappa,r}\, R \\ &\leq C_{\kappa,M,R}\,, \end{aligned}$$

   for some $C_{\kappa,M,R} > 0$, provided (8) is satisfied.

3. The second part of Proposition 2.7 then gives with $\zeta = r$ and $||T_K^{-r} u_t||_{\mathcal{H}_K} \leq CR(\eta t)^{1/2}$, with probability not less than $1-\delta$

$$\begin{aligned} ||f_t||_{\mathcal{H}_K} &\leq C_1 \log(12/\delta)\left(\frac{1}{\sqrt{n}}(\eta t)^{1/2+\nu/2} + 1 + \frac{\eta t}{n}\right) + C_2(\eta t)^{1/2}\left(\frac{\log(4/\delta)}{\sqrt{n}} + (\eta t)^{-r}\right) + C_3(\eta t)^{1/2-r} \\ &\leq C_4 + C_2(\eta t)^{1/2}\left(\frac{\log(4/\delta)}{\sqrt{n}} + (\eta t)^{-r}\right) \\ &\leq C_5 + C_2 \log(4/\delta)\frac{(\eta t)^{1/2}}{\sqrt{n}} \\ &\leq C_6\,, \end{aligned}$$

   for some $C_6 > 0$, depending on κ, M, R.

□

We now come to the main result of this subsection.

Proposition 2.9 (Second Stage GD Variance)**.** *Suppose Assumptions 2.2, 3.1 and 2.2 are satisfied. Let $\eta < 1/\kappa^2$, $T \geq 3$ and define*

$$\mathcal{B}(1/\eta T) := \left(\frac{2\eta T}{n} + \sqrt{\frac{\eta T \mathcal{N}(1/\eta T)}{n}} + \frac{\eta T}{N^{\frac{\alpha}{2}}} + 1\right)\,. \tag{10}$$

1. **If $f_\rho \in Ran(S_K)$:** *The GD variance satisfies with probability not less than $1-\delta$ with respect to the data D*

$$\mathbb{E}_{\hat{D}|D}[||S_K(\bar{\hat{f}}_T - \bar{f}_T)||_{L^2}] \leq C \log(4/\delta)\log(T)\frac{\sqrt{\eta T}}{N^{\frac{\alpha}{2}}}\mathcal{B}(1/\eta T)\,,$$

   for some $C < \infty$, depending on $\kappa, \gamma, L, \alpha$.

2. **If** $f_\rho \notin Ran(S_K)$**:** *Let us define*

$$\varphi(\eta s) = (\eta s)^{\frac{1}{2}\max\{\nu, 1-2r\}} \ . \tag{11}$$

With probability not less than $1-\delta$ *with respect to the data* D

$$\mathbb{E}_{\hat{D}|D}[||S_K(\bar{\hat{f}}_T - \bar{f}_T)||_{L^2}] \le C\log(4/\delta)\log(T)\frac{\sqrt{\eta T}\mathcal{B}(1/\eta T)}{N^{\frac{\alpha}{2}}}(1+\varphi(\eta T)) \ ,$$

for some $C < \infty$*, depending on* $\kappa, \gamma, L, \alpha, r$*.*

Proof of Proposition 2.9. By (2.3) we may write

$$\begin{aligned} T_K^{\frac{1}{2}}(\bar{\hat{f}}_T - \bar{f}_T) &= \frac{2\eta}{T}\sum_{t=\lfloor T/2\rfloor+1}^{T}\sum_{s=0}^{t-1} T_K^{\frac{1}{2}}(Id - \eta T_{\hat{\mathbf{x}}})^{t-1-s}\hat{\xi}_s \\ &= \frac{2\eta}{T}\sum_{t=\lfloor T/2\rfloor+1}^{T}\sum_{s=0}^{t-1} A\cdot B_{t,s}\,\hat{\xi}_s \ , \end{aligned}$$

where for $\lambda > 0$ we introduce

$$\begin{aligned} A &:= T_K^{\frac{1}{2}}(T_{\hat{\mathbf{x}}} + \lambda Id)^{-1} \\ B_{t,s} &:= (T_{\hat{\mathbf{x}}} + \lambda Id)(Id - \eta T_{\hat{\mathbf{x}}})^{t-1-s} \ . \end{aligned}$$

(1) Bounding the operator A**:** This follows directly from Lemma 1.4. Indeed,

$$\mathbb{E}_{\hat{D}|D}[||T_K^{\frac{1}{2}}(T_{\hat{\mathbf{x}}} + \lambda Id)^{-1}||] \le C_{\kappa,\gamma,L,\alpha}\left(\frac{\mathcal{A}_{\mathbf{x}}(\lambda)}{\lambda} + \frac{1}{\lambda^{\frac{3}{2}}N^{\frac{\alpha}{2}}} + \frac{1}{\sqrt{\lambda}}\right) ,$$

where by Blanchard and Mücke (2018, Proposition 5.3) the term $\mathcal{A}_{\mathbf{x}}(\lambda)$ satisfies with probability at least $1-\delta$ with respect to the data D

$$\begin{aligned} \mathcal{A}_{\mathbf{x}}(\lambda) &= ||(T_K + \lambda Id)^{-\frac{1}{2}}(T_K - T_{\mathbf{x}})|| &(12)\\ &\le 2\log(2/\delta)\left(\frac{2}{n\sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{n}}\right) . &(13) \end{aligned}$$

Hence, since $1 \le 2\log(2/\delta)$ for any $\delta \in (0,1]$ we obtain with probability at least $1-\delta$

$$\mathbb{E}_{\hat{D}|D}[||T_K^{\frac{1}{2}}(T_{\hat{\mathbf{x}}} + \lambda Id)^{-1}||] \le C_{\kappa,\gamma,L,\alpha}\log(2/\delta)\frac{\mathcal{B}(\lambda)}{\sqrt{\lambda}} \ . \tag{14}$$

(2) Bounding the operators $B_{t,s}$**:** We write

$$||B_{t,s}|| \le ||T_{\hat{\mathbf{x}}}(Id - \eta T_{\hat{\mathbf{x}}})^{t-1-s}|| + \lambda||(Id - \eta T_{\hat{\mathbf{x}}})^{t-1-s}|| \ .$$

Denoting $\sigma_1 \ge \sigma_2 \ge ...$ the sequence of eigenvalues of $T_{\hat{\mathbf{x}}}$, we have for any $s = 0, ..., t-1$, $t = \lfloor \frac{T}{2} \rfloor, ..., T$ the upper bound

$$||(Id - \eta T_{\hat{\mathbf{x}}})^{t-1-s}|| \le \sup_j |(1-\eta\sigma_j)^{t-1-s}| \le 1 \ ,$$

since $\eta < 1/\kappa^2$.

For bounding the first term note that for $s = t-1$ we have

$$||T_{\hat{\mathbf{x}}}(Id - \eta T_{\hat{\mathbf{x}}})^{t-1-s}|| = ||T_{\hat{\mathbf{x}}}|| \leq \kappa^2 \ .$$

If $0 \leq s < t-1$ we use the inequality $1 + \sigma \leq e^{\sigma}$ for any $\sigma \geq -1$. Then a short calculation gives[1]

$$\begin{aligned} ||T_{\hat{\mathbf{x}}}(Id - \eta T_{\hat{\mathbf{x}}})^{t-1-s}|| &\leq \sup_j |\sigma_j (1 - \eta\sigma_j)^{t-1-s}| \\ &\leq \sup_j \sigma_j e^{-\eta(t-1-s)\sigma_j} \\ &\leq \frac{1}{e\eta(t-1-s)} \ . \end{aligned}$$

Thus, combining the above findings yields

$$||B_{t,t-1}|| \leq \kappa^2 + \lambda \ , \tag{15}$$

and for $0 \leq s < t-1$

$$||B_{t,s}|| \leq \frac{1}{e\eta(t-1-s)} + \lambda \ . \tag{16}$$

(3) Bounding the noise variables $\hat{\xi}_s$: Recall that $\hat{\xi}_s := \hat{\xi}_s^{(1)} + \hat{\xi}^{(2)}$ with

$$\hat{\xi}_s^{(1)} := (T_{\mathbf{x}} - T_{\hat{\mathbf{x}}}) f_s \ , \qquad \hat{\xi}^{(2)} := g_{\hat{\mathbf{z}}} - g_{\mathbf{z}} \ .$$

