# Data splitting improves statistical performance in overparameterized regimes

**Nicole Mücke**
TU Braunschweig
nicole.muecke@tu-braunschweig.de

**Enrico Reiss**
University of Potsdam
enreiss@uni-potsdam.de

**Jonas Rungenhagen**
University of Potsdam
jrungenh@uni-potsdam.de

**Markus Klein**
University of Potsdam
mklein@math.uni-potsdam.de

## Abstract

While large training datasets generally offer improvement in model performance, the training process becomes computationally expensive and time consuming. Distributed learning is a common strategy to reduce the overall training time by exploiting multiple computing devices. Recently, it has been observed in the single machine setting that overparameterization is essential for benign overfitting in ridgeless regression in Hilbert spaces. We show that in this regime, data splitting has a regularizing effect, hence improving statistical performance and computational complexity at the same time. We further provide a unified framework that allows to analyze both the finite and infinite dimensional setting. We numerically demonstrate the effect of different model parameters.

## 1 INTRODUCTION

Modern machine learning applications often involve learning statistical models of great complexity and datasets of massive size become increasingly available. However, while increasing the size of the training datasets generally offers improvement in model performance, the training process is very computation-intensive and thus time-consuming. Indeed, hardware architectures have physical limits in terms of storage, memory, processing speed and communication. A central challenge is thus to design efficient large-scale algorithms.

Distributed learning and parallel computing is a common and simple approach to deal with large datasets. The $n$ observations are evenly split to $M$ machines (or *local nodes, workers*), each having access to only a subset of $n/M$ training samples. Each machine performs local computations to fit a model and transmits it to a central node for merging. This simple *divide and conquer* approach having been proposed in e.g. Mann et al. (2009) for striking best balance between accuracy and communication is highly communication efficient: Only one communication step is performed to only one central node[1].

The field of distributed learning has gained increasing attention in different regimes in the last years, with the aim of establishing conditions for the distributed estimator to be consistent or minimax optimal, see e.g. Chen and Xie (2014), Mackey et al. (2011), Xu et al. (2019), Fan et al. (2019), Shi et al. (2018), Battey et al. (2018), Fan et al. (2021), Bao and Xiong (2021). We give a more detailed overview over approaches that are most closely related to our approach. For a general overview we refer to Bekkerman et al. (2011) and the recent review Gao et al. (2021).

The learning properties of distributed (kernel) ridge regression are well understood. The authors in Zhang et al. (2015) show optimal learning rates with appropriate regularization, if 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

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

[1] This approach is also called *centralized learning.*

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. This is somewhat counterintuitive as smooth signals are easier to reconstruct. To overcome this issue, the authors Chang et al. (2017) utilize additional unlabeled data, leading to a slight improvement.

These works have been extended to more general spectral regularization algorithms for nonparametric least square regression in (reproducing kernel) Hilbert spaces in Guo et al. (2017), Mücke and Blanchard (2018), including gradient descent (Lin and Zhou, 2018) and stochastic gradient descent (Lin and Cevher, 2018).

Finally, we mention Zhang et al. (2013), Dobriban and Sheng (2021), Rosenblatt and Nadler (2016) who study averaged empirical risk minimization in the underparameterized regime, the latter in the high dimensional limit.

We consider distributed ridgeless regression over Hilbert spaces with (local) overparameterization. This setting has been investigated recently in e.g. Bartlett et al. (2020), Chinot and Lerasle (2020), Shang (2021), Muthukumar et al. (2020) in the single machine context with the aim of establishing conditions when *benign* or *harmless* overfitting occurs. This serves as a proxy to understand neural network learning where the phenomenon of benign overfitting was first observed (Bartlett et al., 2021; Belkin, 2021). Indeed, wide networks that are trained with gradient descent can be accurately approximated by linear functions in a Hilbert space. Our results are a step towards understanding the statistical effects in distributed settings in deep learning.

**Contributions.** We provide a unified framework that allows to simultaneously analyze the finite and infinite dimensional distributed ridgeless regression problem. All our bounds are optimal.

We show that in the presence of overparameterization the number of data splits has a regularizing effect that trades off bias and variance. While overparameterization induces an additional bias, averaging reduces variance sufficiently. Hence, data splitting improves statistical accuracy (for an increasing number of splits until the optimal number is achieved) and scales to large data sets at once. Our approach fits into the line of *communication efficient* distributed algorithms and is easy to implement.

To precisely quantify the interplay of statistical accuracy, computational complexity and signal strength we work in a general random-effects model. We find that the numerical speed up[2] is high for low signal strength and improves efficiency. A similar phenomenon is observed in Sheng and Dobriban (2020) for distributed ridge regression. In addition, we do not observe a saturation effect for the number of machines as described above for kernel ridge regression.

The spectral properties of the covariance operator also highly impact the learning properties of distributed ridgeless regression. The spectral decay needs to be sufficiently fast for a high statistical accuracy. Note that this is known for the single machine setting from Bartlett et al. (2020).

**Organization.** In Section 2 we define the mathematical framework needed to present our main results in Section 3. Section 4 is devoted to a discussion with a more detailed comparison to related work. Some numerical illustrations can be found in Section 5 while the Appendix contains all proofs and additional material.

**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 $\Gamma \in \mathcal{L}(\mathcal{H})$ we denote by $\Gamma^T$ the adjoint operator and for compact $\Gamma$ by $(\lambda_j(\Gamma))_j$ the sequence of eigenvalues. We let $[n] = \{1, ..., n\}$. 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 ridgeless regression and state the main assumptions on our model.

### 2.1 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

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

where $\epsilon \in \mathbb{R}$ is a noise variable. The true regression parameter $\beta^* \in \mathcal{H}$ minimizes the least squares test risk, i.e.

$$\mathcal{R}(\beta^*) = \min_{\beta \in \mathcal{H}} \mathcal{R}(\beta) \, , \qquad \mathcal{R}(\beta) := \mathbb{E}[(y - \langle \beta, x \rangle)^2] \, ,$$

where the expectation is taken with respect to the joint distribution of the pair $(x, y) \in \mathcal{H} \times \mathbb{R}$. This framework

[2] In the sense that the optimal number of data splits is large and hence allows more parallelization.

covers many common supervised learning tasks, e.g. learning in reproducing kernel Hilbert spaces (Rosasco and Villa, 2015).

For our analysis we need to impose some distributional assumptions. To this end, we recall that a positive definite operator $\Gamma \in \mathcal{L}(\mathcal{H})$ is *trace class* (and hence compact), if

$$\mathrm{Tr}(\Gamma) = \sum_{j\in\mathbb{N}} \lambda_j(\Gamma) < \infty \ ,$$

see e.g. Reed (2012).

**Definition 2.1** (Hilbert space valued subgaussian random variable)**.** *Let $z$ be a random variable in $\mathcal{H}$ and let $\Gamma : \mathcal{H} \to \mathcal{H}$ be a bounded, linear and self-adjoint positive definite trace class operator. Given some $\sigma > 0$ we say that $z$ is $(\sigma^2, \Gamma)$-subgaussian if for all $v \in \mathcal{H}$ one has*

$$\mathbb{E}\Big[e^{\langle v, z - \mathbb{E}[z]\rangle}\Big] \leq e^{\frac{\sigma^2}{2}\langle \Gamma v, v\rangle} \ .$$

Note that (taking $\mathcal{H} = \mathbb{R}$) this definition includes the special case of real valued variables. On $\mathcal{H}$, we define the *covariance operator* $\Sigma$ by $\Sigma u := \mathbb{E}[\langle u, x\rangle x]$, where $\mathbb{E}$ denotes expectation w.r.t. the marginal distribution. We assume

**Assumption 2.2.** *1. $\mathbb{E}[x] = 0$ and $\mathbb{E}[||x||^2] < \infty$.*

*2. $x$ is $(\sigma_x^2, \Sigma)$-subgaussian and has independent components.*

*3. The covariance $\Sigma$ possesses an orthonormal basis of eigenvectors $v_j$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq ...$ (counted according to multiplicity).*

*4. Conditionally on $x$, the noise $\varepsilon$ in equation* (1) *is centered and $(\tau^2, id)$-subgaussian, where $id$ denotes the identity on $\mathbb{R}$.*

Note that 1. and 3. imply that $\Sigma$ is trace class (and also positive and self-adjoint). Indeed, this easily follows from

$$\mathbb{E}[||x||^2] = \sum_{j\in\mathbb{N}} \langle v_j, \Sigma v_j\rangle = \sum_{j\in\mathbb{N}} \lambda_j < \infty \ ,$$

where $(v_j)_j$ is an orthormal basis of eigenvectors.

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

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

following the above model (1), with i.i.d. noise $\varepsilon = (\varepsilon_1, ..., \varepsilon_n) \in \mathbb{R}^n$. 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 $x_j$).

### 2.2 Distributed Ridgeless Regression

In the distributed setting, our data are evenly divided into $M$ local disjoint subsets

$$D = D_1 \cup ... \cup D_M$$

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

In addition to the above distributional assumptions we require:

**Assumption 2.3.** *Let $m \in [M]$. Almost surely, the projection of the local data $\mathbf{X}_m$ on the space orthogonal to any eigenvector of $\Sigma$ spans a space of dimension $\frac{n}{M}$.*

More precisely, recall that the data matrix $\mathbf{X}_m$ is built up from $n/M$ *row vectors* $x_k \in \mathcal{H}$. The above assumption means that those row vectors almost surely are in *general position*: Only with zero probability the orthogonal projections of those vectors are linearly dependent in each hyperplane $H_j := \{x \in \mathcal{H}; \langle x, v_j\rangle = 0\}$ orthogonal to the eigenvector $v_j$ of $\Sigma$. In particular, data vectors $x_j$ are collinear to some $v_j$ with zero probability.

Note that Assumptions 2.2 and 2.3 are satisfied if $x, y$ are jointly gaussian with zero mean and $\mathrm{rank}(\Sigma) > n/M$.

We define the *local minimum norm estimator* $\hat\beta_m$ as the solution to the optimization problem

$$\min_{\beta\in\mathcal{H}} \ ||\beta||^2 \quad \text{such that}$$
$$||\mathbf{X}_m\beta - \mathbf{Y}_m||^2 = \min_{\tilde\beta\in\mathcal{H}} ||\mathbf{X}_m\tilde\beta - \mathbf{Y}_m||^2 \ .$$

It is well known that $\hat\beta_m$ has a closed form expression (see Engl et al. (1996)) given by

$$\hat\beta_m = \mathbf{X}_m^T(\mathbf{X}_m\mathbf{X}_m^T)^\dagger \mathbf{Y}_m \ , \tag{2}$$

where $(\mathbf{X}_m\mathbf{X}_m^T)^\dagger$ denotes the pseudoinverse of the bounded linear operator $\mathbf{X}_m\mathbf{X}_m^T$.

In the case that $\dim(\mathcal{H}) = d < \frac{n}{M}$ and $\mathbf{X}_m$ has rank $d$, there is a unique solution to the normal equations. However, under Assumption 2.3 we find many local interpolating solutions $\beta \in \mathcal{H}$ to the normal equations with $\mathbf{X}_m\beta = \mathbf{Y}_m$.

The final estimator is defined as the uniform average

$$\bar\beta_M = \frac{1}{M}\sum_{j=1}^{M} \beta_m \ . \tag{3}$$

We aim at finding optimal bounds for the excess risk

$$\mathcal{R}(\bar{\beta}_M) - \mathcal{R}(\beta^*) = ||\Sigma^{\frac{1}{2}}(\bar{\beta}_M - \beta^*)||^2 ,$$

in high probability, as a function of the number of local nodes $M$ and under various model assumptions.

## 3 MAIN RESULTS

In this section we state our main results. We first derive a general upper bound and consider the infinite and finite dimensional settings in more detail. We complete our presentation with matching lower bounds.

### 3.1 A General Error Bound

Before stating our error bounds we briefly describe the underlying error decomposition in bias and variance. For an estimator $\hat{\beta} \in \mathcal{H}$ let us define the *bias* by

$$\widehat{\mathrm{Bias}}(\hat{\beta}) := ||\Sigma^{1/2}(\mathbb{E}_\epsilon[\hat{\beta}] - \beta^*)||^2$$

and the *variance* as

$$\widehat{\mathrm{Var}}(\hat{\beta}) := \mathbb{E}_\epsilon\Big[||\Sigma^{1/2}(\hat{\beta} - \mathbb{E}_\epsilon[\hat{\beta}])||^2\Big] ,$$

where $\mathbb{E}_\varepsilon[\cdot]$ denotes the conditional expectation given the input data. We then have the following preliminary bound for the excess risk whose full proof is given in Appendix A.

**Lemma 3.1.** *Let $\bar{\beta}_M$ be defined by* (3) *and denote by $\hat{\Sigma}_m = \frac{M}{n}X_m^T X_m$ the local empirical covariance operator. The excess risk can be bounded almost surely by*

$$\mathbb{E}_\epsilon\Big[||\Sigma^{\frac{1}{2}}(\bar{\beta}_M - \beta^*)||^2\Big] = \widehat{\mathrm{Bias}}(\bar{\beta}_M) + \widehat{\mathrm{Var}}(\bar{\beta}_M) ,$$

*where*

$$\widehat{\mathrm{Bias}}(\bar{\beta}_M) \leq \frac{1}{M}\sum_{m=1}^{M}\Big|\Big\langle \beta^*, (\Sigma - \hat{\Sigma}_m)\beta^*\Big\rangle\Big| ,$$

$$\widehat{\mathrm{Var}}(\bar{\beta}_M) \leq \frac{8\tau^2}{M^2}\sum_{m=1}^{M}\mathrm{Tr}\Big[\big(X_m^\dagger\big)^T \Sigma X_m^\dagger\Big] .$$

We are interested in finding conditions such that bias and variance (and thus the excess risk) converge to zero with high probability. To this end, we also take the hardness of the learning problem into account. This can be quantified via a classical a-priori assumption on the minimizer $\beta^*$.

**Assumption 3.2** (General random-effects model)**.** *Let $\Theta \in \mathcal{L}(\mathcal{H})$ be compact. Let $\beta^*$ be randomly sampled (independently of $\varepsilon$) with mean $\mathbb{E}_{\beta^*}[\beta^*] = 0$ and covariance $\mathbb{E}_{\beta^*}[\beta^*(\beta^*)^T] = \Theta$.*

This assumption is a slight generalization of the classical concept of a *source condition* in inverse problems (Mathé and Pereverzev, 2003) and learning in (reproducing kernel) Hilbert spaces (Bauer et al., 2007; Blanchard and Mücke, 2018; Lin et al., 2020); see also Richards et al. (2020), Sheng and Dobriban (2020) for the context of (distributed) high dimensional ridge(less) regression. We give some specific examples in Assumptions 3.12, 3.6 below.

