Journal of Machine Learning Research 21 (2020) 1-52 Submitted 4/19; Revised 3/20; Published 4/20

WONDER: Weighted One-shot Distributed Ridge Regression in High Dimensions

Edgar Dobriban DOBRIBAN@WHARTON.UPENN.EDU
Wharton Statistics Department
University of Pennsylvania
Philadelphia, PA 19104, USA

Yue Sheng YUESHENG@SAS.UPENN.EDU
Graduate Group in Applied Mathematics and Computational Science
University of Pennsylvania
Philadelphia, PA 19104, USA

Editor: John Shawe-Taylor

Abstract

In many areas, practitioners need to analyze large data sets that challenge conventional single-machine computing. To scale up data analysis, distributed and parallel computing approaches are increasingly needed. Here we study a fundamental and highly important problem in this area: How to do ridge regression in a distributed computing environment? Ridge regression is an extremely popular method for supervised learning, and has several optimality properties, thus it is important to study. We study one-shot methods that construct weighted combinations of ridge regression estimators computed on each machine. By analyzing the mean squared error in a high-dimensional random-effects model where each predictor has a small effect, we discover several new phenomena.

Infinite-worker limit: The distributed estimator works well for very large numbers of machines, a phenomenon we call "infinite-worker limit".

Optimal weights: The optimal weights for combining local estimators sum to more than unity, due to the downward bias of ridge. Thus, all averaging methods are suboptimal.

We also propose a new Weighted ONe-shot DistributEd Ridge regression algorithm (WONDER). We test WONDER in simulation studies and using the Million Song Dataset as an example. There it can save at least 100x in computation time, while nearly preserving test accuracy.

Keywords: distributed learning, ridge regression, high-dimensional statistics, random matrix theory

1. Introduction

Computers have changed all aspects of our world. Importantly, computing has made data analysis more convenient than ever before. However, computers also pose limitations and challenges for data science. For instance, hardware architecture is based on a model of a universal computer—a Turing machine—but in fact has physical limitations of storage, memory, processing speed, and communication bandwidth over a network. As large data sets become more and more common in all areas of human activity, we need to think carefully about working with these limitations.

How can we design methods for data analysis (statistics and machine learning) that scale to large data sets? A general approach is *distributed and parallel computing*. Roughly speaking, the data is divided up among computing units, which perform most of the computation locally, and synchronize by passing relatively short messages. While the idea is simple, a good implementation can be hard and nontrivial. Moreover, different problems have different inherent needs in terms of local computation and global communication resources. For instance, in statistical problems with high levels of noise, simple one-shot schemes like averaging estimators computed on local data sets can sometimes work well.

In this paper, we study a fundamental problem in this area. We are interested in linear regression, which is arguably one of the most important problems in statistics and machine learning. A popular method for this model is *ridge regression* (aka Tikhonov regularization), which regularizes the estimates using a quadratic penalty to improve estimation and prediction accuracy. We aim to understand how to do ridge regression in a distributed computing environment. We are also interested in the important *high-dimensional* setting, where the number of features can be very large. In fact our approach allows the dimension and sample size to have any ratio. We also work in a random-effects model where each predictor has a small effect on the outcome, which is the model for which ridge regression is best suited.

We consider the simplest and most fundamental method, which performs ridge regression locally on each data set housed on the individual machines or other computing units, sends the estimates to a global datacenter (or parameter server), and then constructs a final one-shot estimator by taking a linear combination of the local estimates. As mentioned, such methods are sometimes near-optimal, and it is therefore well-justified to study them. We will later give several additional justifications for our work.

However, in contrast to existing work, we introduce a completely new mathematical approach to the problem, which has never been used for studying distributed ridge regression before. Specifically, we leverage and further develop sophisticated recent techniques from random matrix theory and free probability theory in our analysis. This enables us to make important contributions, that were simply unattainable using more "traditional" mathematical approaches.

To give a sense of our results, we provide a brief discussion here. We have a data set consisting of n datapoints, for instance 1000 heart disease patients. Each datapoint has an outcome y_j, such as blood pressure, and features x_j, such as age, height, electronic health records, lab results, and genetic variables. Our goal is to predict the outcome of interest (i.e., blood pressure) for new patients based on their features, and to estimate the relationship of the outcome to the features.

The samples are distributed across several sites, for instance patients from different countries are housed in different data centers. We will refer to the sites as "machines", though they may actually be other computing entities, such as entire computer networks or data centers. In many important settings, it can be impossible to share the data across the different sites, for instance due to logistical or privacy reasons.

Therefore, we assume that each site has a subset of the samples. Our approach is to train ridge regression on this local data. As usual, we can arrange the local data set (say on the i-th machine) into a feature matrix X_i, where each row contains a sample (i.e., datapoint), and an outcome vector Y_i where each entry is an outcome. We compute the

local ridge regression estimates

$$\hat{\beta}_i = (X_i^\top X_i + \lambda_i I_p)^{-1} X_i^\top Y_i,$$

where λ_i are some regularization parameters. We then aggregate them by a weighted combination, constructing the final *one-shot distributed ridge estimator* (where k is the number of sites)

$$\hat{\beta}_{dist} = \sum_{i=1}^{k} w_i \hat{\beta}_i.$$

The important questions here are:

1. How does this work?
2. How to tune the parameters? (such as the regularization parameters and weights)

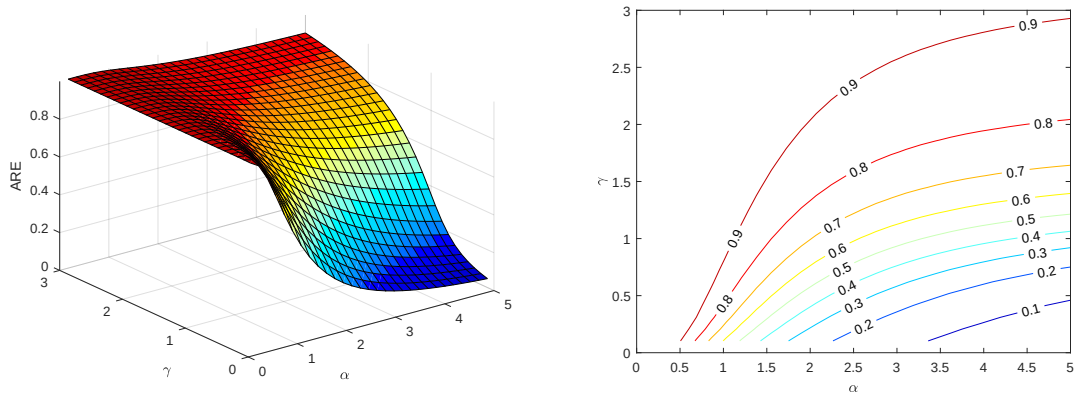
Question 1 is of interest because we wish to know when one-shot methods are a good approach, and when they are not. For this we need to understand the performance as a function of the key problem parameters, such as the signal strength, sample size, and dimension. For question 2, the challenge is posed by the constraints of the distributed computing environment, where standard methods for parameter tuning such as cross-validation may be expensive.

In this work we are able to make several crucial contributions to these questions. We work in an asymptotic setting where n, p grow to infinity at the same rate, which effectively gives good results for any n, p. We study a linear-random effects model, where each regressor has a small random effect on the outcome. This is a good model for the applications where ridge regression is used, because ridge does not assume sparsity, and has optimality properties in certain dense random effects models. Importantly, this analysis does *not* assume any sparsity in a high-dimensional setting. Sparsity has been one of the biggest driving forces in statistics and machine learning in the last 20 years. Our work is in a different line of work, and shows that meaningful results are available without sparsity.

We find the limiting mean squared error of the one-shot distributed ridge estimator. This enables us to characterize the optimal weights and tuning parameters, as well as the *relative efficiency* compared to centralized ridge regression, meaning the ratio of the risk of usual ridge to the distributed estimator. This can precisely pinpoint the computation-accuracy tradeoff achieved via one-shot distributed estimation. See Figure 1 for an illustration.

As a consequence of our detailed and precise risk analysis, we make several qualitative discoveries that we find quite striking:

1. **Efficiency depends strongly on signal strength.** The statistical efficiency of the one-shot distributed ridge estimator depends strongly on signal strength. The efficiency is generally high (meaning distributed ridge regression works well) when the signal strength is low.
2. **Infinite-worker limit.** The one-shot distributed estimator does not lose all efficiency compared to the ridge estimator even in the limit of *infinitely many machines*. Somewhat surprisingly, this suggests that simple one-shot weighted combination methods



(a) Surface

(b) Contour

Figure 1: Efficiency loss due to one-shot distributed learning. This plot shows the relative mean squared error of centralized ridge regression compared to optimally weighted one-shot distributed ridge regression. This quantity is at most unity, and the larger, the "better" distributed ridge works. Specifically, the model is asymptotic, and we show the dependence of the Asymptotic Relative Efficiency (ARE) on the aspect ratio $\gamma = \lim p/n$ (where n is sample size and p is dimension) and on the signal strengh $\alpha = \sqrt{\mathbb{E}\|\beta\|^2}$, in the *infinite-worker limit* when we distribute our data over many machines. We show (a) surface and (b) contour plots of the ARE. See the text for details.

for distributed ridge regression can work well even for very large numbers of machines. The statement that this can be achieved by communication-efficient methods is nontrivial. This finding is clearly important from a practical perspective.