Applying Lemma 1.2 gives for some $c_\alpha < \infty$ the bound

$$\mathbb{E}_{\hat{D}|D}[||\xi^{(2)}||_{\mathcal{H}_K}] = \mathbb{E}_{\hat{D}|D}[||g_{\hat{\mathbf{z}}} - g_{\mathbf{z}}||_{\mathcal{H}_K}] \leq c_\alpha L M \frac{\gamma^\alpha}{N^{\frac{\alpha}{2}}} \ .$$

- $f_\rho \in Ran(S_K)$**:** Lemma 1.3 and Proposition 2.9 gives with probability at least $1-\delta$

  $$\mathbb{E}_{\hat{D}|D}[||\xi_s^{(1)}||_{\mathcal{H}_K}] = \mathbb{E}_{\hat{D}|D}[||(T_{\hat{\mathbf{x}}} - T_{\mathbf{x}}) f_s||_{\mathcal{H}_K}] \leq \frac{C}{N^{\frac{\alpha}{2}}} \ ,$$

  with $C = c'_\alpha \gamma^\alpha \kappa L (2||w_\rho||_{\mathcal{H}_K} + 1)$, for some $c'_\alpha < \infty$. Combining both bounds finally leads to

  $$\mathbb{E}_{\hat{D}|D}[||\hat{\xi}_s||_{\mathcal{H}_K}] \leq \frac{c''_\alpha}{N^{\frac{\alpha}{2}}} \ , \tag{17}$$

  where $c''_\alpha = 2\max\{C, c_\alpha \gamma^\alpha L M\}$ and holding with probability at least $1-\delta$.

- $f_\rho \notin Ran(S_K)$**:** In this case we apply Lemma 1.3 and Corollary 2.8 to get with probability at least $1-\delta$ with respect to the data D

  $$\mathbb{E}_{\hat{D}|D}[||\xi_s^{(1)}||_{\mathcal{H}_K}] \leq \kappa c'_\alpha \gamma^\alpha L \log(6/\delta) \frac{\varphi(\eta s)}{N^{\frac{\alpha}{2}}} \ ,$$

  for some $c'_\alpha < \infty$ and where φ is defined in (11). Hence, with probability at least $1-\delta$ with respect to the data D one has

  $$\mathbb{E}_{\hat{D}|D}[||\hat{\xi}_s||_{\mathcal{H}_K}] \leq \frac{\tilde{c}''_\alpha}{N^{\frac{\alpha}{2}}}(1 + \varphi(\eta s)) \ , \tag{18}$$

  for some $c''_\alpha < \infty$.

We complete the proof now by collecting the above findings. Let us write

$$||T_K^{\frac{1}{2}}(\bar{\hat{f}}_T - \bar{f}_T)||_{\mathcal{H}_K} \leq \underbrace{\frac{2\eta}{T} \sum_{t=\lfloor T/2 \rfloor + 1}^{T} \sum_{s=0}^{t-2} ||A|| \cdot ||B_{t,s}|| \cdot ||\hat{\xi}_s||_{\mathcal{H}_K}}_{\mathcal{I}_1} + \underbrace{\frac{2\eta}{T} \sum_{t=\lfloor T/2 \rfloor + 1}^{T} ||A|| \cdot ||B_{t,t-1}|| \cdot ||\hat{\xi}_{t-1}||_{\mathcal{H}_K}}_{\mathcal{I}_2} \ .$$

We again distinguish between the two cases:

[1] The function $h(\sigma) = \sigma e^{-c\sigma}$, $c > 0$, achieves it's maximum at $\sigma = 1/c$.

- $f_\rho \in Ran(S_K)$**:** From (14), (15) and (17) we obtain with $\lambda \leq \kappa^2$

$$\mathbb{E}_{\hat{D}|D}[\mathcal{I}_2] \leq \eta \tilde{C}_{\kappa,\gamma,L,\alpha} \log(2/\delta) \frac{\mathcal{B}(\lambda)}{\lambda^{\frac{1}{2}} N^{\frac{\alpha}{2}}} \,, \tag{19}$$

for some $\tilde{C}_{\kappa,\gamma,L,\alpha} < \infty$. Additionally, by (14), (16), (17) and Lemma 2.14 we find with $\lambda = (\eta T)^{-1}$

$$\begin{aligned}\mathbb{E}_{\hat{D}|D}[\mathcal{I}_1] &\leq \tilde{C}'_{\kappa,\gamma,L,\alpha} \log(2/\delta) \frac{\mathcal{B}(\lambda)}{\lambda^{\frac{1}{2}} N^{\frac{\alpha}{2}}} \frac{2\eta}{T} \sum_{t=\lfloor T/2 \rfloor + 1}^{T} \sum_{s=0}^{t-2} \left(\frac{1}{e\eta(t-1-s)} + \lambda \right) \\ &\leq 4\tilde{C}'_{\kappa,\gamma,L,\alpha} \log(2/\delta) \log(T) \frac{\sqrt{\eta T}\mathcal{B}(1/\eta T)}{N^{\frac{\alpha}{2}}} \,, \end{aligned} \tag{20}$$

for some $\tilde{C}'_{\kappa,\gamma,L,\alpha} < \infty$.

Combining (19) and (20) gives with $\lambda = (\eta T)^{-1}$

$$\mathbb{E}_{\hat{D}|D}[||S_K(\bar{\hat{f}}_T - \bar{f}_T)||_{L^2}] \leq C''_{\kappa,\gamma,L,\alpha} \log(2/\delta) \log(T) \frac{\sqrt{\eta T}\mathcal{B}((\eta T)^{-1})}{N^{\frac{\alpha}{2}}} \,,$$

with probability at least $1-\delta$, for some $C''_{\kappa,\gamma,L,\alpha} < \infty$.

- $f_\rho \notin Ran(S_K)$**:** From (14), (15), (18) we find

$$\begin{aligned}\mathbb{E}_{\hat{D}|D}[\mathcal{I}_2] &\leq \tilde{C}_{\kappa,\gamma,L,\alpha} \log(2/\delta) \frac{\mathcal{B}(\lambda)}{\sqrt{\lambda} N^{\frac{\alpha}{2}}} \frac{2\eta}{T} \sum_{t=\lfloor T/2 \rfloor + 1}^{T} (1 + \varphi(\eta t)) \\ &\leq 2\tilde{C}_{\kappa,\gamma,L,\alpha} \log(2/\delta) \frac{\mathcal{B}(\lambda)}{\sqrt{\lambda} N^{\frac{\alpha}{2}}} \eta (1 + \bar{\varphi}(\eta T)) \,. \end{aligned}$$

for some $\tilde{C}_{\kappa,\gamma,L,\alpha} < \infty$ and where by Lemma 2.14 for some $C_r < \infty$

$$\bar{\varphi}(\eta T) := \frac{2}{T} \sum_{t=\lfloor T/2 \rfloor + 1}^{T} \varphi(\eta t) \leq C_r \varphi(\eta T) \,. \tag{21}$$

By (14), (16), (18) and Lemma 2.14 we get with $\lambda = (\eta T)^{-1}$ and $\eta < 1/\kappa^2$

$$\begin{aligned}\mathbb{E}_{\hat{D}|D}[\mathcal{I}_1] &\leq \tilde{C}_{\kappa,\gamma,L,\alpha} \log(2/\delta) \frac{\sqrt{\eta T}\mathcal{B}(1/\eta T)}{N^{\frac{\alpha}{2}}} \frac{2\eta}{T} \sum_{t=\lfloor T/2 \rfloor + 1}^{T} \sum_{s=0}^{t-2} \left(\frac{1}{e\eta(t-1-s)} + \lambda \right) (1 + \varphi(\eta s)) \\ &\leq 2\tilde{C}_{\kappa,\gamma,L,\alpha} \log(2/\delta) \frac{\sqrt{\eta T}\mathcal{B}(1/\eta T)}{N^{\frac{\alpha}{2}}} (4\log(T) + C'_r \log(T) \varphi(\eta T)) \,, \end{aligned}$$

for some $C'_r < \infty$.

