

# Distributed SGD in overparameterized Linear Regression

Anonymous authors

Paper under double-blind review

## Abstract

We consider distributed learning using constant stepsize SGD over several devices, each sending a final model update to a central server. In a final step, the local estimates are aggregated. We prove in the setting of overparameterized linear regression general upper bounds with matching lower bounds and derive learning rates for specific data generating distributions. We show that the excess risk is of order of the variance provided the number of local nodes grows not too large with the global sample size.

We further compare distributed SGD with distributed ridge regression and provide an upper bound of the excess SGD-risk in terms of the excess RR-risk for a certain range of the sample size.

## 1 INTRODUCTION

Deep neural networks possess powerful generalization properties in various machine learning applications, despite being overparameterized. It is generally believed that the optimization algorithm itself, e.g., stochastic gradient descent (SGD), implicitly regularizes such overparameterized models. This regularizing effect due to the choice of the optimization algorithm is often referred to as *implicit regularization*. A refined understanding of this phenomenon was recently gained in the setting of linear regression (to be considered as a reasonable approximation of neural network learning) for different variants of SGD. Constant stepsize SGD (with last iterate or tail-averaging) is investigated in Jain et al. (2018), in Dieuleveut & Bach (2016) in an RKHS framework and also in Mücke et al. (2019) with additional mini-batching, see also Mücke & Reiss (2020) for a more general analysis in Hilbert scales. In Zou et al. (2021b;a) it is shown that benign overfitting also occurs for SGD. Multi-pass SGD is analyzed in Lin et al. (2016); Jain et al. (2016); Lin & Rosasco (2017); Zou et al. (2022) while last iterate bounds can be found in Jain et al. (2019); Wu et al. (2022); Varre et al. (2021).

Despite the attractive statistical properties of all these SGD variants, the complexity of computing regression estimates prevents it from being routinely used in large-scale problems. More precisely, the time complexity and space complexity of SGD and other regularization methods in a standard implementation scale as  $\mathcal{O}(n^\alpha)$ ,  $\alpha \in [2, 3]$ . Such scalings are prohibitive when the sample size  $n$  is large.

Distributed learning (DL) based on a divided-and-conquer approach is an effective way to analyze large scale data that can not be handled by a single machine. In this paper we study a distributed learning strategy in linear regression (including both underparameterized and overparameterized regimes) via (tail-) averaged stochastic gradient descent with constant stepsize (DSGD). The approach is quite simple and communication efficient: The training data is distributed across several computing nodes where on each a local SGD is run. In a final step, these local estimates are aggregated (a.k.a. *one-shot SGD*). Local SGD has become state of the art in large scale distributed learning, showing a linear speed-up in the number of workers for convex problems, see e.g. McDonald et al. (2009); Zinkevich et al. (2010); Dieuleveut & Patel (2019); Stich (2018); Spiridonoff et al. (2021) and references therein.

The field of DL has gained increasing attention in statistical learning theory with the aim of deriving conditions under which minimax optimal rates of convergence can be guaranteed, see e.g. Chen & Xie (2014), Mackey et al. (2011), Xu et al. (2019), Fan et al. (2019), Shi et al. (2018), Battay et al. (2018), Fan et al. (2021), Bao & Xiong (2021). Indeed, the learning properties of DL in regression settings over Hilbert

spaces are widely well understood. The authors in Zhang et al. (2015) analyze distributed (kernel) ridge regression and show optimal learning rates with appropriate regularization, provided the number of machines increases sufficiently slowly with the sample size, though under restrictive assumptions on the eigenfunctions of the kernel integral operator. This has been alleviated in Lin et al. (2017). However, in these works the number of machines *saturates* if the target is very smooth, meaning that large parallelization seems not possible in this regime.

An extension of these works to more general spectral regularization algorithms for nonparametric least square regression in (reproducing kernel) Hilbert spaces is given in Guo et al. (2017), Mücke & Blanchard (2018), including gradient descent (Lin & Zhou, 2018) and stochastic gradient descent (Lin & Cevher, 2018). The recent work Tong (2021) studies DL for functional linear regression.

We finally mention the work of Mücke et al. (2022), where distributed ordinary least squares (DOLS) in overparameterized linear regression is studied, i.e. one-shot OLS without any explicit or implicit regularization. It is shown that the number of workers acts as a regularization parameter itself.

**Contributions.** We analyze the performance of DSGD with constant stepsize in overparameterized linear regression and provide upper bounds with matching lower bounds for the excess risk under suitable noise assumptions. Our results show that optimal rates of convergence can be achieved if the number of local nodes grows sufficiently slowly with the sample size. The excess risk as a function of data splits remains constant until a certain threshold is reached. This threshold depends on the structural assumptions imposed on the problem, i.e. on the eigenvalue decay of the Hessian and the coefficients of the true regression parameter.

We additionally perform a comparison between DSGD and DRR, showing that the excess risk of DSGD is upper bounded by the excess risk of DRR under an assumption on the sample complexity (SC) of DSGD, depending on the same structural assumptions. We show that the SC of DSGD remains within constant factors of the SC of DRR.

Our analysis extends known results in this direction from Zou et al. (2021b;a) for the single machine case to the distributed learning setting and from DOLS in Mücke et al. (2022) to SGD with implicit regularization.

**Organization.** In Section 2 we define the mathematical framework needed to present our main results in Section 3, where we provide a theoretical analysis of DSGD with a discussion of our results. In Section 4 we compare DSGD with DRR while Section 5 is devoted to showing some numerical illustrations. The proofs are deferred to the Appendix.

**Notation.** By  $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$  we denote the space of bounded linear operators between real Hilbert spaces  $\mathcal{H}_1, \mathcal{H}_2$ . We write  $\mathcal{L}(\mathcal{H}, \mathcal{H}) = \mathcal{L}(\mathcal{H})$ . For  $\mathbf{A} \in \mathcal{L}(\mathcal{H})$  we denote by  $\mathbf{A}^T$  the adjoint operator. By  $\mathbf{A}^\dagger$  we denote the pseudoinverse of  $\mathbf{A}$  and for  $w \in \mathcal{H}$  we write  $\|w\|_{\mathbf{A}}^2 := \|\mathbf{A}^{\frac{1}{2}}w\|^2$  for an PSD operator  $\mathbf{A}$ .

We let  $[n] = \{1, \dots, n\}$  for every  $n \in \mathbb{N}$ . For two positive sequences  $(a_n)_n, (b_n)_n$  we write  $a_n \lesssim b_n$  if  $a_n \leq cb_n$  for some  $c > 0$  and  $a_n \simeq b_n$  if both  $a_n \lesssim b_n$  and  $b_n \lesssim a_n$ .

## 2 SETUP

In this section we provide the mathematical framework for our analysis. More specifically, we introduce distributed SGD and state the main assumptions on our model.

### 2.1 SGD and linear regression

We consider a linear regression model over a real separable Hilbert space  $\mathcal{H}$  in random design. More precisely, we are given a random covariate vector  $x \in \mathcal{H}$  and a random output  $y \in \mathbb{R}$  following the model

$$y = \langle w^*, x \rangle + \epsilon, \tag{1}$$

where  $\epsilon \in \mathbb{R}$  is a noise variable. We will impose some assumptions on the noise model in Section 3. The true regression parameter  $w^* \in \mathcal{H}$  minimizes the least squares test risk, i.e.

$$L(w^*) = \min_{w \in \mathcal{H}} L(w), \quad L(w) := \frac{1}{2} \mathbb{E}[(y - \langle w, x \rangle)^2],$$

where the expectation is taken with respect to the joint distribution  $\mathbb{P}$  of the pair  $(x, y) \in \mathcal{H} \times \mathbb{R}$ . More specifically, we let  $w^*$  be the minimum norm element in the set of all minimizers of  $L$ .

To derive an estimator  $\hat{w} \in \mathcal{H}$  for  $w^*$  we are given an i.i.d. dataset

$$D := \{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathcal{H} \times \mathbb{R},$$

following the above model equation 1, i.e.,

$$\mathbf{Y} = \mathbf{X}w^* + \varepsilon,$$

with i.i.d. noise  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n) \in \mathbb{R}^n$ . The corresponding random vector of outputs is denoted as  $\mathbf{Y} = (y_1, \dots, y_n)^T \in \mathbb{R}^n$  and we arrange the data  $x_j \in \mathcal{H}$  into a *data matrix*  $\mathbf{X} \in \mathcal{L}(\mathcal{H}, \mathbb{R}^n)$  by setting  $(\mathbf{X}v)_j = \langle x_j, v \rangle$  for  $v \in \mathcal{H}, 1 \leq j \leq n$ . If  $\mathcal{H} = \mathbb{R}^d$ , then  $\mathbf{X}$  is a  $n \times d$  matrix (with row vectors  $x_j$ ). We are particularly interested in the overparameterized regime, i.e. where  $\dim(\mathcal{H}) > n$ .

In the classical setting of stochastic approximation with constant stepsize, the SGD iterates are computed by the recursion

$$w_{t+1} = w_t - \gamma(\langle w_t, x_t \rangle - y_t)x_t, \quad t = 1, \dots, n,$$

with some initialization  $w_1 \in \mathcal{H}$  and where  $\gamma > 0$  is the stepsize. The tail average of the iterates is denoted by

$$\bar{w}_{\frac{n}{2}:n} := \frac{1}{n - n/2} \sum_{t=n/2+1}^n w_t, \quad (2)$$

and where we denote by  $\bar{w}_n := \bar{w}_{0:n}$  the full (uniform) average.

Various forms of SGD (with iterate averaging, tail averaging, multi passes) in the setting of overparameterized linear regression has been analyzed recently in Zou et al. (2021b), Wu et al. (2022), Zou et al. (2022), respectively. In particular, the phenomenon of *benign overfitting* is theoretically investigated in these works. It could be shown that benign overfitting occurs in this setting, i.e. the SGD estimator fits training data very well and still generalizes.

We are interested in this phenomenon for localized SGD, i.e. when our training data is distributed over several computing devices.

## 2.2 Local SGD

In the distributed setting, our data are evenly divided into  $M \in \mathbb{N}$  local disjoint subsets

$$D = D_1 \cup \dots \cup D_M$$

of size  $|D_m| = \frac{n}{M}$ , for  $m = 1, \dots, M$ . To each local dataset we associate a *local design matrix*  $\mathbf{X}_m \in \mathcal{L}(\mathcal{H}, \mathbb{R}^{\frac{n}{M}})$  (build with local row vectors  $x_j^{(m)}$ ) with local output vector  $\mathbf{Y}_m \in \mathbb{R}^{\frac{n}{M}}$  and a local noise vector  $\varepsilon_m \in \mathbb{R}^{\frac{n}{M}}$ .

The local SGD iterates are defined as

$$w_{t+1}^{(m)} = w_t^{(m)} - \gamma \left( \left\langle w_t^{(m)}, x_t^{(m)} \right\rangle - y_t \right) x_t^{(m)},$$

for  $t = 1, \dots, \frac{n}{M}$  and  $m = 1, \dots, M$ . The averaged local iterates  $\bar{w}_{\frac{n}{M}}^{(m)}$  are computed according to equation 2. We are finally interested in the uniform average of the local SGD iterates, building a global estimator:

$$\bar{\bar{w}}_M := \frac{1}{M} \sum_{m=1}^M \bar{w}_{\frac{n}{M}}^{(m)}.$$

Distributed learning in overparameterized linear regression is studied in Mücke et al. (2022) for the ordinary least squares estimator (OLS), i.e. without any implicit or explicit regularization and with local interpolation. It is shown that local overfitting is harmless and regularization is done by the number of data splits.

We aim at finding optimal bounds for the excess risk

$$\mathbb{E}[L(\bar{w}_M)] - L(w^*) ,$$

of distributed SGD (DSGD) with potential local overparameterization and as function of the number of local nodes  $M$  and under various model assumptions, to be given in the next section.

### 3 MAIN RESULTS

In this section we present our main results. To do so, we first impose some model assumptions.

**Definition 3.1.**

1. We define the second moment of  $x \sim \mathbb{P}_x$  to be the operator  $\mathbf{H} : \mathcal{H} \rightarrow \mathcal{H}$ , given by

$$\mathbf{H} := \mathbb{E}[x \otimes x] = \mathbb{E}[\langle \cdot, x \rangle x] .$$

2. The fourth moment operator  $\mathbf{M} : \mathcal{L}(\mathcal{H}) \rightarrow \mathcal{L}(\mathcal{H})$  is defined by

$$\mathbf{M} := \mathbb{E}[x \otimes x \otimes x \otimes x] ,$$

with  $\mathbf{M}(\mathbf{A})(w) = \mathbb{E}[\langle x, \mathbf{A}x \rangle \langle w, x \rangle x]$ , for all  $w \in \mathcal{H}$ .

3. The covariance operator of the gradient noise at  $w^*$  is defined as  $\mathbf{\Sigma} : \mathcal{H} \rightarrow \mathcal{H}$ ,

$$\mathbf{\Sigma} := \mathbb{E}[\langle (w^*, x) - y \rangle^2 x \otimes x] .$$

**Assumption 3.2** (Second Moment Condition). *We assume that  $\mathbb{E}[y^2|x] < \infty$  almost surely. Moreover, we assume that the trace of  $\mathbf{H}$  is finite, i.e.,  $\text{Tr}[\mathbf{H}] < \infty$ .*

**Assumption 3.3** (Fourth Moment Condition). *We assume there exists a positive constant  $\tau > 0$  such that for any PSD operator  $\mathbf{A}$ , we have*

$$\mathbf{M}(\mathbf{A}) \preceq \tau \text{Tr}[\mathbf{H}\mathbf{A}]\mathbf{H} .$$

Note that this assumption holds if  $\mathbf{H}^{-1}x$  is sub-Gaussian, being a standard assumption in least squares regression, see e.g. Bartlett et al. (2020), Zou et al. (2021b), Tsigler & Bartlett (2020).

**Assumption 3.4** (Noise Condition). *Assume that*

$$\sigma^2 := \|\mathbf{H}^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{H}^{-\frac{1}{2}}\| < \infty .$$

This assumption on the noise is standard in the literature about averaged SGD, see e.g. Zou et al. (2021b), Dieuleveut & Bach (2016).

We introduce some further notation involving the second moment operator  $\mathbf{H}$ : We denote the eigendecomposition as

$$\mathbf{H} = \sum_{j=1}^{\infty} \lambda_j v_j \otimes v_j ,$$

where the  $\lambda_1 \geq \lambda_2 \geq \dots$  are the eigenvalues of  $\mathbf{H}$  and the  $v_j$ 's are the corresponding eigenvectors. For  $k \geq 1$ , we let

$$\mathbf{H}_{0:k} := \sum_{j=1}^k \lambda_j v_j \otimes v_j , \quad \mathbf{H}_{k:\infty} := \sum_{j=k+1}^{\infty} \lambda_j v_j \otimes v_j .$$

Similarly,

$$\mathbf{I}_{0:k} = \sum_{j=1}^k v_j \otimes v_j, \quad \mathbf{I}_{k:\infty} := \sum_{j=k+1}^{\infty} v_j \otimes v_j.$$

A short calculation shows that for all  $w \in \mathcal{H}$  we have

$$\|w\|_{\mathbf{H}_{0:k}^\dagger}^2 = \sum_{j=1}^k \frac{\langle w, v_j \rangle^2}{\lambda_j}, \quad \|w\|_{\mathbf{H}_{k:\infty}}^2 = \sum_{j=k+1}^{\infty} \lambda_j \langle w, v_j \rangle^2.$$

We finally set

$$V_k(n, M) := \frac{k}{n} + \gamma^2 \frac{n}{M^2} \sum_{j=k+1}^{\infty} \lambda_j^2. \quad (3)$$

### 3.1 Upper Bound

We now present an upper bound for the averaged local SGD iterates. The proof relies on a bias-variance decomposition and is given in Appendix A.1.

**Theorem 3.5** (DSGD Upper Bound). *Suppose Assumptions 3.2, 3.3 and 3.4 are satisfied and let  $\gamma < \frac{1}{\tau \text{Tr}[\mathbf{H}]}$ ,  $w_1 = 0$ . The excess risk for the averaged local SGD estimate satisfies*

$$\mathbb{E}[L(\bar{w}_M)] - L(w^*) \leq 2\text{Bias}(\bar{w}_M) + 2\text{Var}(\bar{w}_M),$$

where

$$\text{Bias}(\bar{w}_M) \leq \frac{M^2}{\gamma^2 n^2} \|w^*\|_{\mathbf{H}_{0:k^*}^\dagger}^2 + \|w^*\|_{\mathbf{H}_{k^*:\infty}}^2 + \frac{2\tau M^2 (\|w^*\|_{\mathbf{I}_{0:k^*}}^2 + \gamma \frac{n}{M} \|w^*\|_{\mathbf{H}_{k^*:\infty}}^2)}{\gamma n (1 - \gamma \tau \text{Tr}[\mathbf{H}])} \cdot V_{k^*}(n, M)$$

and

$$\text{Var}(\bar{w}_M) \leq \frac{\sigma^2}{1 - \gamma \tau \text{Tr}[\mathbf{H}]} \cdot V_{k^*}(n, M),$$

with  $k^* = \max\{k : \lambda_k \geq \frac{M}{\gamma n}\}$ .

The excess risk is upper bounded in terms of the bias and variance. Both terms crucially depend on the *effective dimension*  $k^* = \max\{k : \lambda_k \geq \frac{M}{\gamma n}\}$ , dividing the full Hilbert space  $\mathcal{H}$  into two parts. On the part associated to the first largest  $k^*$  eigenvalues, the bias may decay faster than on the remaining tail part that is associated to the smaller eigenvalues, see Zou et al. (2021b) in the context of single machine SGD, Bartlett et al. (2020); Tsigler & Bartlett (2020), in the context of single machine ridge regression and Mücke et al. (2022) for distributed ordinary least squares.