3. **Decoupling.** When the features are uncorrelated, the problem of choosing the optimal regularization parameters *decouples* over the different machines. We can choose them in a locally optimal way, and they are also globally optimal. We emphasize that this is a very delicate result, and is not true in general for correlated features. Moreover, this discovery is also important in practice, because it gives conditions under which we can choose the regularization parameters separately for each machine, thus saving valuable computational resources.

4. **Optimal weights do not sum to unity.** Our work uncovers unexpected properties of the optimal weights. Naively, one may think that the weights need to sum to unity, meaning that we need a weighted average. However, it turns out the optimal weights sum to more than unity, because of the negative bias of the ridge estimator. This means that *any type of averaging method is suboptimal*. We characterize the optimal weights and under certain conditions find their explicit analytic form.

Based on these results, we propose a new Weighted ONe-shot DistributEd Ridge regression algorithm (WONDER). We also confirm these results in detailed simulation studies and on an empirical data example, using the Million Song Dataset. Here WONDER can be used over 100-way splits of the data with 5% loss of prediction accuracy.

We also emphasize that some aspects of our work can help practitioners directly (e.g., our new algorithm), while others are developed for deepening our understanding of the nature of the problem. We discuss the practical implications of our work in Section 4.5.

The paper is structured as follows. We discuss some related work in Section 1.1. We start with finite sample results in Section 2. We provide asymptotic results for features with an arbitrary covariance structure in Section 3. We consider the special case of an identity covariance in Section 4. In Section 5 we provide an explicit algorithm for optimally weighted one-shot distributed ridge. We also study in detail the properties of the estimation error, relative efficiency (including minimax properties in Section 4.6), tuning parameters (and decoupling), as well as optimal weights, including answers to the questions above. We provide numerical simulations throughout the paper, and additional ones in Section 6, along with an example using an empirical data set. The code for our paper is available at `github.com/dobriban/dist_ridge`.

1.1. Related work

Here we discuss some related work. Historically, distributed and parallel computation has first been studied in computer science and optimization (see e.g., Bertsekas and Tsitsiklis, 1989; Lynch, 1996; Blelloch and Maggs, 2010; Boyd et al., 2011; Rauber and Rünger, 2013; Koutris et al., 2018). However, the problems studied there are quite different from the ones that we are interested in. Those works often focus on problems where correct answers are required within numerical precision, e.g., 16 bits of accuracy. However, when we have noisy data sets, such as in statistics and machine learning, numerical precision is neither needed nor usually possible. We may only hope for 3-4 bits of accuracy, and thus the problems are different.

The area of distributed statistics and machine learning has attracted increasing attention only relatively recently, see for instance Mcdonald et al. (2009); Zhang et al. (2012, 2013c); Li et al. (2013); Zhang et al. (2013b,a); Chen and Xie (2014); Mackey et al. (2011); Zhang et al. (2015); Braverman et al. (2016); Jordan et al. (2019); Rosenblatt and Nadler (2016); Smith et al. (2018); Banerjee et al. (2019); Zhao et al. (2016); Xu et al. (2018); Fan et al. (2019); Lin et al. (2017); Lee et al. (2017); Volgushev et al. (2019); Shang and Cheng (2017); Battey et al. (2018); Zhu and Lafferty (2018); Chen et al. (2019, 2018); Wang et al. (2019); Shi et al. (2018); Duan et al. (2018); Liu et al. (2018); Cai and Wei (2020), and the references therein. See Huo and Cao (2018) for a review. We can only discuss the most closely related papers due to space limitations.

Zhang et al. (2013c) study the MSE of averaged estimation in empirical risk minimization. Later Zhang et al. (2015) study divide and conquer *kernel ridge regression*, showing that the partition-based estimator achieves the statistical minimax rate over all estimators, when the number of machines is not too large. These results are very general, however they are not as explicit or precise as our results. In addition they consider fixed dimensions, whereas we study increasing dimensions under random effects models. Lin et al. (2017) improve the above results, removing certain eigenvalue assumptions on the kernel, and sharpening the rate.

Guo et al. (2017) study regularization kernel networks, and propose a debiasing scheme that can improve the behavior of distributed estimators. This work is also in the same

framework as those above (general kernel, fixed dimension). Xu et al. (2018) propose a distributed General Cross-Validation method to choose the regularization parameter.

Rosenblatt and Nadler (2016) consider averaging in distributed learning in fixed and high-dimensional M-estimation, without studying regularization. Lee et al. (2017) study sparse linear regression, showing that averaging debiased lasso estimators can achieve the optimal estimation rate if the number of machines is not too large. A related work is Battey et al. (2018), which also includes hypothesis testing under more general sparse models. These last two works are on a different problem (sparse regression), whereas we study ridge regression in random-effects models.

2. Finite Sample Results

We start our study of distributed ridge regression by a finite sample analysis of estimation error in linear models. Consider the standard linear model

$$Y = X\beta + \varepsilon. \tag{1}$$

Here $Y \in \mathbb{R}^n$ is the n-dimensional continuous outcome vector of n independent samples (e.g., the blood pressure level of n patients, or the amount of time spent on an activity by n internet users), X is the $n \times p$ design matrix containing the values of p features for each sample (e.g., demographical and genetic variables of each patient). Moreover, $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is the p-dimensional vector of unknown regression coefficients.

Our goals are to predict the outcome variable for future samples, and also to estimate the regression coefficients. The outcome vector is affected by the random noise $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$. We assume that the coordinates of ε are independent random variables with mean zero and variance σ^2.

The *ridge regression* (or Tikhonov regularization) estimator is one of the most popular methods for estimation and prediction in linear models. Recall that the ridge estimator of β is

$$\hat{\beta}(\lambda) = (X^\top X + n\lambda I_p)^{-1} X^\top Y,$$

where λ is a tuning parameter. This estimator has many justifications. It shrinks the coefficients of the usual ordinary least squares estimator, which can lead to improved estimation and prediction. When the entries of β and ε are i.i.d. Gaussian, and for suitable λ, it is the posterior mean of β given the outcomes, and hence is a Bayes optimal estimator for any quadratic loss function, including estimation and prediction error.

Suppose now that we are in a distributed computation setting. The samples are distributed across k different sites or machines. For instance, the data of users from a particular country may be stored in a separate datacenter. This may happen due to memory or storage limitations of individual data storage facilities, or may be required by data usage agreements. As mentioned, for simplicity we call the sites "machines".

We can write the partitioned data as

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix}, \; Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_k \end{bmatrix}.$$

Thus the i-th machine contains n_i samples whose features are stored in the $n_i \times p$ matrix X_i and also the corresponding $n_i \times 1$ outcome vector Y_i.

Since the ridge regression estimator is a widely used gold standard method, we would like to understand how we can approximate it in a distributed setting. Specifically, we will focus on one-shot *weighting* methods, where we perform ridge regression locally on each subset of the data, and then aggregate the regression coefficients by a weighted sum. There are several reasons to consider weighting methods:

1. This is a practical method with *minimal communication cost*. When communication is expensive, it is imperative to develop methods that minimize communication cost. In this case, one-shot weighting methods are attractive, and so it is important to understand how they work. In a well-known course on scalable machine learning, Alex Smola calls such methods "idiot-proof" (Smola, 2012), meaning that they are straightforward to implement (unlike some of the more sophisticated methods).

2. Averaging (which is a special case of one-shot weighting) has already been studied in several works on distributed ridge regression (e.g., Zhang et al. (2015); Lin et al. (2017)), and much more broadly in distributed learning, see the related work section for details. Such methods are *known to be rate-optimal* under certain conditions.

3. However, in our setting, we are able to discover several *new phenomena* about one-shot weighting. For instance, we can quantify in a much more nuanced way the accuracy loss compared to centralized ridge regression.

4. Weighting may serve as a useful *initialization to iterative methods*. In practical distributed learning problems, iterative optimization algorithms such as distributed gradient descent or ADMM (Boyd et al., 2011) may be used. However, there are examples where the first step of the iterative method has *worse* performance than a simple averaging (Pourshafeie et al., 2018). Therefore, we can imagine hybrid or warm start methods that use weighting as an initialization. This also suggests that studying one-shot weighting is important.

Therefore, we define *local* ridge estimators for each data set X_i, Y_i, with regularization parameter λ_i as

$$\hat{\beta}_i(\lambda_i) = (X_i^\top X_i + n_i \lambda_i I_p)^{-1} X_i^\top Y_i.$$

We consider combining the local ridge estimators at a central server via a one-step weighted summation. We will find the optimally weighted one-shot distributed estimator

$$\hat{\beta}_{dist}(w) = \sum_{i=1}^{k} w_i \hat{\beta}_i.$$

Note that, unlike ordinary least squares (OLS), the local ridge estimators are always well-defined, i.e. n_i can be smaller than p. Also, for the distributed OLS estimator averaging local OLS solutions, it is natural to require $\sum_i w_i = 1$, because this ensures unbiasedness (Dobriban and Sheng, 2018). However, the ridge estimators are biased, so it is not clear if we should put any constraints on the weights. In fact we will find that the optimal weights

typically do not sum to unity. These features distinguish our work from prior art, and lead to some surprising consequences.