For bounding the variance we follow the approach in Chinot and Lerasle (2020), Bartlett et al. (2020) and choose an index $k \in \mathbb{N}$ and split the spectrum of $\Sigma$ accordingly. For a suitable choice of $k$ (called *effective dimension*) it will be crucial to control two notions of the *effective ranks*, see e.g. Koltchinskii and Lounici (2017); Bartlett et al. (2020)

**Definition 3.3** (Effective Ranks)**.** *For $k \geq 0$ with $\lambda_{k+1} > 0$ we define*

$$r_k(\Sigma) := \frac{\sum_{j>k}\lambda_j(\Sigma)}{\lambda_{k+1}(\Sigma)} , \quad R_k(\Sigma) = \frac{\Big(\sum_{j>k}\lambda_j(\Sigma)\Big)^2}{\sum_{j>k}\lambda_j^2(\Sigma)} .$$

**Definition 3.4** (Effective Dimension)**.** *Let $a > 1$ and $M \in [n]$. Define the* effective dimension *as*

$$k^* = k^*_{\frac{n}{M}} := \min\Big\{k \geq 0 \ : \ r_k(\Sigma) \geq a\frac{n}{M}\Big\} ,$$

*where the minimum of the empty set is defined as $\infty$.*

Our main result gives an upper bound for the bias and variance in terms of the source condition, effective ranks and effective dimension.

**Theorem 3.5.** *Suppose Assumptions 2.2, 2.3 and 3.2 are satisfied and let $\delta \in (0, 1]$. There exists a universal constant $c_1 > 0$ such that for all $\frac{n}{M} \geq \frac{1}{c_1}\log(2/\delta)$, with probability at least $1 - \delta$*

$$\mathbb{E}_{\beta^*}[\widehat{\mathrm{Bias}}(\bar{\beta}_M)] \leq \frac{4\sigma_x}{c_1}\log^{\frac{1}{2}}\left(\frac{2M}{\delta}\right)\mathrm{Tr}[\Sigma\Theta]\sqrt{\frac{M}{n}} .$$

*Additionally, there exist $c_2 > 1$ such that, if*

$$k^*_{\frac{n}{M}} \leq \frac{n}{c_2 M} ,$$

*with probability at least $1 - 7Me^{-\frac{n}{c_2 M}}$*

$$\widehat{\mathrm{Var}}(\bar{\beta}_M) \leq 8c_2\tau^2\left(\frac{k^*_{\frac{n}{M}}}{n} + \frac{n}{M^2}\,\frac{1}{R_{k^*_{\frac{n}{M}}}(\Sigma)}\right) . \tag{4}$$

Theorem 3.5 reveals that the excess risk of the averaged local interpolants converges to zero if

$$\mathrm{Tr}[\Sigma\Theta]\sqrt{\frac{M_n}{n}} \to 0 , \quad \frac{k^*_{\frac{n}{M_n}}}{n} \to 0 ,$$

$$\frac{n}{M_n^2}\;\frac{1}{R_{k^*_{\frac{n}{M_n}}}(\Sigma)} \to 0 \;,$$

for $M_n \leq n$. This imposes restrictions on the decay of the eigenvalues of $\Sigma$. Moreover, the convergence of the bias depends on the prior assumption on $\beta^*$.

In the following two subsections we discuss the infinite dimensional and finite dimensional cases in more detail.

### 3.2 Infinite Dimension

We refine the excess risk bound under more specific assumptions on $\beta^*$ and the spectral decay of the covariance.

**Source Condition.** The a-priori assumption on $\beta^*$ from Assumption 3.2 can be expressed via an increasing *source function* $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ by setting $\Theta = \Phi(\Sigma)$[3], describing how coefficients of $\beta^*$ vary along the eigenvectors of $\Sigma$, see e.g. Richards et al. (2020). Recall that the bias in Theorem 3.5 depends on

$$\mathrm{Tr}[\Sigma\Theta] = \mathrm{Tr}[\Sigma\Phi(\Sigma)] = \sum_{j=1}^{\infty} \lambda_j \Phi(\lambda_j) \;.$$

Thus, the bias is finite if the map $x \mapsto x\Phi(x)$ is non-decreasing while the sequence of eigenvalues $(\lambda_j)_{j\in\mathbb{N}}$ is decreasing.

**Assumption 3.6** (Source Condition)**.** *Assume that $\Phi(x) = x^\alpha$, for $\alpha \geq 0$.*

This particular choice of source function goes under the name *Hölder-type source condition* and is a standard assumption in inverse problems Mathé and Pereverzev (2003) and nonparametric regression (Bauer et al., 2007; Blanchard and Mücke, 2018; Lin et al., 2020). Indeed, it has a direct characterization in terms of *smoothness*, where a larger exponent $\alpha$ corresponds to a smoother regression function. In this regard, this assumption also quantifies the *easiness* of the learning problem: Larger values of $\alpha$ indicate an easier problem, as smoother functions are easier to recover.

**Eigenvalue Decay.** Finally, to control the variance in Theorem 3.5 we impose a specific spectral assumption for the covariance:

**Assumption 3.7.** *Assume that $\lambda_j(\Sigma) = j^{-(1+\varepsilon_n)}$ for some positive sequence $(\varepsilon_n)_{n\in\mathbb{N}}$ with $M_n \lesssim \varepsilon_n n$.*

Polynomially decaying eigenvalues are a common assumption in ridgeless regression. Indeed, it is shown for the single machine setting in Bartlett et al. (2020) that under this assumption, the excess risk of the least-norm interpolant converges to zero and *benign overfitting* occurs.

Our main result in this section is a refined upper bound for the excess risk under the two additional assumptions made above. The proof is given in Appendix A.2.

**Proposition 3.8.** *In addition to all assumptions of Theorem 3.5, suppose that Assumptions 3.7, 3.6 hold. Set*

$$C_{\alpha,n} = \frac{1}{\alpha}\mathbf{1}\{\alpha > 0\} + \frac{1}{\varepsilon_n}\mathbf{1}\{\alpha = 0\}$$

*and assume that $\frac{n}{M} \geq \frac{1}{c_1^2}\log(2/\delta)$. With probability at least $1 - \delta - 7Me^{-\frac{n}{c_2 M}}$ we have*

$$\begin{aligned}&\mathbb{E}_{\beta^*,\epsilon}\Big[||\Sigma^{\frac{1}{2}}(\bar{\beta}_M - \beta^*)||^2\Big] \\ &\leq c_3\sigma_x \log^{\frac{1}{2}}\left(\frac{2M}{\delta}\right) C_{\alpha,n}\sqrt{\frac{M}{n}} + c_4\tau^2\frac{\varepsilon_n}{M} \;,\end{aligned}$$

*for some $c_3 > 0$, $c_4 > 0$.*

The dependence of our error approximations on the number of machines reveals an interesting accuracy-complexity trade-off. Indeed, data splitting has a regularizing effect, where the number of local nodes $M$ acts as an explicit regularization parameter: The bias term is increasing as $\sqrt{M}$ while the variance is decreasing as $1/M$.

The source condition controls the bias: The smoother the solution, i.e. the larger $\alpha > 0$, the smaller the bias. Notably, we observe a *phase transition* to the case where $\alpha = 0$ (low smoothness, harder problem). The bias is multiplied by a factor $1/\varepsilon_n$ for a sequence $(\varepsilon_n)_n$ possibly tending to zero and hence grows with $n$ while for $\alpha > 0$ the factor is $1/\alpha$ that is constant in $n$ and decreasing with $\alpha$.

Eigenvalue decay, reflected in the sequence $(\varepsilon_n)_n$ controls the variance: Ideally, we want $\varepsilon_n \to 0$ to achieve fast decay of the variance. However, even increasing $(\varepsilon_n)_n$ is possible as long as we ensure that $\varepsilon_n/M_n \to 0$.

Balancing both terms allows to establish learning rates for different smoothness regimes (see Appendix A.2):

**Corollary 3.9** (Learning rate high smoothness)**.** *Suppose all assumptions of Proposition 3.8 are satisfied and let $\alpha > 0$. For*

$$\frac{1}{\sqrt{n}} \lesssim \varepsilon_n \lesssim n \;, \tag{5}$$

[3] Recall that the spectral Theorem (Reed, 2012) defines the operator $\Phi(\Sigma)$ via a functional calculus. In particular, since $\Sigma$ is trace class and self-adjoint, we may define $\Phi(\Sigma) := \sum_j \Phi(\lambda_j)\langle\cdot, v_j\rangle v_j$, where $(v_j)_j$ is an orthormal basis of eigenvectors.

*the value*

$$M_n = C_{\tau,\sigma_x}\left(\alpha\varepsilon_n\sqrt{n}\right)^{2/3} \tag{6}$$

*with* $C_{\tau,\sigma_x} = \left(\frac{c_4\tau^2}{c_3\sigma_x}\right)^{2/3}$ *trades-off bias and variance and with the same probability as above, we have*

$$\begin{aligned}&\mathbb{E}_{\beta^*,\epsilon}\left[||\Sigma^{\frac{1}{2}}(\bar{\beta}_{M_n} - \beta^*)||^2\right]\\ &\leq \frac{C'_{\tau,\sigma_x}}{\alpha^{2/3}}\log^{\frac{1}{2}}\left(\frac{4M_n}{\delta}\right)\left(\frac{\varepsilon_n}{n}\right)^{1/3},\end{aligned} \tag{7}$$

*for some* $C'_{\tau,\sigma_x} > 0$.

**Corollary 3.10** (Learning rate low smoothness)**.** *Suppose all assumptions of Proposition 3.8 are satisfied and let* $\alpha = 0$*. For*

$$\frac{1}{\sqrt{n}} \lesssim \varepsilon_n^2 \lesssim n\,, \tag{8}$$

*the value*

$$M_n = C_{\tau,\sigma_x}\left(\varepsilon_n^2\sqrt{n}\right)^{2/3}$$

*with* $C_{\tau,\sigma_x} = \left(\frac{c_4\tau^2}{c_3\sigma_x}\right)^{2/3}$ *trades-off bias and variance and with the same probability as above, we have*

$$\begin{aligned}&\mathbb{E}_{\beta^*,\epsilon}\left[||\Sigma^{\frac{1}{2}}(\bar{\beta}_{M_n} - \beta^*)||^2\right]\\ &\leq C'_{\tau,\sigma_x}\log^{\frac{1}{2}}\left(\frac{4M_n}{\delta}\right)\left(\frac{1}{\varepsilon_n n}\right)^{1/3},\end{aligned}$$

*for some* $C'_{\tau,\sigma_x} > 0$.

### 3.3 Finite Dimension

In this section we investigate the finite dimensional setting in more detail and assume $dim(\mathcal{H}) = d < \infty$. To highlight the effects of all characteristics effecting model performance, we make two particularly simple structural assumptions. More specifically, we assume the covariance $\Sigma$ to follow a *strong and weak features model*:

**Assumption 3.11** (Strong-weak-features model)**.** *Let* $F \in [d]$ *and* $\rho_1 \geq \rho_2 > 0$*. Suppose that* $\lambda_j(\Sigma) = \rho_1$ *for all* $j \in [F]$ *and* $\lambda_j(\Sigma) = \rho_2$ *for all* $F+1 \leq j \leq d$*. Without loss of generality, we assume that* $||\Sigma|| = 1$*, i.e.* $\rho_1 = 1$.

Elements in the eigenspace associated to the larger eigenvalue $\rho_1$ are called *strong features* while elements in the eigenspace associated to the smaller eigenvalue are called *weak features*, see e.g. Richards et al. (2020).

Furthermore, we work in a standard *random-effects model*, see Sheng and Dobriban (2020), Dobriban and Wager (2018), Dicker and Erdogdu (2017).

**Assumption 3.12** (Random-effects model)**.** *Define the* signal-to-noise-ratio *as* $\mathrm{SNR} = \mathbb{E}[||\beta^*||^2]/\tau^2$*. The coordinates of* $\beta^*$ *are independent, have zero mean and variance* $\frac{\mathrm{SNR}}{d}$*, i.e.* $\Theta = \frac{\mathrm{SNR}}{d}\,Id_d$.

The next result presents an upper bound for the excess risk under both assumptions. The proof is provided in Appendix A.3.

**Proposition 3.13.** *In addition to all assumptions of Theorem 3.5, suppose Assumptions 3.12, 3.11 hold. Suppose that the weak features satisfy* $\rho_2 d \leq F$ *and*

$$cF \leq \frac{n}{M} \leq \frac{1}{a}(d-F)\,, \tag{9}$$

*for some* $a, c > 1$*. If* $\frac{n}{M} \geq \frac{1}{c_1}\log(2/\delta)$*, then with probability at least* $1-\delta-7Me^{-\frac{n}{c_2 M}}$*, the excess risk satisfies for some* $c > 0$

$$\begin{aligned}&\mathbb{E}_{\beta^*,\epsilon}\left[||\Sigma^{\frac{1}{2}}(\bar{\beta}_M - \beta^*)||^2\right]\\ &\leq c\log^{\frac{1}{2}}\left(\frac{4M}{\delta}\right)\Bigg(\frac{\mathrm{SNR}\cdot F}{d}\sqrt{\frac{M}{n}} +\\ &+ \frac{F}{n} + \frac{1}{M^2}\,\frac{n}{d-F}\Bigg).\end{aligned} \tag{10}$$

*Here,* $c$ *depends on* $\sigma_x$ *and* $\tau$.

The assumption $\rho_2 d \leq F$ controls the bias (see Lemma A.7 for details) and ensures that the strength $\rho_2$ of the weak features is small enough relative to the dimension $d$ and consequently they do not contribute much, while the amount $F$ of strong features is sufficiently high. The bias is further determined by the signal-to-noise ratio SNR and the ratio $F/d \leq 1$ of the number of strong features to the dimension: The bias is small, if both quantities are small.

The above result shows how the various model parameters determine statistical accuracy: As above, we observe that data splitting has a regularizing effect: The bias term is increasing as $\sqrt{M}$ while averaging significantly reduces the variance that decreases as $1/M^2$. Minimizing the rhs in (10) in $M$ allows to trade-off these different contributions.

**Corollary 3.14** (Optimal number of nodes)**.** *Suppose all assumptions of Proposition 3.13 are satisfied. The optimal number*[4] *of local nodes* $M_n$ *is given by*

$$M_n = \left(\frac{4dn^{3/2}}{\mathrm{SNR}\cdot F(d-F)}\right)^{2/5}. \tag{11}$$

*The excess risk satisfies with probability at least* $1-$

[4] The optimal number is defined as the minimizer of the right hand side in (10).

$\delta - 7M_n e^{-\frac{n}{c_2 M_n}}$

$$\mathbb{E}_{\beta^*,\epsilon}\left[||\Sigma^{\frac{1}{2}}(\bar{\beta}_{M_n} - \beta^*)||^2\right] \leq 5c\log^{\frac{1}{2}}\left(\frac{4M_n}{\delta}\right)\left(\frac{F}{n} + \frac{1}{M_n^2}\frac{n}{d-F}\right), \quad (12)$$

*for some $c > 0$.*

The optimal number $M_n$ of local nodes grows with the sample size and the the numerical speed up is high for a low SNR (recall that the larger $M_n$, the more computational savings). A similar phenomenon is observed in Sheng and Dobriban (2020) for distributed ridge regression in finite dimension where it is shown that distributed ridge regression works well when the signal strength is low.

Note that (9) puts a restriction on the growth of $d$ and $F$ with $n$.