Our Theorem 3.5 reveals that the excess risk converges to zero if

$$\|w^*\|_{\mathbf{H}_{k^*:\infty}}^2 \rightarrow 0, \quad \frac{2M^2}{\gamma^2 n^2} \|w^*\|_{\mathbf{H}_{0:k^*}^\dagger}^2 \rightarrow 0$$

and  $V_{k^*}(n, M) \rightarrow 0$  as  $n \rightarrow \infty$ . This requires the eigenvalues of  $\mathbf{H}$  to decay sufficiently fast and to choose the number of local nodes  $M = M_n$  to be a sequence of  $n$ . Note that we have to naturally assume  $M_n \lesssim n$ . In Subsection 3.3 we provide two specific examples of data distributions with specific choices for  $(M_n)_{n \in \mathbb{N}}$  such that the above conditions are met, granting not only convergence but also providing explicit rates of convergence.

### 3.2 Lower Bound

Before we state the lower bounds for the excess risk of the DSGD estimator we need to impose some assumptions.

**Assumption 3.6** (Fourth Moment Lower Bound). *We assume there exists a positive constant  $\theta > 0$  such that for any PSD operator  $\mathbf{A}$ , we have*

$$\mathbf{M}(\mathbf{A}) - \mathbf{H}\mathbf{A}\mathbf{H} \succeq \theta \text{Tr}[\mathbf{H}\mathbf{A}]\mathbf{H}.$$

**Assumption 3.7** (Well-Specified Noise). *The second moment operator  $\mathbf{H}$  is strictly positive definite with  $\text{Tr}[\mathbf{H}] < \infty$ . Moreover, the noise  $\epsilon$  in equation 1 is independent of  $x$  and satisfies*

$$\epsilon \sim \mathcal{N}(0, \sigma_{noise}^2).$$

We now come to the main result whose proof can be found in Appendix A.2.

**Theorem 3.8** (DSGD Lower Bound). *Suppose Assumptions 3.6 and 3.7 are satisfied. Assume  $w_1 = 0$ . The excess risk of the DSGD estimator satisfies*

$$\mathbb{E}[L(\bar{w}_M)] - L(w^*) \geq \frac{M(M-1)}{100\gamma^2 n^2} \left( \|w^*\|_{\mathbf{H}_{0:k^*}^\dagger}^2 + \frac{\gamma^2 n^2}{M^2} \|w^*\|_{\mathbf{H}_{k^*:\infty}}^2 \right) + \frac{\sigma_{noise}^2}{100} \cdot V_{k^*}(n, M),$$

where  $V_{k^*}(n, M)$  is defined in equation 3.

The lower bound for the excess risk also decomposes into a bias part (first term) and a part associated to the variance (second term). Comparing the bias with the upper bound for the bias from Theorem 3.5 shows that both are of the same order. Comparing the variances reveals that they are of the same order if

$$\frac{2\tau M^2 (\|w^*\|_{\mathbf{H}_{0:k^*}}^2 + \gamma \frac{n}{M} \|w^*\|_{\mathbf{H}_{k^*:\infty}}^2)}{\gamma n (1 - \gamma\tau \text{Tr}[\mathbf{H}])} \lesssim 1.$$

In the next section, we will provide specific conditions and examples when this is satisfied.

### 3.3 Fast Rates of convergence for specific distributions

We now consider two particular cases of data distributions, namely the *spiked covariance model* (with local overparameterization) and the case where the eigenvalues of the second moment operator  $\mathbf{H}$  decay *polynomially*. These are standard assumptions for the model, see e.g. Tsigler & Bartlett (2020); Zou et al. (2021b); Mücke et al. (2022). In both cases, we determine a range of the number of local nodes  $M_n$  depending on the global sample size such that the bias is dominated by the variance. The final error is then of the order of the variance, if the number of local nodes grows sufficiently slowly with the sample size. The optimal<sup>1</sup> number exactly balances bias and variance.

**Corollary 3.9** (Spiked Covariance Model). *Suppose all assumptions of Theorem 3.5 are satisfied. Assume that  $\|w^*\| \leq R$  for some  $R > 0$  and  $\mathbf{H} \in \mathbb{R}^{d \times d}$ . Let  $d = \left(\frac{n}{M}\right)^q$  for some  $q > 1$  and  $\tilde{d} = \left(\frac{n}{M}\right)^r < d$  for some  $0 < r \leq 1$ . Suppose the spectrum of  $\mathbf{H}$  satisfies*

$$\lambda_j = \begin{cases} \frac{1}{\tilde{d}} & : j \leq \tilde{d} \\ \frac{1}{d-\tilde{d}} & : \tilde{d} + 1 \leq j \leq d. \end{cases}$$

If

$$M_n \leq \sqrt{\frac{\gamma(1-2\gamma\tau)n}{R^2}}$$

<sup>1</sup>Optimal in the sense of the maximal possible number of local nodes that balances bias and variance.

then for any  $n$  sufficiently large, we have

$$\mathbb{E}[L(\bar{w}_{M_n})] - L(w^*) \leq c \frac{1}{\gamma M_n} \left(\frac{M_n}{n}\right)^\nu,$$

where  $\nu = \min\{1 - r, q - 1\}$  and for some  $c < \infty$ , depending on  $\tau, \gamma, \sigma$ .

Choosing the maximum number of local nodes  $M_n \simeq \sqrt{n}$  gives the fast rate of order

$$\mathbb{E}[L(\bar{w}_{M_n})] - L(w^*) \lesssim \left(\frac{1}{n}\right)^{\frac{\nu+1}{2}}.$$

for the excess risk.

**Corollary 3.10** (Polynomial Decay). *Suppose all assumptions of Theorem 3.5 are satisfied with  $\gamma < \min\left\{1, \frac{1}{\tau \text{Tr}[\mathbf{H}]}\right\}$ . Assume that  $\|w^*\| \leq R$  for some  $R > 0$ . Suppose the spectrum<sup>2</sup> of  $\mathbf{H}$  satisfies for some  $r > 0$*

$$\lambda_j = j^{-(1+r)}.$$

If

$$M_n \leq \left(\frac{\gamma}{R^2}\right)^{\frac{1+r}{2+r}} \cdot (\gamma n)^{\frac{1}{2+r}},$$

then for any  $n$  sufficiently large, we have

$$\mathbb{E}[(\bar{w}_M)] - L(w^*) \leq c \frac{\gamma}{M_n} \left(\frac{M_n}{n}\right)^{\frac{r}{1+r}},$$

for some  $c < \infty$ , depending on  $\tau, \gamma, \sigma$ .

Choosing the maximum number of local nodes  $M_n \simeq n^{\frac{1}{2+r}}$  gives the fast rate of order

$$\mathbb{E}[L(\bar{w}_{M_n})] - L(w^*) \lesssim \left(\frac{1}{n}\right)^{\frac{r+1}{r+2}}.$$

for the excess risk.

### 3.4 Discussion

**Comparison to single machine SGD.** We compare the DSGD algorithm with the single machine SGD algorithm, i.e. when  $M = 1$ . For this case, we recover the results from Zou et al. (2021b) under the same assumptions. Our Corollaries 3.9, 3.10 show that the excess risk is dominated by the variance as long as  $M$  grows sufficiently slowly with the sample size. But we can say even more: In the spiked covariance model, if  $M_n \simeq n^\beta$  for  $\beta \in [0, 1/2]$ , we see that DSGD performs as good as single machine SGD, provided  $\nu \leq 1$ . Indeed, a direct comparison shows that

$$\frac{1}{\gamma M_n} \left(\frac{M_n}{n}\right)^\nu \simeq \frac{1}{\gamma n^\beta} \left(\frac{n^\beta}{n}\right)^\nu \simeq \frac{1}{\gamma} \left(\frac{1}{n}\right)^\nu,$$

for any  $\beta \in [0, 1/2]$  and  $\nu \leq 1$ . Recall that all our bounds are of optimal order, hence the relative efficiency remains of constant order until the critical threshold for  $M_n$  is reached.

However, if  $M_n$  is larger than the threshold, i.e. if  $\beta \in (1/2, 1]$ , then the bias term is dominating. In this case, the excess risk is of order

$$\frac{2M^2}{n^2} \|w^*\|_{\mathbf{H}_{0:k^*}^\dagger}^2 + \|w^*\|_{\mathbf{H}_{k^*:\infty}}^2 \simeq \left(\frac{M_n}{n}\right)^{2-r} + \left(\frac{M_n}{n}\right)^q \simeq \left(\frac{n^\beta}{n}\right)^{2-r} + \left(\frac{n^\beta}{n}\right)^q,$$

<sup>2</sup>Note that the choice  $\lambda_j = j^{-(1+r)}$  ensures that  $\text{Tr}[\mathbf{H}] < \infty$ .

being larger than the variance, see the proof of Corollary 3.9, Appendix A.3.

The same observations can be made for the setting in Corollary 3.10 when the eigenvalues are polynomially decaying. If we let  $M_n \simeq n^\beta$  with  $\beta \in [0, 1/(2+r)]$ , then the variance dominates and for all  $r > 0$ , the test error satisfies

$$\frac{1}{\gamma M_n} \left( \frac{M_n}{n} \right)^{\frac{r}{r+1}} \simeq \frac{1}{\gamma n^\beta} \left( \frac{n^\beta}{n} \right)^{\frac{r}{r+1}} \simeq \frac{1}{\gamma} \left( \frac{1}{n} \right)^{\frac{r}{r+1}}.$$

We refer to Section 5 and Section C for some numerical experiments.

**Comparison to distributed learning in RKHSs.** We emphasize that all our results above hold for a constant stepsize  $0 < \gamma < \min\left\{1, \frac{1}{\tau \text{Tr}[\mathbf{H}]}\right\}$ . In particular,  $\gamma$  does not depend on the number  $M$  of local nodes. This result is line with the results for regularized distributed learning over reproducing kernel Hilbert spaces, see Zhang et al. (2015); Lin et al. (2017); Mücke & Blanchard (2018) and references therein. In this setting it is shown for a large class of spectral regularization methods<sup>3</sup> that the optimal regularization parameter  $\lambda$  that leads to minimax optimal bounds, depends on the global sample size only and is of order  $n^{-\alpha}$ ,  $\alpha \in (0, 1]$ . In particular, this parameter is chosen as in the single machine machine setting and each local subproblem is underregularized. This leads to a roughly constant bias (unchanged by averaging) in the distributed setting, an increase in variance but averaging reduces the variance sufficiently to obtain optimal excess risk bounds. The same phenomenon occurs in our DSGD setting. On each local node the same stepsize  $\gamma$  as for the  $M = 1$  case is applied.

**Comparison to distributed ordinary least squares (DOLS).** We also compare our results with those recently obtained in Mücke et al. (2022) for DOLS in random design linear regression. The general observation in this work is that in the presence of overparameterization, the number of local nodes acts as a regularization parameter, balancing bias and variance. Recall that this is in contrast to what we observe for DSGD due to the implicit regularization. The optimal number of splits  $M_{opt}^{OLS}$  depends on structural assumptions, i.e. eigenvalue decay and decay of the Fourier coefficients of  $w^*$  (a.k.a. *source condition*).

For the spiked covariance model, the optimal number  $M_n^{OLS}$  of DOLS is of order

$$M_n^{OLS} \simeq \left( \frac{dn^{3/2}}{d \cdot \bar{d}} \right)^{2/5} \simeq n^{\frac{3-2r}{5-2r}},$$

see Corollary 3.14 in Mücke et al. (2022). Comparing with our maximum number for  $M_n \simeq n^{1/2}$  from our Corollary 3.9 we observe that  $M_n^{OLS} \lesssim M_n^{SGD}$  if  $\frac{1}{2} \leq r \leq 1$ , i.e., DSGD allows for more parallelization in this regime.

For polynomially decaying eigenvalues  $\lambda_j \sim j^{1+r}$ ,  $r > 0$ , the optimal number of data splits in Corollary 3.9 (Mücke et al., 2022) scales as  $M_n^{OLS} \simeq n^{1/3}$ . Compared to our result from Corollary 3.10 we have

$$M_n^{SGD} \simeq n^{\frac{1}{2+r}} \lesssim n^{1/3}$$

for all  $r \geq 1$ . Thus, DOLS seems to allow more data splits under optimality guarantees for fast polynomial decay, i.e. large  $r$ .

## 4 COMPARISON OF SAMPLE COMPLEXITY OF DSGD AND DRR

In this section we compare the distributed tail-averaged SGD estimator with the distributed Ridge Regression (RR) estimator (see Zhang et al. (2015); Lin et al. (2017); Mücke & Blanchard (2018); Sheng & Dobriban (2020) or Tsigler & Bartlett (2020) for RR in the single machine case). Recall that RR reduces to ordinary least-squares (OLS) if the regularization parameter is set to zero. As a special case, we compare our results to local OLS from Mücke et al. (2022) and analyze the benefit of implicit regularization of local SGD in the presence of local overparameterization.

<sup>3</sup>This class contains, among others, gradient descent and accelerated methods like Heavy ball and Nesterov, ridge regression or PCA.



We recall that for any  $m \in [M]$ ,  $\lambda \geq 0$ , the local RR estimates are defined by

$$\hat{w}_m^{\text{RR}}(\lambda) = \mathbf{X}_m^T (\mathbf{X}_m \mathbf{X}_m^T + \lambda)^{-1} \mathbf{Y}_m.$$

The average is

$$\bar{w}_n^{\text{RR}}(\lambda) = \frac{1}{M} \sum_{m=1}^M \hat{w}_m^{\text{RR}}.$$

We aim at showing that the excess risk of DSGD is upper bounded by the excess risk of DRR under suitable assumptions on the sample complexity. To this end, we first derive a lower bound for DRR to compare with. The proof follows by combining Proposition B.3 and Proposition B.5 with Lemma B.2.

**Assumption 4.1.** *The variable  $\mathbf{H}^{-1}x$  is sub-Gaussian and has independent components.*

Similarly to the bounds for DSGD, our bounds for DRR depend on the effective dimension

$$k_{\text{RR}}^* := \min \left\{ k : \lambda_{k+1} \leq \frac{M \left( \lambda + \sum_{j>k} \lambda_j \right)}{bn} \right\},$$

for  $\lambda > 0$  and some  $b > 1$ .

**Theorem 4.2** (Lower Bound Distributed RR). *Suppose Assumption 4.1 holds and that  $\mathbf{H}$  is strictly positive definite with  $\text{Tr}[\mathbf{H}] < \infty$ . Assume that  $k_{\text{RR}}^* \leq \frac{n}{c'M}$  for some  $c' > 1$ . There exist constants  $b, c > 1$  such that the excess risk of the averaged RR estimator satisfies*

$$\begin{aligned} \mathbb{E}[L(\bar{w}_n^{\text{RR}}(\lambda))] - L(w^*) &\geq \|w^*\|_{\mathbf{H}_{k_{\text{RR}}^*}^*}^2 + \frac{M^2 \left( \lambda + \sum_{j>k_{\text{RR}}^*} \lambda_j \right)^2}{cn^2} \cdot \|w^*\|_{\mathbf{H}_{0:k_{\text{RR}}^*}^{-1}}^2 \\ &\quad + \frac{\sigma^2}{c} \left( \frac{k_{\text{RR}}^*}{n} + \frac{n}{M^2} \cdot \frac{\sum_{j>k_{\text{RR}}^*} \lambda_j^2}{\left( \lambda + \sum_{j>k_{\text{RR}}^*} \lambda_j \right)^2} \right). \end{aligned}$$

We do our risk comparison particularly for tail-averaged DSGD and derive a bias-improved upper bound. The proof is given in Section B.2 and is an extension of Lemma 6.1 in Zou et al. (2021a) to DSGD.

**Theorem 4.3** (Upper Bound Tail-averaged DSGD). *Suppose Assumption 3.7 is satisfied. Let  $\bar{\bar{w}}_{M_n}$  denote the tail-averaged distributed estimator with  $n$  training samples and assume  $\gamma < 1/\text{Tr}[H]$ . For arbitrary  $k_1, k_2 \in [d]$*

$$\mathbb{E}[L(\bar{\bar{w}}_M)] - L(w^*) = \text{Bias}(\bar{\bar{w}}_M) + \text{Var}(\bar{\bar{w}}_M)$$

with

$$\text{Bias}(\bar{\bar{w}}_M) \leq \frac{c_b M^2}{\gamma^2 n^2} \cdot \left\| \exp\left(-\frac{n}{M} \gamma \mathbf{H}\right) w^* \right\|_{\mathbf{H}_{0:k_1}^{-1}}^2 + \|w^*\|_{\mathbf{H}_{k_1:\infty}}^2,$$

$$\text{Var}(\bar{\bar{w}}_M) \leq c_v (1 + R^2) \cdot \sigma^2 \left( \frac{k_2}{n} + \frac{n\gamma^2}{M^2} \cdot \sum_{j>k_2} \lambda_j^2 \right),$$

for some universal constants  $c_b, c_v > 0$ .

To derive the risk comparison we fix a sample size  $n_{\text{RR}}$  and  $n_{\text{SGD}}$  for DRR and tail-averaged DSGD, resp., and derive conditions on the sample complexities such that individually, the bias and variance of DSGD is upper bounded by the bias and variance of DRR, respectively. Combining then both of the above theorems finally leads to the risk comparison result. A detailed computation is given in Section B.3.