Throughout the paper, we will frequently use the notations $\widehat{\Sigma} = n^{-1}X^\top X$ and $\widehat{\Sigma}_i = n_i^{-1}X_i^\top X_i$. A stepping stone to our analysis is the following key result.

Theorem 1 (Finite sample risk and optimal weights) *Consider the distributed ridge regression problem described above. Suppose we have a data set with n datapoints (samples), each with an outcome and p features. The data set is distributed across k sites. Each site has a subset X_i, Y_i of the data, with the $n_i \times p$ matrix X_i of features of n_i samples, and the corresponding outcomes Y_i. We compute the local ridge regression estimator $\hat{\beta}_i(\lambda_i) = (X_i^\top X_i + n_i\lambda_i I_p)^{-1}X_i^\top Y_i$ with fixed regularization parameters $\lambda_i > 0$ on each data set. We send the local estimates to a central location, and combine them via a weighted sum, i.e., $\hat{\beta}_{dist}(w) = \sum_{i=1}^k w_i\hat{\beta}_i$.*

Under the linear regression model (1)*, the optimal weights that minimize the mean squared error of the distributed estimator are*

$$w^* = (A + R)^{-1}v,$$

where the quantities v, A, R are defined below.

1. *v is a k-dimensional vector with i-th coordinate $\beta^\top Q_i\beta$, where $Q_i = (\widehat{\Sigma}_i + \lambda_i I_p)^{-1}\widehat{\Sigma}_i$ are $p \times p$ matrices.*
2. *A is a $k \times k$ matrix with (i, j)-th entry $\beta^\top Q_i Q_j \beta$.*
3. *R is a $k \times k$ diagonal matrix with i-th diagonal entry $n_i^{-1}\sigma^2 \operatorname{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-2}\widehat{\Sigma}_i]$.*

The mean squared error of the optimally weighted *distributed ridge regression estimator $\hat{\beta}_{dist}$ with k sites equals*

$$\mathrm{MSE}^*_{dist}(k) = \mathbb{E}\|\hat{\beta}_{dist}(w^*) - \beta\|^2 = \|\beta\|^2 - v^\top(A + R)^{-1}v,$$

See Appendix F.1 for the proof. The argument proceeds via a direct calculation, recognizing that *finding the optimal weights for combining the local estimators $\hat{\beta}_i$ can be viewed as a k-parameter regression problem* of β on $\hat{\beta}_i$, for $i = 1, \ldots, k$.

This result quantifies the mean squared error of the optimally weighted distributed ridge estimator *for fixed regularization parameters λ_i*. Later we will study how to choose the regularization parameters optimally. The result also gives an exact formula for the optimal weights. However, the optimal weights depend on the unknown regression coefficients β, and are thus not directly usable in practice. Instead, our approach is to make stronger assumptions on β under which we can develop estimators for the weights.

Computational efficiency. We take a short detour here to discuss computational efficiency. Here by computational efficiency we mean the total time consumption. Computing one ridge regression estimator $(X^\top X + \lambda I_p)^{-1}X^\top Y$ for a fixed regularization parameter λ and $n \times p$ design matrix X can be done in time $O(np\min(n, p))$ by first computing the SVD of X. This automatically gives the ridge estimator for all values of λ.

How much time can we save by distributing the data? Suppose first that $n \geq p$, in which case the total time consumption is $O(np^2)$. Computing ridge locally on the i-th

machine takes $O(n_i p \min(n_i, p))$ time. Suppose next that we distribute equally to k of machines, and we also have $n_i = n/k \geq p$. Then the time consumption is reduced to $O((n/k)p^2) = O(np^2/k)$. In this case we can say that the total time consumption decreases *proportionally to the number of machines*. This shows the benefit of parallel data processing.

On the other extreme, if $n \leq p$, then $n_i = n/k \leq p$, the total time consumption is reduced from $O(n^2 p)$ to $O((n/k)^2 p) = O(n^2 p/k^2)$. This shows that the total time consumption decreases *quadratically* in the number of machines (albeit of course the constant is much worse). If we are in an intermediate case where $n \geq p$ and $n_i = n/k \leq p$, then the time decreases at a rate between linear and quadratic.

2.1. Addressing reader concerns

At this stage, our readers may have several concerns about our approach. We address some concerns in turn below.

1. Does it make sense to average ridge estimators, which can be biased?

   A possible concern is that we are working with biased estimators. Would it make sense to debias them first, before weighting? A similar approach has been used for sparse regression, with the debiased Lasso estimators (Lee et al., 2017; Battey et al., 2018). However, our results *allow the regularization parameters to be arbitrarily close to zero*, which leads to least squares estimators, with an inverse or pseudoinverse $(X_i^\top X_i)^\dagger$. These are the "natural" debiasing estimators for ridge regression. For OLS, these are exactly unbiased, while for pseudoinverse, they are approximately so. Hence our approach allows nearly unbiased estimators, and we automatically discover when this is the optimal method.

2. Is it possible to improve the weighted sum of local ridge estimators $\hat{\beta}_i$ in trivial ways?

   One-shot weighting is merely a heuristic. If it were possible to improve it in a simple way, then it would make sense to study those methods instead of weighting. However, we are not aware of such methods. For instance, one possibility is to try and add the constant vector into the regression on the global parameter server, because this may help reduce the bias. In simulation studies, we have observed that this approach does not usually lead to a perceptible decrease in MSE. Specifically we have found that under the simulation setting common throughout the paper, the MSEs with and without a constant term are close (see Appendix A for details).

3. Asymptotics under Linear Random-effects Models

The finite sample results obtained so far can be hard to interpret, and do not allow us to directly understand the performance of the optimal one-shot distributed estimator. Therefore, we will consider an asymptotic setting that leads to more insightful results.

Recall that our basic linear model is $Y = X\beta + \varepsilon$, where the error ε is random. Next, we also assume that a *random-effects model* holds. We assume β is random—independently of ε—with coordinates that are themselves independent random variables with mean zero and variance $p^{-1}\sigma^2\alpha^2$. Thus, each feature contributes a small random amount to the outcome. Ridge regression is designed to work well in such a setting, and has several optimality

properties in variants of this model. The parameters are now $\theta = (\sigma^2, \alpha^2)$: the *noise level* σ^2 and the *signal-to-noise ratio* α^2 respectively. This parametrization is standard and widely used (e.g. Searle et al. (2009); Dicker and Erdogdu (2017); Dobriban and Wager (2018)).

To get more insight into the performance of ridge regression in a distributed environment, we will take an asymptotic approach. Notice from Theorem 1 that the mean squared error depends on the data only through simple functionals of the sample covariance matrices $\widehat{\Sigma}$ and $\widehat{\Sigma}_i$, such as

$$\beta^\top(\widehat{\Sigma}_i + \lambda_i I_p)^{-1}\widehat{\Sigma}_i\beta, \quad \beta^\top(\widehat{\Sigma}_i + \lambda_i I_p)^{-1}\widehat{\Sigma}_i(\widehat{\Sigma}_j + \lambda_j I_p)^{-1}\widehat{\Sigma}_j\beta, \quad \mathrm{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-2}\widehat{\Sigma}_i].$$

When the coordinates of β are i.i.d., the means of the quadratic functionals become proportional to the *traces* of functions of the sample covariance matrices. This motivates us to adopt models from *asymptotic random matrix theory*, where the asymptotics of such quantities are a central topic.

We begin by introducing some key concepts from random matrix theory (RMT) which will be used in our analysis. We will focus on "Marchenko-Pastur" (MP) type sample covariance matrices, which are fundamental and popular in statistics (see e.g., Bai and Silverstein (2010); Anderson (2003); Paul and Aue (2014); Yao et al. (2015)). A key concept is the spectral distribution, which for a $p \times p$ symmetric matrix A is the distribution F_A that places equal mass on all eigenvalues $\lambda_i(A)$ of Σ. This has cumulative distribution function (CDF) $F_A(x) = p^{-1}\sum_{i=1}^p \mathbf{1}(\lambda_i(A) \leq x)$. A central result in the area is the Marchenko-Pastur theorem, which states that eigenvalue distributions of sample covariance matrices converge (Marchenko and Pastur, 1967; Bai and Silverstein, 2010). We state the required assumptions below:

Assumption 1 *Consider the following conditions:*

1. *The $n \times p$ design matrix X is generated as $X = Z\Sigma^{1/2}$ for an $n \times p$ matrix Z with i.i.d. entries (viewed as coming from an infinite array), satisfying $\mathbb{E}[Z_{ij}] = 0$ and $\mathbb{E}[Z_{ij}^2] = 1$, and a deterministic $p \times p$ positive semidefinite population covariance matrix Σ.*

2. *The sample size n grows to infinity proportionally with the dimension p, i.e. $n, p \to \infty$ and $p/n \to \gamma \in (0, \infty)$.*

3. *The sequence of spectral distributions $F_\Sigma := F_{\Sigma,n,p}$ of $\Sigma := \Sigma_{n,p}$ converges weakly to a limiting distribution H supported on $[0, \infty)$, called the population spectral distribution.*