**Example 3.15** (When does the error converges to zero?)**.** *We consider now a* high dimensional *and* infinite-worker *limit $M_n \to \infty$. More specifically, we let $n \to \infty$, $d_n \simeq n^\alpha$ for some $\alpha \geq 1$ while we assume $F = const.$. This requires that the strength of the weak features also needs to be sufficiently small (depending on $n$), i.e. $\rho_2 \leq F/n^\alpha$. Note that for these choices,* (9) *is satisfied for any $n$ large enough. Then, Corollary 3.14 gives $M_n \simeq n^{3/5}$ and*

$$\mathbb{E}_{\beta^*,\epsilon}\left[||\Sigma^{\frac{1}{2}}(\bar{\beta}_{M_n} - \beta^*)||^2\right] \leq 5c\log^{\frac{1}{2}}\left(\frac{4M_n}{\delta}\right)\frac{F}{n}\,.$$

We finally compare our findings from Proposition 3.13 with Dar et al. (2021) for the spiked covariance model in the single machine case, i.e. $M = 1$. Eq. (29) in Dar et al. (2021) provides a bound for the noise test error of order $\frac{F}{n} + \frac{n}{d-F}$, leading to sufficient and necessary conditions for harmless interpolation of noise, namely: $F \lesssim n \lesssim d - F$. Our bound in (10) generalizes this to the setting $M > 1$, with harmless local interpolation if

$$F \lesssim \frac{n}{M} \lesssim d - F\,.$$

We then get for the variance

$$\frac{F}{n} + \frac{1}{M^2}\,\frac{n}{d-F} \lesssim \frac{1}{M}\,,$$

recovering the result of Dar et al. (2021) for $M = 1$ and showing a reduction of variance by a factor of $1/M$ for averaging.

### 3.4 Lower Bound

Finally, we give a matching lower bound for the excess risk for the distributed estimator with the optimal choice of local nodes. All proofs of this section are provided in Appendix A.4.

The derivation of our result requires a lower bound for the noise variance:

**Assumption 3.16.** *The conditional noise variance is almost surely bounded below by some constant $\sigma^2 > 0$, i.e. $\mathbb{E}[\varepsilon^2|x] \geq \sigma^2$.*

We start with a general lower bound for the excess risk in terms of the effective ranks and the effective dimension.

**Theorem 3.17.** *Suppose Assumptions 3.16,2.2, 2.3 are satisfied. With probability at least $1 - 10Me^{-\frac{1}{c}\frac{n}{M}}$*

$$\mathbb{E}_\varepsilon[||\Sigma^{1/2}(\bar{\beta}_M - \beta^*)||^2] \geq c_a\sigma^2\left(\frac{k^*_{\frac{n}{M}}}{n} + \frac{n}{M^2}\frac{1}{R_{k^*}(\Sigma)}\right), \quad (13)$$

*for some $c_a > 0$.*

Note that the lower bound for the excess risk is of the order of the variance bound (4). We emphasize that the optimal number $M_n$ of splits is derived by trading-off bias and variance. Hence, for this value, the bound (13) is optimal. We give now the explicit optimal rates in the special settings from Sections 3.2, 3.3.

**Corollary 3.18** (Optimal rate infinite dimension)**.** *Suppose all Assumptions of Theorem 3.17 and Corollary 3.9 are satisfied. Let $\varepsilon_n \lesssim 1$ for $n$ sufficiently large and recall the definition of $M_n$ from* (6)*. With probability at least $1 - 10M_n e^{-\frac{1}{c}\frac{n}{M_n}}$, the excess risk is lower bounded by*

$$\mathbb{E}_\varepsilon[||\Sigma^{1/2}(\bar{\beta}_{M_n} - \beta^*)||^2] \geq \frac{\tilde{C}_{\tau,\sigma_x,\sigma}}{\alpha^{2/3}}\left(\frac{\varepsilon_n}{n}\right)^{1/3},$$

*for some $\tilde{C}_{\tau,\sigma_x,\sigma} > 0$. Hence, under the Assumptions of Corollary 3.9, the rate of convergence is optimal (up to a log-factor) as it matches the upper bound* (7)*. Note that we also obtain the optimal bound from Corollary 3.10 in the low smoothness regime (see Appendix A.4).*

**Corollary 3.19** (Optimal rate finite dimension)**.** *Recall the strong-weak-features model from Section 3.3 and suppose Assumption 3.11 is satisfied. There is a constant $c_1 > 0$ such that for any $0 \leq F \leq \frac{n}{Mc_1} \leq d - F$, with probability at least $1 - 10e^{-\frac{n}{Mc_1}}$*

$$\mathbb{E}_\varepsilon[||\Sigma^{1/2}(\bar{\beta}_M - \beta^*)||^2] \geq \sigma^2 c_3\left(\frac{F}{n} + \frac{n}{M^2(d-F)}\right),$$

*for some $c_3 > 0$. Moreover, under the assumptions of Corollary 3.14, this lower bound matches the upper bound for the optimal $M_n$ and hence is optimal (up to a log-factor).*

### 3.5 Efficiency

In addition to the non-asymptotic bounds on bias and variance we are interested in the possible gain in efficiency of data splitting compared to the single machine setting. To this end, let us introduce the ratio of the excess risks for the single machine estimator $\bar{\beta}_1$ and the distributed estimator $\bar{\beta}_M$, $M > 1$.

**Definition 3.20.** *We define the* relative prediction efficiency *by*

$$\widehat{\mathrm{Eff}}(M) := \frac{\mathbb{E}_\epsilon\big[||\Sigma^{1/2}(\bar{\beta}_1 - \beta^*)||^2\big]}{\mathbb{E}_\epsilon\big[||\Sigma^{1/2}(\bar{\beta}_M - \beta^*)||^2\big]} \ .$$

We consider the setting of Section 3.2. Note that we obtain for the single machine setting with probability at least $1 - \delta - 7e^{-\frac{n}{c_2}}$

$$\varepsilon_n \lesssim \mathbb{E}_{\beta^*,\epsilon}\Big[||\Sigma^{\frac{1}{2}}(\bar{\beta}_1 - \beta^*)||^2\Big] \lesssim \max\bigg\{\varepsilon_n, \frac{C_{\alpha,n}}{\sqrt{n}}\bigg\} \ ,$$

with

$$C_{\alpha,n} = \frac{1}{\alpha}\mathbf{1}\{\alpha > 0\} + \frac{1}{\varepsilon_n}\mathbf{1}\{\alpha = 0\} \ .$$

This follows from Proposition 3.8 and Corollary 3.18 (in particular (22), (23)). This risk bound is optimal if $\frac{C_{\alpha,n}}{\sqrt{n}} \lesssim \varepsilon_n$.

Similarly, denoting by $M_{opt}$ the optimal number of splits from Corollaries 3.9, 3.10, with probability at least $1 - \delta - 7M_{opt}e^{-\frac{n}{c_2 M_{opt}}}$

$$\frac{\varepsilon_n}{M_{opt}} \lesssim \mathbb{E}_{\beta^*,\epsilon}\Big[||\Sigma^{\frac{1}{2}}(\bar{\beta}_{M_{opt}} - \beta^*)||^2\Big] \lesssim \frac{\varepsilon_n}{M_{opt}} \ .$$

As a result, optimal data splitting leads to a linear increase in efficiency:

**Corollary 3.21.** *Let all assumptions of Corollaries 3.9, 3.10, 3.18 be satisfied and assume that* $\frac{C_{\alpha,n}}{\sqrt{n}} \lesssim \varepsilon_n$. *Then, with probability at least* $1 - \delta - 7M_{opt}e^{-\frac{n}{c_2 M_{opt}}}$

$$\widehat{\mathrm{Eff}}(M_{opt}) \simeq M_{opt} \ .$$

## 4 DISCUSSION

**Double Descent.** The phenomenon of *double descent* describes the shape of risk curves in the context of modern high-complexity learning. With increasing complexity, the risk initially decreases, attains a minimum and then increases until the interpolation threshold is reached and where the training data are fitted perfectly. Increasing the complexity even further, the risk decreases a second and final time, see Belkin et al. (2020), Belkin et al. (2019) in the context of least squares.

In Theorem B.6 we prove a general lower bound in finite dimension under more general distributional assumptions, i.e.

$$\begin{aligned}&\mathbb{E}\Big[||\Sigma^{1/2}(\bar{\beta}_M - \beta^*)||^2\Big] \\ &\quad \geq \frac{\tilde{\tau}^2}{M}\,\frac{\min\{d, \frac{n}{M}\}}{\max\{d, \frac{n}{M}\} + 1 - \min\{d, \frac{n}{M}\}} \ .\end{aligned}$$

This bound reveals that double descent occurs with a peak at the local interpolation thresholds, i.e. at $d = \frac{n}{M}$ with height at least $\tilde{\tau}^2 \frac{d}{M}$. We show this phenomenon in Fig. 2 on the MSDYear dataset.

**Comparison to averaged ordinary least squares.** To understand the regularizing effect of the number of data-splits we compare our approach to averaged ordinary least squares (AOLS) for $d < n$, i.e. (2), (3) in the *underparameterized* regime (Rosenblatt and Nadler, 2016).

It is known that under Gaussian design, the risk of the OLS estimator is given by $\frac{d}{n-1-d}$, provided $d < n - 1$ (Breiman and Freedman, 1983). Since OLS is *unbiased*, AOLS is unbiased, too, and the risk behaves fundamentally different as a function of $M$. In particular, there is no trade-off between bias and variance and hence, data splitting has no regularizing effect. Even worse: Since in the distributed setting there are locally less samples, $n/M$, we observe a blow up in the local variance. Averaging reduces this by a factor of $1/M$, giving $\frac{d}{M(n/M-1-d)}$ for the risk. Hence, the relative prediction efficiency (see Def. 3.20) is

$$\begin{aligned}\widehat{\mathrm{Eff}}(M) &= \frac{d}{n-1-d}\Big/\frac{d}{M(n/M-1-d)} \\ &= \frac{n}{n-1-d} - M\frac{d+1}{n-1-d} \ ,\end{aligned}$$

i.e. linearly decreasing in $M$. In other words, by parallelizing, we trade accuracy for speed, see Fig. 4 in Section C.1. This is opposed to the overparametrized regime, where we observe an additional bias and hence an *increase* in efficiency until the optimal number of splits $M_{opt}$ is achieved.

**Comparison to distributed Ridge Regression.** The learning properties of the distributed ridgeless estimator also changes with additional regularization as for distributed (kernel) ridge regression (DRR). This setting is extensively investigated in kernel learning e.g. Zhang et al. (2015), Lin et al. (2017), Mücke and Blanchard (2018). In this setup, the averaged estimator suffers *no loss* in accuracy, i.e. has constant risk, if

appropriately regularized and the number of machines grows sufficiently slowly with the sample size, see Fig. 4 in Section C.1.
The work Sheng and Dobriban (2020) investigates DRR in the high dimensional limit and finds that the efficiency is generally high when the signal strength is low. Note that we observe a similar phenomenon in Corollary 3.14 through the signal-to-noise-ratio SNR. A low SNR increases the optimal number $M_n$. Moreover, the authors show that even in the limit of many machines, DRR does not lose all efficiency.

## 5 NUMERICAL ILLUSTRATION

In this section we present some numerical examples, illustrating our main findings. Additional numerical results are presented in Appendix C.

**Simulated data.** We illustrate the findings of Section 3.3 in the *strong-weak-features model.* In a first experiment we generate $n = 200$ i.i.d. training points $x_j \sim \mathcal{N}(0, \Sigma)$, with $d = 600$, $\rho_1 = 1$, $\rho_2 = 10^{-4}$. The target $\beta^*$ is simulated according to Assumption 3.12 with SNR $= 0.1$. We illustrate the effect of the number $F$ of strong features on the optimal number of data splits. The left plot in Fig. 1 shows the regularizing effect of data splitting. Interestingly, for fixed $F$, efficiency increases until the optimal number of splits is achieved. The optimal number of splits decreases as $F$ increases.

In a second experiment, we investigate the interplay of the spectral gap $\rho_1 - \rho_2$ and the optimal splits. The strength $\rho_2$ of weak features varies between $10^{-3}$ and $10^{-1}$. The right plot in Fig. 1 plots the test error for different values of the spectral gap for an increasing number of machines. We clearly observe the regularizing effect of data splitting in the presence of overparameterization. Moreover, the optimal number of splits decreases as the spectral gap increases.

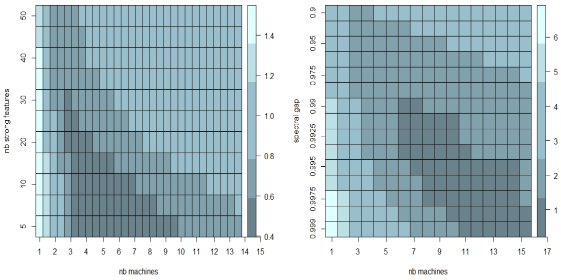

Figure 1: **Left:** Interplay between number of strong features and number of machines. **Right:** Interplay between spectral gap and optimal number of machines.

**Real data.** We utilize the million song dataset Bertin-Mahieux et al., consisting of 463, 715 training samples, $n_{test} = 51,630$ test samples and $d = 90$ features. To illustrate the effect of splitting we elaborate two different settings: The left plot in Fig. 3 shows data splitting in the presence of global overparameterization. We subsampled $n = 45$ training samples and report the average test error with 100 repetitions. We observe a better accuracy with splitting. In the second setting in Fig. 3, the total sample size is larger than the number of parameters. As long as there is local underparameterization, the test error increases. However, after a certain number of splits $M = n/d = 15$, local overparameterization appears and the test error starts to decrease.

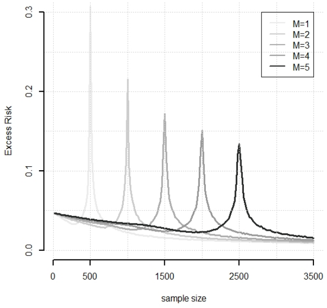

Figure 2: Double descent for MSDYear dataset with $d = 90$ features. We observe peaks whenever $d = \frac{n}{M}$, as Theorem B.6 predicts.

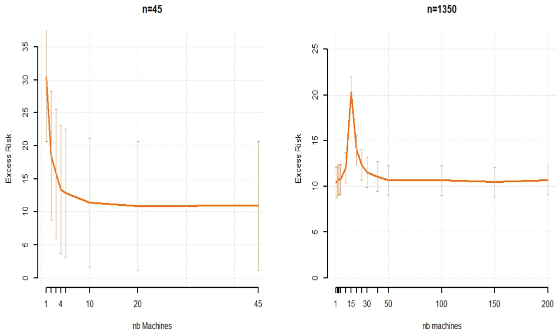

Figure 3: MSDYear dataset. **Left:** Data splitting reduces the test error in the presence of overparameterization. **Right:** The test error has a peak for $M = n/d$ and decreases as local overparameterization increases.

### Acknowledgements

E.R. acknowledges support from the Potsdam Graduate School. We thank all reviewers who gave useful comments and constructive feedback, improving readability of the final version.