Combining the bounds for $\mathcal{I}_1$ and $\mathcal{I}_2$ finally gives with $\eta < 1/\kappa^2$, $1 \leq \log(T)$

$$\mathbb{E}_{\hat{D}|D}[||S_K(\bar{\hat{f}}_T - \bar{f}_T)||_{L^2}] \leq \tilde{C} \log(4/\delta) \frac{\sqrt{\eta T}\mathcal{B}(1/\eta T)}{N^{\frac{\alpha}{2}}} (\log(T) + \log(T)\varphi(\eta T)) \,.$$

for some $\tilde{C} < \infty$, depending on $\kappa, \gamma, L, \alpha, r$ and holding with probability at least $1-\delta$.

□

2.4 Main Result Second Stage Tail-Averaged Gradient Descent

We now derive the final error for the excess risk of the second-stage tail-ave GD estimator for tackling distribution regression. Recall that we have the decomposition

$$\bar{\hat{f}}_T - f_\rho = \underbrace{(\bar{\hat{f}}_T - \bar{f}_T)}_{GD\ Variance\ 2.\ stage} + \underbrace{(\bar{f}_T - f_\rho)}_{Error\ GD\ first\ stage} .$$

Our main results follows then immediately from Proposition 2.5 and Proposition 2.9.

Theorem 2.10 (Excess Risk Second-Stage tail-ave GD; Part I)**.** *Suppose Assumptions 2.2, 2.2 are satisfied. Let additionally Assumptions 3.2 and 3.3 hold. Let $T \in \mathbb{N}$ and denote*

$$B_T = \max\{2M, ||S_K \bar{u}_T - f_\rho||_\infty\} .$$

Let further $\delta \in (0,1]$, $\lambda = (\eta T)^{-1}$, assume $0 < r \leq 1$ and recall the definition of $\mathcal{B}(1/\eta T)$ in (10) and of φ in (11). With probability not less than $1-\delta$, the excess risk for the second-stage tail-averaged Gradient Descent satisfies:

1. **If** $1/2 \leq r \leq 1$**:**

$$\begin{aligned}\mathbb{E}_{\hat{D}|D}[||S_K \bar{\hat{f}}_T - f_\rho||_{L^2}] \leq\ & C_1 \log(24/\delta) \frac{1}{\sqrt{n}} \left(M\sqrt{\mathcal{N}(\lambda)} + \frac{||S_K \bar{u}_T - f_\rho||_{L^2}}{\sqrt{\lambda}} + \frac{B_T}{\sqrt{n\lambda}} \right) \\ & + C_2 \lambda\ ||(T_K + \lambda)^{-1/2} \bar{u}_T||_{\mathcal{H}_K} + C_3 ||S_K \bar{u}_T - f_\rho||_{L^2} \\ & + C_4 \log(8/\delta) \log(T) \frac{\sqrt{\eta T}}{N^{\frac{\alpha}{2}}} \mathcal{B}(1/\eta T) ,\end{aligned}$$

   for some constants $C_1 > 0$, $C_2 > 0$, $C_3 > 0$, $C_4 > 0$.

2. **If** $0 < r \leq 1/2$**:**

$$\begin{aligned}\mathbb{E}_{\hat{D}|D}[||S_K \bar{\hat{f}}_T - f_\rho||_{L^2}] \leq\ & C_1 \log(24/\delta) \frac{1}{\sqrt{n}} \left(M\sqrt{\mathcal{N}(\lambda)} + \frac{||S_K \bar{u}_T - f_\rho||_{L^2}}{\sqrt{\lambda}} + \frac{B_T}{\sqrt{n\lambda}} \right) \\ & + C_2 \lambda\ ||(T_K + \lambda)^{-1/2} \bar{u}_T||_{\mathcal{H}_K} + C_3 ||S_K \bar{u}_T - f_\rho||_{L^2} \\ & + C_4 \log(8/\delta) \log(T) \frac{\sqrt{\eta T}\mathcal{B}(1/\eta T)}{N^{\frac{\alpha}{2}}} (1 + \varphi(\eta T)) ,\end{aligned}$$

   for some constants $C_1 > 0$, $C_2 > 0$, $C_3 > 0$, $C_4 > 0$.

Theorem 2.11 (Excess Risk Second-Stage tail-ave GD; Part II)**.** *Suppose Assumptions 2.2, 2.2 are satisfied. Let additionally Assumptions 3.2 and 3.3 hold. Let $T \in \mathbb{N}$ and denote*

$$B_T = \max\{2M, ||S_K \bar{u}_T - f_\rho||_\infty\} .$$

Let further $\delta \in (0,1]$, $\lambda = (\eta T)^{-1}$, assume that $r \geq 1$ and recall the definition of $\mathcal{B}(1/\eta T)$ in (10). Then with probability not less than $1-\delta$, the excess risk for the second-stage tail-averaged Gradient Descent satisfies

$$\begin{aligned}\mathbb{E}_{\hat{D}|D}[||S_K \bar{\hat{f}}_T - f_\rho||_{L^2}] \leq\ & C_1' \log(24/\delta) \frac{1}{\sqrt{n}} \left(M\sqrt{\mathcal{N}(\lambda)} + \frac{||S_K \bar{u}_T - f_\rho||_{L^2}}{\sqrt{\lambda}} + \frac{B_T}{\sqrt{n\lambda}} \right) \\ & + C_2' \lambda^{1/2}\ ||T_K^{-r} \bar{u}_T||_{\mathcal{H}_K} \left(\frac{\log(4/\delta)}{\sqrt{n}} + \lambda^\zeta \right) + C_3' ||S_K \bar{u}_T - f_\rho||_{L^2} \\ & + C_4' \log(8/\delta) \log(T) \frac{\sqrt{\eta T}}{N^{\frac{\alpha}{2}}} \mathcal{B}(1/\eta T) ,\end{aligned}$$

for some constants $C_1' > 0$, $C_2' > 0$, $C_3' > 0$, $C_4' > 0$.

Corollary 2.12 (Rate of Convergence Second-Stage Tail-Ave GD; mis-specified Case)**.** *Suppose all Assumptions of Proposition 2.5 are satisfied. Assume additionally that $r \leq 1/2$ and*

$$n \geq 64e\kappa^2 \log^2(24/\delta)(\eta t)\log((\eta T)^\nu)\ .$$

The excess risk for second stage tail-averaged Gradient Descent satisfies in the mis-specified case with probability not less than $1-\delta$:

1. *If $2r+\nu > 1$, $\eta_n T_n = \left(\frac{R^2}{M^2}n\right)^{\frac{1}{2r+\nu}}$ and $N_n \geq \log^{2/\alpha}(n)\left(\frac{R^2 n}{M^2}\right)^{\frac{2+\nu}{\alpha(2r+\nu)}}$, then*

$$\mathbb{E}_{\hat{D}|D}[||S_K \bar{\hat{f}}_{T_n} - f_\rho||_{L^2}] \leq C\ \log(24/\delta) R\left(\frac{M^2}{R^2 n}\right)^{\frac{r}{2r+\nu}}\ ,$$

   for some $C < \infty$, provided n is sufficiently large.

2. *If $2r+\nu \leq 1$, $\eta_n T_n = \frac{R^2 n}{M^2 \log^K(n)}$ for some $K > 1$ and $N_n \geq \log^{2/\alpha}(n)\left(\frac{R^2 n}{M^2\log^K(n)}\right)^{\frac{2+\nu}{\alpha}}$, then*

$$\mathbb{E}_{\hat{D}|D}[||S_K \bar{\hat{f}}_{T_n} - f_\rho||_{L^2}] \leq C'\ \log(6/\delta) R\left(\frac{M^2\log^K(n)}{R^2 n}\right)^{r}\ ,$$

   for some $C' < \infty$, provided n is sufficiently large.

Proof of Corollary 2.12. Assume $0 < r \leq 1/2$.