Then, the Marchenko-Pastur theorem states that with probability 1, the spectral distribution $F_{\widehat{\Sigma}}$ of the sample covariance matrix $\widehat{\Sigma}$ also converges weakly (in distribution) to a limiting distribution $F_\gamma := F_\gamma(H)$ supported on $[0, \infty)$ (Marchenko and Pastur, 1967; Bai and Silverstein, 2010). The limiting distribution is determined uniquely by a fixed-point equation for its *Stieltjes transform*, which is defined for any distribution G supported on $[0, \infty)$ as

$$m_G(z) := \int_0^\infty \frac{1}{t - z} dG(t), \quad z \in \mathbb{C} \setminus \mathbb{R}^+.$$

With this notation, the Stieltjes transform of the spectral measure of $\widehat{\Sigma}$ satisfies

$$m_{\widehat{\Sigma}}(z) = p^{-1}\operatorname{tr}[(\widehat{\Sigma} - zI_p)^{-1}] \to_{a.s.} m_{F_\gamma}(z), \quad z \in \mathbb{C} \setminus \mathbb{R}^+,$$

where $m_{F_\gamma}(z)$ is the Stieltjes transform of F. In addition, we denote by $m'(z)$ the derivative of the Stieltjes transform. Then, it is also known that

$$p^{-1}\operatorname{tr}[(\widehat{\Sigma} - zI_p)^{-2}] \to_{a.s.} m'_{F_\gamma}(z).$$

The results stated above can be expressed in a different, and perhaps slightly more modern language, using *deterministic equivalents* (Serdobolskii, 2007; Hachem et al., 2007; Couillet et al., 2011; Dobriban and Sheng, 2018). For instance, the Marchenko-Pastur law is a consequence of the following result. For any z where it is well-defined, consider the resolvent $(\widehat{\Sigma} - zI_p)^{-1}$. This random matrix is *equivalent* to a deterministic matrix $(x_p\Sigma - zI_p)^{-1}$ for a certain scalar $x_p = x(\Sigma, n, p, z)$, and we write

$$(\widehat{\Sigma} - zI_p)^{-1} \asymp (x_p\Sigma - zI_p)^{-1}.$$

Here two sequences of $n \times n$ matrices A_n, B_n (not necessarily symmetric) of growing dimensions are *equivalent*, and we write

$$A_n \asymp B_n$$

if

$$\lim_{n\to\infty} \operatorname{tr}\left[C_n(A_n - B_n)\right] = 0$$

almost surely, for any sequence C_n of $n \times n$ deterministic matrices (not necessarily symmetric) with bounded trace norm, i.e., such that $\limsup \|C_n\|_{tr} < \infty$ (Dobriban and Sheng, 2018). Informally, any linear combination of the entries of A_n can be approximated by the entries of B_n. This also can be viewed as a kind of *weak convergence* in the matrix space equipped with an inner product (trace). From this, it also follows that the traces of the two matrices are equivalent, from which we can recover the MP law.

In Dobriban and Sheng (2018), we collected some useful properties of the calculus of deterministic equivalents. In this work, we use those properties extensively. We also develop and use a new *differentiation rule for the calculus of deterministic equivalents* (see Appendix B).

We are now ready to study the asymptotics of the risk. We express the limits of interest in *two equivalent forms*, one in terms of *population quantities* (such as the limiting spectral distribution H of Σ), and one in terms of *sample quantities* (such as the limiting spectral distribution F_γ of $\widehat{\Sigma}$). Moreover, we will denote by T a random variable distributed according to H, so that $\mathbb{E}_H[g(T)]$ denotes the mean of $g(T)$ when T is a random variable distributed according to the limit spectral distribution H.

The key to obtaining the results based on population quantities is that the quadratic forms involving β have asymptotic equivalents that only depend on α^2, σ^2, based on the concentration of quadratic forms. Specifically, we have

$$\beta^\top A\beta \approx \frac{1}{p}\sigma^2\alpha^2 \cdot \operatorname{tr}(A)$$

for suitable matrices A (see the proof of Theorem 2 for details). The key to the results based on sample quantities is the MP law and the calculus of deterministic equivalents.

Theorem 2 (Asymptotics for distributed ridge, arbitrary subsample size) *In the linear random-effects model under Assumption 1, suppose in addition that the eigenvalues of Σ are uniformly bounded away from zero and infinity, and that the entries of Z have a finite $(8+\varepsilon)$-th moment for some $\varepsilon > 0$. Suppose moreover that the local sample sizes n_i grow proportionally to p, so that $p/n_i \to \gamma_i > 0$.*

Then the optimal weights for distributed ridge regression, and its mean square error, converge to definite limits. Recall from Theorem 1 that we have the formulas $w^* = (A+R)^{-1}v$ and $MSE^*_{dist} = \|\beta\|^2 - v^\top (A+R)^{-1}v$ for the optimal finite sample weights and risk, and thus it is enough to find the limit of v, A and R. These have the following limits:

1. *With probability one, we have the convergence $v \to V \in \mathbb{R}^k$. The i-th coordinate of the limit V has the following two equivalent forms, in terms of population and sample quantities, respectively:*

$$V_i = \sigma^2\alpha^2 \mathbb{E}_H \frac{x_i T}{x_i T + \lambda_i} = \sigma^2\alpha^2 \left[1 - \lambda_i m_{F_{\gamma_i}}(-\lambda_i)\right].$$

Recall that H is the limiting population spectral distribution of Σ, and T is a random variable distributed according to H. Among the empirical quantities, F_{γ_i} is the limiting empirical spectral distribution of $\widehat{\Sigma}_i$ and $x_i := x_i(H, \lambda_i, \gamma_i) > 0$ is the unique solution of the fixed point equation

$$1 - x_i = \gamma_i \left[1 - \lambda_i \int_0^\infty \frac{dH(t)}{x_i t + \lambda_i}\right] = \gamma_i \left[1 - \mathbb{E}_H \frac{\lambda_i}{x_i T + \lambda_i}\right].$$

It is part of the theorem's claim that there is such an x_i.

2. *With probability one, $A \to \mathcal{A} \in \mathbb{R}^{k\times k}$. For $i \neq j$, the (i,j)-th entry of $\mathcal{A}$ is, in terms of the population spectral distribution H,*

$$\mathcal{A}_{ij} = \sigma^2\alpha^2 \mathbb{E}_H \frac{x_i x_j T^2}{(x_i T + \lambda_i)(x_j T + \lambda_j)}.$$

The i-th diagonal entry of $\mathcal{A}$ is, in terms of population and sample quantities, respectively,

$$\begin{aligned}\mathcal{A}_{ii} &= \sigma^2\alpha^2 \left[1 - \mathbb{E}_H \frac{2\lambda_i x_i T + \lambda_i^2}{(x_i T + \lambda_i)^2} + \frac{\lambda_i^2 \gamma_i x_i \left(\mathbb{E}_H \frac{T}{(x_i T + \lambda_i)^2}\right)^2}{1 + \gamma_i \lambda_i \mathbb{E}_H \frac{T}{(x_i T + \lambda_i)^2}}\right] \\ &= \sigma^2\alpha^2 \left[1 - 2\lambda_i m_{F_{\gamma_i}}(-\lambda_i) + \lambda_i^2 m'_{F_{\gamma_i}}(-\lambda_i)\right].\end{aligned}$$

3. *With probability one, the diagonal matrix R converges, $R \to \mathcal{R} \in \mathbb{R}^{k\times k}$, where of course $\mathcal{R}$ is also diagonal. The i-th diagonal entry of $\mathcal{R}$ is, in terms of population and sample quantities, respectively,*

$$\mathcal{R}_{ii} = \sigma^2 \left[\frac{x_i \mathbb{E}_H \frac{T}{(x_i T + \lambda_i)^2}}{1 + \lambda_i \gamma_i \mathbb{E}_H \frac{T}{(x_i T + \lambda_i)^2}}\right] = \sigma^2 \left[\gamma_i m_{F_{\gamma_i}}(-\lambda_i) - \gamma_i \lambda_i m'_{F_{\gamma_i}}(-\lambda_i)\right].$$

The limiting weights and mean square error are then

$$\mathcal{W}_k^* = (\mathcal{A} + \mathcal{R})^{-1} V$$

and

$$\mathcal{M}_k = \sigma^2 \alpha^2 - V^\top (\mathcal{A} + \mathcal{R})^{-1} V.$$

See Appendix F.2 for the proof. The statement may look complicated, but the formulas simplify considerably in the uncorrelated case $\Sigma = I_p$, on which we will focus later. Moreover, these limiting formulas are also fundamental for developing consistent estimators for the optimal weights. To develop an algorithm for the practically common general covariance case, the following theorem is crucial.