**Theorem 4.4** (Comparison DSGD with DRR). *Let  $\bar{w}_{M_{n_{\text{SGD}}}}$  denote the tail-averaged distributed estimator with  $n_{\text{SGD}}$  training samples. Let further  $\bar{w}_{n_{\text{RR}}}^{\text{RR}}(\lambda)$  denote the distributed RR estimator with  $n_{\text{RR}}$  training samples and with regularization parameter  $\lambda \geq 0$ . Suppose all assumptions from Theorems 4.2 ,4.3 are satisfied. There exist constants  $b, c > 1$  and  $0 < L_{\lambda, \gamma} \leq L'_{\lambda, \gamma}$  such that for  $C^* := c\left(1 + \frac{\|w^*\|^2}{\sigma^2}\right)$ ,*

$$C^*_\lambda := \lambda + \sum_{j > k_{\text{RR}}^*} \lambda_j, \quad \gamma < \min\left\{\frac{1}{\text{Tr}[H]}, \frac{1}{\sqrt{c}C^*_\lambda}\right\} \quad (4)$$

and

$$L_{\lambda, \gamma} \cdot n_{\text{RR}} \leq n_{\text{SGD}} \leq L'_{\lambda, \gamma} \cdot n_{\text{RR}}$$

the excess risks of DSGD and DRR satisfy

$$\mathbb{E}[L(\bar{w}_M)] - L(w^*) \leq \mathbb{E}[L(\bar{w}_{n_{\text{RR}}}^{\text{RR}}(\lambda))] - L(w^*). \quad (5)$$

The constants  $L_{\lambda, \gamma}, L'_{\lambda, \gamma}$  are explicitly given by

$$L_{\lambda, \gamma} = \max\left\{C^*, \frac{\sqrt{c(1 - \gamma\lambda_{k_{\text{RR}}^*})}}{\gamma C^*_\lambda}\right\}, \quad L'_{\lambda, \gamma} = \frac{1}{C^*\gamma^2(C^*_\lambda)^2}.$$

Note that in the above Theorem, assumption equation 4 on the stepsize ensures that  $0 < L_{\lambda, \gamma} \leq L'_{\lambda, \gamma}$ . Our result shows that DSGD performs better than DRR/ DOLS if the sample complexity (SC) of SGD differs from the SC of RR/OLS by no more than a constant. This constant depends on the amount of regularization  $\lambda$ , the stepsize  $\gamma$  and the tail behavior of the eigenvalues of the Hessian. We refer to Section B.3.2 for a more detailed discussion. Our bound slightly differs from Zou et al. (2021a) for the case  $M = 1$  in two respects: We scale our SC such that the constant in equation 5 is equal to one while Zou et al. (2021a) show that both risks are of the same order (with a constant larger than one). Second, we also show that the SC of DSGD is upper bounded by a factor of the SC of DRR/DOLS while Zou et al. (2021a) only derive a lower bound. However, we remark that  $n_{\text{SGD}}$  in our Theorem is larger than  $n_{\text{RR}}$  as the constant  $L_{\lambda, \gamma} \geq 1$ . A look onto Figure 1 reveals that optimally tuned DSGD may perform better than optimally tuned DRR even with the same or smaller sample size for certain problem instances. This suggests that our bound may be refined.

## 5 NUMERICAL EXPERIMENTS

We illustrate our theoretical findings with experiments on simulated and real data. The reader may find additional experiments in Section C.

**Simulated Data.** In a first experiment in Figure 1 (left) we analyze the test error of DSGD as a function of the local nodes  $M$ . We generate  $n = 500$  i.i.d. training data with  $x_j \sim \mathcal{N}(0, \mathbf{H})$  with mildly overparameterization  $d = 700$ . The target  $w^*$  satisfies three different decay conditions  $w_j^* = j^{-\alpha}$ ,  $\alpha \in \{0, 1, 10\}$ . The eigenvalues of  $\mathbf{H}$  follow a polynomial decay  $\lambda_j = j^{-2}$ . The local nodes satisfy  $M_n = n^\beta$ ,  $\beta \in \{0, 0.1, \dots, 0.9\}$ . According to Corollary 3.10 we see that a fast decay of  $w_j^*$  (i.e. a smaller norm  $\|w^*\|$ ) allows for more parallelization until the test error blows up.

In a second experiment we compare the sample complexity of optimally tuned tail-averaged DSGD and DRR for different sources  $w^*$ , see Figures 1 (right), 2. Here, the data are generated as above with  $d = 200$ ,  $\lambda_j = j^{-2}$  and  $w_j^* = j^{-\alpha}$ ,  $\alpha \in \{0, 1, 10\}$ . The number of local nodes is fixed at  $M_n = n^{1/3}$  for each  $n \in \{100, \dots, 6000\}$ . For this problem instance, DSGD may perform even better than DRR for sparse targets ( $\alpha = 10$ ), i.e., DSGD achieves the same accuracy as DRR with less samples in this regime. For less sparse targets  $\alpha = 1$ , the sample complexities of DSGD and DRR are comparable while for non-sparse targets ( $\alpha = 0$ ), DRR outperforms DSGD.

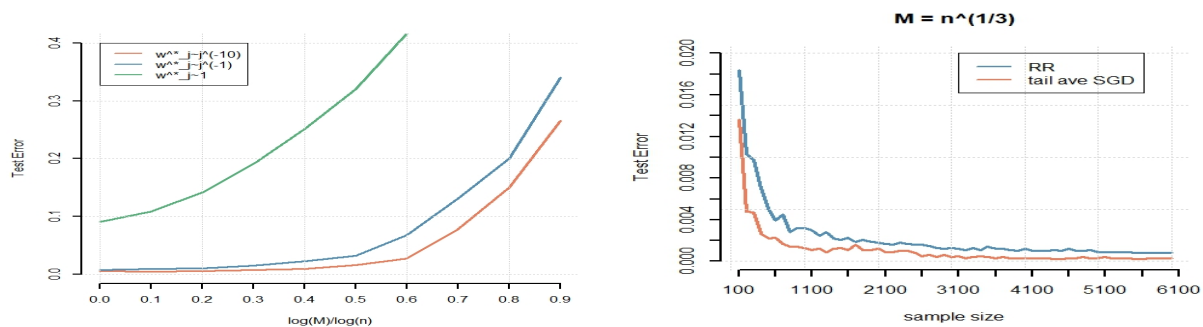


Figure 1: **Left:** Test error for DSGD with  $\lambda_j = j^{-2}$  for different sources  $w^*$  as a function of  $M$ . **Right:** Comparison of optimally tuned tail-ave DSGD with DRR with  $\lambda_j = j^{-2}$ ,  $w_j^* = j^{-10}$ ,  $M_n = n^{1/3}$ .

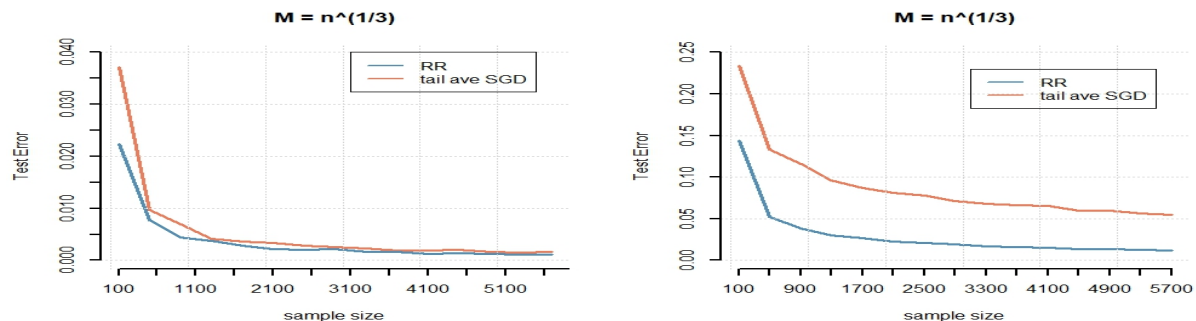


Figure 2: Comparison of optimally tuned tail-ave DSGD with DRR with  $\lambda_j = j^{-2}$  for different sources  $w^*$ , with  $\lambda_j = j^{-2}$  and  $M_n = n^{1/3}$ . **Left:**  $w_j^* = j^{-1}$  **Right:**  $w_j^* = 1$ .

**Real Data.** To analyze the performance of DSGD on real data, we considered the classification problem of the Gisette data set<sup>4</sup>, containing pictures of the digits four and nine. We used the first 3000 samples of the original train data set for training and the second 3000 samples for evaluation. The feature dimension of one picture is  $d = 5000$ . Hyper-parameters had been fine-tuned on the validation data set to achieve the best performance. The first experiment in Figure 3(left) again analyzes the test error of DSGD as a function of the local nodes  $M$ . Because the feature dimension is quite large, the optimal stepsize is small ( $\gamma \sim 10^{-10}$ ). Theorem 3.8 therefore explains why in our example the bias-term and thus the test error grows rather quickly with the number of local nodes. In Figure 3(right) we compare DRR with tail- and full-averaged DSGD. We observe that DRR slightly outperforms DSG. According to Theorem 4.4, we need sparsity for  $w^*$  so that DSGD can keep up with DRR. This might be not the case for the Gisette data set.

## 6 Summary

We analyzed the performance of distributed constant stepsize (tail-) averaged SGD for linear regression in an overparameterized regime. We find that the relative efficiency as a function of the number of workers remains largely unchanged until a certain threshold is reached. This threshold depends on the structural assumptions imposed by the problem at hand (eigenvalue decay of the Hessian  $\mathbf{H}$  and the norm of the target  $w^*$ ). This is in contrast to distributed OLS without any implicit or explicit regularization with local overparameterization, where the number of workers itself acts as a regularization parameter, see Figure 4 in Appendix C.

We also compared the sample complexity of DSGD and DRR and find that the sample complexity of DSGD remains within constant factors of the sample complexity of DRR. For some problem instances, tail-averaged

<sup>4</sup><http://archive.ics.uci.edu/ml/datasets/Gisette>

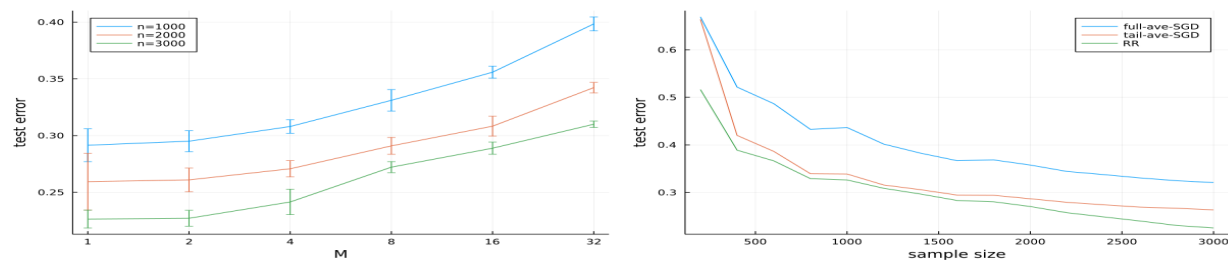


Figure 3: **Left:** Test error for DSGD with  $n = 1000, 2000, 3000$  and different  $M$ . **Right:** Comparison of DSGD with DRR for  $M_n = n^{1/4}$ .

SGD may outperform DRR, i.e., achieves the same or better accuracy with less samples. Our bound is not sharp and may be improved in future research.

## References

- Yajie Bao and Weijia Xiong. One-round communication efficient distributed m-estimation. In *International Conference on Artificial Intelligence and Statistics*, pp. 46–54. PMLR, 2021. 1
- Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020. 3, 3.1, B.1.3
- Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *Annals of statistics*, 46(3):1352, 2018. 1
- Xueying Chen and Min-ge Xie. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, pp. 1655–1684, 2014. 1
- 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. 1, 3
- Aymeric Dieuleveut and Kumar Kshitij Patel. Communication trade-offs for local-sgd with large step size. *Advances in Neural Information Processing Systems*, 32, 2019. 1
- Jianqing Fan, Dong Wang, Kaizheng Wang, and Ziwei Zhu. Distributed estimation of principal eigenspaces. *Annals of statistics*, 47(6):3009, 2019. 1
- Jianqing Fan, Yongyi Guo, and Kaizheng Wang. Communication-efficient accurate statistical estimation. *Journal of the American Statistical Association*, pp. 1–11, 2021. 1
- Zheng-Chu Guo, Shao-Bo Lin, and Ding-Xuan Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7):074009, 2017. 1
- P. Jain, S.M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging and model misspecification. *arXiv:1610.03774v3*, 2016. 1, A.1.1
- Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, Venkata Krishna Pillutla, and Aaron Sidford. A markov chain theory approach to characterizing the minimax optimality of stochastic gradient descent (for least squares). In *37th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science*, 2018. 1
- Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of sgd information theoretically optimal. In *Conference on Learning Theory*, pp. 1752–1755. PMLR, 2019. 1
- Junhong Lin and Volkan Cevher. Optimal distributed learning with multi-pass stochastic gradient methods. In *International Conference on Machine Learning*, pp. 3092–3101. PMLR, 2018. 1

- Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research* 18, 2017. 1
- Junhong Lin, Raffaello Camoriano, and Lorenzo Rosasco. Generalization properties and implicit regularization for multiple passes sgm. *International Conference on Machine Learning*, 2016. 1
- Shao-Bo Lin and Ding-Xuan Zhou. Distributed kernel-based gradient descent algorithms. *Constructive Approximation*, 47(2):249–276, 2018. 1
- Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. *The Journal of Machine Learning Research*, 18(1):3202–3232, 2017. 1, 3.4, 4
- Lester Mackey, Ameet Talwalkar, and Michael I Jordan. Divide-and-conquer matrix factorization. *Advances in neural information processing systems*, 24, 2011. 1
- Ryan Mcdonald, Mehryar Mohri, Nathan Silberman, Dan Walker, and Gideon Mann. Efficient large-scale distributed training of conditional maximum entropy models. *Advances in neural information processing systems*, 22, 2009. 1
- Nicole Mücke and Gilles Blanchard. Parallelizing spectrally regularized kernel algorithms. *The Journal of Machine Learning Research*, 19(1):1069–1097, 2018. 1, 3.4, 4
- Nicole Mücke and Enrico Reiss. Stochastic gradient descent in hilbert scales: Smoothness, preconditioning and earlier stopping. *stat*, 1050:18, 2020. 1
- Nicole Mücke, Gergely Neu, and Lorenzo Rosasco. Beating sgd saturation with tail-averaging and mini-batching. In *Advances in Neural Information Processing Systems*, pp. 12568–12577, 2019. 1
- Nicole Mücke, Enrico Reiss, Jonas Rungenhagen, and Markus Klein. Data-splitting improves statistical performance in overparameterized regimes. In *International Conference on Artificial Intelligence and Statistics*, pp. 10322–10350. PMLR, 2022. 1, 1, 2.2, 3.1, 3.3, 3.4, 4
- Yue Sheng and Edgar Dobriban. One-shot distributed ridge regression in high dimensions. In *International Conference on Machine Learning*, pp. 8763–8772. PMLR, 2020. 4
- Chengchun Shi, Wenbin Lu, and Rui Song. A massive data framework for m-estimators with cubic-rate. *Journal of the American Statistical Association*, 113(524):1698–1709, 2018. 1
- Artin Spiridonoff, Alex Olshevsky, and Yannis Paschalidis. Communication-efficient sgd: From local sgd to one-shot averaging. *Advances in Neural Information Processing Systems*, 34:24313–24326, 2021. 1
- Sebastian U Stich. Local sgd converges fast and communicates little. In *International Conference on Learning Representations*, 2018. 1
- Hongzhi Tong. Distributed least squares prediction for functional linear regression. *Inverse Problems*, 38(2):025002, 2021. 1
- Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020. 3, 3.1, 3.3, 4, B.1
- Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of sgd for least-squares in the interpolation regime. *Advances in Neural Information Processing Systems*, 34: 21581–21591, 2021. 1
- Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Last iterate risk bounds of sgd with decaying stepsize for overparameterized linear regression. In *International Conference on Machine Learning*, pp. 24280–24314. PMLR, 2022. 1, 2.1
- Ganggang Xu, Zuofeng Shang, and Guang Cheng. Distributed generalized cross-validation for divide-and-conquer kernel ridge regression and its asymptotic optimality. *Journal of computational and graphical statistics*, 28(4):891–908, 2019. 1

Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research*, 16(1):3299–3340, 2015. 1, 3.4, 4

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex Smola. Parallelized stochastic gradient descent. *Advances in neural information processing systems*, 23, 2010. 1

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, Dean P Foster, and Sham Kakade. The benefits of implicit regularization from sgd in least squares problems. *Advances in Neural Information Processing Systems*, 34:5456–5468, 2021a. 1, 1, 4, 4, 1, 2, B.1.2, B.1.3, B.2

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Benign overfitting of constant-stepsizes sgd for linear regression. In *Conference on Learning Theory*, pp. 4633–4635. PMLR, 2021b. 1, 1, 2.1, 3, 3, 3.1, 3.3, 3.4, A.1.1, A.1.1, A.1.1, A.1.2, A.2.1, A.2.1, A.2.2

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. Risk bounds of multi-pass sgd for least squares in the interpolation regime. *arXiv preprint arXiv:2203.03159*, 2022. 1, 2.1

**Notation.** By  $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$  we denote the space of bounded linear operators between real Hilbert spaces  $\mathcal{H}_1, \mathcal{H}_2$  with operator norm  $\|\cdot\|$ . We write  $\mathcal{L}(\mathcal{H}, \mathcal{H}) = \mathcal{L}(\mathcal{H})$ . For  $\mathbf{A} \in \mathcal{L}(\mathcal{H})$  we denote by  $\mathbf{A}^T$  the adjoint operator. For two PSD operators on  $\mathcal{H}$  we write  $\mathbf{A} \preceq \mathbf{B}$  if  $\langle (\mathbf{A} - \mathbf{B})v, v \rangle \geq 0$  for all  $v \in \mathcal{H}$ . We further let  $\langle \mathbf{A}, \mathbf{B} \rangle_{op} = \text{Tr}[\mathbf{A}^T \mathbf{B}]$ .

## A PROOFS SECTION 3 (BOUNDS FOR DSGD)

### A.1 Proofs Upper Bound

#### A.1.1 Bias-Variance Decomposition

We will use an iterative bias-variance-decomposition which has been extensively studied before in the non distributed case (see Jain et al. (2016), Zou et al. (2021b)). First we need a couple of definitions.