## References

Yajie Bao and Weijia Xiong. One-round communication efficient distributed m-estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 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. 1, 1, 3.1, 3.2, A.1, A.3, A.1, 1, 2, A.9, A.4

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *arXiv preprint arXiv:2103.09177*, 2021. 1

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

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007. 3.1, 3.2

Ron Bekkerman, Mikhail Bilenko, and John Langford. *Scaling up machine learning: Parallel and distributed approaches.* Cambridge University Press, 2011. 1

Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *arXiv e-prints*, pages arXiv–2105, 2021. 1

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116 (32):15849–15854, 2019. 4

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020. 4

Thierry Bertin-Mahieux, Daniel PW Ellis, EE LabROSA, Brian Whitman, and Paul Lamere. The million song dataset. 5

Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018. 3.1, 3.2, A.1

Leo Breiman and David Freedman. How many variables should be entered in a regression equation? *Journal of the American Statistical Association*, 78 (381):131–136, 1983. 4

Xiangyu Chang, Shao-Bo Lin, and Ding-Xuan Zhou. Distributed semi-supervised learning with kernel ridge regression. *The Journal of Machine Learning Research*, 18(1):1493–1514, 2017. 1

Xueying Chen and Min-ge Xie. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, pages 1655–1684, 2014. 1

Geoffrey Chinot and Matthieu Lerasle. Benign overfitting in the large deviation regime. *arXiv e-prints*, pages arXiv–2003, 2020. 1, 3.1

Yehuda Dar, Vidya Muthukumar, and Richard G Baraniuk. A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning. *arXiv preprint arXiv:2109.02355*, 2021. 3.3

Lee H Dicker and Murat A Erdogdu. Flexible results for quadratic forms with applications to variance components estimation. *The Annals of Statistics*, 45(1):386–414, 2017. 3.3

Edgar Dobriban and Yue Sheng. Distributed linear regression by averaging. *The Annals of Statistics*, 49(2):918–943, 2021. 1

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018. 3.3

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996. 2.2

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*, pages 1–11, 2021. 1

Yuan Gao, Weidong Liu, Hansheng Wang, Xiaozhou Wang, Yibo Yan, and Riquan Zhang. A review of distributed statistical inference. *Statistical Theory and Related Fields*, pages 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

David Holzmüller. On the universality of the double descent peak in ridgeless regression. *stat*, 1050:5, 2020. B.2

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017. 3.1

Junhong Lin and Volkan Cevher. Optimal distributed learning with multi-pass stochastic gradient methods. In *International Conference on Machine Learning*, pages 3092–3101. PMLR, 2018. 1

Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020. 3.1, 3.2

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, 4

Lester Mackey, Ameet Talwalkar, and Michael I Jordan. Divide-and-conquer matrix factorization. *Advances in neural information processing systems*, 24, 2011. 1

Gideon Mann, Ryan McDonald, Mehryar Mohri, Nathan Silberman, and Daniel D Walker. Efficient large-scale distributed training of conditional maximum entropy models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, pages 1231–1239, 2009. 1

Peter Mathé and Sergei V Pereverzev. Geometry of linear ill-posed problems in variable hilbert scales. *Inverse problems*, 19(3):789, 2003. 3.1, 3.2

Nicole Mücke and Gilles Blanchard. Parallelizing spectrally regularized kernel algorithms. *The Journal of Machine Learning Research*, 19(1):1069–1097, 2018. 1, 4

Nicole Mücke, Gergely Neu, and Lorenzo Rosasco. Beating sgd saturation with tail-averaging and minibatching. In *Advances in Neural Information Processing Systems*, pages 12568–12577, 2019. 1, A.2

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020. 1

Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel Roy. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. In *International Conference on Machine Learning*, pages 7263–7272. PMLR, 2020. A.1

Stephen Page and Steffen Grünewälder. Ivanov-regularised least-squares estimators over large rkhss and their interpolation spaces. *J. Mach. Learn. Res.*, 20:120–1, 2019. A.1

Michael Reed. *Methods of modern mathematical physics: Functional analysis*. Elsevier, 2012. 2.1, 3, A.1

Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge (less) regression under general source condition. *arXiv preprint arXiv:2006.06386*, 2020. 3.1, 3.2, 3.3

Lorenzo Rosasco and Silvia Villa. Learning with incremental iterative regularization. *Advances in Neural Information Processing Systems 28*, pages 1630–1638, 2015. 2.1

Jonathan D Rosenblatt and Boaz Nadler. On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, 5(4):379–404, 2016. 1, 4

Zong Shang. Benign overfitting without concentration. *arXiv preprint arXiv:2101.00914*, 2021. 1

Yue Sheng and Edgar Dobriban. One-shot distributed ridge regression in high dimensions. In *International Conference on Machine Learning*, pages 8763–8772. PMLR, 2020. 1, 3.1, 3.3, 3.3, 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

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018. A.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 C Duchi, and Martin J Wainwright. Communication-efficient algorithms for statistical optimization. *The Journal of Machine Learning Research*, 14(1):3321–3363, 2013. 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, 4

# Supplementary Material: Data splitting improves statistical performance in overparameterized regimes

## A PROOFS OF SECTION 3

In this section we provide all proofs of our results in Section 3.

### A.1 Proofs of Section 3.1

**Lemma A.1.** *Let $n \in \mathbb{N}$, $\beta \in \mathcal{H}$. Define the* empirical covariance *operator by $\hat{\Sigma} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ and denote by*

$$\tilde{\Pi} := Id - \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{\dagger}\mathbf{X}$$

*the orthogonal projection onto the nullspace of $\mathbf{X}$. We have almost surely*

$$||\Sigma^{1/2}\tilde{\Pi}\beta||^2 \leq \left|\left\langle \beta, (\Sigma - \hat{\Sigma})\beta \right\rangle\right| .$$

*Proof of Lemma A.1.* For the proof we will use the following facts that can be found in e.g. Reed (2012):

(a) For all $\beta \in \mathcal{H}$ it holds: $||\beta||^2 = \mathrm{Tr}[\beta \otimes \beta]$.

(b) The trace is invariant under cyclic permutations: $\mathrm{Tr}[ABC] = \mathrm{Tr}[CAB] = \mathrm{Tr}[BCA]$.

(c) If $A, B, C$ are self-adjoint, then the trace is invariant under any permutation:

$$\mathrm{Tr}[ABC] = \mathrm{Tr}[(ABC)^T] = \mathrm{Tr}[CBA] = \mathrm{Tr}[ACB] .$$

(d) If $A$ has rank one, then $|\mathrm{Tr}[A]| = ||A||$. In particular, $\beta \otimes \beta$ has rank one and $|\mathrm{Tr}[\beta \otimes \beta A]| = |\mathrm{Tr}[A\beta \otimes \beta]| = ||\beta \otimes \beta A||$.

First observe that

$$\begin{aligned} ||\Sigma^{1/2}\tilde{\Pi}\beta^*||^2 &\overset{(a)}{=} \left|\mathrm{Tr}\left[\Sigma^{1/2}\tilde{\Pi}\beta \otimes \Sigma^{1/2}\tilde{\Pi}\beta\right]\right| \\ &= \left|\mathrm{Tr}\left[\Sigma^{1/2}\tilde{\Pi}(\beta \otimes \beta)\tilde{\Pi}\Sigma^{1/2}\right]\right| \\ &\overset{(b)}{=} \left|\mathrm{Tr}\left[\tilde{\Pi}\Sigma\tilde{\Pi}(\beta \otimes \beta)\right]\right| \\ &\overset{(d)}{=} ||\tilde{\Pi}\Sigma\tilde{\Pi}(\beta \otimes \beta)|| =: \bullet . \end{aligned}$$

Since $\tilde{\Pi}$ is an orthogonal projection onto the nullspace of $\mathbf{X}$ we have $||\tilde{\Pi}|| \leq 1$ and

$$\tilde{\Pi}\mathbf{X}^T = 0 , \quad \tilde{\Pi}\hat{\Sigma} = 0 .$$

Hence, we find

$$\begin{aligned}\bullet &= ||\tilde{\Pi}(\Sigma - \hat{\Sigma})\tilde{\Pi}(\beta \otimes \beta)|| \\ &\leq ||\tilde{\Pi}||\; ||(\Sigma - \hat{\Sigma})\tilde{\Pi}(\beta \otimes \beta)|| \\ &\overset{(d)}{=} |\operatorname{Tr}[(\Sigma - \hat{\Sigma})\tilde{\Pi}(\beta \otimes \beta)]| \\ &\overset{(c)}{=} |\operatorname{Tr}[\tilde{\Pi}(\Sigma - \hat{\Sigma})(\beta \otimes \beta)]| \\ &\overset{(d)}{=} ||\tilde{\Pi}(\Sigma - \hat{\Sigma})(\beta \otimes \beta)|| \\ &\leq ||(\Sigma - \hat{\Sigma})(\beta \otimes \beta)|| \\ &\overset{(d)}{=} |\operatorname{Tr}[(\Sigma - \hat{\Sigma})(\beta \otimes \beta)]| \\ &= \left|\left\langle \beta, (\Sigma - \hat{\Sigma})\beta \right\rangle\right| .\end{aligned}$$

□

The next Proposition is useful for bounding the bias in Lemma 3.1. We follow the lines of Negrea et al. (2020), Lemma B.1, where a similar result is shown for gaussian variables. We extend this to the subgaussian setting.

**Proposition A.2.** *Suppose Assumption 2.2 is satisfied and let $\beta \in \mathcal{H}$. There exists a universal constant $c > 0$ such that for any $\delta \geq 2e^{-c^2 n}$, with probability at least $1 - \delta$ we have*

$$\left|\left\langle \beta, (\Sigma - \hat{\Sigma})\beta \right\rangle\right| \leq \frac{4\sigma_x}{c}\; \log^{\frac{1}{2}}(2/\delta)\; \frac{||\Sigma^{\frac{1}{2}}\beta||^2}{\sqrt{n}} .$$

*Proof of Proposition A.2.* Set $B^2 := \langle \Sigma\beta, \beta\rangle$. We then write

$$\begin{aligned}\left|\left\langle \beta, (\Sigma - \hat{\Sigma})\beta \right\rangle\right| &= \left|\left\langle \beta, \hat{\Sigma}\beta \right\rangle - \langle \beta, \Sigma\beta\rangle\right| \\ &= \left|\frac{1}{n}\sum_{j=1}^{n}\langle \beta, (x_j \otimes x_j)\beta\rangle - B^2\right| \\ &= \left|\frac{B^2}{n}\left(\sum_{j=1}^{n}\frac{\langle \beta, x_j\rangle^2}{B^2} - 1\right)\right| . \qquad (14)\end{aligned}$$

We next show that for any $j = 1, ..., n$, the real valued variables $z_j := \frac{\langle \beta, x_j\rangle}{B}$ are $(\sigma_x^2, id)$-subgaussian. Indeed, by Assumption 2.2 and Definition 2.1 we find for all $\alpha \in \mathbb{R}$

$$\begin{aligned}\mathbb{E}[e^{\alpha z_j}] &= \mathbb{E}\left[e^{\left\langle \frac{\alpha}{B}\beta, x_j\right\rangle}\right] \\ &\leq e^{\frac{\sigma_x^2}{2}\left\langle \Sigma\frac{\alpha}{B}\beta, \frac{\alpha}{B}\beta\right\rangle} \\ &= e^{\frac{\sigma_x^2}{2}\alpha^2} . \qquad (15)\end{aligned}$$

For bounding (14) with high probability we use the fact that for any $j = 1, ..., n$ the random variable $z_j^2 - 1$ is $16\sigma_x^2$-subexponential. Indeed, this follows from (15) and results in Vershynin (2018, Section 2) that are condensed in Bartlett et al. (2020, Lemma S.4). Next, Bernstein's inequality for the independent and mean zero subexponential variables $z_1, ..., z_n$ in Vershynin (2018, Theorem 2.8.2) shows that there exists a universal constant $c > 0$ such that for all $t \geq 0$, with probability at least

$$1 - 2\exp\left(-c^2 \min\left\{\frac{t^2}{16\sigma_x^2 n}, \frac{t}{4\sigma_x}\right\}\right)$$

we have

$$\left|\frac{B^2}{n}\left(\sum_{j=1}^{n}\frac{\langle \beta, x_j\rangle^2}{B^2} - 1\right)\right| \leq \frac{B^2}{n}\; t .$$

Assuming that $t \leq 4\sigma_x n$ we find that

$$\min\left\{\frac{t^2}{16\sigma_x^2 n}, \frac{t}{4\sigma_x}\right\} = \frac{t^2}{16\sigma_x^2 n} \; .$$

Setting now $\delta = 2e^{-\frac{c^2}{16\sigma_x^2}\frac{t^2}{n}}$ we finally get

$$\left|\left\langle \beta, (\Sigma - \hat{\Sigma})\beta \right\rangle\right| \leq \frac{4\sigma_x}{c}\frac{||\Sigma^{\frac{1}{2}}\beta||^2}{\sqrt{n}} \log^{\frac{1}{2}}(2/\delta) \; .$$

with probability at least $1-\delta$, for all $\delta \geq 2e^{-c^2 n}$. □

The next result establishes a bound for the single machine variance. This is a first step for bounding the variance from Lemma 3.1 in the distributed setting.

**Proposition A.3** (Bartlett et al. (2020), Lemma 6)**.** *Let $n \in \mathbb{N}$ and suppose Assumption 2.2 is satisfied. Define*

$$\widehat{C} := (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\Sigma\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1} \; . \tag{16}$$

*There are constants $a, c > 1$ such that, if $0 \leq k \leq n/c$, $r_k(\Sigma) \geq a \cdot n$, and $l \leq k$, with probability at least $1 - 7e^{-\frac{n}{c}}$ it holds*

$$\mathrm{Tr}[\widehat{C}] \leq c\left(\frac{l}{n} + n\,\frac{\sum_{j>l}\lambda_j^2}{\left(\sum_{j>k}\lambda_j\right)^2}\right) ,$$

*where $(\lambda_j)_{j\in\mathbb{N}}$ are the eigenvalues of $\Sigma$, arranged in decreasing order.*

**Proposition A.4.** *Let $n \in \mathbb{N}$ and suppose Assumption 2.2 is satisfied. Define $\widehat{C}$ by* (16)*. There exists a universal constant $c > 1$ and a $0 \leq k_n^* \leq \frac{n}{c}$ such that with probability at least $1 - 7e^{-\frac{n}{c}}$ it holds*

$$\mathrm{Tr}[\widehat{C}] \leq c\left(\frac{k_n^*}{n} + n\,\frac{\sum_{j>k_n^*}\lambda_j^2}{\left(\sum_{j>k_n^*}\lambda_j\right)^2}\right) ,$$

*where $(\lambda_j)_{j\in\mathbb{N}}$ are the eigenvalues of $\Sigma$, arranged in decreasing order.*

*Proof of Proposition A.4.* The proof follows from Bartlett et al. (2020, Lemma 6) and Bartlett et al. (2020, Lemma 11). □

Combining now the above results allows to prove Lemma 3.1.