Theorem 3 (Asymptotics for distributed ridge, equal subsample size) *Consider the assumptions and the notations of Theorem 2. We further assume the samples are equally distributed across the local machines, i.e.* $n_1 = n_2 = \cdots = n_k = n/k$ *and* $\gamma_1 = \gamma_2 = \cdots = \gamma_k = k\gamma$. *We use the same tuning parameter* λ *for each local estimator. Then the limiting optimal weights* $\mathcal{W}_k^*$ *and the limiting MSE* $\mathcal{M}_k$ *have the following forms:*

$$\mathcal{W}_k^* = (1, 1, \ldots, 1)^\top \cdot \frac{\sigma^2 \alpha^2 (1 - \lambda m)}{\mathcal{F} + k\mathcal{G}} \quad \text{and} \quad \mathcal{M}_k = \sigma^2 \alpha^2 - \frac{\sigma^4 \alpha^4 (1 - \lambda m)^2 k}{\mathcal{F} + k\mathcal{G}}.$$

Here $\mathcal{F}$ *and* $\mathcal{G}$ *are defined as follows:*

$$\mathcal{F} = \sigma^2 \alpha^2 \frac{k\gamma\lambda^2 (m - \lambda m')^2}{1 - k\gamma + k\gamma\lambda m'} + \sigma^2 k\gamma (m - \lambda m')$$

and

$$\mathcal{G} = \sigma^2 \alpha^2 \left(1 - 2\lambda m + \lambda^2 m' - \frac{k\gamma\lambda^2 (m - \lambda m')^2}{1 - k\gamma + k\gamma\lambda m'}\right)$$

where $m := m_{F_{k\gamma}}(-\lambda)$ *and* $m' := -\frac{dm}{d\lambda}$.

See Appendix F.3 for the proof and an explanation of why we need to assume the samples are uniformly distributed. Based on this theorem, we are able to develop an algorithm which works for arbitrary covariance structures. See Section 5 for the details.

Now we discuss the problem of estimating the optimal weights, which is crucial for developing practical methods. The results in Theorem 3 show that to estimate the weights consistently, if the tuning parameter λ is known, we *only need to estimate* α^2, σ^2 consistently. The reason is that we can use $\mathrm{tr}[(\widehat{\Sigma}_i + \lambda I)^{-1}]/p$ to approximate m, and use $\mathrm{tr}[(\widehat{\Sigma}_i + \lambda I)^{-2}]/p$ to approximate m'.

Estimating these two parameters is a well-known problem, and several approaches have been proposed, for instance restricted maximum likelihood (REML) estimators (Jiang, 1996; Searle et al., 2009; Dicker, 2014; Dicker and Erdogdu, 2016; Jiang et al., 2016), etc. We can use—for instance—results from Dicker and Erdogdu (2017), who showed that the Gaussian MLE is consistent and asymptotically efficient for $\theta = (\sigma^2, \alpha^2)$ even in the non-Gaussian setting of this paper (see Appendix C for a summary).

4. Special Case: Identity Covariance

When the population covariance matrix is the identity, that is $\Sigma = I$, the results simplify considerably. In this case the features are nearly uncorrelated. It is known that the limiting Stieltjes transform $m_{F_\gamma} := m_\gamma$ of $\widehat{\Sigma}$ has the explicit form (Marchenko and Pastur, 1967):

$$m_\gamma(z) = \frac{(z+\gamma-1) + \sqrt{(z+\gamma-1)^2 - 4z\gamma}}{-2z\gamma}. \tag{2}$$

As usual in the area, we use the principal branch of the square root of complex numbers.

4.1. Properties of the estimation error and asymptotic relative efficiency

We can use the closed form expression for the Stieltjes transform to get explicit formulas for the optimal weights. From Theorem 2, we conclude the following simplified result:

Theorem 4 (Asymptotics for isotropic population covariance) *In addition to the assumptions of Theorem 2, suppose that the population covariance matrix $\Sigma = I$. Then the limits of v, A and R have simple explicit forms:*

1. *The i-th coordinate of V is:*

$$V_i = \sigma^2\alpha^2\left[1 - \lambda_i m_{\gamma_i}(-\lambda_i)\right],$$

   where $m_{\gamma_i}(-\lambda_i)$ is the Stieltjes transform given above in equation (2).

2. *The entries of $\mathcal{A}$ are*

$$\mathcal{A}_{ij} = \begin{cases} \sigma^2\alpha^2\left[1 - \lambda_i m_{\gamma_i}(-\lambda_i)\right]\cdot\left[1 - \lambda_j m_{\gamma_j}(-\lambda_j)\right], & \text{for } i \neq j \\ \sigma^2\alpha^2\left[1 - 2\lambda_i m_{\gamma_i}(-\lambda_i) + \lambda_i^2 m'_{\gamma_i}(-\lambda_i)\right], & \text{for } i = j. \end{cases}$$

3. *The i-th diagonal entry of $\mathcal{R}$ is*

$$\mathcal{R}_{ii} = \sigma^2\gamma_i\left[m_{\gamma_i}(-\lambda_i) - \lambda_i m'_{\gamma_i}(-\lambda_i)\right].$$

The limiting optimal weights for combining the local ridge estimators are $\mathcal{W}_k^* = (\mathcal{A}+\mathcal{R})^{-1}V$, and MSE of the optimally weighted distributed estimator is

$$\mathcal{M}_k = \frac{\sigma^2\alpha^2}{1 + \sum_{i=1}^k \frac{V_i^2}{\sigma^2\alpha^2(\mathcal{R}_{ii}+\mathcal{A}_{ii}) - V_i^2}}.$$

See Appendix F.4 for the proof. This theorem shows the surprising fact that the limiting risk *decouples* over the different machines. By this we mean that the limiting risk can be written in a simple form, involving a sum of terms depending on each machine, without any interaction. This seems like a major surprise.

To explain in more detail the decoupling phenomenon, let us study how the local risks are related to the distributed risks. Define $V = V(\gamma, \lambda)$ to be the limiting scalar $V \in \mathbb{R}$ defined above, in the special case $k = 1$. Explicitly, this is the limit of the quantity $\beta^\top Q\beta$,

where $Q = (\widehat{\Sigma} + \lambda I_p)^{-1}\widehat{\Sigma}$, as given in Theorem 1 applied for $k = 1$. Let D be the scalar expression $D(\gamma, \lambda) = \sigma^2\alpha^2(\mathcal{R} + \mathcal{A}) - V$ when $k = 1$. With these notations, the risk $\mathcal{M}_1$ of ridge regression when computed on the entire data set equals

$$\mathcal{M}_1(\gamma, \lambda) = \frac{\sigma^2\alpha^2}{1 + \frac{V(\gamma,\lambda)}{D(\gamma,\lambda)}}.$$

Moreover, the risk of optimally weighted one-shot distributed ridge over k subsets, with arbitrary regularization parameters λ_i, equals

$$\mathcal{M}_k(\gamma_1, \ldots, \gamma_k, \lambda_1, \ldots, \lambda_k) = \frac{\sigma^2\alpha^2}{1 + \sum_{i=1}^k \frac{V_i^2(\gamma_i,\lambda_i)}{D_i(\gamma_i,\lambda_i)}}.$$

Then one can check that we have the following equations connecting the risk computed on the entire data set and the distributed risk:

$$\frac{\sigma^2\alpha^2}{\mathcal{M}_k(\gamma_1, \ldots, \gamma_k, \lambda_1, \ldots, \lambda_k)} - 1 = \sum_{i=1}^k \frac{\sigma^2\alpha^2}{\mathcal{M}_1(\gamma_i, \lambda_i)} - k,$$
$$\mathcal{M}_k(\gamma_1, \ldots, \gamma_k, \lambda_1, \ldots, \lambda_k) = \frac{1}{\sum_{i=1}^k \frac{1}{\mathcal{M}_1(\gamma_i,\lambda_i)} + \frac{1-k}{\sigma^2\alpha^2}}.$$

These equations are precisely what we mean by *decoupling*. The distributed risk can be written as a function of the type $1/(\sum_i 1/x_i + b)$ of the distributed risks. Therefore, there are no "interactions" between the different risk functions. Similar expressions have been obtained for linear regression (Dobriban and Sheng, 2018).

Next, we discuss in more depth why the limiting risk decouples. Mathematically, the key reason is that when $\Sigma = I$, the limit of A_{ij} for $i \neq j$ decouples into a product of two terms. Therefore, the distributed risk function involves a quadratic form with zero *off-diagonal* terms. This is not the case for general population covariance Σ. We provide an explanation via free probability theory in Appendix D.