1. Let $2r+\nu > 1$ and $\eta_n T_n = \left(\frac{R^2}{M^2}n\right)^{\frac{1}{2r+\nu}}$. The first part of Corollary 2.6 gives a rate for the first-stage GD of order

$$||S_K \bar{f}_T - f_\rho||_{L^2} \leq C\ \log(12/\delta) R\left(\frac{M^2}{R^2 n}\right)^{\frac{r}{2r+\nu}}\ ,$$

   provided n is sufficiently large. It remains to bound the term

$$\log(T)\frac{\sqrt{\eta T}\mathcal{B}(1/\eta T)}{N^{\frac{\alpha}{2}}}(1+\varphi(\eta T))$$

   from the second part of Theorem 2.10 for an appropriate choice on N. Note that in this case we have

$$\varphi(\eta T) = (\eta T)^{\frac{1}{2}\max\{\nu, 1-2r\}} = (\eta T)^{\nu/2}\ .$$

   Moreover, by the definition (10), the choice of $\eta_n T_n$ shows that for sufficiently large n

$$\begin{aligned}\mathcal{B}(1/\eta_n T_n) &\lesssim 1 + \frac{2\eta_n T_n}{n} + \sqrt{\frac{(\eta_n T_n)^{\nu+1}}{n}} + \frac{\eta_n T_n}{N^{\alpha/2}} \\ &\lesssim 2 + \left(\frac{R^2}{M^2}n\right)^{\frac{1-2r}{2(2r+\nu)}} + N^{-\alpha/2}\left(\frac{R^2}{M^2}n\right)^{\frac{1}{2r+\nu}}\ .\end{aligned}$$

   Thus, letting $N_n \geq \left(\frac{R^2}{M^2}n\right)^{\frac{1+2r}{\alpha(2r+\nu)}}$ gives for sufficiently large n

$$\begin{aligned}\mathcal{B}(1/\eta_n T_n) &\lesssim 2 + \left(\frac{R^2}{M^2}n\right)^{\frac{1-2r}{2(2r+\nu)}} \\ &\lesssim \left(\frac{R^2}{M^2}n\right)^{\frac{1-2r}{2(2r+\nu)}}\ .\end{aligned}$$

Hence, give these choices,

$$\log(T_n)\frac{\sqrt{\eta_n T_n}\mathcal{B}(1/\eta_n T_n)}{N_n^{\frac{\alpha}{2}}}(1+\varphi(\eta_n T_n)) \lesssim \log(T_n)N^{-\alpha/2}\sqrt{\eta_n T_n}\left(\frac{R^2}{M^2}n\right)^{\frac{1-2r}{2(2r+\nu)}}(\eta_n T_n)^{\nu/2}$$
$$\lesssim N^{-\alpha/2}\log(T_n)\left(\frac{R^2}{M^2}n\right)^{\frac{2-2r+\nu}{2(2r+\nu)}} .$$

Thus, if $N_n \geq \log^{2/\alpha}(n)\left(\frac{R^2}{M^2}n\right)^{\frac{2+\nu}{\alpha(2r+\nu)}}$ gives

$$N^{-\alpha/2}\log(T_n)\left(\frac{R^2}{M^2}n\right)^{\frac{2-2r+\nu}{2(2r+\nu)}} \lesssim R\left(\frac{M^2}{R^2 n}\right)^{\frac{r}{2r+\nu}} .$$

Hence, to obtain the given rate of convergence we need to choose

$$N_n \geq \log^{2/\alpha}(n)\left(\frac{R^2}{M^2}n\right)^{\beta}, \quad \beta = \max\left\{\frac{1+2r}{\alpha(2r+\nu)}, \frac{2+\nu}{\alpha(2r+\nu)}\right\} = \frac{2+\nu}{\alpha(2r+\nu)} .$$

2. Let $2r+\nu \leq 1$, $\eta_n T_n = \frac{R^2 n}{M^2 \log^K(n)}$ for some $K > 1$. We again have to bound the expression

$$\log(T)\frac{\sqrt{\eta T}\mathcal{B}(1/\eta T)}{N^{\frac{\alpha}{2}}}(1+\varphi(\eta T))$$

for a suitable choice of N. Note that we have in this case

$$\varphi(\eta T) = (\eta T)^{\frac{1}{2}\max\{\nu, 1-2r\}} = (\eta T)^{\frac{1}{2}-r} .$$

Moreover, for sufficiently large n

$$\mathcal{B}(1/\eta_n T_n) \lesssim 1 + \frac{2\eta_n T_n}{n} + \sqrt{\frac{(\eta_n T_n)^{\nu+1}}{n}} + \frac{\eta_n T_n}{N^{\alpha/2}}$$
$$\lesssim 2 + \left(\frac{R^2 n}{M^2 \log^K(n)}\right)^{\nu/2} + N^{-\alpha/2}\frac{R^2 n}{M^2 \log^K(n)} .$$

Thus, if $N_n \geq \left(\frac{R^2 n}{M^2 \log^K(n)}\right)^{\frac{2-\nu}{\alpha}}$ we have

$$N_n^{-\alpha/2}\frac{R^2 n}{M^2 \log^K(n)} \lesssim \left(\frac{R^2 n}{M^2 \log^K(n)}\right)^{\nu/2}$$

and therefore

$$\mathcal{B}(1/\eta_n T_n) \lesssim \left(\frac{R^2 n}{M^2 \log^K(n)}\right)^{\nu/2} .$$

We thus obtain for sufficiently large n

$$\log(T_n)\frac{\sqrt{\eta_n T_n}\mathcal{B}(1/\eta_n T_n)}{N_n^{\frac{\alpha}{2}}}(1+\varphi(\eta_n T_n)) \lesssim \log(T_n)N_n^{-\alpha/2}\left(\frac{R^2 n}{M^2 \log^K(n)}\right)^{1+\nu/2-r} .$$

Hence, with $N_n \geq \log^{2/\alpha}(n)\left(\frac{R^2 n}{M^2 \log^K(n)}\right)^{\frac{2+\nu}{\alpha}}$ we find

$$\log(T_n)\frac{\sqrt{\eta_n T_n}\mathcal{B}(1/\eta_n T_n)}{N_n^{\frac{\alpha}{2}}}(1+\varphi(\eta_n T_n)) \lesssim \left(\frac{M^2 \log^K(n)}{R^2 n}\right)^{r} .$$

□

Corollary 2.13 (Rate of Convergence Second-Stage Tail-Ave GD; well-specified Case)**.** *Suppose all Assumptions of Proposition 2.5 are satisfied. Assume additionally that $r \geq 1/2$ and*

$$n \geq 64e\kappa^2 \log^2(24/\delta)(\eta T)^{1+\nu} .$$

Let $\eta_n T_n = \left(\frac{R^2}{M^2} n\right)^{\frac{1}{2r+\nu}}$ and $N_n \geq \log^{2/\alpha}(n)\left(\frac{R^2 n}{M^2}\right)^{\frac{2r+1}{\alpha(2r+\nu)}}$. The excess risk for second stage tail-averaged Gradient Descent satisfies in the well-specified case with probability not less than $1-\delta$

$$\mathbb{E}_{\hat{D}|D}[||S_K \bar{\hat{f}}_{T_n} - f_\rho||_{L^2}] \leq C \ \log(24/\delta) R \left(\frac{M^2}{R^2 n}\right)^{\frac{r}{2r+\nu}} ,$$

for some $C < \infty$, provided n is sufficiently large.

Proof of Corollary 2.13. The proof follows the same lines as the proof of Corollary 2.12 and can be obtained from standard calculations. □

2.5 Additional Material

Lemma 2.14. *Let φ be defined by* (11)*.*

1. *Let $\bar{\varphi}(\eta T)$ be defined by* (21)*. For some $C_r \in \mathbb{R}_+$ we have the bound*

$$\bar{\varphi}(\eta T) \leq C_r \varphi(\eta T) .$$

2. *For some $C_r' \in \mathbb{R}_+$ we have the bound*

$$\frac{2}{T} \sum_{t=\lfloor T/2 \rfloor + 1}^{T} \sum_{s=0}^{t-2} \frac{\varphi(\eta s)}{t-1-s} \leq C_r' \log(T) \varphi(\eta T) .$$

3. *With $\lambda = (\eta T)^{-1}$ we have*

$$\frac{2\eta}{T} \sum_{t=\lfloor T/2 \rfloor + 1}^{T} \sum_{s=0}^{t-2} \left(\frac{1}{e\eta(t-1-s)} + \lambda \right) \leq 4 \log(T) .$$

4. *With $\lambda = (\eta T)^{-1}$ we have for some $C_r' < \infty$*

$$\frac{2\eta}{T} \sum_{t=\lfloor T/2 \rfloor + 1}^{T} \sum_{s=0}^{t-2} \left(\frac{1}{e\eta(t-1-s)} + \lambda \right) (1 + \varphi(\eta s)) \leq 4 \log(T) + C_r' \log(T) \varphi(\eta T) .$$

Proof of Lemma 2.14. 1. Here we use the fact that for any $\alpha > 0$, $1 \leq S \leq T$

$$\sum_{t=S}^{T} t^\alpha \leq \int_S^{T+1} t^\alpha \, dt \leq \frac{2^{\alpha+1}}{\alpha+1} T^{\alpha+1} .$$