*Proof of Lemma 3.1.* We first derive a bound for the bias. Linearity of the expectation and (1) yields

$$\mathbb{E}_\varepsilon[\bar{\beta}_M] = \frac{1}{M}\sum_{m=1}^{M} X_m^T(X_m X_m^T)^\dagger \mathbb{E}_\varepsilon[Y_m] = \frac{1}{M}\sum_{m=1}^{M} X_m^T(X_m X_m^T)^\dagger X_m \beta^* \; , \tag{17}$$

since, conditionally on the inputs $\mathbf{X}$, the noise is centered. Hence

$$\beta^* - \mathbb{E}_\epsilon[\bar{\beta}_M] = \frac{1}{M}\sum_{m=1}^{M} \tilde{\Pi}_m \beta^* \; ,$$

where we denote by

$$\tilde{\Pi}_m := Id - \mathbf{X}_m^T(\mathbf{X}_m\mathbf{X}_m^T)^\dagger \mathbf{X}_m$$

the orthogonal projection onto the nullspace of $\mathbf{X}_m$. Convexity and Lemma A.1 allow to deduce

$$\begin{aligned}\widehat{\mathrm{Bias}}(\bar\beta_M) &= ||\Sigma^{1/2}(\mathbb{E}_\epsilon[\bar\beta_M]-\beta^*)||^2\\ &\leq \frac{1}{M}\sum_{m=1}^M ||\Sigma^{1/2}\tilde\Pi_m\beta^*||^2\\ &\leq \frac{1}{M}\sum_{m=1}^M \Big|\Big\langle \beta^*, (\Sigma-\hat\Sigma_m)\beta^*\Big\rangle\Big| \,.\end{aligned}$$

Next, we derive a bound for the variance. By definition of the variance, (2) and (17) we find

$$\begin{aligned}\widehat{\mathrm{Var}}(\bar\beta_M) &= \mathbb{E}_\epsilon\Big[||\Sigma^{1/2}(\bar\beta_M-\mathbb{E}_\epsilon[\bar\beta_M])||_2^2\Big]\\ &= \mathbb{E}_\varepsilon\Big[||\Sigma^{1/2}\Big(\frac{1}{M}\sum_{m=1}^M \hat\beta^{(m)} - X_m^T(X_mX_m^T)^\dagger X_m\beta^*\Big)||^2\Big]\\ &= \mathbb{E}_\varepsilon\Big[||\Sigma^{1/2}\Big(\frac{1}{M}\sum_{m=1}^M X_m^T(X_mX_m^T)^\dagger(Y_m-X_m\beta^*)\Big)||^2\Big]\\ &= \mathbb{E}_\varepsilon\Big[||\Sigma^{1/2}\Big(\frac{1}{M}\sum_{m=1}^M X_m^T(X_mX_m^T)^\dagger\varepsilon_m\Big)||^2\Big]\\ &= \mathbb{E}_\varepsilon\Big[||\Sigma^{1/2}\Big(\frac{1}{M}\sum_{m=1}^M X_m^\dagger\varepsilon_m\Big)||^2\Big] \,.\end{aligned}$$

In the last step we use $X_m^T(X_mX_m^T)^\dagger = X_m^\dagger$. Recall that for any $\beta\in\mathcal{H}$ we may write $||\beta||^2 = \mathrm{Tr}[\beta\otimes\beta]$. Hence,

$$\begin{aligned}&||\Sigma^{1/2}\Big(\frac{1}{M}\sum_{m=1}^M X_m^\dagger\varepsilon_m\Big)||^2\\ &\mathrm{Tr}\left[\left(\frac{1}{M}\sum_{m=1}^M \Sigma^{1/2}X_m^\dagger\varepsilon_m\right)\otimes\left(\frac{1}{M}\sum_{m'=1}^M \Sigma^{1/2}X_{m'}^\dagger\varepsilon_{m'}\right)\right]\\ &= \frac{1}{M^2}\sum_{m,m'=1}^M \mathrm{Tr}\Big[\Sigma^{1/2}X_m^\dagger\varepsilon_m\otimes\varepsilon_{m'}(X_{m'}^\dagger)^T\Sigma^{1/2}\Big] \,.\end{aligned}$$

By linearity of the trace and independence, taking the expectation gives $\mathbb{E}_\varepsilon[\varepsilon_m\otimes\varepsilon_{m'}]=0$ for any $m\neq m'$ and the sum reduces to

$$\begin{aligned}\mathbb{E}_\varepsilon\Big[||\Sigma^{1/2}\Big(\frac{1}{M}\sum_{m=1}^M X_m^\dagger\varepsilon_m\Big)||^2\Big] &= \frac{1}{M^2}\sum_{m=1}^M \mathbb{E}_\varepsilon\Big[\,\mathrm{Tr}\,\Big[\Sigma^{1/2}X_m^\dagger\varepsilon_m\otimes\varepsilon_m(X_m^\dagger)^T\Sigma^{1/2}\Big]\Big]\\ &= \frac{1}{M^2}\sum_{m=1}^M \mathbb{E}_\varepsilon\Big[||\Sigma^{1/2}X_m^\dagger\varepsilon_m||^2\Big]\\ &= \frac{1}{M^2}\sum_{m=1}^M \mathbb{E}_\varepsilon\Big[\langle\varepsilon_m, C_m\varepsilon_m\rangle\Big] \,, \qquad (18)\end{aligned}$$

where we set

$$C_m := \left(X_m^\dagger\right)^T\Sigma X_m^\dagger \,.$$

To proceed, we apply a conditional subgaussian version of the Hanson-Wright inequality taken from Page and Grünewälder (2019, Lemma 35). This gives almost surely conditional on the data $X_m$, for all $t\geq 0$, with probability at least $1-e^{-t}$ (w.r.t. the noise)

$$\begin{aligned}\langle\varepsilon_m, C_m\varepsilon_m\rangle &\leq \tau^2\,\mathrm{Tr}[C_m] + 2\tau^2 t||C_m|| + 2\tau^2\sqrt{t^2||C_m||^2 + t\,\mathrm{Tr}[C_m^2]}\\ &\leq 4\tau^2\,\mathrm{Tr}[C_m]\,(t+1) \,,\end{aligned}$$

where we use that $||C_m|| \leq \mathrm{Tr}[C_m]$ and $\mathrm{Tr}[C_m^2] \leq ||C_m|| \, \mathrm{Tr}[C_m] \leq \mathrm{Tr}[C_m]^2$. From Blanchard and Mücke (2018, Lemma C.1) we obtain after integration for the conditional expectation

$$\mathbb{E}_\varepsilon[\langle \varepsilon_m, C_m \varepsilon_m \rangle] \leq 8\tau^2 \, \mathrm{Tr}[C_m] \; .$$

Inserting the last bound into (18) finally gives almost surely

$$\widehat{\mathrm{Var}}(\bar{\beta}_M) \leq \frac{8\tau^2}{M^2} \sum_{m=1}^{M} \mathrm{Tr}[C_m] \; .$$

□

Finally, we give the proof of our main result, a general upper bound for distributed ridgeless regression.

*Proof of Theorem 3.5.* We start with bounding the bias term.

**Bounding the Bias.** Recall that by Lemma 3.1 we have almost surely

$$\widehat{\mathrm{Bias}}(\bar{\beta}_M) \leq \frac{1}{M} \sum_{m=1}^{M} \left| \left\langle \beta^*, (\Sigma - \hat{\Sigma}_m)\beta^* \right\rangle \right| \; .$$

Proposition A.2 gives for all $\delta \geq 2e^{-c^2 \frac{n}{M}}$, with probability at least $1 - \delta$

$$\left| \left\langle \beta^*, (\Sigma - \hat{\Sigma}_m)\beta^* \right\rangle \right| \leq \frac{4\sigma_x}{c} \; \log^{\frac{1}{2}}(2/\delta) \; ||\Sigma^{\frac{1}{2}}\beta^*||^2 \sqrt{\frac{M}{n}} \; ,$$

for some universal constant $c > 0$. Performing now a union bound and invoking Assumption 3.2 finally gives with probability at least $1 - \delta$

$$\mathbb{E}_{\beta^*}[\widehat{\mathrm{Bias}}(\bar{\beta}_M)] \leq \frac{4\sigma_x}{c} \; \log^{\frac{1}{2}}(2M/\delta) \; \mathrm{Tr}[\Sigma\Theta] \sqrt{\frac{M}{n}} \; ,$$

where we use that

$$\mathbb{E}_{\beta^*}[||\Sigma^{\frac{1}{2}}\beta^*||^2] = \mathrm{Tr}[\mathbb{E}_{\beta^*}[\Sigma\beta^* \otimes \beta^*]] = \mathrm{Tr}[\Sigma\Theta] \; .$$

**Bounding the Variance.** Applying Lemma 3.1 once more we have almost surely

$$\widehat{\mathrm{Var}}(\bar{\beta}_M) \leq \frac{8\tau^2}{M^2} \sum_{m=1}^{M} \mathrm{Tr}[C_m] \; .$$

With Lemma A.4 together with a union bound we get with probability at least $1 - 7Me^{-\frac{n}{c_2 M}}$

$$\begin{aligned}
\widehat{\mathrm{Var}}(\bar{\beta}_M) &\leq \frac{8c_2\tau^2}{M^2} \sum_{m=1}^{M} \left( \frac{M}{n} k^*_{n/M} + \frac{n}{M} \; \frac{\sum_{j > k^*_{n/M}} \lambda_j^2}{\left( \sum_{j > k^*_{n/M}} \lambda_j \right)^2} \right) \\
&= 8c_2\tau^2 \left( \frac{k^*_{n/M}}{n} + \frac{n}{M^2} \; \frac{\sum_{j > k^*_{n/M}} \lambda_j^2}{\left( \sum_{j > k^*_{n/M}} \lambda_j \right)^2} \right) ,
\end{aligned}$$

for some constant $c_2 > 1$ and $0 \leq k^*_{n/M} \leq \frac{n}{c_2 M}$. □

## A.2 Proofs of Section 3.2

This section establishes a refined upper bound for the excess risk in the infinite dimensional setting under the specific Assumptions 3.6, 3.7. We start with a preliminary Lemma that is needed to estimate the variance.

**Lemma A.5.** *Suppose all assumptions of Theorem 3.5 are satisfied. Assume that $\lambda_j(\Sigma) = j^{-(1+\varepsilon_n)}$ for a positive sequence $(\varepsilon_n)_{n\in\mathbb{N}}$. We have*

1. $\frac{k^*_{\frac{n}{M}}}{n} \leq a\,\frac{\varepsilon_n}{M}$.
2. *For any $n$ sufficiently large,* $\frac{n}{M^2}\,\frac{1}{R_{k^*_{\frac{n}{M}}}(\Sigma)} \leq \frac{6}{a}\,\frac{\varepsilon_n}{M}$. *If $M \lesssim n\varepsilon_n$, then $k^*_{\frac{n}{M}} \gtrsim 1$.*

*Proof of Lemma A.5.*

1. We follow the lines of Bartlett et al. (2020), Proof of Theorem 31, by lower bounding the effective rank $r_k(\Sigma)$. With Lemma 14 in Mücke et al. (2019) we may write

$$\begin{aligned} r_k(\Sigma) &= (k+1)^{1+\varepsilon_n} \sum_{j>k} j^{-(1+\varepsilon_n)} \\ &\geq (k+1)^{1+\varepsilon_n} \int_{k+1}^{\infty} t^{-(1+\varepsilon_n)} \\ &= \frac{k+1}{\varepsilon_n}\,. \end{aligned}$$

   By Definition 3.4, the effective dimension $k^*_{\frac{n}{M}}$ is the smallest number satisfying $r_{k^*_{\frac{n}{M}}}(\Sigma) \geq a\frac{n}{M}$. Hence, $k^*_{\frac{n}{M}} \leq a\varepsilon_n\frac{n}{M}$.

2. A short calculation shows that

$$R_k(\Sigma) \geq \frac{k}{\varepsilon_n^2}\left(1 - \frac{1}{k+1}\right)^{2\varepsilon_n} \quad , \quad r_k(\Sigma) \leq \frac{2k}{\varepsilon_n} e^{\varepsilon_n}\,.$$

   Following the arguments in the proof of Theorem 31 in Bartlett et al. (2020) we find also in the distributed setting that $k^*_{\frac{n}{M}} \geq \frac{a\varepsilon}{3}\frac{n}{M}$ for sufficiently large $n$. Hence,

$$R_{k^*_{\frac{n}{M}}}(\Sigma) \geq \frac{a\varepsilon_n}{6}\frac{n}{M}\,.$$

□

The second preliminary Lemma will help to bound the bias.

**Lemma A.6.** *Suppose Assumption 3.7 is satisfied. Let $\alpha \geq 0$. Then*

$$\mathrm{Tr}[\Sigma^{1+\alpha}] \;=\; \sum_{j=1}^{\infty}\left(\frac{1}{j}\right)^{(1+\alpha)(1+\varepsilon_n)} \;\leq\; \frac{1}{\alpha + \varepsilon_n(1+\alpha)} \;\leq\; \begin{cases} \alpha = 0 & : \frac{1}{\varepsilon_n} \\ \alpha > 0 & : \frac{1}{\alpha} \end{cases}\,.$$

*Proof of Lemma A.6.* Let $\beta = (1+\alpha)(1+\varepsilon_n)$. The infinite sum can easily be bounded by an integral

$$\sum_{j=1}^{\infty}\left(\frac{1}{j}\right)^{\beta} \leq \int_1^{\infty} t^{-\beta}dt = \frac{1}{\beta - 1}\,,$$

see e.g. Lemma 14 in Mücke et al. (2019). □

Combining now Lemma A.5 and Lemma A.6 with Theorem 3.5 gives the main result in this section.

*Proof of Proposition 3.8.* For bounding the bias we apply Lemma A.6

$$\mathrm{Tr}[\Sigma\Theta] = \mathrm{Tr}[\Sigma^{1+\alpha}] = \sum_{j\in\mathbb{N}} \lambda_j^{1+\alpha}(\Sigma) \leq \frac{1}{\alpha}\nVdash\{\alpha > 0\} + \frac{1}{\varepsilon_n}\nVdash\{\alpha = 0\} =: C_{\alpha,n} \ .$$

Combining this with Lemma A.5, Lemma 3.1 and Theorem 3.5 leads to

$$\begin{aligned}
&\mathbb{E}_{\beta^*,\epsilon}\Big[||\Sigma^{\frac{1}{2}}(\bar{\beta}_M - \beta^*)||^2\Big] \\
&\leq \frac{4\sigma_x}{c_1}\log^{\frac{1}{2}}\left(\frac{2M}{\delta}\right)\mathrm{Tr}[\Sigma^{1+\alpha}]\sqrt{\frac{M}{n}} + 8c_2\tau^2\left(a\,\frac{\varepsilon_n}{M} + \frac{6}{a}\,\frac{\varepsilon_n}{M}\right) \\
&\leq \frac{4\sigma_x}{c_1}C_{\alpha,n}\log^{\frac{1}{2}}\left(\frac{2M}{\delta}\right)\sqrt{\frac{M}{n}} + 8c_2c_a\tau^2\frac{\varepsilon_n}{M} \ ,
\end{aligned}$$

holding with probability at least $1-\delta-7Me^{-\frac{n}{c_2M}}$, for any $\delta \geq 2e^{-c_1^2\frac{n}{M}}$. Here, we set $c_a = \max\{a, 6/a\}$. The result follows with $c_3 = 4/c_1$ and $c_4 = 8c_2c_a$. □

*Proof of Corollary 3.9 and Corollary 3.10.* We determine the maximum number of local nodes by balancing bias and variance. To this end, firstly note that $1 \leq \log^{\frac{1}{2}}\left(\frac{4M_n}{\delta}\right)$. Setting now

$$A := c_3\sigma_x\frac{C_{\alpha,n}}{\sqrt{n}} \ , \quad B := c_4\tau^2\varepsilon_n \ ,$$

we find that

$$A\sqrt{M} = \frac{B}{M} \quad \Longleftrightarrow \quad M = \left(\frac{B}{A}\right)^{2/3} .$$

Hence, the value

$$M_n := C_{\tau,\sigma_x}\left(\frac{\varepsilon_n\sqrt{n}}{C_{\alpha,n}}\right)^{2/3} , \quad C_{\tau,\sigma_x} = \left(\frac{c_4\tau^2}{c_3\sigma_x}\right)^{2/3}$$

trades off bias and variance and the excess risk is bounded as

$$\begin{aligned}
\mathbb{E}_{\beta^*,\epsilon}\Big[||\Sigma^{\frac{1}{2}}(\bar{\beta}_{M_n} - \beta^*)||^2\Big] &\leq 2c_4\tau^2\log^{\frac{1}{2}}\left(\frac{4M_n}{\delta}\right)\frac{\varepsilon_n}{M_n} \\
&= C'_{\tau,\sigma_x}\log^{\frac{1}{2}}\left(\frac{4M_n}{\delta}\right)\left(\frac{C^2_{\alpha,n}\varepsilon_n}{n}\right)^{1/3} ,
\end{aligned}$$

where $C'_{\tau,\sigma_x} = \frac{2c_4\tau^2}{C_{\tau,\sigma_x}}$. □

### A.3 Proofs of Section 3.3

In this section we provide the proofs for our results in finite dimension with $dim(\mathcal{H}) = d < \infty$ from Section 3.3. We start with two preliminary Lemmata.