An important consequence of the decoupling is that *we can optimize the individual risks over the tuning parameters λ_i separately.*

Proposition 5 (Optimal regularization (tuning) parameters) *Under the assumptions of Theorem 4, the optimal regularization (tuning) parameters λ_i that minimize the local MSEs also minimize the distributed risk $\mathcal{M}_k$. They have the form*

$$\lambda_i = \frac{\gamma_i}{\alpha^2}, \quad i = 1, 2, \ldots, k.$$

Moreover, the risk $\mathcal{M}_k$ of distributed ridge regression with optimally tuned regularization parameters is

$$\mathcal{M}_k = \frac{\sigma^2\alpha^2}{1 + \sum_{i=1}^k \left[\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)} - 1\right]},$$

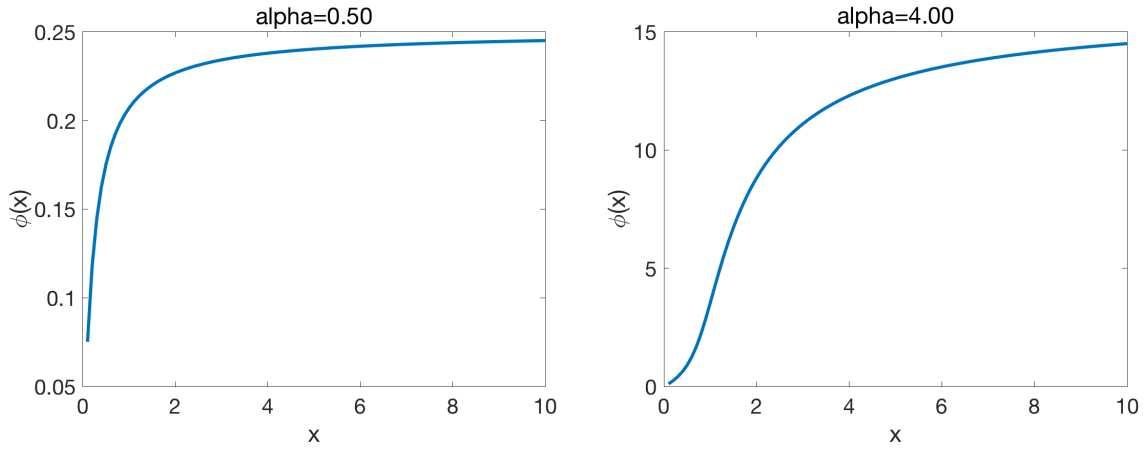


Figure 2: Plots of the optimal risk function ϕ as a function of the aspect ratio γ (denoted by x in the plots), for different signal strength parameters α.

See Appendix F.5 for the proof.

The main goal of our paper is to study the behavior of the one-shot distributed ridge estimator and compare it with the centralized estimator. It is helpful to first understand the properties of the *optimal risk* function $\phi(\gamma) := \gamma m_\gamma(-\gamma/\alpha^2)$. The optimal risk function equals the optimally tuned global risk $\mathcal{M}_1$ up to a factor σ^2. It has the explicit form

$$\phi(\gamma) = \gamma m_\gamma(-\gamma/\alpha^2) = \frac{-\gamma/\alpha^2 + \gamma - 1 + \sqrt{(-\gamma/\alpha^2 + \gamma - 1)^2 + 4\gamma^2/\alpha^2}}{2\gamma/\alpha^2}.$$

Proposition 6 (Properties of the optimal risk function) *The optimal risk function $\phi(\gamma)$ has the following properties:*

1. **Monotonicity:** *$\phi(\gamma)$ is an increasing function of $\gamma \in [0, \infty)$ with $\lim_{\gamma \to 0+} \phi(\gamma) = 0$ and $\lim_{\gamma \to +\infty} \phi(\gamma) = \alpha^2$.*

2. **Concavity:** *When $\alpha \leq 1, \phi(\gamma)$ is a concave function of $\gamma \in [0, \infty)$. When $\alpha > 1$, $\phi(\gamma)$ is convex for small γ (close to 0), and concave for large γ.*

See Appendix F.6 for the proof. See also Figure 2 for plots of ϕ for different α, which show its monotonicity and convexity properties. The aspect ratio γ characterizes the dimensionality of the problem. It makes sense that $\phi(\gamma)$ is increasing, since the regression problem should become more difficult as the dimension increases. For the second property, the concavity of the function means that it grows very fast to approach its limit. When the signal-to-noise ratio α^2 is small, the risk is concave, so it grows fast with the dimension. But when the signal-to-noise ratio becomes large, the risk will grow much slower at the beginning. Here the phase transition happens at $\alpha^2 = 1$. This gives insight into the effect of the signal-to-noise ratio on the regression problem.

To compare the distributed and centralized estimators, we will study their (asymptotic) relative efficiency (ARE), which is the (limit of the) ratio of their mean squared errors. Here we assume each estimator is optimally tuned. This quantity, which is at most unity,

captures the loss of efficiency due to the distributed setting. An ARE close to 1 is "good", while an ARE close to 0 is "bad". From the results above, it follows that the ARE has the form

$$ARE = \frac{\mathcal{M}_1}{\mathcal{M}_k} = \frac{\gamma m_\gamma(-\gamma/\alpha^2)}{\alpha^2}\left[1 + \sum_{i=1}^{k}\left(\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)} - 1\right)\right] \leq 1.$$

We have the following properties of the ARE.

Theorem 7 (Properties of the asymptotic relative efficiency (ARE)) *The asymptotic relative efficiency (ARE) has the following properties:*

1. **Worst case is equally distributed data***: For fixed k, α^2 and γ, the ARE attains its minimum when the samples are* equally distributed *across k machines, i.e. $\gamma_1 = \gamma_2 = \cdots = \gamma_k = k\gamma$. We denote the minimal value by $\psi(k, \gamma, \alpha^2)$. That is*

$$\min_{\gamma_1,\ldots,\gamma_k} ARE = \psi(k, \gamma, \alpha^2) := \frac{\gamma m_\gamma(-\gamma/\alpha^2)}{\alpha^2}\left(1 - k + \frac{\alpha^2}{\gamma m_{k\gamma}(-k\gamma/\alpha^2)}\right).$$

2. **Adding more machines leads to efficiency loss***: For fixed α^2 and γ, $\psi(k, \gamma, \alpha^2)$ is a decreasing function on $k \in [1, \infty)$ with $\lim_{k\to 1+} \psi(k, \gamma, \alpha^2) = 1$ and* infinite-worker limit

$$\lim_{k\to\infty} \psi(k, \gamma, \alpha^2) = h(\alpha^2, \gamma) < 1.$$

Here we can view ψ as a continuous function of k for convenience, although originally it is only well-defined for $k \in \mathbb{N}$. We emphasize that the infinite-worker limit tells us how much efficiency we have for a very large number of machines. It is a nontrivial result that this quantity is strictly positive.

3. **Form of the infinite-worker limit***: As a function of α^2 and γ, $h(\alpha^2, \gamma)$ has the explicit form*

$$h(\alpha^2, \gamma) = \frac{-\gamma/\alpha^2 + \gamma - 1 + \sqrt{(-\gamma/\alpha^2 + \gamma - 1)^2 + 4\gamma^2/\alpha^2}}{2\gamma}\left(1 + \frac{\alpha^2}{\gamma(1+\alpha^2)}\right).$$

4. **Edge cases of the infinite-worker limit***: For fixed α^2, $h(\alpha^2, \gamma)$ is an increasing function of $\gamma \in [0, \infty)$ with limit*

$$\lim_{\gamma\to 0} h(\alpha^2, \gamma) = \frac{1}{1+\alpha^2}, \quad \lim_{\gamma\to\infty} h(\alpha^2, \gamma) = 1.$$

On the other hand, for fixed γ, $h(\alpha^2, \gamma)$ is a decreasing function of $\alpha^2 \in [0, \infty)$ with limit

$$\lim_{\alpha^2\to 0} h(\alpha^2, \gamma) = 1, \quad \lim_{\alpha^2\to\infty} h(\alpha^2, \gamma) = \begin{cases} 1 - \frac{1}{\gamma^2}, & \gamma > 1, \\ 0, & 0 < \gamma \leq 1. \end{cases}$$

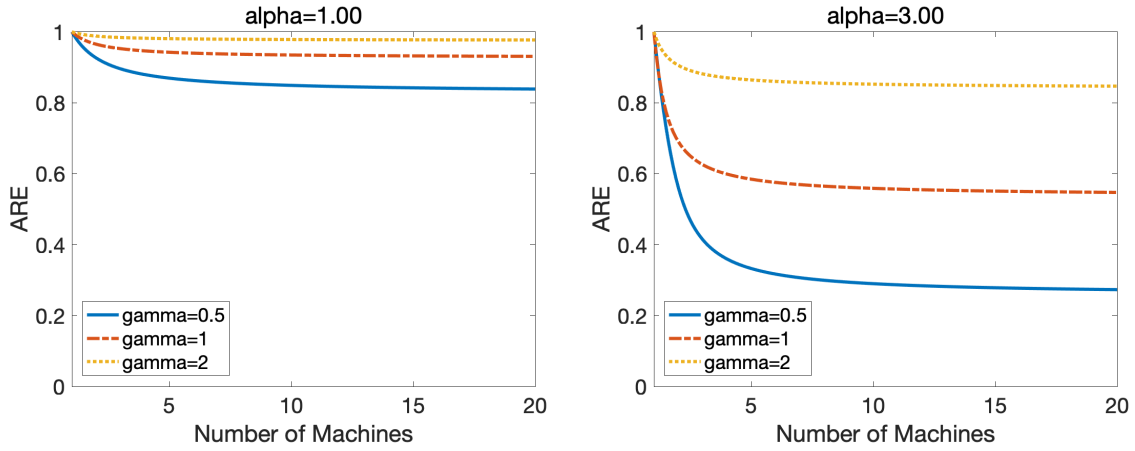


Figure 3: Plots of the asymptotic relative efficiency ψ when the data set are evenly distributed, for different α and γ. See Theorem 7 for the properties of the ARE.

See Appendix F.7 for the proof. See Figure 3 for some plots of the evenly distributed ARE ψ for various α and γ and Figure 1 for the surface and contour plots of $h(\alpha^2, \gamma)$. The efficiency loss tends to be larger (ARE is smaller) when the signal-to-noise ratio α^2 is larger. The plots confirm the theoretical result that the efficiency always decreases with the number of machines. Relatively speaking, the distributed problem becomes easier and easier as the dimension increases, compared to the aggregated problem (i.e., the ARE increases in γ for fixed parameters). This can be viewed as a blessing of dimensionality.