-) **Centered local iterates:** Set  $\eta_t^{(m)} := w_t^{(m)} - w^*$  and

$$\bar{\eta}_n^{(m)} := \frac{M}{n} \sum_{t=1}^{n/M} \eta_t^{(m)}, \quad \bar{\eta}_M := \frac{1}{M} \sum_{m=1}^M \bar{\eta}_n^{(m)}.$$

-) **Local bias:** For  $m = 1, \dots, M$  we set  $b_1^{(m)} = w_1 - w^*$ ,

$$b_t^{(m)} := (\mathbf{I} - \gamma x_t^{(m)} \otimes x_t^{(m)}) b_{t-1}^{(m)}, \quad t = 2, \dots, \frac{n}{M}$$

$$\bar{b}_n^{(m)} := \frac{M}{n} \sum_{t=1}^{n/M} b_t^{(m)}, \quad \bar{b}_M := \frac{1}{M} \sum_{m=1}^M \bar{b}_n^{(m)}.$$

-) **Local variance:** For  $m = 1, \dots, M$  we set  $v_1^{(m)} = 0$  and

$$v_t^{(m)} := (\mathbf{I} - \gamma x_t^{(m)} \otimes x_t^{(m)}) v_{t-1}^{(m)} + \gamma \epsilon_t^{(m)} x_t^{(m)}, \quad t = 2, \dots, \frac{n}{M},$$

$$\bar{v}_n^{(m)} := \frac{M}{n} \sum_{t=1}^{n/M} v_t^{(m)}, \quad \bar{v}_M := \frac{1}{M} \sum_{m=1}^M \bar{v}_n^{(m)},$$

where we let  $\epsilon_t^{(m)} := y_t^{(m)} - \langle x_t^{(m)}, w^* \rangle$ .

Note that for any  $m = 1, \dots, M$  and  $t \geq 1$  one has

$$\mathbb{E}[b_{t+1}^{(m)}] = \mathbb{E}[\mathbb{E}[b_{t+1}^{(m)} | b_t^{(m)}]] = \mathbb{E}[\mathbb{E}[(\mathbf{I} - \gamma x_{t+1}^{(m)} \otimes x_{t+1}^{(m)}) b_t^{(m)} | b_t^{(m)}]] = (\mathbf{I} - \gamma \mathbf{H}) \mathbb{E}[b_t^{(m)}]. \quad (6)$$

Moreover, from B.4 in Zou et al. (2021b), we find

$$\mathbb{E}[v_{t+1}^{(m)}] = (\mathbf{I} - \gamma H) \mathbb{E}[v_t^{(m)}] = (\mathbf{I} - \gamma H)^t \mathbb{E}[v_1^{(m)}] = 0. \quad (7)$$

It is easy to see that  $\eta_t^{(m)} = b_t^{(m)} + v_t^{(m)}$  and therefore

$$\bar{\eta}_M = \bar{b}_M + \bar{v}_M. \quad (8)$$

**Lemma A.1.** *Define*

$$\text{Bias}(\bar{w}_M) := \frac{1}{2} \langle \mathbf{H}, \mathbb{E}[\bar{b}_M \otimes \bar{b}_M] \rangle_{op}, \quad \text{Var}(\bar{w}_M) := \frac{1}{2} \langle \mathbf{H}, \mathbb{E}[\bar{v}_M \otimes \bar{v}_M] \rangle_{op}.$$

a) *We have the following decomposition for the excess risk,*

$$\mathbb{E}[L(\bar{w}_M)] - L(w^*) \leq \left( \sqrt{\text{Bias}(\bar{w}_M)} + \sqrt{\text{Var}(\bar{w}_M)} \right)^2.$$

b) *Suppose the model noise  $\epsilon_t^{(m)}$  is well-specified, i.e.,  $\epsilon_t^{(m)} := y_t^{(m)} - \langle x_t^{(m)}, w^* \rangle$  and  $x_t^{(m)}$  are independent and  $\mathbb{E}[\epsilon_t^{(m)}] = 0$ , then we have the following equality for the excess risk,*

$$\mathbb{E}[L(\bar{w}_M)] - L(w^*) = \text{Bias}(\bar{w}_M) + \text{Var}(\bar{w}_M).$$

*Proof of Lemma A.1.* The proof strategy is similar to the non distributed case (see Zou et al. (2021b), Lemma B2 and Lemma C1). For completeness we included it here.

a) By definition of the excess risk we have

$$\begin{aligned} L(\bar{w}_M) - L(w^*) &= \frac{1}{2} \int_{\mathcal{H}} \langle \bar{w}_M - w^*, x \rangle^2 \mathbb{P}_{\mathbf{x}}(d\mathbf{x}) \\ &= \frac{1}{2} \langle \mathbf{H}(\bar{w}_M - w^*), \bar{w}_M - w^* \rangle \\ &= \frac{1}{2} \|\mathbf{H}^{\frac{1}{2}}(\bar{w}_M - w^*)\|^2 \\ &= \frac{1}{2} \|\bar{b} + \bar{v}\|_{\mathbf{H}}^2, \end{aligned}$$

where we used (8) for the last equality. Using Cauchy-Schwarz inequality we obtain

$$\begin{aligned} \mathbb{E}[L(\bar{w}_M) - L(w^*)] &\leq \left( \sqrt{\frac{1}{2} \mathbb{E} \|\bar{b}\|_{\mathbf{H}}^2} + \sqrt{\frac{1}{2} \mathbb{E} \|\bar{v}\|_{\mathbf{H}}^2} \right)^2 \\ &= \left( \sqrt{\frac{1}{2} \langle \mathbf{H}, \mathbb{E}[\bar{b}_M \otimes \bar{b}_M] \rangle_{op}} + \sqrt{\frac{1}{2} \langle \mathbf{H}, \mathbb{E}[\bar{v}_M \otimes \bar{v}_M] \rangle_{op}} \right)^2 \end{aligned}$$

b) Set  $P_t^{(m)} = \mathbf{I} - \gamma x_t^{(m)} \otimes x_t^{(m)}$ . Note that we have

$$b_t^{(m)} = \prod_{k=1}^t P_k^{(m)} b_0^{(m)}, \quad v_t^{(m)} = \gamma \sum_{i=1}^t \prod_{j=i+1}^t \epsilon_i^{(m)} P_j^{(m)} x_i^{(m)}.$$

By assumption, we therefore have for all  $s, t \leq n/M$  and  $m, m' \leq M$ ,

$$\begin{aligned} \mathbb{E}[b_s^{(m)} \otimes v_t^{(m')}] &= \gamma \mathbb{E} \left[ \prod_{k=1}^s P_k^{(m)} b_0^{(m)} \otimes \sum_{i=1}^t \prod_{j=i+1}^t \epsilon_i^{(m')} P_j^{(m')} x_i^{(m')} \right] \\ &= \gamma \sum_{i=1}^t \mathbb{E} \left[ \prod_{k=1}^s P_k^{(m)} b_0^{(m)} \otimes \prod_{j=i+1}^t P_j^{(m')} x_i^{(m')} \right] \mathbb{E}[\epsilon_i^{(m')}] = 0. \end{aligned}$$

This implies

$$\mathbb{E}[\bar{b}_M \otimes \bar{v}_M] = 0. \tag{9}$$



From (8) we therefore have

$$\mathbb{E}[\bar{\eta}_M \otimes \bar{\eta}_M] = \mathbb{E}[\bar{b}_M \otimes \bar{b}_M] + \mathbb{E}[\bar{v}_M \otimes \bar{v}_M] \quad (10)$$

Finally, by definition of the excess risk we have

$$\begin{aligned} \mathbb{E}[L(\bar{w}_M) - L(w^*)] &= \frac{1}{2} \mathbb{E} \left[ \int_{\mathcal{H}} \langle \bar{w}_M - w^*, x \rangle^2 \mathbb{P}_{\mathbf{x}}(d\mathbf{x}) \right] \\ &= \frac{1}{2} \mathbb{E}[\langle \mathbf{H}(\bar{w}_M - w^*), \bar{w}_M - w^* \rangle] \\ &= \frac{1}{2} \langle \mathbf{H}, \mathbb{E}[\bar{\eta}_M \otimes \bar{\eta}_M] \rangle_{op} \end{aligned} \quad (11)$$

$$= \text{Bias}(\bar{w}_M) + \text{Var}(\bar{w}_M), \quad (12)$$

where we used (10) for the last equality. □

### A.1.2 Upper Bound

For the non distributed case Zou et al. (2021b) (see Lemma B.11 and Lemma B.6 ) already established upper bounds. More precisely we have for the local bias and variance term:

**Proposition A.2.** *Set  $k^* = \max\{k : \lambda_k \geq \frac{M}{n\gamma}\}$ . If the step size satisfies  $\gamma < 1/(\tau \text{tr}(\mathbf{H}))$ , we have for every  $m \in [M]$ :*

a) *Under Assumption 3.2 and 3.3, it holds that*

$$\begin{aligned} \text{Bias}\left(\bar{w}_{\frac{n}{M}}^{(m)}\right) &:= \frac{1}{2} \langle \mathbf{H}, \mathbb{E}\left[b_t^{(m)} \otimes b_t^{(m)}\right] \rangle_{op} \\ &\leq \frac{M^2}{\gamma^2 n^2} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{0:k^*}^{-1}}^2 + \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2 \\ &\quad + \frac{2\tau M^2 \left( \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{I}_{0:k^*}}^2 + \frac{n}{M} \gamma \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2 \right)}{\gamma n (1 - \gamma \tau \text{tr}(\mathbf{H}))} \cdot \left( \frac{k^*}{n} + \frac{n}{M^2} \gamma^2 \sum_{i>k^*} \lambda_i^2 \right). \end{aligned}$$

b) *Under Assumptions 3.2 - 3.4, it holds that*

$$\text{Var}\left(\bar{w}_{\frac{n}{M}}^{(m)}\right) := \frac{1}{2} \langle \mathbf{H}, \mathbb{E}\left[\bar{v}_n^{(m)} \otimes \bar{v}_n^{(m)}\right] \rangle_{op} \leq \frac{\sigma^2}{1 - \gamma \tau \text{tr}(\mathbf{H})} \left( \frac{k^* M}{n} + \gamma^2 \frac{n}{M} \cdot \sum_{i>k^*} \lambda_i^2 \right).$$

**Lemma A.3.** *Set  $k^* = \max\{k : \lambda_k \geq \frac{M}{n\gamma}\}$ . If the step size satisfies  $\gamma < 1/(\tau \text{tr}(\mathbf{H}))$ , we have for every  $m \in [M]$ :*

a) *Under Assumption 3.2 and 3.3, it holds that*

$$\begin{aligned} \text{Bias}(\bar{w}_M) &\leq \frac{M^2}{\gamma^2 n^2} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{0:k^*}^{-1}}^2 + \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2 \\ &\quad + \frac{2\tau M^2 \left( \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{I}_{0:k^*}}^2 + \frac{n}{M} \gamma \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2 \right)}{\gamma n (1 - \gamma \tau \text{tr}(\mathbf{H}))} \cdot \left( \frac{k^*}{n} + \frac{n}{M^2} \gamma^2 \sum_{i>k^*} \lambda_i^2 \right). \end{aligned}$$

b) Under Assumptions 3.2 - 3.4 , it holds that

$$\text{Var}(\bar{w}_M) \leq \frac{\sigma^2}{1 - \gamma\tau \text{tr}(\mathbf{H})} \left( \frac{k^*}{n} + \gamma^2 \frac{n}{M^2} \cdot \sum_{i>k^*} \lambda_i^2 \right).$$

*Proof of Lemma A.3 . a)* For the Bias-term we simply use

$$\text{Bias}(\bar{w}_M) = \frac{1}{2} \mathbb{E} \left\| \bar{b} \right\|_{\mathbf{H}}^2 = \frac{1}{2} \mathbb{E} \left\| \frac{1}{M} \sum_{m=1}^M b_{M_n}^{(m)} \right\|_{\mathbf{H}}^2 \leq \frac{1}{M} \sum_{m=1}^M \frac{1}{2} \mathbb{E} \left\| \bar{b}_n^{(m)} \right\|_{\mathbf{H}}^2 = \frac{1}{M} \sum_{m=1}^M \text{Bias} \left( \bar{w}_{\frac{n}{M}}^{(m)} \right). \quad (13)$$

Taking the bound of the local Bias-term  $\text{Bias} \left( \bar{w}_{\frac{n}{M}}^{(m)} \right)$  from A.2, proves the claim.

b) First we split the expectation operator as follows

$$\begin{aligned} & \mathbb{E} [\bar{v}_M \otimes \bar{v}_M] \\ &= \frac{1}{M^2} \sum_{m, m'=1}^M \mathbb{E} [\bar{v}_n^{(m)} \otimes \bar{v}_n^{(m')}] \\ &= \frac{1}{M^2} \sum_{m=1}^M \mathbb{E} [\bar{v}_n^{(m)} \otimes \bar{v}_n^{(m)}] + \frac{1}{M^2} \sum_{m \neq m'} \mathbb{E} [\bar{v}_n^{(m)} \otimes \bar{v}_n^{(m')}] \\ &=: I_1 + I_2. \end{aligned} \quad (14)$$

Now we prove that the second operator  $I_2$  is equal zero. First rewrite  $I_2$  as

$$I_2 = \frac{1}{M^2} \sum_{m \neq m'} \frac{M^2}{n^2} \sum_{s, t=0}^{\frac{n}{M}-1} \mathbb{E} [v_t^{(m)} \otimes v_s^{(m')}].$$

Therefore it is enough to to prove  $\mathbb{E} [v_t^{(m)} \otimes v_s^{(m')}] = 0$  for any  $m \neq m'$ . Since we assume our data sets to be independent we have  $\mathbb{E} [v_t^{(m)} \otimes v_s^{(m')}] = \mathbb{E} [\langle \cdot, v_t^{(m)} \rangle] \mathbb{E} [v_s^{(m')}] = 0$ , where the last equality follows from (7). This proves  $I_2 = 0$ . To sum up we have from (14) for the variance term,

$$\begin{aligned} \text{Var}(\bar{w}_M) &= \frac{1}{2} \langle \mathbf{H}, \mathbb{E} [\bar{v}_M \otimes \bar{v}_M] \rangle_{op} \\ &= \frac{1}{M^2} \sum_{m=1}^M \frac{1}{2} \langle \mathbf{H}, \mathbb{E} [\bar{v}_n^{(m)} \otimes \bar{v}_n^{(m)}] \rangle_{op} \\ &= \frac{1}{M^2} \sum_{m=1}^M \text{Var} \left( \bar{w}_{\frac{n}{M}}^{(m)} \right). \end{aligned} \quad (15)$$

Using the bound of the local variance term from A.2 completes the proof.  $\square$

*Proof of Theorem 3.5.* Using lemma A.1 a) we have

$$\mathbb{E} [L(\bar{w}_M)] - L(w^*) \leq 2\text{Bias}(\bar{w}_M) + 2\text{Var}(\bar{w}_M).$$

The claim now follows from lemma A.3.  $\square$

## A.2 Proofs Lower Bound

### A.2.1 Lower Bound Bias

**Proposition A.4** (Lower Bound Bias). *Suppose Assumptions 3.2 and 3.6 are satisfied and let  $\gamma < \frac{1}{\|\mathbf{H}\|}$ . Recall the definition of  $\text{Bias}(\bar{w}_M)$  in Lemma A.1. The bias of the distributed SGD estimator satisfies the lower bound*

$$\text{Bias}(\bar{w}_M) \geq \frac{M(M-1)}{100\gamma^2 n^2} \left( \|w_1 - w^*\|_{\mathbf{H}_{0:k^*}}^2 + \frac{\gamma^2 n^2}{M^2} \|w_1 - w^*\|_{\mathbf{H}_{k^*:\infty}}^2 \right).$$

*Proof of Proposition A.4.* From the definition of the bias in Lemma A.1, we have

$$\begin{aligned} \text{Bias}(\bar{w}_M) &= \frac{1}{2} \left\langle \mathbf{H}, \mathbb{E} \left[ \bar{b}_M \otimes \bar{b}_M \right] \right\rangle_{op} \\ &= \frac{1}{2M^2} \sum_{m_1=1}^M \sum_{m_2=1}^M \left\langle \mathbf{H}, \mathbb{E} \left[ \bar{b}_n^{(m_1)} \otimes \bar{b}_n^{(m_2)} \right] \right\rangle_{op} \\ &= \frac{1}{2M^2} \sum_{m=1}^M \left\langle \mathbf{H}, \mathbb{E} \left[ \bar{b}_n^{(m)} \otimes \bar{b}_n^{(m)} \right] \right\rangle_{op} + \frac{1}{2M^2} \sum_{m_1 \neq m_2}^M \left\langle \mathbf{H}, \mathbb{E} \left[ \bar{b}_n^{(m_1)} \otimes \bar{b}_n^{(m_2)} \right] \right\rangle_{op}. \end{aligned} \quad (16)$$

We show that the first term in the above decomposition can be lower bounded by zero. Indeed, from (C.2) and (C.4) in Zou et al. (2021b) we have for all  $m = 1, \dots, M$  the local lower bound

$$\begin{aligned} \left\langle \mathbf{H}, \mathbb{E} \left[ \bar{b}_n^{(m)} \otimes \bar{b}_n^{(m)} \right] \right\rangle_{op} &\geq \frac{M^2}{n^2} \sum_{t=1}^{\frac{n}{M}} \sum_{k=t}^{\frac{n}{M}} \left\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbb{E} \left[ b_t^{(m)} \otimes b_t^{(m)} \right] \right\rangle_{op} \\ &\geq \frac{M^2}{\gamma n^2} \left\langle \mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^{\frac{n}{2M}}, \mathbf{S}_{\frac{n}{2M}}^{(m)} \right\rangle_{op}, \end{aligned}$$

where we set

$$\mathbf{S}_{\frac{n}{2M}}^{(m)} := \sum_{t=1}^{\frac{n}{2M}} \mathbb{E} \left[ b_t^{(m)} \otimes b_t^{(m)} \right].$$

Setting  $\mathbf{B}_1 = b_1^{(m)} \otimes b_1^{(m)} = (w_1 - w^*) \otimes (w_1 - w^*)$  and applying Lemma C.4 from Zou et al. (2021b) gives then for all  $m = 1, \dots, M$

$$\begin{aligned} \mathbf{S}_{\frac{n}{2M}}^{(m)} &\succeq \underbrace{\frac{\theta}{4} \text{Tr} \left[ (\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^{\frac{n}{2M}}) \mathbf{B}_1 \right]}_{PSD} \cdot \underbrace{((\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^{\frac{n}{2M}})) + \sum_{t=1}^{\frac{n}{M}} (\mathbf{I} - \gamma \mathbf{H})^t \cdot \mathbf{B}_1 \cdot (\mathbf{I} - \gamma \mathbf{H})^t}_{PSD} \\ &\succeq 0. \end{aligned}$$