**Lemma A.7.** *Suppose Assumption 3.12 holds and let $d\rho_2 \leq F$. Then*

$$\mathrm{Tr}[\Sigma\Theta] \leq \frac{2\cdot\mathrm{SNR}}{d}\,F \ .$$

*Proof of Lemma A.7.* By Assumption 3.12 with $\Theta = \frac{\mathrm{SNR}}{d}\, Id_d$ and since $d\rho_2 \leq F$ we easily obtain

$$\begin{aligned}
\mathrm{Tr}[\Sigma\Theta] &= \frac{\mathrm{SNR}}{d}\ \mathrm{Tr}[\Sigma] \\
&= \frac{\mathrm{SNR}}{d}\ \left(\sum_{j=1}^{F}\rho_1 + \sum_{j=F+1}^{d}\rho_2\right) \\
&= \frac{\mathrm{SNR}}{d}\ (F\rho_1 + (d-F)\rho_2) \\
&= \frac{\mathrm{SNR}}{d}\ ((\rho_1-\rho_2)F + d\rho_2) \\
&\leq \frac{\mathrm{SNR}}{d}\ (\rho_1-\rho_2+1)F \\
&\leq \frac{2\mathrm{SNR}}{d}\ F\ .
\end{aligned}$$

In the last step we use that $\rho_1 = 1$ and $-\rho_2 \leq 0$. □

**Lemma A.8.** *Suppose all Assumptions of Theorem 3.5 are satisfied. If additionally Assumption 3.11 holds and*

$$cF \leq \frac{n}{M} \leq \frac{1}{a}(d-F)\ ,$$

*for some $a, c > 1$, then with probability at least $1 - 7Me^{-\frac{n}{cM}}$*

$$\widehat{\mathrm{Var}}(\bar\beta_M) \leq 8c\tau^2\left(\frac{F}{n} + \frac{1}{M^2}\frac{n}{d-F}\right)\ .$$

*Proof of Lemma A.8.* Applying Lemma 3.1 and Proposition A.3 with $k = l = F$ gives with probability at least $1 - 7e^{-\frac{n}{cM}}$

$$\begin{aligned}
\widehat{\mathrm{Var}}(\bar\beta_M) &\leq \frac{8\tau^2}{M^2}\sum_{m=1}^{M}\mathrm{Tr}\left[\left(X_m^\dagger\right)^T\Sigma X_m^\dagger\right] \\
&\leq c\frac{8\tau^2}{M^2}\sum_{m=1}^{M}\left(\frac{FM}{n} + \frac{n}{M}\ \frac{\sum_{j>F}\lambda_j^2}{\left(\sum_{j>F}\lambda_j\right)^2}\right) \\
&= c\frac{8\tau^2}{M^2}\sum_{m=1}^{M}\left(\frac{FM}{n} + \frac{n}{M}\ \frac{(d-F)\rho_2^2}{(d-F)^2\rho_2^2}\right) \\
&= 8c\tau^2\left(\frac{F}{n} + \frac{1}{M^2}\frac{n}{d-F}\right)\ ,
\end{aligned}$$

provided $r_F(\Sigma) \geq a \cdot \frac{n}{M}$ and $0 \leq F \leq \frac{n}{cM}$. Finally, note that by Assumption 3.11

$$r_F(\Sigma) = \frac{1}{\lambda_{F+1}(\Sigma)}\sum_{j=F+1}^{d}\lambda_j(\Sigma) = d - F\ .$$

Hence,

$$r_F(\Sigma) \geq a \cdot \frac{n}{M} \Longleftrightarrow \frac{n}{M} \leq \frac{1}{a}(d-F)\ .$$

□

*Proof of Proposition 3.13.* The proof follows directly from Lemma A.7, Lemma A.8, Lemma 3.1 and Theorem 3.5. Hence, if $\log(2/\delta) \leq c_1^2 \frac{n}{M}$, we find with probability at least $1 - \delta - 7Me^{-\frac{n}{cM}}$

$$\begin{aligned}\mathbb{E}_{\beta^*,\epsilon}\Big[||\Sigma^{\frac{1}{2}}(\bar{\beta}_M - \beta^*)||^2\Big] &\leq \frac{4\sigma_x}{c_1}\log^{\frac{1}{2}}\left(\frac{2M}{\delta}\right)\frac{2\cdot \mathrm{SNR}}{d}\;F\sqrt{\frac{M}{n}} + 8c\tau^2\left(\frac{F}{n} + \frac{1}{M^2}\frac{n}{d-F}\right)\\ &\leq \tilde{c}\log^{\frac{1}{2}}\left(\frac{4M}{\delta}\right)\left(\frac{\mathrm{SNR}\cdot F}{d}\;\sqrt{\frac{M}{n}} + \frac{F}{n} + \frac{1}{M^2}\frac{n}{d-F}\right),\end{aligned}$$

since $1 \leq \log^{\frac{1}{2}}\left(\frac{4M}{\delta}\right)$ for all $\delta \in (0,1]$. Setting $\tilde{c} = 8\max\{\frac{\sigma_x}{c_1}, c\tau^2\}$ proves our result. □

*Proof of Corollary 3.14.* We need to determine the minimum of the function $h : \mathbb{R}_+ \to \mathbb{R}_+$, given by

$$h(M) = C_1\sqrt{M} + \frac{C_2}{M^2} + C_3\;, \quad C_1 > 0\;, C_2 > 0\;, C_3 > 0\;.$$

A short calculation shows that the optimum is achieved at

$$M_{opt} = \left(\frac{4C_2}{C_1}\right)^{2/5}\;,$$

with value

$$h(M_{opt}) = 5C_2\left(\frac{C_1}{4C_2}\right)^{4/5} + C_3 = 5C_2\frac{1}{M_{opt}^2} + C_3\;.$$

Setting now

$$C_1 := \frac{\mathrm{SNR}\cdot F}{d\sqrt{n}}\;, \quad C_2 := \frac{n}{d-F}\;, \quad C_3 = \frac{F}{n}$$

gives for the optimal number of local nodes

$$M_{opt} = M_n = \left(\frac{4dn^{3/2}}{\mathrm{SNR}\cdot F(d-F)}\right)^{2/5}\;,$$

and by Proposition 3.13

$$\begin{aligned}\mathbb{E}_{\beta^*,\epsilon}\Big[||\Sigma^{\frac{1}{2}}(\bar{\beta}_{M_n} - \beta^*)||^2\Big] &\leq c\log^{\frac{1}{2}}\left(\frac{4M_n}{\delta}\right)\left(5C_2\left(\frac{C_1}{4C_2}\right)^{4/5} + C_3\right)\\ &\leq 5c\log^{\frac{1}{2}}\left(\frac{4M_n}{\delta}\right)\left(\frac{F}{n} + \frac{1}{M_n^2}\frac{n}{d-F}\right)\\ &\leq 2c\log^{\frac{1}{2}}\left(\frac{4M_n}{\delta}\right)\left(\frac{F}{n} + \frac{n}{d-F}\left(\frac{\mathrm{SNR}\cdot F(d-F)}{dn^{3/2}}\right)^{4/5}\right),\end{aligned}$$

with probability at least $1 - \delta - 7M_n e^{-\frac{n}{cM_n}}$. □

### A.4 Proofs of Section 3.4

We first recall a lower bound for the variance in the single machine setting.

**Proposition A.9** (Lemma 10 and Lemma 11 in Bartlett et al. (2020))**.** *Define*

$$\widehat{C} := (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\Sigma\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\;.$$

*There exists a constant $c > 0$ such that for any $0 \leq k \leq n/c$ and any $a > 1$ with probability at least $1 - 10e^{-n/c}$, if $r_k(\Sigma) \geq a\, n$, then*

$$\mathrm{Tr}[\widehat{C}] \geq \frac{1}{ca} \min_{l \leq k} \left( \frac{l}{n} + \frac{a^2 n \sum_{j>l} \lambda_j^2}{(\lambda_{k+1} r_k(\Sigma))^2} \right) .$$

*Moreover, for*

$$k^* := \min\{k : r_k(\Sigma) \geq a\, n\}$$

*and if $k^* < \infty$, then*

$$\min_{l \leq k^*} \left( \frac{l}{a\, n} + \frac{a\, n \sum_{j>l} \lambda_j^2}{(\lambda_{k+1} r_k(\Sigma))^2} \right) = \frac{k^*}{a\, n} + \frac{a\, n}{R_{k^*}(\Sigma)} .$$

*Proof of Theorem 3.17.* From Lemma 3.1 and its proof, in particular by (18), and by Assumption 3.16 we may lower bound the excess risk by the variance and find

$$\mathbb{E}_\varepsilon[||\Sigma^{1/2}(\bar{\beta}_M - \beta^*)||^2] \geq \frac{1}{M^2} \sum_{m=1}^{M} \mathbb{E}_\varepsilon \Big[ \mathrm{Tr} \left[ \Sigma^{1/2} X_m^\dagger \varepsilon_m \otimes \varepsilon_m (X_m^\dagger)^T \Sigma^{1/2} \right] \Big] \tag{19}$$

$$= \frac{1}{M^2} \sum_{m=1}^{M} \mathrm{Tr} \left[ \Sigma^{1/2} X_m^\dagger \mathbb{E}_\varepsilon[\varepsilon_m \otimes \varepsilon_m] (X_m^\dagger)^T \Sigma^{1/2} \right] \tag{20}$$

$$\geq \frac{\sigma^2}{M^2} \sum_{m=1}^{M} \mathrm{Tr} \left[ C_m \right] , \tag{21}$$

where $C_m = (\mathbf{X}_m^\dagger)^T \Sigma \mathbf{X}_m^\dagger$. Recall that by definition of $k^*_{\frac{n}{M}}$ from Definition 3.4 we have $r_{k^*_{\frac{n}{M}}}(\Sigma) \geq a \frac{n}{M}$. Hence, we may apply Proposition A.9 and obtain with probability at least $1 - 10Me^{-\frac{1}{c}\frac{n}{M}}$

$$\begin{aligned}
\mathbb{E}_\varepsilon[||\Sigma^{1/2}(\bar{\beta}_M - \beta^*)||^2] &\geq \frac{\sigma^2}{caM^2} \sum_{m=1}^{M} \left( \frac{M k^*_{\frac{n}{M}}}{n} + \frac{a^2 n}{M} \frac{\sum_{j>k^*} \lambda_j^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2} \right) \\
&= \frac{\sigma^2}{ca} \left( \frac{k^*_{\frac{n}{M}}}{n} + \frac{a^2 n}{M^2} \frac{\sum_{j>k^*} \lambda_j^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2} \right) \\
&\geq c_a \sigma^2 \left( \frac{k^*_{\frac{n}{M}}}{n} + \frac{a^2 n}{M^2} \frac{\sum_{j>k^*} \lambda_j^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2} \right) ,
\end{aligned}$$

where we set $c_a := \frac{1}{ca}$ and use that $a > 1$. □

*Proof of Corollary 3.18.* The proof of Lemma A.5 shows that $r_k(\Sigma) \geq \frac{k+1}{\varepsilon_n}$, for any $k$. A similar calculation gives as upper bound

$$\begin{aligned}
r_k(\Sigma) &= (k+1)^{1+\varepsilon_n} \sum_{j>k} j^{-(1+\varepsilon_n)} \\
&\leq \int_k^\infty t^{-(1+\varepsilon_n)} dt \\
&= \frac{(k+1)^{1+\varepsilon_n}}{\varepsilon_n k^{\varepsilon_n}} \\
&\leq \frac{(2k)^{1+\varepsilon_n}}{\varepsilon_n k^{\varepsilon_n}} \\
&\leq \frac{4k}{\varepsilon_n} ,
\end{aligned}$$

where we use that $1 \leq k$ and $2^{1+\varepsilon_n} \leq 4$, since $\varepsilon_n \leq 1$ for $n$ sufficiently large. In particular,

$$a\frac{n}{M} \leq \frac{k^*+1}{\varepsilon_n} \leq r_{k^*}(\Sigma) \leq \frac{4k^*}{\varepsilon_n} \, . \tag{22}$$

Thus, $k^* \geq \frac{a}{4}\frac{n}{M}\varepsilon_n$.

**High Smoothness $\alpha > 0$.** Moreover, the definition of $M_n$ in Corollary 3.9 gives

$$\frac{k^*_{\frac{n}{M_n}}}{n} \geq \frac{a}{4}\frac{\varepsilon_n}{M_n} = \frac{a}{4C_{\tau,\sigma_x}}\alpha^{-2/3}\left(\frac{1}{n}\right)^{1/3} .$$

Hence, applying Theorem 3.17 gives with probability at least $1 - 10Me^{-\frac{1}{c}\frac{n}{M}}$

$$\begin{aligned}
\mathbb{E}_\varepsilon[||\Sigma^{1/2}(\bar\beta_{M_n} - \beta^*)||^2] &\geq c_a\sigma^2\left(\frac{k^*_{\frac{n}{M_n}}}{n} + \frac{a^2 n}{M_n^2}\frac{\sum_{j>k^*}\lambda_j^2}{(\lambda_{k^*+1}r_{k^*}(\Sigma))^2}\right) \\
&\geq c_a\sigma^2\frac{k^*_{\frac{n}{M_n}}}{n} \\
&\geq \frac{\tilde C_{\tau,\sigma_x,\sigma}}{\alpha^{2/3}}\left(\frac{\varepsilon_n}{n}\right)^{1/3} ,
\end{aligned} \tag{23}$$

with $\tilde C_{\tau,\sigma_x,\sigma} = \frac{ac_a\sigma^2}{4C_{\tau,\sigma_x}}$.