Hence,

$$\frac{2}{T} \sum_{t=\lfloor T/2 \rfloor + 1}^{T} t^\alpha \leq \frac{2^{\alpha+2}}{\alpha+1} T^\alpha .$$

2. Observe that for $\alpha \geq 0$

$$\sum_{s=0}^{t-1} \frac{s^\alpha}{t-1-s} \leq 4 t^\alpha \log(t) .$$

Thus, by the first part of the Lemma we find

$$\begin{aligned}\frac{2}{T}\sum_{t=\lfloor T/2\rfloor+1}^{T}\sum_{s=0}^{t-2}\frac{\varphi(\eta s)}{t-1-s} &\leq \frac{8C_r}{T}\sum_{t=\lfloor T/2\rfloor+1}^{T}\log(t)\varphi(\eta t)\\ &\leq 4C_r\log(T)\bar{\varphi}(\eta T)\\ &\leq C_r'\log(T)\varphi(\eta T)\ .\end{aligned}$$

3. Note that for any $t \geq 3$

$$\sum_{s=0}^{t-2}\frac{1}{(t-1-s)} \leq 4\ log(t)$$

$$\begin{aligned}\frac{2\eta}{T}\sum_{t=\lfloor T/2\rfloor+1}^{T}\sum_{s=0}^{t-2}\Big(\frac{1}{e\eta(t-1-s)}+\lambda\Big) &= \frac{2\lambda\eta}{T}\sum_{t=\lfloor T/2\rfloor+1}^{T}\sum_{s=0}^{t-2}1+\frac{2}{eT}\sum_{t=\lfloor T/2\rfloor+1}^{T}\sum_{s=0}^{t-2}\frac{1}{t-1-s}\\ &\leq \lambda\eta T+\frac{8}{eT}\sum_{t=\lfloor T/2\rfloor+1}^{T}log(t)\\ &\leq \lambda\eta T+2\log(T)\ .\end{aligned}$$

The result follows by setting $\lambda=(\eta T)^{-1}$ and with $1\leq 2\log(T)$.

4. This follows immediately from the other parts of the Lemma.

□

3 Results for Tail-Averaged SGD

This section is devoted to providing our final error bound for the second-stage SGD algorithm. Here, we write

$$\mathbb{E}_{\hat{D}|D}[||S_K\bar{h}_T-f_\rho||_{L^2}] \leq \underbrace{\mathbb{E}_{\hat{D}|D}[||S_K\bar{f}_T-f_\rho||_{L^2}]}_{2.\ stage\ \ GD}+\underbrace{\mathbb{E}_{\hat{D}|D}[S_K(\bar{\hat{h}}_T-\bar{\hat{f}}_T)||_{L^2}]}_{2.\ stage\ \ SGD\ \ variance}\ . \tag{22}$$

3.1 Bounding Second Stage SGD Variance

A short calculation shows that the second stage SGD variance can be rewritten as

$$\hat{h}_{t+1}-\hat{f}_{t+1}=(Id-\eta\hat{T}_{t+1})(\hat{h}_t-\hat{f}_t)+\eta\hat{\zeta}_{t+1}$$

where we set $J_t:=\{b(t-1)+1,...,bt\}$ and define

$$\hat{T}_t:=\frac{1}{b}\sum_{i\in J_t}K_{\mu_{\hat{x}_{j_i}}}\otimes K_{\mu_{\hat{x}_{j_i}}}\ ,\quad \hat{g}_t:=\frac{1}{b}\sum_{i\in J_t}y_{j_i}K_{\mu_{\hat{x}_{j_i}}}$$

and

$$\hat{\zeta}_t:=(T_{\hat{\mathbf{x}}}-\hat{T}_t)\hat{f}_t+(\hat{g}_t-g_{\hat{\mathbf{x}}})\ .$$

This gives

$$\mathbb{E}_{J_t}[\hat{\zeta}_t|\hat{D},D]=0$$

and by Lemma 6 in Mücke et al. (2019) we find

$$\mathbb{E}_{J_t}[\hat{\zeta}_t\otimes\hat{\zeta}_t|\hat{D},D]\preceq\frac{1}{b}\Big(\kappa^4\sup_t||\hat{f}_t||^2_{\mathcal{H}_K}+M^2\Big)T_{\hat{\mathbf{x}}}\ . \tag{23}$$

As a preliminary step we need to bound the norm of the second stage GD updates.

Proposition 3.1. *Suppose Assumptions 2.1, 2.2, 3.1, 3.2 and 3.3 are satisfied and let $\eta < 1/\kappa^2$.*

1. **If $f_\rho \in Ran(S_K)$:** *Assume that*
$$n \geq 64e\kappa^2 \log^2(12/\delta)(\eta t)^{1+\nu} . \tag{24}$$
Then
$$\mathbb{E}_{\hat{D}|D}[||\hat{f}_{t+1}||^2_{\mathcal{H}_K}] \leq C_{\alpha,\kappa,M,R}\left(\frac{\eta^2(t+1)^2}{N^\alpha} + 1\right) ,$$
with probability at least $1-\delta$ w.r.t. the data D, for some $C_{\alpha,\kappa,M,R} < \infty$.

2. **If $f_\rho \notin Ran(S_K)$:** *Assume*
$$n \geq 64\kappa^2 \log(12/\delta)(\eta t)\log((\eta t)^\nu) .$$
With probability at least $1-\delta$ w.r.t. the data D we have
$$\mathbb{E}_{\hat{D}|D}[||\hat{f}_{t+1}||^2_{\mathcal{H}_K}] \leq C'_{\alpha,\kappa,M,R}\frac{\eta^2(t+1)^2}{N^\alpha}\big(1+\varphi^2(\eta t)\big) ,$$
for some $C'_{\alpha,\kappa,M,R} < \infty$ and where
$$\varphi(\eta t) = (\eta t)^{\frac{1}{2}\max\{\nu, 1-2r\}} . \tag{25}$$

Proof of Proposition 3.1. We split
$$||\hat{f}_t||^2_{\mathcal{H}_K} \leq 2||\hat{f}_t - f_t||^2_{\mathcal{H}_K} + 2||f_t||^2_{\mathcal{H}_K} . \tag{26}$$
According to (6) we have
$$\hat{f}_{t+1} - f_{t+1} = \eta \sum_{s=0}^{t} (Id - \eta T_{\hat{\mathbf{x}}})^{t-s}\hat{\xi}_s ,$$
where ξ_s is defined in (5). We proceed by using convexity to obtain
$$\begin{aligned}
||\hat{f}_{t+1} - f_{t+1}||^2 &= \eta^2(t+1)^2\left|\left|\frac{1}{t+1}\sum_{s=0}^{t}(Id - \eta T_{\hat{\mathbf{x}}})^{t-s}\hat{\xi}_s\right|\right|^2_{\mathcal{H}_K} \\
&\leq \eta^2(t+1)\sum_{s=0}^{t}\left|\left|(Id - \eta T_{\hat{\mathbf{x}}})^{t-s}\hat{\xi}_s\right|\right|^2_{\mathcal{H}_K} \\
&\leq \eta^2(t+1)\sum_{s=0}^{t}||\hat{\xi}_s||^2_{\mathcal{H}_K} .
\end{aligned}$$
For bounding the noise variables we follow the proof of Proposition 2.9 and distinguish between the two cases:

- $Pf_\rho \in Ran(S_K)$**:** By Lemma 1.3, Lemma 1.4 and Corollary 2.8 we have
$$\mathbb{E}_{\hat{D}|D}[||\hat{\xi}_s||^2_{\mathcal{H}_K}] \leq \frac{c_\alpha}{N^\alpha} ,$$
for some $c_\alpha < \infty$ and holding with probability at least $1-\delta$ w.r.t. the data D, provided (24) is satisfied. Thus,
$$\mathbb{E}_{\hat{D}|D}[||\hat{f}_{t+1} - f_{t+1}||^2_{\mathcal{H}_K}] \leq c_\alpha\frac{\eta^2(t+1)^2}{N^\alpha}$$
in this case. Combining with (26) and Corollary 2.8 once more leads to
$$\mathbb{E}_{\hat{D}|D}[||\hat{f}_{t+1}||^2_{\mathcal{H}_K}] \leq \tilde{c}_\alpha\left(\frac{\eta^2(t+1)^2}{N^\alpha} + 1\right) ,$$
with probability at least $1-\delta$ w.r.t. the data D, for some $\tilde{c}_\alpha < \infty$.