Hence,

$$\begin{aligned} \frac{1}{2M^2} \sum_{m=1}^M \left\langle \mathbf{H}, \mathbb{E} \left[ \bar{b}_n^{(m)} \otimes \bar{b}_n^{(m)} \right] \right\rangle_{op} &\geq \frac{1}{2M^2} \cdot \frac{M^2}{\gamma n^2} \sum_{m=1}^M \left\langle \mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^{\frac{n}{2M}}, \mathbf{S}_{\frac{n}{2M}}^{(m)} \right\rangle_{op} \\ &\geq 0. \end{aligned} \quad (17)$$

We now bound the second term in equation 16. Note that by independence of the local nodes and with equation 6 we may write for any fixed  $m_1 \neq m_2$

$$\begin{aligned} \mathbb{E}[\bar{b}_n^{(m_1)} \otimes \bar{b}_n^{(m_2)}] &= \frac{M^2}{n^2} \sum_{t=1}^{\frac{n}{M}} \sum_{k=1}^{\frac{n}{M}} \mathbb{E}[b_t^{(m_1)}] \otimes \mathbb{E}[b_k^{(m_2)}] \\ &= \frac{M^2}{n^2} \sum_{t=1}^{\frac{n}{M}} \sum_{k=1}^{\frac{n}{M}} (\mathbf{I} - \gamma \mathbf{H})^t \cdot \mathbf{B}_1 \cdot (\mathbf{I} - \gamma \mathbf{H})^k. \end{aligned}$$

Hence,

$$\begin{aligned} \frac{1}{2M^2} \sum_{m_1 \neq m_2}^M \left\langle \mathbf{H}, \mathbb{E}[\bar{b}_n^{(m_1)} \otimes \bar{b}_n^{(m_2)}] \right\rangle_{op} &= \frac{1}{2M^2} \frac{M^2}{n^2} \sum_{m_1 \neq m_2}^M \sum_{t=1}^{\frac{n}{M}} \sum_{k=1}^{\frac{n}{M}} \left\langle \mathbf{H}, (\mathbf{I} - \gamma \mathbf{H})^t \cdot \mathbf{B}_1 \cdot (\mathbf{I} - \gamma \mathbf{H})^k \right\rangle_{op} \\ &= \frac{M(M-1)}{2\gamma n^2} \left\langle \sum_{k=1}^{\frac{n}{M}} (\mathbf{I} - \gamma \mathbf{H})^k (\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^{\frac{n}{M}+1}), \mathbf{B}_1 \right\rangle_{op} \\ &= \frac{M(M-1)}{2\gamma^2 n^2} \left\langle (\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^{\frac{n}{M}+1})^2 \mathbf{H}^{-1}, \mathbf{B}_1 \right\rangle_{op}. \end{aligned}$$

Following now the lines of the proof of Lemma C.5 in Zou et al. (2021b) (adapted to our local setting) gives

$$\frac{1}{2M^2} \sum_{m_1 \neq m_2}^M \left\langle \mathbf{H}, \mathbb{E}[\bar{b}_n^{(m_1)} \otimes \bar{b}_n^{(m_2)}] \right\rangle_{op} \geq \frac{M(M-1)}{100\gamma^2 n^2} \left( \|w_1 - w^*\|_{\mathbf{H}_{0:k^*}^\dagger}^2 + \frac{\gamma^2 n^2}{M^2} \|w_1 - w^*\|_{\mathbf{H}_{k^*:\infty}}^2 \right).$$

Combining now the last bound with equation 17 and equation 16 finally gives

$$\text{Bias}(\bar{w}_M) \geq \frac{M(M-1)}{100\gamma^2 n^2} \left( \|w_1 - w^*\|_{\mathbf{H}_{0:k^*}^\dagger}^2 + \frac{\gamma^2 n^2}{M^2} \|w_1 - w^*\|_{\mathbf{H}_{k^*:\infty}}^2 \right).$$

□

## A.2.2 Lower Bound Variance

**Proposition A.5** (Lower Bound Variance). *Suppose Assumptions 3.2 and 3.6 are satisfied and let  $\frac{n}{M} \geq 500$ ,  $\gamma < \frac{1}{\|\mathbf{H}\|}$ . Recall the definition of  $\text{Var}(\bar{w}_M)$  in Lemma A.1. The variance of the distributed SGD estimator satisfies the lower bound*

$$\text{Var}(\bar{w}_M) \geq \frac{\sigma_{noise}^2}{100} \cdot \left( \frac{k^*}{n} + \frac{\gamma^2 n}{M^2} \sum_{j>k^*} \lambda_j^2 \right).$$

*Proof of Proposition A.5.* From the definition of the variance in Lemma A.1, we have

$$\begin{aligned} \text{Var}(\bar{w}_M) &= \frac{1}{2} \left\langle \mathbf{H}, \mathbb{E}[\bar{v}_M \otimes \bar{v}_M] \right\rangle_{op} \\ &= \frac{1}{2M^2} \sum_{m_1=1}^M \sum_{m_2=1}^M \left\langle \mathbf{H}, \mathbb{E}[\bar{v}_n^{(m_1)} \otimes \bar{v}_n^{(m_2)}] \right\rangle_{op} \\ &= \frac{1}{2M^2} \sum_{m=1}^M \left\langle \mathbf{H}, \mathbb{E}[\bar{v}_n^{(m)} \otimes \bar{v}_n^{(m)}] \right\rangle_{op} + \frac{1}{2M^2} \sum_{m_1 \neq m_2}^M \left\langle \mathbf{H}, \mathbb{E}[\bar{v}_n^{(m_1)} \otimes \bar{v}_n^{(m_2)}] \right\rangle_{op}. \end{aligned} \quad (18)$$

We first lower bound the first term. By Eq. (C.3) and Lemma C.3 in Zou et al. (2021b) (adapted to our local setting) we obtain

$$\begin{aligned} \frac{1}{2M^2} \sum_{m=1}^M \left\langle \mathbf{H}, \mathbb{E} \left[ \bar{v}_n^{(m)} \otimes \bar{v}_n^{(m)} \right] \right\rangle_{op} &\geq \frac{1}{2M^2} \frac{M^2}{n^2} \sum_{m=1}^M \sum_{t=0}^{\frac{n}{M}-1} \sum_{k=t}^{\frac{n}{M}-1} \left\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbb{E} \left[ v_t^{(m)} \otimes v_t^{(m)} \right] \right\rangle_{op} \\ &\geq \frac{\sigma_{noise}^2}{100M^2} \sum_{m=1}^M \left( \frac{M}{n} k^* + \frac{\gamma^2 n}{M} \sum_{j>k^*} \lambda_j^2 \right) \\ &= \frac{\sigma_{noise}^2}{100} V_{k^*}(n, M), \end{aligned}$$

where

$$V_{k^*}(n, M) := \left( \frac{k^*}{n} + \frac{\gamma^2 n}{M^2} \sum_{j>k^*} \lambda_j^2 \right).$$

To derive the final bound we argue that the second term in equation 18 is zero. Indeed, by independence of the local nodes we may write for any  $m_1 \neq m_2$  with equation 7

$$\begin{aligned} \mathbb{E} \left[ v_t^{(m_1)} \otimes v_k^{(m_2)} \right] &= \mathbb{E} \left[ v_t^{(m_1)} \right] \otimes \mathbb{E} \left[ v_k^{(m_2)} \right] \\ &= (\mathbf{I} - \gamma \mathbf{H})^t (v_0^{(m_1)} \otimes v_0^{(m_2)}) (\mathbf{I} - \gamma \mathbf{H})^k \\ &= 0, \end{aligned}$$

since  $v_0^{(m)} = 0$  for all  $m = 1, \dots, M$ . Hence,

$$\frac{1}{2M^2} \sum_{m_1 \neq m_2}^M \left\langle \mathbf{H}, \mathbb{E} \left[ \bar{v}_n^{(m_1)} \otimes \bar{v}_n^{(m_2)} \right] \right\rangle_{op} = 0.$$

this finishes the proof.  $\square$

### A.3 Proofs Rates of Convergence

*Proof of Corollary 3.9.* Let the sequence  $M_n \leq \sqrt{\frac{\gamma(1-2\gamma\tau)n}{R^2}}$ . By definition of  $k^*$  we know that  $k^* = \tilde{d} = \left(\frac{n}{M_n}\right)^r$  and hence  $\lambda_{k^*} = \left(\frac{M_n}{n}\right)^r$ . We first bound the bias from Theorem 3.5. Since  $\|w^*\|_2 \leq R$  by assumption, we find

$$\|w^*\|_{\mathbf{H}_{0:k^*}}^2 \leq \frac{\|w^*\|_2^2}{\lambda_{k^*}} \leq R^2 \left(\frac{n}{M_n}\right)^r. \quad (19)$$

Similarly, since  $\frac{n}{M_n} \rightarrow \infty$  as  $n \rightarrow \infty$ , there exists  $n_0 \in \mathbb{N}$  such that

$$\|w^*\|_{\mathbf{H}_{k^*:\infty}}^2 \leq \frac{R^2}{\left(\frac{n}{M_n}\right)^q - \left(\frac{n}{M_n}\right)^r} \leq c_{n_0} R^2 \left(\frac{M_n}{n}\right)^q, \quad (20)$$

for any  $n \geq n_0$  and some  $c_{n_0} < \infty$ . Using that  $\text{Tr}[\mathbf{H}] = 2$  and  $\|w^*\|_{\mathbf{I}_{0:k^*}}^2 \leq R^2$ , we find for all  $n \geq n_0$ , for some  $n_0 \in \mathbb{N}$ , that

$$\frac{2\tau M^2 (\|w^*\|_{\mathbf{I}_{0:k}}^2 + \gamma \frac{n}{M} \|w^*\|_{\mathbf{H}_{k:\infty}}^2)}{\gamma n (1 - \gamma\tau \text{Tr}[\mathbf{H}])} \leq 4 \max\{1, c_{n_0}\} \frac{\tau R^2}{1 - 2\gamma\tau} \frac{M_n^2}{\gamma n}.$$

Note that we also use that  $M_n \leq n$  and hence  $\left(\frac{M_n}{n}\right)^{q-1} \leq 1$ , since  $q > 1$ . Since

$$M_n \leq \sqrt{\frac{\gamma(1-2\gamma\tau)n}{R^2}}$$

we have

$$\frac{\tau R^2}{1 - 2\gamma\tau} \frac{M_n^2}{\gamma n} \leq 1$$

and hence

$$\frac{2\tau M_n^2 (\|w^*\|_{\mathbf{I}_{0:k}}^2 + \gamma \frac{n}{M} \|w^*\|_{\mathbf{H}_{k:\infty}}^2)}{\gamma n (1 - \gamma\tau \text{Tr}[\mathbf{H}])} \leq 4 \max\{1, c_{n_0}\}. \quad (21)$$

We further observe that by the definition of the spectrum of  $\mathbf{H}$

$$\sum_{j>k^*} \lambda_j^2 = \sum_{j=\tilde{d}}^d \frac{1}{d - \tilde{d}} = \frac{1}{\left(\frac{n}{M_n}\right)^q - \left(\frac{n}{M_n}\right)^r} \leq c_{n_0} \left(\frac{M_n}{n}\right)^q,$$

for any  $n$  sufficiently large, by using the argumentation as above. Hence,

$$\begin{aligned} V_{k^*}(n, M_n) &:= \frac{k^*}{n} + \gamma^2 \frac{n}{M_n^2} \sum_{j=k^*+1}^{\infty} \lambda_j^2 \\ &\leq \max\{1, c_{n_0}\} \frac{1}{M_n} \cdot \left( \left(\frac{M_n}{n}\right)^{1-r} + \gamma^2 \left(\frac{M_n}{n}\right)^{q-1} \right). \end{aligned} \quad (22)$$

Combining (19), (20), (21) and (22), we find for the bias term

$$\begin{aligned} \text{Bias}(M_n) &\leq \frac{R^2}{\gamma^2 (n/M_n)^2} \left(\frac{n}{M_n}\right)^r + c_{n_0} R^2 \left(\frac{M_n}{n}\right)^q + 4 \max\{1, c_{n_0}\} V_{k^*}(n, M_n) \\ &\leq \max\{1, c_{n_0}\} R^2 \left( \frac{1}{\gamma^2} \left(\frac{M_n}{n}\right)^{2-r} + \left(\frac{M_n}{n}\right)^q \right) + \end{aligned} \quad (23)$$

$$4 \max\{1, c_{n_0}\}^2 \frac{1}{M_n} \cdot \left( \left(\frac{M_n}{n}\right)^{1-r} + \gamma^2 \left(\frac{M_n}{n}\right)^{q-1} \right). \quad (24)$$

We now turn to the bound of the variance term. From equation 22 we have

$$\text{Var}(M_n) \leq \max\{1, c_{n_0}\} \left( \frac{\sigma^2}{1 - \gamma\tau \text{Tr}[\mathbf{H}]} \right) \cdot \frac{1}{M_n} \cdot \left( \left(\frac{M_n}{n}\right)^{1-r} + \gamma^2 \left(\frac{M_n}{n}\right)^{q-1} \right).$$

Combining the bounds for bias and variance leads to the total error bound

$$\begin{aligned} \mathbb{E}[L(\bar{w}_{M_n})] - L(w^*) &\leq \\ 2\tilde{c}_{n_0} \cdot c_{\gamma,\tau,\sigma} \cdot &\left( R^2 \left( \frac{1}{\gamma^2} \left(\frac{M_n}{n}\right)^{2-r} + \left(\frac{M_n}{n}\right)^q \right) + \frac{1}{M_n} \cdot \left( \left(\frac{M_n}{n}\right)^{1-r} + \gamma^2 \left(\frac{M_n}{n}\right)^{q-1} \right) \right), \end{aligned}$$

with

$$c_{\gamma,\tau,\sigma} := 1 + \frac{\sigma^2}{1 - \gamma\tau \text{Tr}[\mathbf{H}]}, \quad \tilde{c}_{n_0} = 4 \max\{1, c_{n_0}\}^2,$$

holding for any  $n$  sufficiently large. We proceed by further simplifying the right hand side of the above inequality. Since  $\tau \geq 1$  and  $1 - \gamma\tau \text{Tr}[\mathbf{H}] < 1$ , the assumption on  $M_n$  implies that

$$M_n^2 \leq \frac{n\gamma}{R^2},$$

further implying that

$$\frac{R^2}{\gamma} \left(\frac{M_n}{n}\right)^{2-r} \leq \frac{1}{M_n} \left(\frac{M_n}{n}\right)^{1-r}$$

and

$$R^2 \left( \frac{M_n}{n} \right)^q \leq \frac{\gamma}{M_n} \left( \frac{M_n}{n} \right)^{q-1}.$$

As a result, applying Theorem 3.5, the excess risk can be bounded by

$$\begin{aligned} \mathbb{E}[L(\bar{w}_{M_n})] - L(w^*) &\leq 4\tilde{c}_{n_0} \cdot c_{\gamma, \tau, \sigma} \cdot \frac{1}{M_n} \left( \left( \frac{1}{\gamma} + 1 \right) \left( \frac{M_n}{n} \right)^{1-r} + (\gamma + \gamma^2) \left( \frac{M_n}{n} \right)^{q-1} \right) \\ &\leq 4\tilde{c}_{n_0} \cdot c_{\gamma, \tau, \sigma} \cdot \frac{1}{\gamma M_n} \left( \left( \frac{M_n}{n} \right)^{1-r} + \left( \frac{M_n}{n} \right)^{q-1} \right). \end{aligned}$$

In the last step we use that  $\gamma < \frac{1}{2r} < \frac{1}{2}$ . □

*Proof of Corollary 3.10.* Assume the sequence  $(M_n)_n$  satisfies  $M_n/n \rightarrow 0$  as  $n \rightarrow \infty$ . We use Theorem 3.5 to bound the excess risk and find estimates for bias and variance. By the definition of  $k^*$  we have

$$k^* = \max \left\{ k \in \mathbb{N} : k \leq \left( \frac{\gamma n}{M_n} \right)^{\frac{1}{1+r}} \right\} = \left\lfloor \left( \frac{\gamma n}{M_n} \right)^{\frac{1}{1+r}} \right\rfloor.$$

Hence, there exists  $n_0 \in \mathbb{N}$  such that for all  $n \geq n_0$

$$c_{n_0} \left( \frac{\gamma n}{M_n} \right)^{\frac{1}{1+r}} \leq k^* \leq C_{n_0} \left( \frac{\gamma n}{M_n} \right)^{\frac{1}{1+r}},$$

for some constants  $0 < c_{n_0} \leq C_{n_0}$ . Therefore,

$$\lambda_{k^*} = (k^*)^{-(1+r)} \leq \left( \frac{1}{c_{n_0}} \right)^{1+r} \cdot \frac{M_n}{\gamma n}$$

and

$$\frac{1}{\lambda_{k^*}} = (k^*)^{1+r} \leq C_{n_0}^{1+r} \cdot \frac{n}{\gamma M_n}.$$

We therefore get for the first two terms of the bias

$$\frac{M_n^2}{\gamma^2 n^2} \cdot \|w^*\|_{\mathbf{H}_{0:k^*}^\dagger}^2 \leq \frac{R^2 M_n^2}{\gamma^2 n^2 \lambda_{k^*}} \tag{25}$$

$$\leq C_{n_0}^{1+r} \frac{R^2}{\gamma} \cdot \frac{M_n}{n}. \tag{26}$$

and

$$\|w^*\|_{\mathbf{H}_{k^*:\infty}}^2 \leq R^2 \lambda_{k^*} \leq R^2 \left( \frac{1}{c_{n_0}} \right)^{1+r} \cdot \frac{M_n}{\gamma n}. \tag{27}$$