**Low Smoothness $\alpha = 0$.** The result in this regime follows from the same arguments as above by inserting the definition of $M_n$ in Corollary 3.10 in the above equations. Indeed, one easily finds with (22) and (23)

$$\begin{aligned}
\mathbb{E}_\varepsilon[||\Sigma^{1/2}(\bar\beta_{M_n} - \beta^*)||^2] &\geq c_a\sigma^2\frac{k^*_{\frac{n}{M_n}}}{n} \\
&\geq \tilde C_{\tau,\sigma_x,\sigma}\left(\frac{1}{\varepsilon_n n}\right)^{1/3} ,
\end{aligned}$$

for some $\tilde C_{\tau,\sigma_x,\sigma} > 0$. □

**Lemma A.10** (Lower Bound Variance Single Machine)**.** *Define*

$$\widehat C := (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\Sigma\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1} .$$

*Suppose Assumption 3.11 is satisfied and let $q < 1/\sqrt{2}$. Assume further that $n \leq d - F$ and that for $n$ sufficiently large*

$$\frac{\rho_2}{\rho_1}(d-F) \leq n\left(\frac{1}{\sqrt{q}} - 2\right) . \tag{24}$$

*There is a constant $c_1 > 0$ such that for any $0 \leq F \leq n/c_1$, with probability at least $1 - 10e^{-n/c_1}$*

$$\mathrm{Tr}\Big[\widehat C\Big] \geq \frac{\min\{q, 1/9\}}{c_1}\left(\frac{F}{n} + \frac{n}{d-F}\right) .$$

*Proof of Lemma A.10.* The proof of Lemma 10 in Bartlett et al. (2020) in conjunction with our Assumption 3.11 show that with probability at least $1 - 10e^{-n/c_1}$

$$\begin{aligned}
\mathrm{Tr}\Big[\widehat C\Big] &\geq \frac{1}{c_1 n}\sum_{i=1}^{d}\left(1 + \frac{1}{n}\sum_{j=F+1}^{d}\frac{\lambda_j}{\lambda_i} + \frac{\lambda_{F+1}}{\lambda_i}\right)^{-2} \\
&= \frac{1}{c_1 n}\sum_{i=1}^{F}\left(1 + \frac{1}{n}\sum_{j=F+1}^{d}\frac{\lambda_j}{\lambda_i} + \frac{\lambda_{F+1}}{\lambda_i}\right)^{-2} + \frac{1}{c_1 n}\sum_{i=F+1}^{d}\left(1 + \frac{1}{n}\sum_{j=F+1}^{d}\frac{\lambda_j}{\lambda_i} + \frac{\lambda_{F+1}}{\lambda_i}\right)^{-2} && (25) \\
&= \frac{F}{c_1 n}\left(1 + \frac{d-F}{n}\frac{\rho_2}{\rho_1} + \frac{\rho_2}{\rho_1}\right)^{-2} + \frac{d-F}{c_1 n}\left(2 + \frac{d-F}{n}\right)^{-2} . && (26)
\end{aligned}$$

We lower bound the first term in (25). To this end, we let $q < 1/\sqrt{2}$ and we show that

$$\left(1 + \frac{d-F}{n}\frac{\rho_2}{\rho_1} + \frac{\rho_2}{\rho_1}\right)^{-2} \geq q \; , \tag{27}$$

provided

$$\frac{\rho_2}{\rho_1}(d-F) \leq n\left(\frac{1}{\sqrt{q}} - 2\right) ,$$

for any $n$ sufficiently large. Indeed, since $\frac{\rho_2}{\rho_1} < 1$, this assumption implies

$$\frac{d-F}{n} \leq \frac{\rho_1}{\rho_2}\left(\frac{1}{\sqrt{q}} - 2\right) \leq \frac{\rho_1}{\rho_2}\left(\frac{1}{\sqrt{q}} - 1 - \frac{\rho_2}{\rho_1}\right) ,$$

being equivalent to (27). Hence,

$$\frac{F}{c_1 n}\left(1 + \frac{d-F}{n}\frac{\rho_2}{\rho_1} + \frac{\rho_2}{\rho_1}\right)^{-2} \geq \frac{q}{c_1}\;\frac{F}{n} \; . \tag{28}$$

To lower bound the second term in (25), recall that we assume $n \leq d - F$, implying $2n + (d-F) \leq 3(d-F)$. Hence,

$$\frac{1}{\left(2 + \frac{d-F}{n}\right)^2} = \frac{n^2}{\left(2n + (d-F)\right)^2} \geq \frac{n^2}{9(d-F)^2} \; . \tag{29}$$

Thus,

$$\frac{d-F}{c_1 n}\left(2 + \frac{d-F}{n}\right)^{-2} \geq \frac{d-F}{c_1 n}\;\frac{n^2}{9(d-F)^2} = \frac{1}{9c_1}\;\frac{n}{d-F} \; . \tag{30}$$

Combining (28) with (30) and (25) finally gives with probability at least $1 - 10e^{-n/c_1}$

$$\mathrm{Tr}\left[\widehat{C}\right] \geq \frac{q}{c_1}\;\frac{F}{n} + \frac{1}{9c_1}\;\frac{n}{d-F} \geq \frac{\min\{q, 1/9\}}{c_1}\left(\frac{F}{n} + \frac{n}{d-F}\right) .$$

□

**Corollary A.11** (Optimal rate finite dimension)**.** *Recall the strong-weak-features model from Section 3.3 and suppose Assumption 3.11 is satisfied. There is a constant* $c_1 > 0$ *such that for any* $0 \leq F \leq \frac{n}{Mc_1} \leq d - F$*, with probability at least* $1 - 10e^{-\frac{n}{Mc_1}}$

$$\mathbb{E}_\varepsilon[||\Sigma^{1/2}(\bar\beta_M - \beta^*)||^2] \geq \sigma^2 c_3 \left(\frac{F}{n} + \frac{n}{M^2(d-F)}\right) ,$$

*for some* $c_3 > 0$*. Moreover, under the assumptions of Corollary 3.14, this lower bound matches the upper bound for the optimal* $M_n$ *and hence is optimal (up to a log-factor).*

*Proof of Corollary A.11.* We apply Lemma A.10 to the local variances and need to ensure, that all assumptions are satisfied. Note that the conditions $\rho_2 d \leq F \leq \frac{n}{Mc_1}$ and $\rho_1 = 1$ imply

$$\rho_2(d-F) \leq \rho_2 d \leq \frac{n}{Mc_1}$$

and hence (24) is satisfied for some well chosen $q$. Applying (19) and Lemma A.10 shows with probability at least $1 - 10e^{-n/Mc_1}$

$$\begin{aligned}\mathbb{E}_\varepsilon[||\Sigma^{1/2}(\bar\beta_M - \beta^*)||^2] &\geq \frac{\sigma^2}{M^2}\sum_{m=1}^M \mathrm{Tr}\left[(\mathbf{X}_m^\dagger)^T \Sigma \mathbf{X}_m^\dagger\right] \\ &\geq \frac{\min\{q, 1/9\}}{c_1}\;\frac{\sigma^2}{M^2}\sum_{m=1}^M\left(\frac{MF}{n} + \frac{n}{M(d-F)}\right) \\ &= \frac{\sigma^2\min\{q, 1/9\}}{c_1}\left(\frac{F}{n} + \frac{n}{M^2(d-F)}\right) ,\end{aligned}$$

for some $q < 1/\sqrt{2}$. □

# B ADDITIONAL RESULTS IN FINITE DIMENSION

In this section we collect some additional results. We first analyze the finite dimensional setting under a more general *source condition* and investigate the impact of the hardness of the problem on the number of optimal machines. In addition, we give a general lower bound in finite dimension under general distributional assumptions.

## B.1 General source condition in the strong-weak-features model

In this section we analyze the setting of Section 3.3 under a more general prior assumption. Here, the covariance of $\beta^*$ will have a specific structure, described by a *source function* $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$, with $t \mapsto t\Phi(t)$ non-decreasing.

**Assumption B.1** (Source Condition)**.** *Assume that* $\beta^* \sim \mathcal{N}(0, \frac{R^2}{d}\Phi(\Sigma))$*, for some* $R > 0$*. Note that* $R^2$ *can be intepreted as the* expected signal strength*.*

**Lemma B.2.** *Suppose Assumption 3.11 is satisfied. Let* $\rho_2\Phi(\rho_2)d \leq F$*. Then*

$$\mathrm{Tr}[\Sigma\Phi(\Sigma))] \ \leq \ (\rho_1\Phi(\rho_1) + 1)F \ .$$

*Proof of Lemma B.2.* We write

$$\begin{aligned}
\mathrm{Tr}[\Sigma\Phi(\Sigma))] &= \sum_{j=1}^{F}\rho_1\Phi(\rho_1) + \sum_{j=F+1}^{d}\rho_2\Phi(\rho_2) \\
&= (\rho_1\Phi(\rho_1) - \rho_2\Phi(\rho_2))F + \rho_2\Phi(\rho_2)d \\
&\leq (\rho_1\Phi(\rho_1) - \rho_2\Phi(\rho_2) + 1)F \\
&\leq (\rho_1\Phi(\rho_1) + 1)F \ .
\end{aligned}$$

□

**Proposition B.3.** *In addition to all assumptions of Theorem 3.5, suppose that Assumptions B.1 and 3.11 are satisfied. Assume that* $\rho_2\Phi(\rho_2)d \leq F$*. For any* $\delta \geq 2e^{-c_1^2\frac{n}{M}}$*, with probability at least* $1 - \delta - 7Me^{-\frac{n}{c_2 M}}$ *we have*

$$\mathbb{E}_{\beta^*,\epsilon}\left[||\Sigma^{\frac{1}{2}}(\bar{\beta}_M - \beta^*)||^2\right] \leq c_3 C_{\rho_1}\log^{\frac{1}{2}}\left(\frac{2M}{\delta}\right)\frac{R^2F}{d}\sqrt{\frac{M}{n}} + c_4\ \Delta^{-1}(\rho_1,\rho_2)\ \frac{n}{M^2}\ \frac{1}{F} \ ,$$

*where* $C_{\rho_1} = \rho_1\Phi(\rho_1) + 1$ *and* $\Delta(\rho_1, \rho_2) := (\rho_1 - \rho_2)^2$ *and for some* $c_1, c_2, c_3, c_4 > 0$*.*

*Proof of Proposition B.3.* We combine Theorem 3.5, Lemma A.8 and Lemma B.2. This gives with probability at least $1 - \delta - 7Me^{-\frac{n}{c_2 M}}$

$$\begin{aligned}
\mathbb{E}_{\beta^*,\epsilon}\left[||\Sigma^{\frac{1}{2}}(\bar{\beta}_M - \beta^*)||^2\right] &\leq \frac{4\sigma_x}{c_1}\log^{\frac{1}{2}}\left(\frac{2M}{\delta}\right)\mathrm{Tr}[\Sigma\Theta]\sqrt{\frac{M}{n}} + 16c_2\tau^2\ \frac{1}{(\rho_1-\rho_2)^2}\ \frac{n}{M^2}\ \frac{1}{F} \\
&\leq \frac{4\sigma_x}{c_1}(\rho_1\Phi(\rho_1) + 1)\frac{R^2F}{d}\log^{\frac{1}{2}}\left(\frac{2M}{\delta}\right)\sqrt{\frac{M}{n}} + 16c_2\tau^2\ \frac{1}{(\rho_1-\rho_2)^2}\ \frac{n}{M^2}\ \frac{1}{F} \ .
\end{aligned}$$

The results follows by setting $C_{\rho_1} := \rho_1\Phi(\rho_1)+1$, $\Delta(\rho_1, \rho_2) := \Delta := (\rho_1-\rho_2)^2$, $c_3 := 4\sigma_x/c_1$ and $c_4 := 16c_2\tau^2$. □

**Corollary B.4** (Optimal number of machines)**.** *Suppose all assumptions of Proposition B.3 are satisfied. Let* $(\rho_{2,n})_n$ *be decreasing and* $\rho_{2,n}\Phi(\rho_{2,n})d_n \leq F_n$*. Denote* $\Delta_n := (\rho_1 - \rho_{2,n})^2$ *and assume*

$$n^{-3/2} \lesssim \frac{d_n}{\Delta_n F_n^2} \lesssim n \ . \tag{31}$$

*The optimal number of local nodes $M_n$ is given by*

$$M_n = A \cdot \left( \frac{d_n n^{3/2}}{R^2 \Delta_n \cdot F_n^2} \right)^{2/5} , \tag{32}$$

*where $A = \left( \frac{4c_4}{c_3 C_{\rho_1}} \right)^{2/5}$. The excess risk satisfies with probability at least $1 - \delta - 7M_n e^{-\frac{n}{c_2 M_n}}$*

$$\mathbb{E}_{\beta^*, \epsilon}\Big[ ||\Sigma^{\frac{1}{2}}(\bar{\beta}_{M_n} - \beta^*)||^2 \Big] \leq c' \log^{\frac{1}{2}} \left( \frac{2M_n}{\delta} \right) \left( \frac{R^2 F_n}{d_n} \right)^{4/5} \left( \frac{1}{F_n \cdot n\Delta_n} \right)^{1/5} , \tag{33}$$

*where $c' = (5c_4)/A^2$.*

*Proof of Corollary B.4.* We need to determine the minimum of the function $h : \mathbb{R}_+ \to \mathbb{R}_+$, given by

$$h(M) = C_1 \sqrt{M} + \frac{C_2}{M^2} , \quad C_1 > 0 \,, C_2 > 0 \,.$$

A short calculation shows that the optimum is achieved at

$$M_{opt} = \left( \frac{4C_2}{C_1} \right)^{2/5} ,$$

with value

$$h(M_{opt}) = 5C_2 \left( \frac{C_1}{4C_2} \right)^{4/5} = 5C_2 \frac{1}{M_{opt}^2} \,.$$

Setting now

$$C_1 := c_3 \; C_{\rho_1} \frac{R^2 F}{d\sqrt{n}} , \quad C_2 := c_4 \; \frac{n}{\Delta_n \cdot F}$$

gives for the optimal number of local nodes

$$M_{opt} = M_n = A \cdot \left( \frac{dn^{3/2}}{R^2 \Delta_n \cdot F^2} \right)^{2/5} , \quad A := \left( \frac{4c_4}{c_3 C_{\rho_1}} \right)^{2/5} ,$$

and

$$\begin{aligned} \mathbb{E}_{\beta^*, \epsilon}\Big[ ||\Sigma^{\frac{1}{2}}(\bar{\beta}_{M_n} - \beta^*)||^2 \Big] &\leq 5c_4 \log^{\frac{1}{2}} \left( \frac{2M_n}{\delta} \right) \frac{n}{F \cdot \Delta_n M_n^2} \\ &= c' \log^{\frac{1}{2}} \left( \frac{2M_n}{\delta} \right) \left( \frac{R^2 F}{d} \right)^{4/5} \left( \frac{1}{F \cdot n\Delta_n} \right)^{1/5} , \end{aligned}$$

where $c' = (5c_4)/A^2$. Moreover, for our bounds to be meaningful we have to require that $1 \lesssim M_n \lesssim n$. This is satisfied if

$$n^{-3/2} \lesssim \frac{d_n}{\Delta_n F_n^2} \lesssim n \,.$$

□

The two conditions

1. $n^{-3/2} \lesssim \frac{d_n}{\Delta_n F_n^2} \lesssim n$
2. $\rho_{2,n} \Phi(\rho_{2,n}) d_n \leq F_n$

from Corollary B.4 determine the number of optimal splits and the learning rate of the distributed minimum norm interpolant. In particular, the a-priori assumption on $\beta^*$ through the source function $\Phi$ has an influence on the possible number of splits and hence on the efficiency of averaging. We discuss three special examples in more detail below. In all cases, we exclusively focus on the overparameterized regime where $n \lesssim d_n$ and $1 \lesssim F_n \lesssim d_n$. Suppose that

$$d_n \simeq n^\gamma \,, \quad \gamma > 1 \;\text{ and }\; F_n \simeq n^\delta \,, \quad 0 \leq \delta \leq \gamma \,.$$

Condition $(II)$ from above sets now restrictions on the decay of the strength of the weak features.