- $Pf_\rho \notin Ran(S_K)$**:** From Corollary 2.8, Lemma 1.2 and Lemma 1.3 we obtain with probability at least $1-\delta$ with respect to the data D

$$\mathbb{E}_{\hat{D}|D}[||\hat{\xi}_s||^2_{\mathcal{H}_K}] \leq \log^2(6/\delta)\frac{\tilde{c}'_\alpha}{N^\alpha}(1+\varphi(\eta s))^2 \ ,$$

for some $c'_\alpha < \infty$. Thus,

$$\mathbb{E}_{\hat{D}|D}[||\hat{f}_{t+1} - f_{t+1}||^2_{\mathcal{H}_K}] \leq 2\tilde{c}'_\alpha \frac{\eta^2(t+1)}{N^\alpha}\sum_{s=0}^{t}(1+\varphi^2(\eta s))$$

and by (26), since φ is non-decreasing in s

$$\begin{aligned}\mathbb{E}_{\hat{D}|D}[||\hat{f}_{t+1}||^2_{\mathcal{H}_K}] &\leq 2\tilde{c}'_\alpha \frac{\eta^2(t+1)}{N^\alpha}\sum_{s=0}^{t}(1+\varphi^2(\eta s)) + c_{\kappa,M,R}\varphi^2(\eta t) \\ &\leq \tilde{c}''_{\alpha,\kappa,M,R}\frac{\eta^2(t+1)^2}{N^\alpha}\left(1+\frac{1}{t+1}\sum_{s=0}^{t}\varphi^2(\eta s)\right) \\ &\leq \tilde{c}''_{\alpha,\kappa,M,R}\frac{\eta^2(t+1)^2}{N^\alpha}\big(1+\varphi^2(\eta t)\big) \ ,\end{aligned}$$

for some $\tilde{c}''_{\alpha,\kappa,M,R} < \infty$ and with probability at least $1-\delta$ with respect to the data D. □

Proposition 3.2 (Second Stage SGD Variance)**.** *Suppose Assumptions 2.2 and 3.1 are satisfied and let $\eta\kappa^2 < 1/4$, $\nu \in (0,1]$. Assume further that $Trace[T^\nu_{\hat{\mathbf{x}}}] \leq C_\nu$ almost surely for some $C_\nu \in \mathbb{R}_+$. The second stage SGD variance satisfies with probability at least $1-\delta$ w.r.t. the data D*

$$\begin{aligned}\mathbb{E}_{\hat{D}|D}[||T_K^{1/2}(\bar{\hat{h}}_T - \bar{\hat{f}}_T)||_{\mathcal{H}_K}] \leq \tilde{C}_{\nu,\kappa,M} 6\log(4/\delta)\sqrt{\frac{\eta}{b}(\eta T)^{\nu-1}}\left(1+\mathbb{E}_{\hat{D}|D}\left[||\hat{f}_T||^2_{\mathcal{H}_K}\right]^{1/2}\right) \\ \left(\frac{\eta T}{\sqrt{n}} + c_\alpha\gamma^\alpha LM\frac{\sqrt{\eta T}}{N^{\frac{\alpha}{2}}}+1\right)^{1/2} \ ,\end{aligned}$$

for some $\tilde{C}_{\nu,\kappa,M} < \infty$.

Proof of Proposition 3.2. Hölder's inequality allows us to write for any $\lambda > 0$

$$\mathbb{E}_{\hat{D}|D}[||T_K^{1/2}(\bar{\hat{h}}_T - \bar{\hat{f}}_T)||_{\mathcal{H}_K}] \leq \left[\mathbb{E}_{\hat{D}|D}[||T_K^{1/2}(T_{\hat{\mathbf{x}}}+\lambda)^{-1/2}||^2]\right]^{\frac{1}{2}}\left[\mathbb{E}_{\hat{D}|D}[||(T_{\hat{\mathbf{x}}}+\lambda)^{1/2}(\bar{\hat{h}}_T - \bar{\hat{f}}_T)||^2_{\mathcal{H}_K}]\right]^{\frac{1}{2}} . \tag{27}$$

For bounding the first term let us firstly observe that by Lemma 1.1 with probability at least $1-\delta$

$$\begin{aligned}\mathbb{E}_{\hat{D}|D}[||T_K^{1/2}(T_{\hat{\mathbf{x}}}+\lambda)^{-1/2}||^2] &\leq ||T_K(T_{\hat{\mathbf{x}}}+\lambda)^{-1}|| \\ &\leq 6\log(2/\delta)\frac{1}{\lambda\sqrt{n}} + \frac{c_\alpha\gamma^\alpha LM}{\sqrt{\lambda}N^{\frac{\alpha}{2}}} + 1 \ .\end{aligned} \tag{28}$$

For bounding the second term we write

$$||(T_{\hat{\mathbf{x}}}+\lambda)^{1/2}(\bar{\hat{h}}_T - \bar{\hat{f}}_T)||^2_{\mathcal{H}_K} = ||T^{1/2}_{\hat{\mathbf{x}}}(\bar{\hat{h}}_T - \bar{\hat{f}}_T)||^2_{\mathcal{H}_K} + \lambda||\bar{\hat{h}}_T - \bar{\hat{f}}_T||^2_{\mathcal{H}_K} \ .$$

Applying Proposition 5 in Mücke et al. (2019) with $\sigma^2 = \frac{1}{b}\mathbb{E}_{\hat{D}|D}\left[\kappa^4||\hat{f}_T||^2 + M^2\right]$ then gives with $\lambda = (\eta T)^{-1}$ and for any $\nu \in (0,1]$

$$\begin{aligned}\mathbb{E}_{\hat{D}|D}[||(T_{\hat{\mathbf{x}}}+\lambda)^{1/2}(\bar{\hat{h}}_T - \bar{\hat{f}}_T)||^2_{\mathcal{H}_K}] &\leq C\frac{\eta}{b}(\eta T)^{\nu-1}\mathbb{E}_{\hat{D}|D}\left[\left(\kappa^4||\hat{f}_T||^2 + M^2\right)Trace[T^\nu_{\hat{\mathbf{x}}}]\right] \\ &\leq \tilde{C}_{\nu,\kappa,M}\frac{\eta}{b}(\eta T)^{\nu-1}\left(\mathbb{E}_{\hat{D}|D}[||\hat{f}_T||^2_{\mathcal{H}_K}]+1\right) \ ,\end{aligned}$$

for some $\tilde{C}_{\nu,\kappa,M} < \infty$. Combining this with (28) and (27) finally leads to the result. □

From Proposition 3.1 and Proposition 3.2 we immediately obtain:

Corollary 3.3 (Second Stage SGD Variance)**.** *In addition to the Assumptions from Proposition 3.2 suppose that Assumptions 3.2, 3.3 are satisfied.*

1. **If** $Pf_\rho \in Ran(S_K)$**:** *Assume that*
$$n \geq 64e\kappa^2 \log^2(12/\delta)(\eta t)^{1+\nu} .$$
Then
$$\mathbb{E}_{\hat{D}|D}[||T_K^{1/2}(\bar{\hat{h}}_T - \bar{\hat{f}}_T)||_{\mathcal{H}_K}] \leq C_{\nu,\kappa,\gamma,M,\alpha,L} \left(1 + \frac{\eta T}{\sqrt{n}} + \frac{\sqrt{\eta T}}{N^{\frac{\alpha}{2}}}\right)^{1/2} \sqrt{\frac{\eta}{b}(\eta T)^{\nu-1}} \left(1 + \frac{\eta T}{N^{\frac{\alpha}{2}}}\right),$$
with probability at least $1-\delta$ *w.r.t. the data* D*, for some* $C_{\nu,\kappa,\gamma,M,\alpha,L} < \infty$*.*

2. **If** $Pf_\rho \notin Ran(S_K)$**:** *Assume that*
$$n \geq 64e\kappa^2 \log^2(24/\delta)(\eta t)\log((\eta T)^\nu) .$$
Then, with probability at least $1-\delta$ *w.r.t. the data* D *we have*
$$\mathbb{E}_{\hat{D}|D}[||T_K^{1/2}(\bar{\hat{h}}_T - \bar{\hat{f}}_T)||_{\mathcal{H}_K}] \leq \tilde{C}_{\nu,\kappa,\gamma,M,\alpha,L} \log(6/\delta) \left(1 + \frac{\eta T}{\sqrt{n}} + \frac{\sqrt{\eta T}}{N^{\frac{\alpha}{2}}}\right)^{1/2} \sqrt{\frac{\eta}{b}(\eta T)^{\nu-1}} \left(1 + \varphi(\eta T)\frac{\eta T}{N^{\frac{\alpha}{2}}}\right),$$
for some $\tilde{C}_{\nu,\kappa,\gamma,M,\alpha,L} < \infty$ *and where* φ *is defined in* (25)*.*

3.2 Main Result Second Stage Tail-Averaged SGD

Combining now (22) with Proposition 3.2, Theorem 3.4 and Theorem 2.11 finally leads to our main results.