We now bound the last term of the bias. To this end, we apply a well known bound for sums over decreasing functions, i.e.,

$$\sum_{j \geq k} f(j) \leq \int_k^\infty f(x) dx.$$

This gives

$$\sum_{j > k^*} \lambda_j^2 \leq \int_{k^*}^\infty x^{-2(r+1)} dx \leq \frac{1}{2r+1} (k^*)^{-(2r+1)} \leq \frac{1}{2r+1} c_{n_0}^{-(2r+1)} \left( \frac{M_n}{\gamma n} \right)^{1+\frac{r}{1+r}}.$$

Thus,

$$\begin{aligned}
V_{k^*}(n, M_n) &= \frac{k^*}{n} + \gamma^2 \frac{n}{M^2} \sum_{j=k^*+1}^{\infty} \lambda_j^2 \\
&\leq \frac{1}{n} C_{n_0} \left( \frac{\gamma n}{M_n} \right)^{\frac{1}{1+r}} + \gamma^2 \frac{n}{M_n^2} \frac{c_{n_0}^{-(2r+1)}}{2r+1} \left( \frac{M_n}{\gamma n} \right)^{1+\frac{r}{1+r}} \\
&\leq c'_{r, n_0} \left( \frac{1}{n} \left( \frac{\gamma n}{M_n} \right)^{\frac{1}{1+r}} + \frac{\gamma}{M_n} \left( \frac{M_n}{\gamma n} \right)^{\frac{r}{1+r}} \right) \\
&\leq 2c'_{r, n_0} \cdot \frac{\gamma}{M_n} \left( \frac{M_n}{\gamma n} \right)^{\frac{r}{1+r}}, \tag{28}
\end{aligned}$$

with

$$c'_{r, n_0} = \max \left\{ C_{n_0}, \frac{c_{n_0}^{-(2r+1)}}{2r+1} \right\}.$$

Moreover,

$$\begin{aligned}
\frac{2\tau M_n^2 \left( \|w^*\|_{\mathbf{I}_{0:k^*}}^2 + \gamma \frac{n}{M_n} \|w^*\|_{\mathbf{H}_{k^*:\infty}}^2 \right)}{\gamma n (1 - \gamma\tau \text{Tr}[\mathbf{H}])} &\leq \frac{2\tau M_n^2}{\gamma n (1 - \gamma\tau \text{Tr}[\mathbf{H}])} \left( R^2 + R^2 \gamma \frac{n}{M_n} \lambda_{k^*} \right) \\
&\leq c''_{n_0} \frac{2\tau M_n^2}{\gamma n (1 - \gamma\tau \text{Tr}[\mathbf{H}])} R^2 \left( 1 + \gamma \frac{n}{M_n} \frac{M_n}{\gamma n} \right) \\
&\leq 2c''_{n_0} \frac{2\tau}{(1 - \gamma\tau \text{Tr}[\mathbf{H}])} \cdot \frac{R^2 M_n^2}{\gamma n},
\end{aligned}$$

with

$$c''_{n_0} = \max \left\{ 1, \left( \frac{1}{c_{n_0}} \right)^{1+r} \right\}.$$

Hence, combining this with equation 28 and choosing

$$M_n \leq \frac{\sqrt{\gamma n}}{R} \tag{29}$$

leads to

$$\begin{aligned}
&\frac{2\tau M_n^2 \left( \|w^*\|_{\mathbf{I}_{0:k^*}}^2 + \gamma \frac{n}{M_n} \|w^*\|_{\mathbf{H}_{k^*:\infty}}^2 \right)}{\gamma n (1 - \gamma\tau \text{Tr}[\mathbf{H}])} \cdot V_{k^*}(n, M_n) \\
&\leq 2c''_{n_0} \frac{2\tau}{(1 - \gamma\tau \text{Tr}[\mathbf{H}])} \cdot \frac{R^2 M_n^2}{\gamma n} \cdot 2c'_{r, n_0} \cdot \frac{\gamma}{M_n} \left( \frac{M_n}{\gamma n} \right)^{\frac{r}{1+r}} \\
&\leq c_{r, n_0} \frac{\tau}{(1 - \gamma\tau \text{Tr}[\mathbf{H}])} \cdot \frac{R^2 M_n^2}{\gamma n} \cdot \frac{\gamma}{M_n} \left( \frac{M_n}{\gamma n} \right)^{\frac{r}{1+r}} \\
&\leq c_{r, n_0} \frac{\tau}{(1 - \gamma\tau \text{Tr}[\mathbf{H}])} \cdot \frac{\gamma}{M_n} \left( \frac{M_n}{\gamma n} \right)^{\frac{r}{1+r}}, \tag{30}
\end{aligned}$$

where  $c_{r, n_0} = 8c''_{n_0} \cdot c'_{r, n_0}$ . Combining (26), (27) and (30), we find for all  $n \geq n_0$

$$\text{Bias}(M_n) \leq \tilde{c}_{r, n_0} \cdot \frac{R^2}{\gamma} \cdot \frac{M_n}{n} + c_{r, n_0} \frac{\tau}{(1 - \gamma\tau \text{Tr}[\mathbf{H}])} \cdot \frac{\gamma}{M_n} \left( \frac{M_n}{\gamma n} \right)^{\frac{r}{1+r}}, \tag{31}$$

where we set

$$\tilde{c}_{r, n_0} = 2 \max \left\{ C_{n_0}^{1+r}, \left( \frac{1}{c_{n_0}} \right)^{1+r} \right\}.$$



We now turn to bounding the variance. Using equation 28 once more, the variance can be bounded by

$$\text{Var}(M_n) \leq \frac{\sigma^2}{1 - \gamma\tau \text{Tr}[\mathbf{H}]} \cdot 2c'_{r,n_0} \cdot \frac{\gamma}{M_n} \left( \frac{M_n}{\gamma n} \right)^{\frac{r}{1+r}}.$$

Combining the bias bound equation 31 with the variance bound, we obtain for the excess risk

$$\begin{aligned} \mathbb{E}[L(\bar{w}_{M_n})] - L(w^*) &\leq c_{r,n_0,\tau,\sigma} \left( \frac{R^2}{\gamma} \cdot \frac{M_n}{n} + \frac{\gamma}{M_n} \left( \frac{M_n}{\gamma n} \right)^{\frac{r}{1+r}} \right). \\ c_{r,n_0,\tau,\sigma} &:= \max \left\{ \tilde{c}_{r,n_0}, 2c_{r,n_0} \cdot c'_{r,n_0} \cdot \frac{\max\{\tau, \sigma^2\}}{1 - \gamma\tau \text{Tr}[\mathbf{H}]} \right\}. \end{aligned}$$

Note that the choice

$$M_n \leq \left( \frac{\gamma}{R^2} \right)^{\frac{1+r}{2+r}} \cdot (\gamma n)^{\frac{1}{2+r}} \quad (32)$$

leads to a dominating variance part, i.e.

$$\frac{R^2}{\gamma} \cdot \frac{M_n}{n} \leq \frac{\gamma}{M_n} \left( \frac{M_n}{\gamma n} \right)^{\frac{r}{1+r}}$$

and

$$\mathbb{E}[L(\bar{w}_{M_n})] - L(w^*) \leq 2c_{r,n_0,\tau,\sigma} \frac{\gamma}{M_n} \left( \frac{M_n}{\gamma n} \right)^{\frac{r}{1+r}}.$$

Note that the choice equation 32 is compatible with the choice equation 29, i.e.,

$$M_n \leq \left( \frac{\gamma}{R^2} \right)^{\frac{1+r}{2+r}} \cdot (\gamma n)^{\frac{1}{2+r}} \leq \frac{\sqrt{\gamma n}}{R},$$

following from the fact that  $r > 0$ , provided that  $n$  is sufficiently large.  $\square$

## B PROOFS SECTION 4 ( COMPARISON OF SAMPLE COMPLEXITY OF DSGD AND DRR)

### B.1 Lower Bound for distributed ridge regression

In this section we derive a lower bound for the distributed RR estimator. We adopt the following notation and assumptions from Tsigler & Bartlett (2020).

- $\mathbf{H}^{-1/2}x$ , where  $x \in \mathbb{R}^d$  is sub-Gaussian with independent components
- $\mathbf{X} = (\sqrt{\lambda_1}z_1, \dots, \sqrt{\lambda_d}z_d)$  with  $z_j$  being sub-Gaussian with independent components
- $\mathbf{A} := \mathbf{X}\mathbf{X}^T + \lambda I_n$ ,  $\mathbf{A}_m := \mathbf{X}_m\mathbf{X}_m^T + \lambda I_n$
- $\mathbf{A}_{-j} = \sum_{i \neq j} \lambda_i z_i z_i^T + \lambda I_n$

Crucial for our analysis is the following quantity, called the *local effective dimension* for the RR problem:

$$k_{\text{RR}}^* := \min \left\{ k : \lambda_{k+1} \leq \frac{M \left( \lambda + \sum_{j>k} \lambda_j \right)}{bn} \right\}. \quad (33)$$

### B.1.1 Bias-Variance Decomposition DRR

**Definition B.1** (Bias and Variance of Distributed RR). *Let*

$$\begin{aligned}\Pi_m(\lambda) &:= (\mathbf{X}_m^T \mathbf{X}_m + \lambda)^{-1} \mathbf{X}_m^T \mathbf{X}_m - Id, \\ \widehat{\text{Bias}}(\bar{w}_n^{\text{RR}}(\lambda)) &:= \left\| \mathbf{H}^{1/2} \left( \frac{1}{M} \sum_{m=1}^M \Pi_m(\lambda) w^* \right) \right\|^2, \\ \widehat{\text{Var}}(\bar{w}_n^{\text{RR}}(\lambda)) &:= \left\| \mathbf{H}^{1/2} \left( \frac{1}{M} \sum_{m=1}^M (\mathbf{X}_m^T \mathbf{X}_m + \lambda)^{-1} \mathbf{X}_m^T \epsilon_m \right) \right\|^2.\end{aligned}$$

We call

$$\text{Bias}(\bar{w}_n^{\text{RR}}(\lambda)) = \mathbb{E} \left[ \widehat{\text{Bias}}(\bar{w}_n^{\text{RR}}(\lambda)) \right]$$

the (expected) bias of the distributed RR estimator and

$$\text{Var}(\bar{w}_n^{\text{RR}}(\lambda)) = \mathbb{E} \left[ \widehat{\text{Var}}(\bar{w}_n^{\text{RR}}(\lambda)) \right]$$

the (expected) variance.

We immediately obtain:

**Lemma B.2.** *The excess risk satisfies*

$$\mathbb{E} \left[ \left\| \mathbf{H}^{1/2} (\bar{w}_n^{\text{RR}}(\lambda) - w^*) \right\|^2 \right] = \text{Bias}(\bar{w}_n^{\text{RR}}(\lambda)) + \text{Var}(\bar{w}_n^{\text{RR}}(\lambda)).$$

*Proof of Lemma B.2.* We split the excess risk as

$$\begin{aligned}\left\| \mathbf{H}^{1/2} (\bar{w}_n^{\text{RR}}(\lambda) - w^*) \right\|^2 &= \left\| \mathbf{H}^{1/2} \left( \frac{1}{M} \sum_{m=1}^M \hat{w}_m^{\text{RR}}(\lambda) - w^* \right) \right\|^2 \\ &= \left\| \mathbf{H}^{1/2} \left( \frac{1}{M} \sum_{m=1}^M (\mathbf{X}_m^T \mathbf{X}_m + \lambda)^{-1} \mathbf{X}_m^T \mathbf{Y}_m - w^* \right) \right\|^2 \\ &= \left\| \mathbf{H}^{1/2} \left( \frac{1}{M} \sum_{m=1}^M (\mathbf{X}_m^T \mathbf{X}_m + \lambda)^{-1} \mathbf{X}_m^T (\mathbf{X}_m w^* + \epsilon_m) - w^* \right) \right\|^2 \\ &= \widehat{\text{Bias}}(\bar{w}_n^{\text{RR}}(\lambda)) + \widehat{\text{Var}}(\bar{w}_n^{\text{RR}}(\lambda)) \\ &\quad + \frac{2}{M^2} \sum_{m=1}^M \sum_{m'=1}^M \langle \mathbf{H} \Pi_m(\lambda) w^*, (\mathbf{X}_m^T \mathbf{X}_m + \lambda)^{-1} \mathbf{X}_m^T \epsilon_m \rangle.\end{aligned}$$

We argue that the expectation with respect to the noise (i.e. conditioned on  $\mathbf{X}$ ) of the last term is equal to zero. Indeed, by linearity and since  $\epsilon_m$  is centered (conditioned on  $\mathbf{X}_m$ ) for all  $m \in [M]$ , we find

$$\begin{aligned}\mathbb{E}_{\epsilon_m} \left[ \langle \mathbf{H} \Pi_m(\lambda) w^*, (\mathbf{X}_m^T \mathbf{X}_m + \lambda)^{-1} \mathbf{X}_m^T \epsilon_m \rangle \right] &= \langle \mathbf{H} \Pi_m(\lambda) w^*, (\mathbf{X}_m^T \mathbf{X}_m + \lambda)^{-1} \mathbf{X}_m^T \mathbb{E}_{\epsilon_m} [\epsilon_m] \rangle \\ &= 0.\end{aligned}$$

Hence,

$$\mathbb{E} \left[ \left\| \mathbf{H}^{1/2} (\bar{w}_n^{\text{RR}}(\lambda) - w^*) \right\|^2 \right] = \mathbb{E} \left[ \widehat{\text{Bias}}(\bar{w}_n^{\text{RR}}(\lambda)) \right] + \mathbb{E} \left[ \widehat{\text{Var}}(\bar{w}_n^{\text{RR}}(\lambda)) \right].$$

□

### B.1.2 Lower Bound of Bias for DRR

**Proposition B.3** (Lower Bound of Bias for local RR). *Assume  $\mathbf{H}$  is strictly positive definite with  $\text{Tr}[\mathbf{H}] < \infty$ . There exist absolute constants  $b > 1$ ,  $c > 1$  such that*

$$\text{Bias}(\bar{w}_n^{\text{RR}}(\lambda)) \geq \frac{M-1}{cM} \cdot \left( \frac{M^2 \left( \lambda + \sum_{j>k_{\text{RR}}^*} \lambda_j \right)^2}{n^2} \cdot \|w^*\|_{\mathbf{H}_{0:k_{\text{RR}}^*}^{-1}}^2 + \|w^*\|_{\mathbf{H}_{k_{\text{RR}}^*:\infty}}^2 \right),$$

where  $k_{\text{RR}}^*$  is defined in equation 33.

For proving this Proposition we need the following Lemma.

**Lemma B.4.** *Let  $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times d}$  be an independent copy of  $\mathbf{X} \in \mathbb{R}^{n \times d}$  and set  $\tilde{\mathbf{A}} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T + \lambda$ . Define further the operator*

$$\mathbf{B} := (I_d - \mathbf{X}^T \mathbf{A}^{-1} \mathbf{X}) \mathbf{H} (I_d - \tilde{\mathbf{X}}^T \tilde{\mathbf{A}}^{-1} \tilde{\mathbf{X}}).$$

1. For any  $i \neq j$ , we have

$$\mathbb{E}_{\mathbf{X}, \tilde{\mathbf{X}}}[\mathbf{B}_{ij}] = 0.$$

2. The diagonal elements satisfy for any  $k$

$$\mathbb{E}_{\mathbf{X}, \tilde{\mathbf{X}}}[\mathbf{B}_{ii}] \geq \frac{1}{c} \cdot \frac{\lambda_i}{\left(1 + \frac{\lambda_i}{\lambda_{k+1}} \cdot \frac{n}{\rho_k}\right)^2},$$

for some absolute constant  $c > 1$  and where we define

$$\rho_k = \frac{\lambda + \sum_{j>k} \lambda_j}{\lambda_{k+1}}.$$

*Proof of Lemma B.4.* Recall that  $\mathbf{H} = \text{diag}\{\lambda_1, \dots, \lambda_d\}$  and

$$\mathbf{X} = (\sqrt{\lambda_1} z_1, \dots, \sqrt{\lambda_d} z_d), \quad \tilde{\mathbf{X}} = (\sqrt{\lambda_1} \tilde{z}_1, \dots, \sqrt{\lambda_d} \tilde{z}_d).$$

1. Let  $i \neq j$ . We expand

$$\begin{aligned} \mathbf{B}_{ij} &= \underbrace{\langle e_i, \mathbf{H} e_j \rangle}_{=0} - \sqrt{\lambda_i} \langle e_j, \mathbf{H} \mathbf{X}^T \mathbf{A}^{-1} z_i \rangle - \sqrt{\lambda_j} \langle e_i, \mathbf{H} \tilde{\mathbf{X}}^T \tilde{\mathbf{A}}^{-1} \tilde{z}_j \rangle + \sqrt{z_i z_j} \langle z_i, \mathbf{A}^{-1} \mathbf{X} \mathbf{H} \tilde{\mathbf{X}}^T \tilde{\mathbf{A}}^{-1} \tilde{z}_j \rangle \\ &= -\lambda_j \sqrt{\lambda_i \lambda_j} \langle z_j, \mathbf{A}^{-1} z_i \rangle - \lambda_i \sqrt{\lambda_i \lambda_j} \langle \tilde{z}_i, \tilde{\mathbf{A}}^{-1} \tilde{z}_j \rangle + \sqrt{\lambda_i \lambda_j} \langle z_i, \mathbf{A}^{-1} \mathbf{X} \mathbf{H} \tilde{\mathbf{X}}^T \tilde{\mathbf{A}}^{-1} \tilde{z}_j \rangle. \end{aligned}$$

We define the map  $F(z_j) := \langle z_j, \mathbf{A}^{-1} z_i \rangle$ . Following the lines of the proof of Lemma C.7 in Zou et al. (2021a) shows that  $\mathbb{E}_{z_j}[F(z_j)] = 0$ . Using similar arguments, the same is true for the second and last term, showing the result.