We also observe a nontrivial *infinite-worker limit*. Even in the limit of many machines, distributed ridge *does not lose all efficiency*. This is in contrast to doing linear regression on each machine, where all efficiency is lost when the local sample sizes are less than the dimension (Dobriban and Sheng, 2018). This is one of the few results in the distributed learning literature where one-step weighting gives nontrivial results for *arbitrary large* k, i.e., we can take $k \to \infty$ and we still obtain nontrivial results. We find this quite remarkable.

Overall, the ARE is generally large, *except* when γ is small and α is large. This is a setting with strong signal and relatively low dimension, which is also the "easiest" setting from a statistical point of view. In this case, perhaps we should use other techniques for distributed estimation, such as iterative methods.

4.2. Properties of the optimal weights

Next, we study properties of the optimal weights. This is important, because choosing them is a crucial practical question. The literature on distributed regression typically considers simple averages of local estimators, for which $\hat{\beta}_{dist} = k^{-1} \sum_{i=1}^{k} \hat{\beta}_i$ (see, e.g. Zhang et al. (2015); Lee et al. (2017); Battey et al. (2018)). In contrast, we will find that the optimal weights *do not sum up to unity*.

Formally, we have the following properties of the optimal weights.

Theorem 8 (Properties of the optimal weights) *The asymptotically optimal weights* $\mathcal{W}_k^* = (\mathcal{A} + \mathcal{R})^{-1} V$ *have the following properties:*

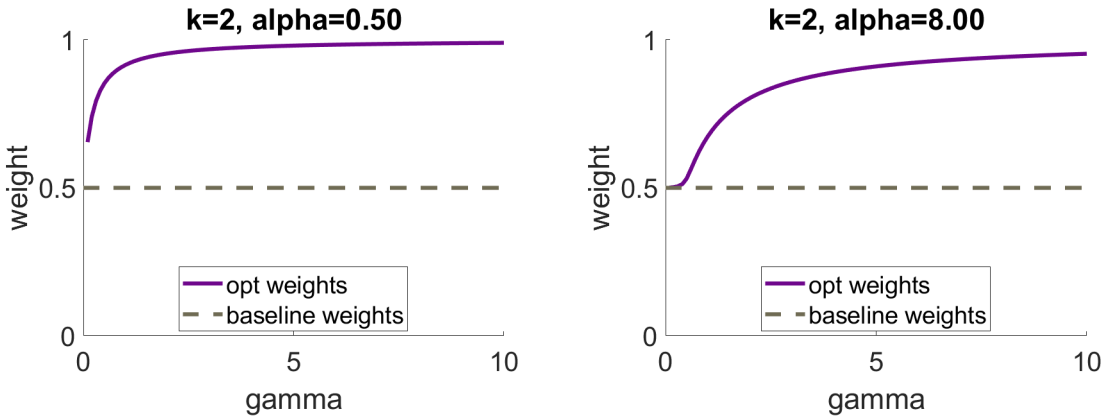


Figure 4: Plots of optimal weights for different α.

1. **Form of the optimal weights**: *The i-th coordinate of $\mathcal{W}_k$ is:*

$$\mathcal{W}_{k,i} = \left(\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)}\right) \cdot \left(\frac{1}{1+\sum_{i=1}^{k}\left[\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)}-1\right]}\right),$$

and the sum of the limiting weights is always greater than or equal to one:

$$\sum_{i=1}^{k} \mathcal{W}_{k,i} \geq 1.$$

When $k \geq 2$, the sum is strictly greater than one.

2. **Evenly distributed optimal weights**: *When the samples are evenly distributed, so that all limiting aspect ratios γ_i are equal, $\gamma_i = k\gamma$, then all $\mathcal{W}_{k,i}$ equal the* optimal weight function *$\mathcal{W}(k,\gamma,\alpha^2)$, which has the form*

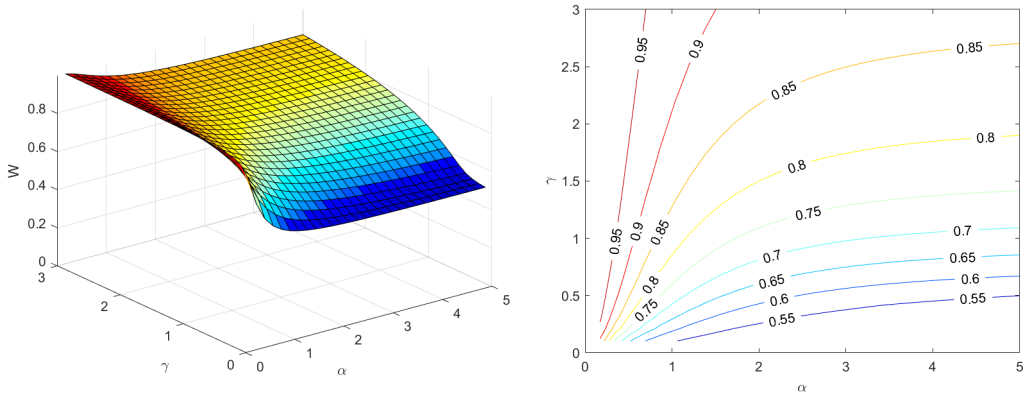
$$\mathcal{W}(k,\gamma,\alpha^2) = \frac{\alpha^2}{\alpha^2 k + (1-k)k\gamma \cdot m_{k\gamma}(-k\gamma/\alpha^2)}.$$

This can also be written in terms of the optimal risk function *$\phi(\gamma,\alpha^2)$ defined above as*

$$\mathcal{W}(k,\gamma,\alpha^2) = \frac{\alpha^2}{\alpha^2 k - (k-1)\phi(k\gamma,\alpha^2)}.$$

3. **Limiting cases**: *For fixed k and α^2, the optimal weight function $\mathcal{W}(k,\gamma,\alpha^2)$ is an increasing function of $\gamma \in [0,\infty)$ with $\lim_{\gamma\to 0+}\mathcal{W}(\gamma) = 1/k$ and $\lim_{\gamma\to\infty}\mathcal{W}(\gamma) = 1$.*

See Appendix F.8 for the proof. See Figures 4 and 5 for some plots of the optimal weight function with $k = 2$. We can see that the optimal weights are usually large, and always greater than $1/k$. When the signal-to-noise ratio α^2 is small, the weight function is concave and increases fast to approach one. In the low dimensional setting where $\gamma \to 0$, the weights tend to the uniform average $1/k$. Hence in this setting we recover the classical uniform



(a) Surface (b) Contour

Figure 5: Surface and contour plots of the optimal weights.

averaging methods, which makes sense, because ridge regression with optimal regularization parameter tends to linear regression in this regime.

How much does optimal weighting help? It is both interesting and important to know this, especially compared to naive uniform weighting, because it allows us to compare our proposed weighting method to the "baseline". See Figure 6. We have plotted the risk of distributed ridge regression for both the optimally weighted version and the simple average, as a function of the regularization parameter. We observe that *optimal weighting can lead to a 30-40% decrease in the risk*. Therefore, our proposed weighting scheme can lead to a substantial benefit.

Why are the weights large, and why do they sum to a quantity greater than one? The short intuitive answer is that ridge regression is *negatively* (or *downward*) biased, and so we must *counter the effect of bias by upweighting*. This also can be viewed as a way of *debiasing*. In different contexts, it is already well known that debiasing can play a kew role in distributed learning (Lee et al. (2017); Battey et al. (2018)). We provide a slightly more detailed intuitive explanation in Appendix E.

4.3. Out-of-sample prediction

So far, we have discussed the estimation problem. In real applications, out-of-sample prediction is also of interest. We consider a test datapoint (x_t, y_t), generated from the same model $y_t = x_t^\top \beta + \varepsilon_t$, where x_t, ε_t are independent of X, ε. We want to use $x_t^\top \hat{\beta}$ to predict y_t, and the out-of-sample prediction error is defined as $\mathbb{E}(y_t - x_t^\top \hat{\beta})^2$. Then we have the following proposition.