Theorem 3.4 (Excess Risk Second-Stage tail-ave GD; Part I)**.** *Suppose Assumptions 2.2, 2.2 are satisfied. Let additionally Assumptions 3.2 and 3.3 hold. Let* $T \in \mathbb{N}$ *and denote*
$$B_T = \max\{2M, ||S_K\bar{u}_T - f_\rho||_\infty\} .$$
Let further $\delta \in (0,1]$*,* $\lambda = (\eta T)^{-1}$*,* $\eta\kappa^2 < 1/4$*, assume* $0 < r \leq 1$ *and recall the definition of* $\mathcal{B}(1/\eta T)$ *in* (10) *and of* φ *in* (11)*. With probability not less than* $1-\delta$*, the excess risk for the second-stage tail-averaged SGD satisfies:*

1. **If** $1/2 \leq r \leq 1$**:**
$$\begin{aligned}\mathbb{E}_{\hat{D}|D}[||S_K\bar{\hat{h}}_T - f_\rho||_{L^2}] \leq\ & C_1 \log(24/\delta)\frac{1}{\sqrt{n}}\left(M\sqrt{\mathcal{N}(\lambda)} + \frac{||S_K\bar{u}_T - f_\rho||_{L^2}}{\sqrt{\lambda}} + \frac{B_T}{\sqrt{n\lambda}}\right) \\ &+ C_2\lambda\, ||(T_K+\lambda)^{-1/2}\bar{u}_T||_{\mathcal{H}_K} + C_3||S_K\bar{u}_T - f_\rho||_{L^2} \\ &+ C_4 \log(8/\delta)\log(T)\frac{\sqrt{\eta T}}{N^{\frac{\alpha}{2}}}\mathcal{B}(1/\eta T) \\ &+ C_5 \log(8/\delta)\sqrt{\frac{\eta}{b}(\eta T)^{\nu-1}}\left(1 + \mathbb{E}_{\hat{D}|D}\left[||\hat{f}_T||^2_{\mathcal{H}_K}\right]^{1/2}\right)\left(\frac{\eta T}{\sqrt{n}} + \frac{\sqrt{\eta T}}{N^{\frac{\alpha}{2}}} + 1\right)^{1/2},\end{aligned}$$
for some constants $C_1 > 0$*,* $C_2 > 0$*,* $C_3 > 0$*,* $C_4 > 0$*,* $C_5 > 0$*.*

2. **If** $0 < r \leq 1/2$**:**
$$\begin{aligned}\mathbb{E}_{\hat{D}|D}[||S_K\bar{\hat{h}}_T - f_\rho||_{L^2}] \leq\ & C_1 \log(24/\delta)\frac{1}{\sqrt{n}}\left(M\sqrt{\mathcal{N}(\lambda)} + \frac{||S_K\bar{u}_T - f_\rho||_{L^2}}{\sqrt{\lambda}} + \frac{B_T}{\sqrt{n\lambda}}\right) \\ &+ C_2\lambda\, ||(T_K+\lambda)^{-1/2}\bar{u}_T||_{\mathcal{H}_K} + C_3||S_K\bar{u}_T - f_\rho||_{L^2} \\ &+ C_4 \log(8/\delta)\log(T)\frac{\sqrt{\eta T}\mathcal{B}(1/\eta T)}{N^{\frac{\alpha}{2}}}(1+\varphi(\eta T)) \\ &+ C_5 \log(8/\delta)\sqrt{\frac{\eta}{b}(\eta T)^{\nu-1}}\left(1 + \mathbb{E}_{\hat{D}|D}\left[||\hat{f}_T||^2_{\mathcal{H}_K}\right]^{1/2}\right)\left(\frac{\eta T}{\sqrt{n}} + \frac{\sqrt{\eta T}}{N^{\frac{\alpha}{2}}} + 1\right)^{1/2},\end{aligned}$$
for some constants $C_1 > 0$*,* $C_2 > 0$*,* $C_3 > 0$*,* $C_4 > 0$*,* $C_5 > 0$*.*

Theorem 3.5 (Excess Risk Second-Stage tail-ave GD; Part II)**.** *Suppose Assumptions 2.2, 2.2 are satisfied. Let additionally Assumptions 3.2 and 3.3 hold. Let $T \in \mathbb{N}$ and denote*

$$B_T = \max\{2M, ||S_K \bar{u}_T - f_\rho||_\infty\} .$$

Let further $\delta \in (0,1]$, $\lambda = (\eta T)^{-1}$, $\eta\kappa^2 < 1/4$, assume that $r \geq 1$ and recall the definition of $\mathcal{B}(1/\eta T)$ in (10). Then with probability not less than $1-\delta$, the excess risk for the second-stage tail-averaged SGD satisfies

$$\begin{aligned}
\mathbb{E}_{\hat{D}|D}[||S_K \bar{\hat{h}}_T - f_\rho||_{L^2}] \leq\ & C_1' \log(24/\delta) \frac{1}{\sqrt{n}} \left(M\sqrt{\mathcal{N}(\lambda)} + \frac{||S_K \bar{u}_T - f_\rho||_{L^2}}{\sqrt{\lambda}} + \frac{B_T}{\sqrt{n\lambda}} \right) \\
& + C_2' \lambda^{1/2}\, ||T_K^{-r} \bar{u}_T||_{\mathcal{H}_K} \left(\frac{\log(4/\delta)}{\sqrt{n}} + \lambda^\zeta \right) + C_3' ||S_K \bar{u}_T - f_\rho||_{L^2} \\
& + C_4' \log(8/\delta) \log(T) \frac{\sqrt{\eta T}}{N^{\frac{\alpha}{2}}} \mathcal{B}(1/\eta T) \\
& + C_5' \log(8/\delta) \sqrt{\frac{\eta}{b}(\eta T)^{\nu-1}} \left(1 + \mathbb{E}_{\hat{D}|D}\left[||\hat{f}_T||^2_{\mathcal{H}_K} \right]^{1/2} \right) \left(\frac{\eta T}{\sqrt{n}} + \frac{\sqrt{\eta T}}{N^{\frac{\alpha}{2}}} + 1 \right)^{1/2} ,
\end{aligned}$$

for some constants $C_1' > 0$, $C_2' > 0$, $C_3' > 0$, $C_4' > 0$, $C_5' > 0$.