2. We expand

$$\begin{aligned} \mathbf{B}_{ii} &= \left\langle \mathbf{H}(e_i - \sqrt{\lambda_i} \mathbf{X}^T \mathbf{A}^{-1} z_i), e_i - \sqrt{\lambda_i} \tilde{\mathbf{X}}^T \tilde{\mathbf{A}}^{-1} \tilde{z}_i \right\rangle \\ &= \underbrace{\langle \mathbf{H} e_i, e_i \rangle}_{=\lambda_i} + \lambda_i \langle \mathbf{H} \mathbf{X}^T \mathbf{A}^{-1} z_i, \tilde{\mathbf{X}}^T \tilde{\mathbf{A}}^{-1} \tilde{z}_i \rangle - \sqrt{\lambda_i} \langle \mathbf{H} e_i, \tilde{\mathbf{X}}^T \tilde{\mathbf{A}}^{-1} \tilde{z}_i \rangle - \sqrt{\lambda_i} \langle \mathbf{H} e_i, \mathbf{X}^T \mathbf{A}^{-1} z_i \rangle \\ &= \lambda_i [1 - \lambda_i (\langle z_i, \mathbf{A}^{-1} z_i \rangle + \langle \tilde{z}_i, \tilde{\mathbf{A}}^{-1} \tilde{z}_i \rangle)] + \lambda_i \langle \mathbf{H} \mathbf{X}^T \mathbf{A}^{-1} z_i, \tilde{\mathbf{X}}^T \tilde{\mathbf{A}}^{-1} \tilde{z}_i \rangle. \end{aligned}$$

Setting

$$a_i := \langle z_i, \mathbf{A}^{-1} z_i \rangle, \quad \tilde{a}_i := \langle \tilde{z}_i, \tilde{\mathbf{A}}^{-1} \tilde{z}_i \rangle$$

we further find that

$$\begin{aligned}
\lambda_i \langle \mathbf{H} \mathbf{X}^T \mathbf{A}^{-1} z_i, \tilde{\mathbf{X}}^T \tilde{\mathbf{A}}^{-1} \tilde{z}_i \rangle &= \lambda_i \sum_{j=1}^d \lambda_j \langle \mathbf{X}^T \mathbf{A}^{-1} z_i \rangle_j \cdot \langle \tilde{\mathbf{X}}^T \tilde{\mathbf{A}}^{-1} \tilde{z}_i \rangle_j \\
&= \lambda_i \sum_{j=1}^d \lambda_j^2 \langle z_j, \mathbf{A}^{-1} z_i \rangle \cdot \langle \tilde{z}_j, \tilde{\mathbf{A}}^{-1} \tilde{z}_i \rangle \\
&= \lambda_i^3 \cdot a_i \cdot \tilde{a}_i + \lambda_i \sum_{j \neq i} \lambda_j^2 \langle z_j, \mathbf{A}^{-1} z_i \rangle \cdot \langle \tilde{z}_j, \tilde{\mathbf{A}}^{-1} \tilde{z}_i \rangle.
\end{aligned}$$

By independence, the last term is non-negative in expectation, i.e.

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X}, \tilde{\mathbf{X}}} \left[ \lambda_i \sum_{j \neq i} \lambda_j^2 \langle z_j, \mathbf{A}^{-1} z_i \rangle \cdot \langle \tilde{z}_j, \tilde{\mathbf{A}}^{-1} \tilde{z}_i \rangle \right] \\
&= \lambda_i \sum_{j \neq i} \lambda_j^2 \mathbb{E}_{\mathbf{X}} [\langle z_j, \mathbf{A}^{-1} z_i \rangle] \cdot \mathbb{E}_{\tilde{\mathbf{X}}} [\langle \tilde{z}_j, \tilde{\mathbf{A}}^{-1} \tilde{z}_i \rangle] \\
&= \lambda_i \sum_{j \neq i} \lambda_j^2 \cdot \mathbb{E}_{\mathbf{X}} [\langle z_j, \mathbf{A}^{-1} z_i \rangle]^2 \\
&\geq 0.
\end{aligned}$$

Hence, for deriving a lower bound in expectation it is sufficient to lower bound the expression

$$\lambda_i \cdot [1 - \lambda_i(a_i + \tilde{a}_i)] + \lambda_i^3 \cdot a_i \cdot \tilde{a}_i = \lambda_i \cdot (1 - \lambda_i a_i) \cdot (1 - \lambda_i \tilde{a}_i).$$

Using independence once more we find

$$\mathbb{E}_{\mathbf{X}, \tilde{\mathbf{X}}} [\mathbf{B}_{ii}] \geq \lambda_i \cdot \mathbb{E}_{\mathbf{X}, \tilde{\mathbf{X}}} [(1 - \lambda_i a_i) \cdot (1 - \lambda_i \tilde{a}_i)].$$

We proceed as in the proof of Lemma C.7 in Zou et al. (2021a). Recall that

$$(1 - \lambda_i a_i) = \frac{1}{1 + \lambda_i \langle z_i, \mathbf{A}_{-i}^{-1} z_i \rangle}$$

and for all  $k$

$$\langle z_i, \mathbf{A}_{-i}^{-1} z_i \rangle \leq c \cdot \frac{n}{\lambda_{k+1} \rho_k},$$

for some  $c > 0$ , with high probability. Concluding as in Zou et al. (2021a) and using independence finishes the proof. □

*Proof of Proposition B.3.* Setting  $w_m(\lambda) = \mathbf{H}^{1/2} \Pi_m(\lambda) w^*$  (see Definition B.1), we decompose the bias as

$$\begin{aligned}
\text{Bias}(\bar{w}_n^{\text{RR}}(\lambda)) &= \mathbb{E} \left[ \widehat{\text{Bias}}(\bar{w}_n^{\text{RR}}(\lambda)) \right] \\
&= \mathbb{E} \left[ \left\| \frac{1}{M} \sum_{m=1}^M w_m(\lambda) \right\|^2 \right] \\
&= \frac{1}{M^2} \mathbb{E} \left[ \text{Tr} \left[ \left( \sum_{m=1}^M w_m(\lambda) \right) \otimes \left( \sum_{m'=1}^M w_{m'}(\lambda) \right) \right] \right] \\
&= \frac{1}{M^2} \sum_{m=1}^M \mathbb{E} [\text{Tr}[w_m(\lambda) \otimes w_m(\lambda)]] + \frac{1}{M^2} \sum_{m \neq m'} \mathbb{E} [\text{Tr}[w_m(\lambda) \otimes w_{m'}(\lambda)]]. \tag{34}
\end{aligned}$$

We aim to find a lower for the above expression. Since

$$\sum_{m=1}^M \mathbb{E}[\text{Tr}[w_m(\lambda) \otimes w_m(\lambda)]] \succeq 0$$

we proceed to lower bound the second term in equation 34 . Setting

$$\mathbf{B}_{m,m'} := \Pi_m(\lambda) \circ \mathbf{H} \circ \Pi_{m'}(\lambda)$$

for  $m, m' \in [M]$  we may write

$$\begin{aligned} \text{Bias}(\bar{w}_n^{\text{RR}}(\lambda)) &\geq \frac{1}{M^2} \sum_{m \neq m'} \mathbb{E}[\text{Tr}[w_m(\lambda) \otimes w_{m'}(\lambda)]] \\ &= \frac{1}{M^2} \sum_{m \neq m'} \mathbb{E}[\langle \mathbf{H} \circ \Pi_m(\lambda) w^*, \Pi_{m'}(\lambda) \rangle] \\ &= \frac{1}{M^2} \sum_{m \neq m'} \mathbb{E}[\langle \mathbf{B}_{m,m'} w^*, w^* \rangle] \\ &= \frac{1}{M^2} \sum_{m \neq m'} \left( \sum_i \mathbb{E}[(\mathbf{B}_{m,m'})_{ii}] (w_i^*)^2 + 2 \sum_{i>j} \mathbb{E}[(\mathbf{B}_{m,m'})_{ij}] w_i^* \cdot w_j^* \right). \end{aligned} \quad (35)$$

We now apply Lemma B.4 and follow the lines of the proof of Theorem C.8 in Zou et al. (2021a) to obtain for every  $k$

$$\begin{aligned} \text{Bias}(\bar{w}_n^{\text{RR}}(\lambda)) &\geq \frac{1}{M^2} \sum_{m \neq m'} \sum_i \mathbb{E}[(\mathbf{B}_{m,m'})_{ii}] (w_i^*)^2 \\ &\geq \frac{1}{cM^2} \sum_{m \neq m'} \sum_i \frac{\lambda_i \cdot (w_i^*)^2}{\left(1 + \frac{\lambda_i}{\lambda_{k+1}} \cdot \frac{n}{M\rho_k}\right)^2} \\ &= \frac{M-1}{cM} \sum_i \frac{\lambda_i \cdot (w_i^*)^2}{\left(1 + \frac{\lambda_i}{\lambda_{k+1}} \cdot \frac{n}{M\rho_k}\right)^2} \\ &\geq \frac{M-1}{cM} \cdot \left( \frac{M^2 \left(\lambda + \sum_{j>k_{\text{RR}}^*} \lambda_j\right)^2}{n^2} \cdot \|w^*\|_{\mathbf{H}_{0:k_{\text{RR}}^*}^{-1}}^2 + \|w^*\|_{\mathbf{H}_{k_{\text{RR}}^*:\infty}}^2 \right), \end{aligned} \quad (36)$$

for some  $c > 1$ . □

### B.1.3 Lower Bound of Variance for DRR

**Proposition B.5** (Lower Bound of Bias for local RR). *Suppose  $k_{\text{RR}}^* < \frac{n}{c'M}$ , for some universal constant  $c' > 1$ . There exist constants  $b, c > 1$  such that*

$$\text{Var}(\bar{w}_n^{\text{RR}}(\lambda)) \geq \frac{\sigma^2}{c} \left( \frac{k_{\text{RR}}^*}{n} + \frac{n}{M^2 \cdot (\lambda + \sum_{j>k_{\text{RR}}^*} \lambda_j)} \cdot \sum_{j>k_{\text{RR}}^*} \lambda_j^2 \right),$$

where  $k_{\text{RR}}^*$  is defined in equation 33.

*Proof of Proposition B.5.* By definition of the variance, we may write

$$\begin{aligned}
\text{Var}(\bar{w}_n^{\text{RR}}(\lambda)) &= \mathbb{E} \left[ \widehat{\text{Var}}(\bar{w}_n^{\text{RR}}(\lambda)) \right] \\
&= \mathbb{E} \left[ \left\| \mathbf{H}^{1/2} \left( \frac{1}{M} \sum_{m=1}^M (\mathbf{X}_m^T \mathbf{X}_m + \lambda)^{-1} \mathbf{X}_m^T \epsilon_m \right) \right\|^2 \right] \\
&= \frac{1}{M^2} \sum_{m,m'=1}^M \mathbb{E}_{\mathbf{X}} \left[ \text{Tr} \left[ \mathbf{H}^{1/2} (\mathbf{X}_m^T \mathbf{X}_m + \lambda)^{-1} \mathbf{X}_m^T \mathbb{E}_{\epsilon_m} [\epsilon_m \otimes \epsilon_{m'}] \mathbf{X}_m (\mathbf{X}_{m'}^T \mathbf{X}_{m'} + \lambda)^{-1} \mathbf{H}^{1/2} \right] \right] \\
&= \frac{\sigma^2}{M^2} \sum_{m=1}^M \mathbb{E}_{\mathbf{X}} \left[ \text{Tr} \left[ \mathbf{H}^{1/2} (\mathbf{X}_m^T \mathbf{X}_m + \lambda)^{-1} \mathbf{X}_m^T \mathbf{X}_m (\mathbf{X}_m^T \mathbf{X}_m + \lambda)^{-1} \mathbf{H}^{1/2} \right] \right] \\
&= \frac{1}{M^2} \sum_{m=1}^M \text{Var}(\hat{w}_m^{\text{RR}}(\lambda)),
\end{aligned}$$

where the local variance is given by

$$\text{Var}(\hat{w}_m^{\text{RR}}(\lambda)) = \sigma^2 \mathbb{E}_{\mathbf{X}} \left[ \text{Tr} \left[ \mathbf{H}^{1/2} (\mathbf{X}_m^T \mathbf{X}_m + \lambda)^{-1} \mathbf{X}_m^T \mathbf{X}_m (\mathbf{X}_m^T \mathbf{X}_m + \lambda)^{-1} \mathbf{H}^{1/2} \right] \right].$$

To lower bound the variance we utilize Theorem C.5 from Zou et al. (2021a) (see also Bartlett et al. (2020)) and obtain

$$\begin{aligned}
\text{Var}(\bar{w}_n^{\text{RR}}(\lambda)) &\geq \frac{\sigma^2}{cM^2} \sum_{m=1}^M \left( \frac{M \cdot k_{\text{RR}}^*}{n} + \frac{n}{M \cdot (\lambda + \sum_{j>k_{\text{RR}}^*} \lambda_j)} \cdot \sum_{j>k_{\text{RR}}^*} \lambda_j^2 \right) \\
&= \frac{\sigma^2}{c} \left( \frac{k_{\text{RR}}^*}{n} + \frac{n}{M^2 \cdot (\lambda + \sum_{j>k_{\text{RR}}^*} \lambda_j)} \cdot \sum_{j>k_{\text{RR}}^*} \lambda_j^2 \right),
\end{aligned}$$

provided  $k_{\text{RR}}^* < \frac{n}{c'M}$ , for some universal constants  $c, c' > 1$ .  $\square$

#### B.1.4 Proof of Theorem 4.2

The proof follows by combining Proposition B.3 and Proposition B.5 with Lemma B.2.

## B.2 Upper Bound Excess Risk Tail-Averaged DSGD

**Theorem B.6** (Upper Bound Tail-averaged DSGD). *Suppose Assumption 3.7 is satisfied. Let  $\bar{w}_{M,n}$  denote the tail-averaged distributed estimator with  $n$  training samples and assume  $\gamma < 1/\text{Tr}[H]$ . For arbitrary  $k_1, k_2 \in [d]$*

$$\mathbb{E}[L(\bar{w}_M)] - L(w^*) = \text{Bias}(\bar{w}_M) + \text{Var}(\bar{w}_M)$$

with

$$\begin{aligned}
\text{Bias}(\bar{w}_M) &\leq \frac{c_b M^2}{\gamma^2 n^2} \cdot \left\| \exp\left(-\frac{n}{M} \gamma \mathbf{H}\right) w^* \right\|_{\mathbf{H}_{0:k_1}^{-1}}^2 + \|w^*\|_{\mathbf{H}_{k_1:\infty}}^2, \\
\text{Var}(\bar{w}_M) &\leq c_v (1 + R^2) \cdot \sigma^2 \left( \frac{k_2}{n} + \frac{n\gamma^2}{M^2} \cdot \sum_{j>k_2} \lambda_j^2 \right),
\end{aligned}$$

for some universal constants  $c_b, c_v > 0$ .

*Proof of Theorem B.6.* Utilizing equation 13 and Lemma 6.1 in Zou et al. (2021a), we have

$$\begin{aligned} \text{Bias}(\bar{w}_M) &= \frac{1}{M} \sum_{m=1}^M \text{Bias}\left(\bar{w}_{\frac{n}{M}}^{(m)}\right) \\ &\leq \frac{1}{M} \sum_{m=1}^M \frac{c_b M^2}{\gamma^2 n^2} \cdot \left\| \exp\left(-\frac{n}{M} \gamma \mathbf{H}\right) \right\|_{\mathbf{H}_{0:k_1}^{-1}}^2 + \|w^*\|_{\mathbf{H}_{k_1:\infty}}^2 \\ &= \frac{c_b M^2}{\gamma^2 n^2} \cdot \left\| \exp\left(-\frac{n}{M} \gamma \mathbf{H}\right) w^* \right\|_{\mathbf{H}_{0:k_1}^{-1}}^2 + \|w^*\|_{\mathbf{H}_{k_1:\infty}}^2, \end{aligned}$$

for some universal constant  $c_b > 0$ .