**Easy Case.** We let $\Phi(t) = t$. Condition $(II)$ can be rewritten as $\rho_{2,n} \lesssim \left(\frac{1}{n}\right)^{\frac{1}{2}(\gamma-\delta)}$. To meet condition $(I)$ we need to distinguish two cases:

1. If $\gamma \leq 2\delta$, we have $\max\{1, 2\delta - 3/2\} < \gamma \leq 2\delta$ and $\delta > 1/2$. In particular, the number of strong features needs to grow at as $F_n \gtrsim \sqrt{n}$.

2. If $\gamma \geq 2\delta$, we have $\max\{1, 2\delta\} < \gamma \leq 2\delta + 1$ and $\delta < \gamma/2$. Here, the number of strong features can not grow faster that $n^{\gamma/2}$.

**Isotropic Case.** We let $\Phi(t) = 1$. Condition $(II)$ can be rewritten as $\rho_{2,n} \lesssim \left(\frac{1}{n}\right)^{\gamma-\delta}$. Compared to the *easy case*, the strength of the weak features $\rho_{2,n}$ needs to decay faster. Condition $(I)$ holds under the same assumptions as in the *easy case.*

**Hard Case.** We let $\Phi(t) = t^{-1}$. Condition $(II)$ reduces to $F_n \simeq d_n$, i.e., the number of strong features needs to grow as fast as the dimension. In this case, the optimal number of machines scales as $M_n \simeq n^{\frac{2}{5}(\frac{3}{2}-\delta)}$. To ensure $1 \lesssim M_n \lesssim n$, the growth of $d_n$ can not be too fast: $\gamma = \delta \in (1, 3/2]$.

### B.2 A universal lower bound

We aim at deriving a lower bound for the distributed ridgeless regression estimator under fairly general distributional assumptions if $dim(\mathcal{H}) = d < \infty$.

**Assumption B.5.** *1. The input $x \in \mathbb{R}^d$ is strongly square integrable: $\mathbb{E}[||x||^2] < \infty$.*

*2. The covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ is invertible.*

*3. $\mathbb{E}[y^2] < \infty$.*

*4. The conditional variance is bounded from below: For some $\tilde{\tau} \geq 0$ we assume $\mathbb{V}[y|x] \geq \tilde{\tau}^2$ almost surely.*

*5. For any $m = 1, ..., M$, the local data matrix $X_m \in \mathbb{R}^{\frac{n}{M}\times d}$ has almost surely full rank, i.e.,* $\text{rank}[X_m] = \min\{\frac{n}{M}, d\}$.

Under these assumptions we have the following lower bound for the ridgeless distributed estimator in finite dimension.

**Theorem B.6** (Lower bound)**.** *Let $\bar{\beta}_M$ be defined by* (3)*. The excess risk satisfies*

$$\mathbb{E}\Big[||\Sigma^{1/2}(\bar{\beta}_M - \beta^*)||^2\Big] \geq \frac{\tilde{\tau}^2}{M}\,\frac{\min\{d, \frac{n}{M}\}}{\max\{d, \frac{n}{M}\} + 1 - \min\{d, \frac{n}{M}\}}\;.$$

Thus, we observe peaks at $d = \frac{n}{M}$ with height at least $\tilde{\tau}^2 \frac{d}{M}$, see Fig. 2.

We consider functions of the form $f_\beta : \mathcal{H} \to \mathbb{R}$, $\beta \in \mathcal{H}$, with $f_\beta(x) := \langle \beta, x\rangle$ and define for any estimator $\hat{\beta} \in \mathcal{H}$ the quantity

$$\tilde{\mathcal{E}} := \mathbb{E}[(f_{\hat{\beta}}(x) - \mathbb{E}_{Y|X}[f_{\hat{\beta}}(x)])^2]\;.$$

One easily verifies that

$$\tilde{\mathcal{E}} \leq \mathbb{E}[\mathcal{R}(\hat{\beta})] - \mathcal{R}(\beta^*)\;. \tag{34}$$

Thus, finding a lower bound for $\tilde{\mathcal{E}}$ leads to a lower bound for the excess risk.

*Proof of Theorem B.6.* Define the centered output variables $\tilde{Y}_m := Y_m - \mathbb{E}_{Y_m|X}[Y_m]$, $m = 1, ..., M$ and set

$$Cov(Y_m, X_m) := \mathbb{E}_{Y_m|X_m}[\tilde{Y}_m \otimes \tilde{Y}_m]\;.$$

We then write

$$\begin{aligned}
\tilde{\mathcal{E}}(X) &:= \mathbb{E}_{Y,x|X}[(f_{\bar{\beta}_0}(x) - \mathbb{E}_{Y|X}[f_{\bar{\beta}_0}(x)])^2] \\
&= \mathbb{E}_{Y,x|X}\left[\left(\left\langle x, \frac{1}{M}\sum_{m=1}^{M} X_m^{\dagger}\tilde{Y}_m\right\rangle\right)^2\right] \\
&= \mathbb{E}_{Y,x|X}\left[\left(\frac{1}{M}\sum_{m=1}^{M}\left\langle x, X_m^{\dagger}\tilde{Y}_m\right\rangle\right)^2\right] \\
&= \frac{1}{M^2}\sum_{m=1}^{M}\sum_{m'=1}^{M}\mathbb{E}_{Y,x|X}\left[\left\langle x, X_m^{\dagger}\tilde{Y}_m\right\rangle\left\langle x, X_{m'}^{\dagger}\tilde{Y}_{m'}\right\rangle\right] .
\end{aligned}$$

Note that by definition of $\tilde{Y}_m$ and linearity we have

$$\mathbb{E}_{Y,x|X}\left[\left\langle x, X_m^{\dagger}\tilde{Y}_m\right\rangle\right] = 0 .$$

Thus, by independence and Assumption B.5 we find

$$\begin{aligned}
\tilde{\mathcal{E}}(X) &= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{Y,x|X}\left[\left\langle x, X_m^{\dagger}\tilde{Y}_m\right\rangle^2\right] \\
&= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{Y,x|X}\left[\left\langle x, X_m^{\dagger}\tilde{Y}_m\right\rangle^2\right] \\
&= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_x\left[\langle (X_m^{\dagger})^T x, Cov(Y_m, X_m)(X_m^{\dagger})^T x\rangle\right] \\
&\geq \frac{\tilde{\tau}^2}{M}\sum_{m=1}^{M}\mathbb{E}_x\left[||(X_m^{\dagger})^T x||^2\right] \\
&= \frac{\tau^2}{M}\sum_{m=1}^{M} Tr\left[(X_m^{\dagger})^T\mathbb{E}_x[x \otimes x]X_m^{\dagger}\right] \\
&= \frac{\tau^2}{M}\sum_{m=1}^{M} Tr\left[(X_m^{\dagger})^T\Sigma X_m^{\dagger}\right] .
\end{aligned} \tag{35}$$

We proceed by introducing the whitened data matrices

$$W_m := X_m\Sigma^{-1/2} .$$

We then distinguish the two cases:

$d \geq b = \frac{n}{M}$: Following the arguments in Holzmüller (2020) (Proof of Theorem 3) shows that

$$\mathbb{E}_X\left[Tr\left[(X_m^{\dagger})^T\Sigma X_m^{\dagger}\right]\right] \geq \mathbb{E}_X\left[Tr\left[(W_m W_m^T)^{-1}\right]\right] \geq \frac{b}{d+1-b} .$$

Combining this with (35) gives by independence

$$\begin{aligned}
\tilde{\mathcal{E}} &= \mathbb{E}_X[\tilde{\mathcal{E}}(X)] \\
&\geq \frac{\tilde{\tau}^2}{M}\frac{b}{d+1-b} \\
&= \frac{\tilde{\tau}^2}{M}\frac{b}{d+1-b} .
\end{aligned}$$

The result follows from (34).

$d \leq b = \frac{n}{M}$: A short calculation shows that

$$Tr\left[(X_m^\dagger)^T \Sigma X_m^\dagger\right] = Tr\left[(W_m^T W_m)^{-1}\right] .$$

Following again Holzmüller (2020) (Proof of Theorem 3) we readily obtain

$$\mathbb{E}_X\left[Tr\left[(X_m^\dagger)^T \Sigma X_m^\dagger\right]\right] = \mathbb{E}_X\left[Tr\left[(W_m^T W_m)^{-1}\right]\right] \geq \frac{d}{b+1-d} .$$

We conclude as above to obtain the result. □

## C ADDITIONAL NUMERICAL RESULTS

### C.1 Comparison with Ordinary Least Squares (underparameterized) and Ridge Regression

We complete our discussion of Section 4 by illustrating the efficiency for OLS in the underparameterized case ($d < n$) and for Ridge Regression, both under Gaussian design.

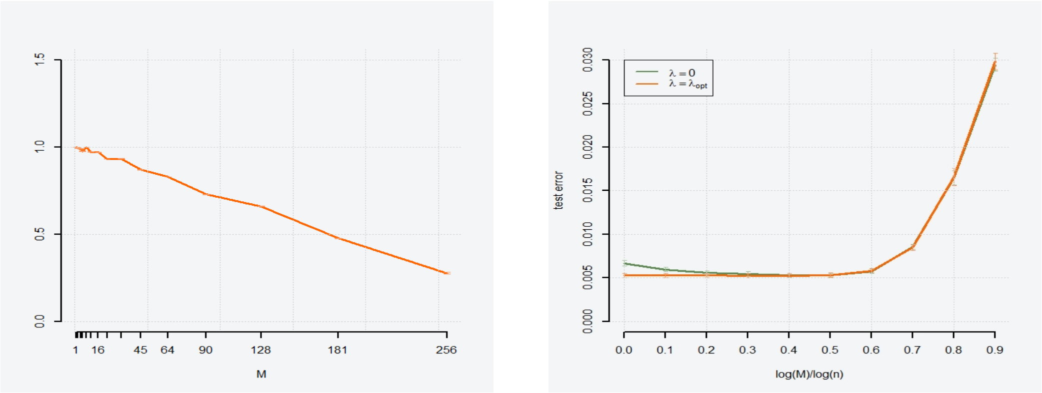

Figure 4: **Left:** Linear loss in efficiency for OLS with $d = 10$, $n = 8000$. **Right:** Comparison of Ridge Regression with optimal regularization with no regularization. We observe a constant accuracy for optimally regularized RR until the number of machines gets too large.

### C.2 Additional Experiment with Simulated Data

In a final experiment we investigate the effect of decay of the eigenvalues on the (normalized) relative prediction efficiency, defined in Definition 3.20. We generate $n = 200$ i.i.d. training points $x_j \sim \mathcal{N}(0, \Sigma)$, with $d = 400$, $\lambda_j(\Sigma) = j^{-(1+\varepsilon)}$, with $\varepsilon = 0.1, 0.5, 1, 1.5$. The target $\beta^*$ is simulated according to Assumption 3.12 with SNR $= 1$. As expected from our main results, faster decay (larger $\varepsilon$) allows larger parallelization, that is, the optimal number of splits (largest efficiency) increases with faster decay.

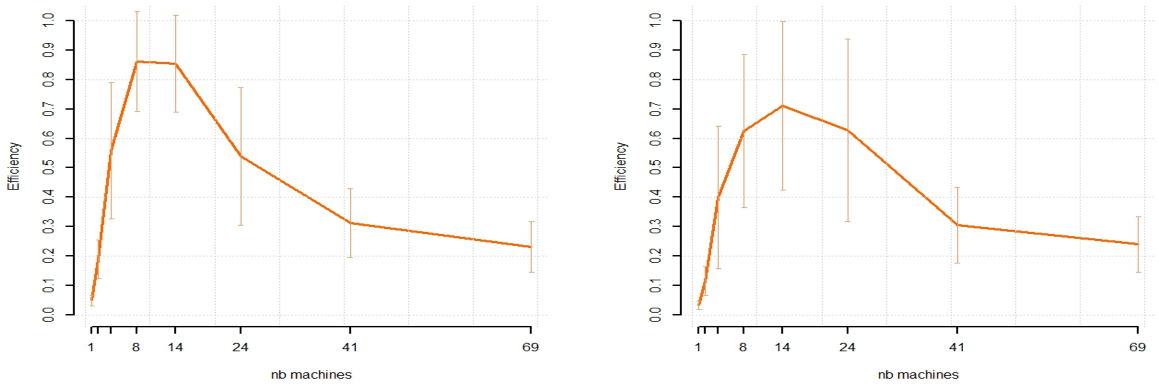

Figure 5: **Left:** $\varepsilon = 0.1$ **Right:** $\varepsilon = 0.5$.

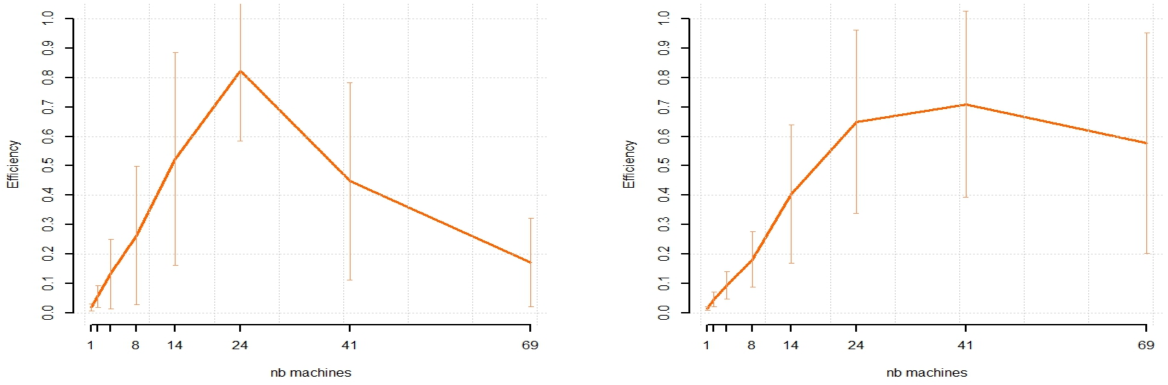

Figure 6: **Left:** $\varepsilon = 1$ **Right:** $\varepsilon = 1.5$.