Corollary 3.6 (Learning Rates Second Stage Ave-SGD Mis-Specified Model)**.** *Suppose all assumptions of Theorem 3.4 and Theorem 3.5 are satisfied. Assume additionally that $r \leq 1/2$, $K > 1$, $\eta_0 < \frac{1}{4\kappa^2}$ and*

$$n \geq 64 e \kappa^2 \log(4/\delta)(\eta T) \log((\eta T)^\nu) .$$

1. *Let $2r + \nu > 1$. Then, for any n sufficiently large, the excess risk satisfies with probability at least $1-\delta$ w.r.t. the data D*

$$\mathbb{E}_{\hat{D}|D}[||S_K \bar{\hat{h}}_{T_n} - f_\rho||_{L^2}]^2 \leq C \log(24/\delta) R \left(\frac{M^2}{R^2 n} \right)^{\frac{r}{2r+\nu}} ,$$

provided $N_n \geq \log^{2/\alpha}(n) \left(\frac{R^2}{M^2} n \right)^{\frac{2+\nu}{\alpha(2r+\nu)}}$ and

- *Multi-pass SGD: $b_n = \sqrt{n}$, $\eta_n = \eta_0$ and $T_n = \left(\frac{R^2 n}{\sigma^2} \right)^{\frac{1}{2r+\nu}}$,*
- *Batch GD: $b_n = n$, $\eta_n = \eta_0$ and $T_n = \left(\frac{R^2 n}{\sigma^2} \right)^{\frac{1}{2r+\nu}}$.*

2. *Let $2r + \nu \leq 1$. Then, for any n sufficiently large, the excess risk satisfies with probability at least $1-\delta$ w.r.t. the data D*

$$\mathbb{E}_{\hat{D}|D}[||S_K \bar{\hat{h}}_{T_n} - f_\rho||_{L^2}]^2 \leq C \log(24/\delta) R \left(\frac{M^2 \log^K(n)}{R^2 n} \right)^r ,$$

provided $N_n \geq \log^{2/\alpha}(n) \left(\frac{R^2 n}{M^2 \log^K(n)} \right)^{\frac{3-2r}{\alpha}}$ and

- *$b_n = 1$, $\eta_n = \left(\frac{M^2 \log^K(n)}{R^2 n} \right)^{2r+\nu}$ and $T_n = \left(\frac{R^2 n}{M^2 \log^K(n)} \right)^{2r+\nu+1}$,*
- *$b_n = \left(\frac{R^2 n}{M^2 \log^K(n)} \right)^{2r+\nu}$, $\eta_n = \eta_0$ and $T_n = \frac{R^2 n}{M^2 \log^K(n)}$,*
- *$b_n = \frac{R^2 n}{M^2 \log^K(n)}$, $\eta_n = \eta_0$ and $T_n = \frac{R^2 n}{M^2 \log^K(n)}$.*

Proof of Corollary 3.6. Here, we combine the results from Corollary 2.12 and Corollary 3.3. We have to show that

$$\underbrace{\left(1+\frac{\eta T}{\sqrt{n}}+\frac{\sqrt{\eta T}}{N^{\frac{\alpha}{2}}}\right)^{1/2}}_{\mathcal{T}_1}\ \underbrace{\left(1+\varphi(\eta T)\frac{\eta T}{N^{\frac{\alpha}{2}}}\right)}_{\mathcal{T}_2}\ \underbrace{\sqrt{\frac{\eta}{b}(\eta T)^{\nu-1}}}_{\mathcal{T}_3}$$

is of optimal order under appropriate choices of all parameters.

1. Let $2r+\nu>1$ and $\eta_n T_n=\left(\frac{R^2}{M^2}n\right)^{\frac{1}{2r+\nu}}$. Given this choices, one easily verifies that the leading order term in $\mathcal{T}_1$ is given by $\left(\frac{\eta_n T_n}{\sqrt{n}}\right)^{1/2}$, provided

$$N_n\geq\left(\frac{n}{\eta_n T_n}\right)^{1/\alpha}\sim\left(\frac{R^2}{M^2}n\right)^{\frac{2r+\nu-1}{\alpha(2r+\nu)}}.$$

Moreover, we have $\varphi(\eta T)=(\eta T)^{\nu/2}$ and the second term $\mathcal{T}_2$ is of order 1 if

$$(\eta_n T_n)^{1+\nu/2}N^{-\frac{\alpha}{2}}\lesssim 1$$

hence if

$$N_n\geq\left(\frac{R^2}{M^2}n\right)^{\frac{2+\nu}{\alpha(2r+\nu)}},$$

for n sufficiently large. Note that we have

$$\max\left\{\frac{2r+\nu-1}{\alpha(2r+\nu)},\frac{2+\nu}{\alpha(2r+\nu)}\right\}=\frac{2+\nu}{\alpha(2r+\nu)}.$$

Finally, we have to determine now appropriate values of η_n, T_n, b_n such that

$$\left(\frac{M^2}{R^2n}\right)^{\frac{2r+\nu-2}{2(2r+\nu)}}\ \frac{\eta_n}{b_n}\left(\frac{R^2}{M^2}n\right)^{\frac{\nu-1}{2r+\nu}}\lesssim R\left(\frac{M^2}{R^2n}\right)^{\frac{2r}{2r+\nu}},$$

that is, if

$$\frac{\eta_n}{b_n}\lesssim R\left(\frac{M^2}{R^2n}\right)^{\frac{1}{2}}.$$

This is surely satisfied by all the given choices.

2. Let $2r+\nu\leq 1$ and $\eta_n T_n=\frac{R^2 n}{M^2\log^K(n)}$ for some $K>1$. Again, the leading order term in $\mathcal{T}_1$ is given by $\left(\frac{\eta_n T_n}{\sqrt{n}}\right)^{1/2}$, provided

$$N_n^{\alpha/2}\geq\sqrt{\frac{n}{\eta_n T_n}}\sim\log^{K/2}(n),$$

or equivalently,

$$N_n\geq\log^{K/\alpha}(n).$$

For bounding $\mathcal{T}_2$ note that $\varphi(\eta T)=(\eta T)^{\frac{1}{2}-r}$. Then, $\mathcal{T}_2$ is of order 1 if

$$N_n\geq\left(\frac{R^2n}{M^2\log^K(n)}\right)^{\frac{3-2r}{\alpha}}.$$

Finally, we have to determine now appropriate values of η_n, T_n, b_n such that

$$\frac{R^2n}{M^2\log^K(n)}\ \frac{\eta_n}{b_n}\left(\frac{R^2n}{M^2\log^K(n)}\right)^{\nu-1}\lesssim R\left(\frac{M^2\log^K(n)}{R^2n}\right)^{2r},$$

that is, if

$$\frac{\eta_n}{b_n} \lesssim R\left(\frac{M^2 \log^K(n)}{R^2 n}\right)^{2r+\nu} .$$

This is surely satisfied by all the given choices.

□

Corollary 3.7 (Learning Rates Second Stage Ave-SGD Well-Specified Model)**.** *Suppose all assumptions of Theorem 3.4 and Theorem 3.5 are satisfied. Assume additionally that* $r \geq \frac{1}{2}$ *and*

$$n \geq 64 e \kappa^2 \log(4/\delta)(\eta T)^{1+\nu} .$$

Let $\eta_0 < \frac{1}{4\kappa^2}$ *and choose* $N_n \geq \log^{2/\alpha}(n)\left(\frac{R^2 n}{\sigma^2}\right)^{\frac{2r+1}{\alpha(2r+\nu)}}$*. Then, for any* n *sufficiently large, the excess risk satisfies with probability at least* $1-\delta$ *w.r.t. the data* D

$$\mathbb{E}_{\hat{D}|D}[||S_K \bar{\hat{h}}_{T_n} - P f_\rho||_{L^2}]^2 \leq C \log(24/\delta) R\left(\frac{\sigma^2}{R^2 n}\right)^{\frac{r}{2r+\nu}} ,$$

for each of the following choices:

1. *One-pass SGD:* $b = 1$*,* $\eta_n = \eta_0 \frac{R^2}{\sigma^2}\left(\frac{\sigma^2}{R^2 n}\right)^{\frac{2r+\nu-1}{2r+\nu}}$ *and* $T_n = \frac{R^2}{\sigma^2} n$*,*
2. *Early stopping and one-pass SGD:* $b = n^{\frac{2r+\nu-1}{2r+\nu}}$*,* $\eta_n = \eta_0$ *and* $T_n = \left(\frac{R^2 n}{\sigma^2}\right)^{\frac{1}{2r+\nu}}$*,*
3. *Batch-GD:* $b = n$*,* $\eta_n = \eta_0$ *and* $T_n = \left(\frac{R^2 n}{\sigma^2}\right)^{\frac{1}{2r+\nu}}$*.*

Proof of Corollary 3.7. The proof follows the same lines as the proof of Corollary 3.6 by standard calculations. □

References

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018.

Andrea Caponnetto and E. De Vito. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2006.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *Ann. Statist.*, 44(4):1363–1399, 08 2016. doi: 10.1214/15-AOS1391.

Zhiying Fang, Zheng-Chu Guo, and Ding-Xuan Zhou. Optimal learning rates for distribution regression. *Journal of Complexity*, 56:101426, 2020.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithm. *arXiv preprint arXiv:1702.07254*, 2017.

Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020.

Nicole Mücke, Gergely Neu, and Lorenzo Rosasco. Beating sgd saturation with tail-averaging and minibatching. In *Advances in Neural Information Processing Systems*, pages 12568–12577, 2019.