For the variance, we utilize equation 15 and Lemma 6.1 in Zou et al. (2021a) once more to obtain

$$\begin{aligned} \text{Var}(\bar{w}_M) &\leq \frac{1}{M^2} \sum_{m=1}^M \text{Var}\left(\bar{w}_{\frac{n}{M}}^{(m)}\right) \\ &\leq c_v \frac{(1+R^2) \cdot \sigma^2}{M^2} \sum_{m=1}^M \left( \frac{k_2 M}{n} + \frac{n \gamma^2}{M} \cdot \sum_{j>k_2} \lambda_j^2 \right) \\ &= c_v (1+R^2) \cdot \sigma^2 \left( \frac{k_2}{n} + \frac{n \gamma^2}{M^2} \cdot \sum_{j>k_2} \lambda_j^2 \right), \end{aligned}$$

for some universal constant  $c_v > 0$ . □

### B.3 Comparing DSGD with DRR

#### B.3.1 Proof of Theorem 4.4

To prove Theorem 4.4 we derive conditions on  $n_{\text{RR}}$  and  $n_{\text{SGD}}$  such that the upper bound for the excess risk of  $\bar{w}_M$  for DSGD from Theorem 4.3 can be upper bounded by the lower bound of  $\bar{w}_n^{\text{RR}}(\lambda)$  for DRR from Theorem 4.2, i.e. such that

$$\begin{aligned} &\frac{c_b M^2}{\gamma^2 n_{\text{SGD}}^2} \cdot \left\| \exp\left(-\frac{n_{\text{SGD}}}{M} \gamma \mathbf{H}\right) w^* \right\|_{\mathbf{H}_{0:k_{\text{RR}}^*}^{-1}}^2 + \|w^*\|_{\mathbf{H}_{k_{\text{RR}}^*:\infty}}^2 \\ &\leq \frac{M^2 \left( \lambda + \sum_{j>k_{\text{RR}}^*} \lambda_j \right)^2}{c n_{\text{RR}}^2} \cdot \|w^*\|_{\mathbf{H}_{0:k_{\text{RR}}^*}^{-1}}^2 + \|w^*\|_{\mathbf{H}_{k_{\text{RR}}^*:\infty}}^2 \end{aligned} \quad (37)$$

and

$$c_v \left( 1 + \frac{\|w^*\|^2}{\sigma^2} \right) \cdot \sigma^2 \left( \frac{k_{\text{RR}}^*}{n_{\text{SGD}}} + \frac{n_{\text{SGD}} \gamma^2}{M^2} \cdot \sum_{j>k_{\text{RR}}^*} \lambda_j^2 \right) \leq \frac{\sigma^2}{c} \left( \frac{k_{\text{RR}}^*}{n_{\text{RR}}} + \frac{n_{\text{RR}}}{M^2} \cdot \frac{\sum_{j>k_{\text{RR}}^*} \lambda_j^2}{(\lambda + \sum_{j>k_{\text{RR}}^*} \lambda_j)^2} \right). \quad (38)$$

We start with equation 38. For

$$c_v \left( 1 + \frac{\|w^*\|^2}{\sigma^2} \right) \cdot \sigma^2 \frac{k_{\text{RR}}^*}{n_{\text{SGD}}} \leq \frac{\sigma^2}{c} \frac{k_{\text{RR}}^*}{n_{\text{RR}}}$$

to hold we need that

$$C^* n_{\text{RR}} \leq n_{\text{SGD}}, \quad C^* := c_v \cdot c \cdot \left( 1 + \frac{\|w^*\|^2}{\sigma^2} \right). \quad (39)$$

To

$$c_v \left( 1 + \frac{\|w^*\|^2}{\sigma^2} \right) \cdot \sigma^2 \frac{n_{\text{SGD}} \gamma^2}{M^2} \cdot \sum_{j>k_{\text{RR}}^*} \lambda_j^2 \leq \frac{\sigma^2 n_{\text{RR}}}{c M^2} \cdot \frac{\sum_{j>k_{\text{RR}}^*} \lambda_j^2}{(\lambda + \sum_{j>k_{\text{RR}}^*} \lambda_j)^2}$$

to hold we need

$$n_{\text{SGD}} \leq \frac{n_{\text{RR}}}{C^* \cdot (C_\lambda^*)^2 \gamma^2}, \quad C_\lambda^* := \lambda + \sum_{j>k_{\text{RR}}^*} \lambda_j. \quad (40)$$

Finally, from equation 37 we need

$$\frac{c_b M^2}{\gamma^2 n_{\text{SGD}}^2} \cdot \left\| \exp\left(-\frac{n_{\text{SGD}}}{M} \gamma \mathbf{H}\right) w^* \right\|_{\mathbf{H}_{0:k_{\text{RR}}^*}^{-1}}^2 \leq \frac{M^2 (C_\lambda^*)^2}{c n_{\text{RR}}^2} \cdot \|w^*\|_{\mathbf{H}_{0:k_{\text{RR}}^*}^{-1}}^2. \quad (41)$$

To ensure this, note that

$$\left\| \exp\left(-\frac{n_{\text{SGD}}}{M} \gamma \mathbf{H}\right) w^* \right\|_{\mathbf{H}_{0:k_{\text{RR}}^*}^{-1}}^2 \leq e^{-\frac{n_{\text{SGD}}}{M} \gamma \lambda_{k_{\text{RR}}^*}} \cdot \|w^*\|_{\mathbf{H}_{0:k_{\text{RR}}^*}^{-1}}^2 \leq (1 - \gamma \lambda_{k_{\text{RR}}^*}) \cdot \|w^*\|_{\mathbf{H}_{0:k_{\text{RR}}^*}^{-1}}^2.$$

Hence, equation 41 is implied if

$$\frac{c_b}{\gamma^2 n_{\text{SGD}}^2} (1 - \gamma \lambda_{k_{\text{RR}}^*}) \leq \frac{(C_\lambda^*)^2}{c n_{\text{RR}}^2},$$

being equivalent to

$$\frac{\sqrt{c c_b (1 - \gamma \lambda_{k_{\text{RR}}^*})}}{\gamma C_\lambda^*} n_{\text{RR}} \leq n_{\text{SGD}}. \quad (42)$$

Combining conditions equation 39, equation 40 and equation 42 we need

$$\max \left\{ C^*, \frac{\sqrt{c c_b (1 - \gamma \lambda_{k_{\text{RR}}^*})}}{\gamma C_\lambda^*} \right\} \cdot n_{\text{RR}} \leq n_{\text{SGD}} \leq \frac{1}{C^* \cdot (C_\lambda^*)^2 \gamma^2} \cdot n_{\text{RR}}.$$

A short calculation shows that the condition

$$\gamma < \min \left\{ \frac{1}{\text{Tr}[H]}, \frac{1}{\sqrt{c C^* C_\lambda^*}} \right\}$$

implies that

$$\max \left\{ C^*, \frac{\sqrt{c c_b (1 - \gamma \lambda_{k_{\text{RR}}^*})}}{\gamma C_\lambda^*} \right\} \leq \frac{1}{C^* \cdot (C_\lambda^*)^2 \gamma^2}.$$

This finishes the proof.

### B.3.2 Discussion

We give a more detail discussion about the sample complexities (SCs) of DSGD and DRR. In particular, we derive conditions under which the SCs are of the same order to ensure that

$$\mathbb{E} \left[ L(\bar{w}_{M_{n_{\text{SGD}}}}) \right] - L(w^*) \leq \mathbb{E} \left[ L(\bar{w}_{n_{\text{RR}}}^{\text{RR}}(\lambda_{n_{\text{RR}}})) \right] - L(w^*).$$

Recall that in order to achieve this bound we would need

$$L_{\lambda_{n_{\text{RR}}}, \gamma} \cdot n_{\text{RR}} \leq n_{\text{SGD}} \leq L'_{\lambda_{n_{\text{RR}}}, \gamma} \cdot n_{\text{RR}},$$

for a suitable choice of the regularization parameter  $\lambda_{n_{\text{RR}}}$  and number of machines  $M_{n_{\text{SGD}}}$ .



To ensure that  $n_{\text{RR}} \lesssim n_{\text{SGD}}$  we need to require that

$$1 \lesssim L_{\lambda_{n_{\text{RR}}}, \gamma} = \max \left\{ C^*, \frac{\sqrt{c(1 - \gamma \lambda_{k_{\text{RR}}^*})}}{\gamma C_{\lambda_{n_{\text{RR}}}^*}} \right\},$$

with  $C_{\lambda}^* := \lambda + \sum_{j > k_{\text{RR}}^*} \lambda_j$ . Recall that

$$\gamma < \min \left\{ \frac{1}{\text{Tr}[H]}, \frac{1}{\sqrt{c} C^* C_{\lambda_{n_{\text{RR}}}^*}} \right\},$$

and that  $1 - \gamma \lambda_{k_{\text{RR}}^*} < 1$ . A short calculation shows that

$$1 \lesssim \frac{\sqrt{c(1 - \gamma \lambda_{k_{\text{RR}}^*})}}{\gamma C_{\lambda_{n_{\text{RR}}}^*}}$$

if

$$\gamma \left( \lambda_{n_{\text{RR}}} + \sum_{j > k_{\text{RR}}^*} \lambda_j \right) \lesssim 1.$$

Furthermore, to ensure that  $n_{\text{SGD}} \lesssim n_{\text{RR}}$  we have to require that

$$L'_{\lambda_{n_{\text{RR}}}, \gamma} = \frac{1}{C^* \gamma^2 (C_{\lambda_{n_{\text{RR}}}^*})^2} \lesssim 1.$$

This is satisfied if

$$1 \lesssim \gamma \cdot C_{\lambda_{n_{\text{RR}}}^*} = \gamma \left( \lambda_{n_{\text{RR}}} + \sum_{j > k_{\text{RR}}^*} \lambda_j \right).$$

We summarize our finding the following

**Corollary B.7.** *Suppose all assumptions of Theorem 4.4 are satisfied. If*

$$\gamma \left( \lambda_{n_{\text{RR}}} + \sum_{j > k_{\text{RR}}^*} \lambda_j \right) \simeq 1 \tag{43}$$

*holds, then the sample complexities of DSGD and DRR are of the same order, i.e.*

$$n_{\text{SGD}} \simeq n_{\text{RR}}$$

and

$$\mathbb{E} \left[ L(\bar{w}_{M_{n_{\text{SGD}}}}) \right] - L(w^*) \leq \mathbb{E} \left[ L(\bar{w}_{n_{\text{RR}}}^{\text{RR}}(\lambda_{n_{\text{RR}}})) \right] - L(w^*).$$

**Example B.8** (Spiked Covariance Model). *We show that condition equation 43 is satisfied in the spiked covariance model from Corollary 3.9 under a suitable choice for  $\lambda_{n_{\text{RR}}}$  and  $M_{n_{\text{SGD}}}$ . Here, we assume that with*

$$M_{n_{\text{SGD}}} = M_{n_{\text{RR}}} \simeq n_{\text{RR}}^{\frac{3-2r}{5-2r}},$$

*for  $1/2 \leq r \leq 1$ , see our discussion in Section 3.4 (comparison with DOLS). A short calculation shows that*

$$k_{\text{RR}}^* \simeq \tilde{d} \simeq \left( \frac{n_{\text{RR}}}{M_{n_{\text{RR}}}} \right)^r \simeq n_{\text{RR}}^{\frac{2r}{5-2r}}.$$

Moreover, for  $\lambda_{n_{\text{RR}}} \simeq n_{\text{RR}}^{-\zeta}$ ,  $\zeta \geq 0$  and  $\gamma = \text{const.}$ , we have

$$\gamma \left( \lambda_{n_{\text{RR}}} + \sum_{j > k_{\text{RR}}^*} \lambda_j \right) \simeq \gamma \left( n_{\text{RR}}^{-\zeta} + 1 \right) \simeq 1.$$

Hence, for a wide range of regularization, the condition equation 43 is met and the SCs of DSGD and DRR in the spiked covariance model are of the same order.

## C FURTHER NUMERICAL EXPERIMENTS

In this Section we collect further experimental results conducted on simulated data from Section 5.

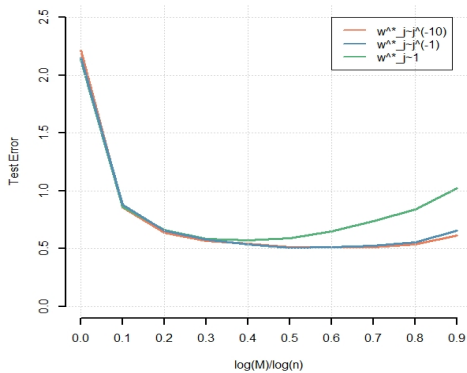


Figure 4: Test error for distributed ridgeless regression with  $\lambda_j = j^{-2}$  for different sources  $w^*$  as a function of  $M = n^\alpha$ ,  $\alpha \in \{0, 0.1, \dots, 0.9\}$ . The number of local nodes acts as a regularization parameter. We generate  $n = 500$  i.i.d. training data with  $x_j \sim \mathcal{N}(0, \mathbf{H})$  with mildly overparameterization  $d = 700$ .

We compare the sample complexity of optimally tuned full-averaged DSGD, tail-averaged DSGD and last-iterate DSGD with optimally tuned DRR for different sources  $w^*$ , see Figures 5, 6 and 6. Here, the data are generated as in Section 5 with  $d = 200$ ,  $\lambda_j = j^{-2}$  and  $w_j^* = j^{-\alpha}$ ,  $\alpha \in \{0, 1, 10\}$ . The number of local nodes is fixed at  $M_n = n^{1/3}$  for each  $n \in \{100, \dots, 6000\}$ .

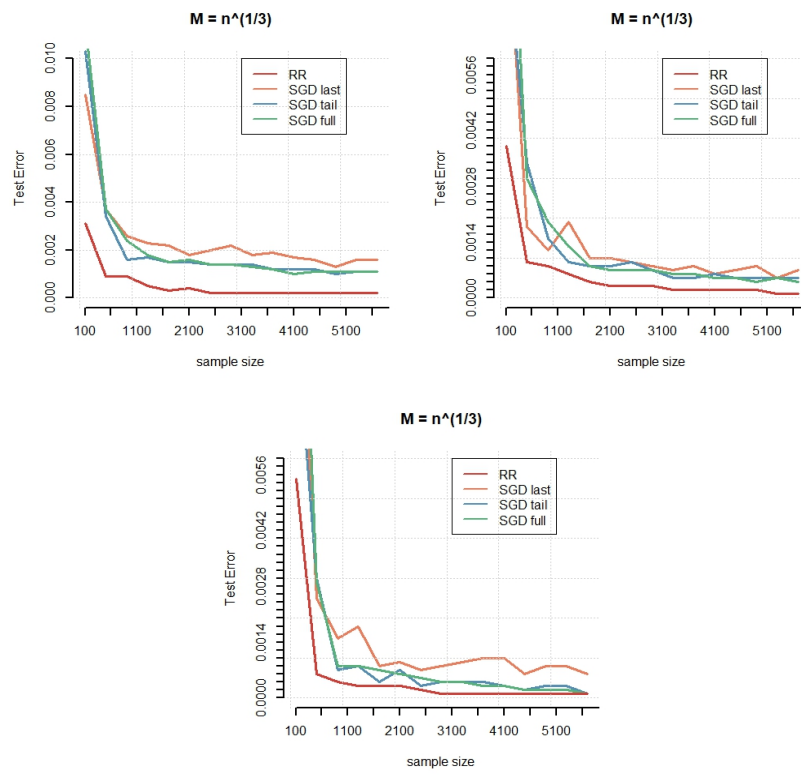


Figure 5: **Left:**  $\lambda_j = j^{-10}, w_j^* = 1$  **Middle:**  $\lambda_j = j^{-10}, w_j^* = j^{-1}$  **Right:**  $\lambda_j = j^{-10}, w_j^* = j^{-10}$

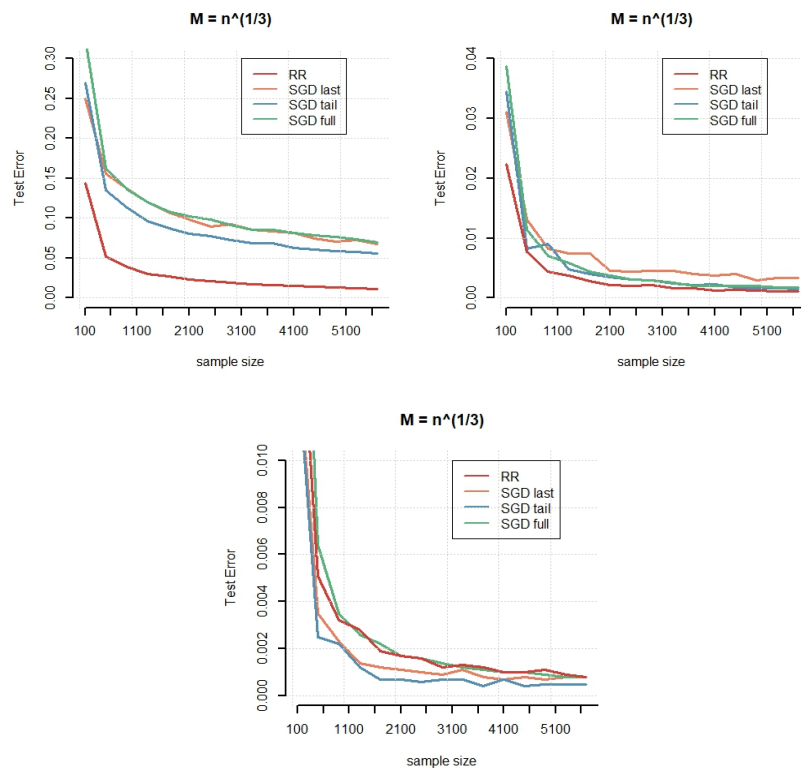


Figure 6: **Left:**  $\lambda_j = j^{-2}$ ,  $w_j^* = 1$  **Middle:**  $\lambda_j = j^{-2}$ ,  $w_j^* = j^{-1}$  **Right:**  $\lambda_j = j^{-2}$ ,  $w_j^* = j^{-10}$

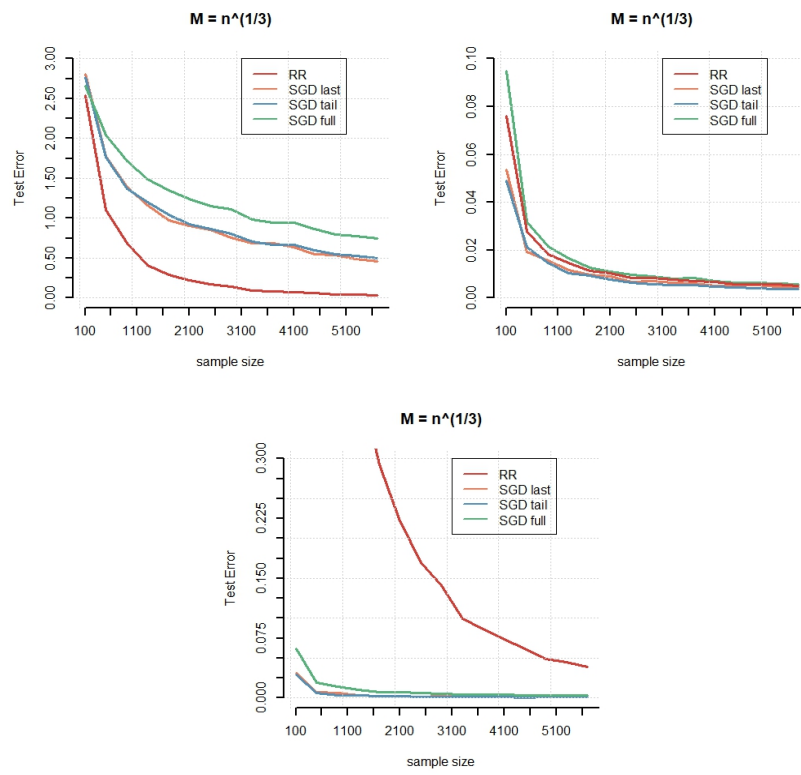


Figure 7: **Left:**  $\lambda_j = j^{-1}, w_j^* = 1$  **Middle:**  $\lambda_j = j^{-1}, w_j^* = j^{-1}$  **Right:**  $\lambda_j = j^{-1}, w_j^* = j^{-10}$