Proposition 9 (Out-of-sample prediction error and relative efficiency) *Under the conditions of Theorem 4, the limiting out-of-sample prediction error of the optimal distributed estimator $\hat{\beta}_{dist}$ is*

$$\mathcal{O}_k = \sigma^2 + \mathcal{M}_k.$$

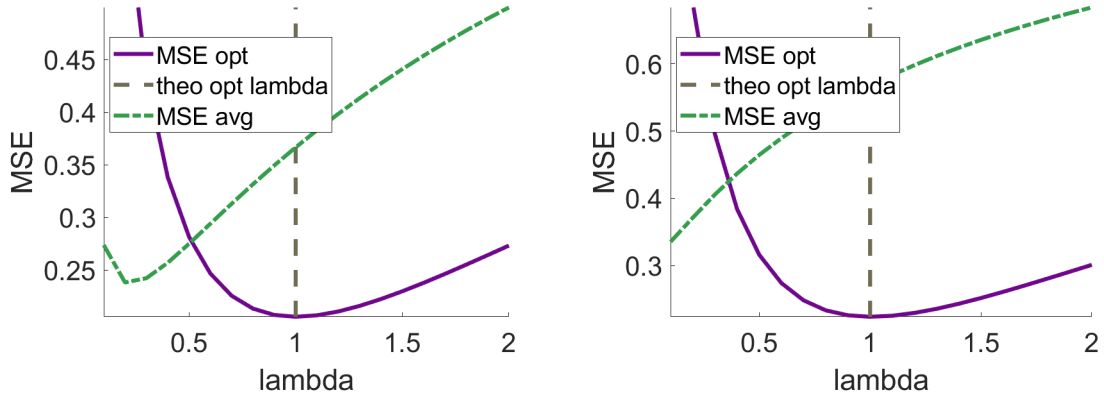


Figure 6: Distributed risk as a function of the regularization parameter. We plot both the risk with optimal weights (MSE opt) and the risk obtained from sub-optimal averaging (MSE avg). We set $\alpha = 1$, $\gamma = 0.17$ and $k = 5, 10$.

Thus, the asymptotic out-of-sample relative efficiency, meaning the ratio of prediction errors, is

$$OE = \frac{\mathcal{O}_1}{\mathcal{O}_k} = \frac{\mathcal{M}_1 + \sigma^2}{\mathcal{M}_k + \sigma^2},$$

and the efficiency for prediction is higher than for estimation $OE \geq ARE$. *Furthermore, when the samples are equally distributed, the relative efficiency has the form*

$$\Psi(k, \gamma, \alpha^2) = \frac{1 + \gamma m_\gamma(-\gamma/\alpha^2)}{1 + \frac{\alpha^2 \gamma m_{k\gamma}(-k\gamma/\alpha^2)}{\alpha^2 + (1-k)\gamma m_{k\gamma}(-k\gamma/\alpha^2)}},$$

and the corresponding infinite-worker limit (taking $k \to \infty$*) is*

$$\mathcal{H}(\alpha^2, \gamma) = \frac{1 + \gamma m_\gamma(-\gamma/\alpha^2)}{1 + \frac{\gamma\alpha^2(1+\alpha^2)}{\alpha^2 + \gamma(1+\alpha^2)}}.$$

See Appendix F.9 for the proof and Figure 7 for some plots. This proposition implies that, for the identity covariance case, the efficiency loss of the distributed estimator in terms of the test error is always less than the loss in terms of the estimation error. When the signal-to-noise ratio α^2 is small, the relative efficiency is always very large and close to 1. This observation can be an encouragement to use our distributed methods for out-of-sample prediction.

4.4. Choosing the regularization parameter

Previous work found that, under certain conditions, the regularization parameters on the individual machines should be chosen as if they had the all samples (Zhang et al., 2015). Our findings are consistent with these results. However, the reasons behind our findings are very different from prior work. The intuition for the previous results is that the *variance* of distributed estimators averages out, while the *bias* does not do so. Therefore, the

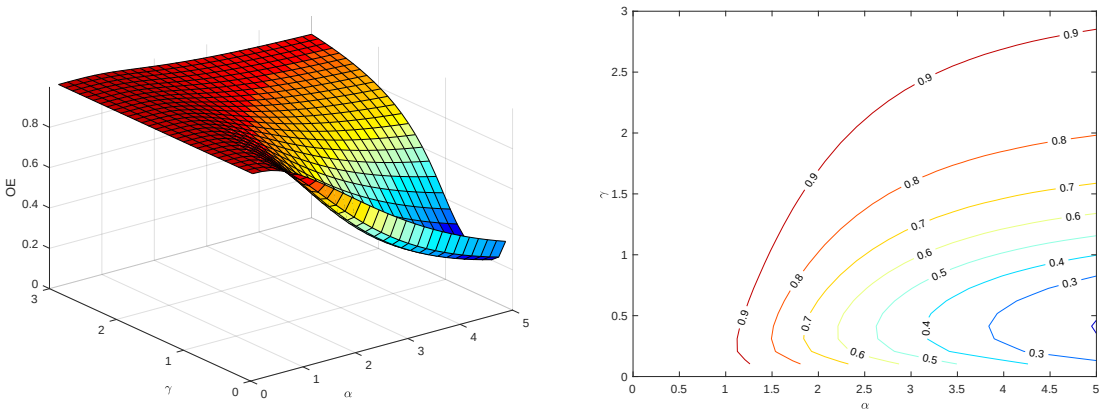


Figure 7: Limit of OE: (a) surface and (b) contour plots of $\mathcal{H}(\alpha^2, \gamma)$.

regularization parameters should be chosen such that the local bias is lower than for locally optimal tuning. This means that we should use smaller regularization parameters locally.

In our case, we find that for isotropic covariance, the optimal risk *decouples across machines*. Hence, the regularization parameters on the machines can be chosen optimally for each machine. Moreover, in our asymptotics the locally optimal choice is a *constant multiple* of the globally optimal choice, namely the multiple in front of the identity matrix in the local ridge estimator $(X_i^\top X_i + n_i\lambda_i I_p)^{-1} X_i^\top Y_i$ should be $\lambda_i = p/(n_i\alpha^2)$ whereas the globally optimal λ is $\lambda = p/(n\alpha^2)$.

Roughly speaking, this derivation reaches the same conclusion as prior work about the choice of regularization parameters, namely that the regularization parameters on the machines should be chosen as if they had the all samples. However, we emphasize that our results are very different, because the optimal weighting procedure has weights summing to *greater than unity*. Moreover, we also consider the proportional-limit case, and the conclusion for regularization parameters only applies to the isotropic case.

4.5. Implications and practical relevance

We discuss some of the implications of our results. Our finite-sample results show that the optimal way to weight the estimators depends on functionals of the unknown parameter β, while the asymptotic results in general depend on the eigenvalues of $\widehat{\Sigma}$ (or Σ). These are unavailable in practice, and hence these results can typically not be used on real data sets. However, since our results are precise and accurate (they capture the *truth* about the problem), we interpret this as saying that *the problem is hard in general*. Meaning that optimal weighting for ridge regression is a challenging statistical problem. In practice that means that we may be content with uniform weighting. It remains to be investigated in future work how much we should up-adjust those equal weights.

The optimal weights become usable in the case of spherical data, when $\Sigma = I$ (or, more accurately, the limiting spectral distribution of Σ is the point mass at unity). In practice, we can get closer to this assumption by using some form of *whitening* on the data, for instance by scaling all variables to the same scale, by estimating Σ over restricted classes, such as assuming block-covariance structures. Alternatively, we can use correlation screening,

where we remove features with high correlation. At this stage, all these approaches are heuristic, but we include them to explain how our results can be relevant in practice. It is a topic of future research to make these ideas more concrete. In the algorithm we proposed in Section 5, we use grid search to find a good tuning parameter under general covariance structures.

On the theoretical side, our results can also be interpreted as a form of *reduction* between statistical problems. *If* we can estimate the quadratic functionals of the unknown regression parameter involved in our weights, *then* we can do optimally weighted ridge regression. In this sense, we reduce distributed ridge regression to the estimation of those quadratic functionals. We think that in the challenging and novel setting of distributed learning, such reductions can be both interesting and potentially useful.

An important question is "Should we use distributed linear or ridge regression?". If we have $n_i \geq p$ and linear regression is defined on each local machine, then we can use either distributed linear (Dobriban and Sheng, 2018) or ridge regression. Linear regression has the advantage that the optimal weights are easy to find. Therefore, if we cannot reasonably reduce to the case $\Sigma = I$, it seems we should use linear regression.

4.6. Minimax optimality of the optimal distributed estimator

To deepen our understanding of the distributed problem, we next show that the optimal distributed ridge estimator is asymptotically rate-minimax. Suppose without loss of generality that the noise level $\sigma^2 = 1$, and let $\mathbb{S}^{p-1}(\alpha) = \{\beta \in \mathbb{R}^p; ||\beta|| = \alpha\}$ denote the sphere of radius $\alpha \geq 0$ in $\mathbb{R}^p$ centered at the origin. Then the minimax risk for estimating β over the sphere $\mathbb{S}^{p-1}(\alpha)$ is

$$r(\alpha) = \inf_{\hat{\beta}} \sup_{\beta \in \mathbb{S}^{p-1}(\alpha)} R(\hat{\beta}, \beta) = \inf_{\hat{\beta}} \sup_{\beta \in \mathbb{S}^{p-1}(\alpha)} \mathbb{E}_\beta ||\hat{\beta} - \beta||^2,$$

where the expectation is over both X and ε. This problem has been well studied by Dicker (2016), who reduced it to the following Bayes problem. Let π be the uniform measure on $\mathbb{S}^{p-1}(\alpha)$. Then the Bayes risk with respect to π is

$$r_B(\alpha) = \inf_{\hat{\beta}} \int_{\mathbb{S}^{p-1}(\alpha)} R(\hat{\beta}, \beta) d\pi(\beta) = \inf_{\hat{\beta}} \mathbb{E}_\pi ||\hat{\beta} - \beta||^2.$$

The Bayes estimator is the posterior mean $\hat{\beta}_{\mathbb{S}^{p-1}(\alpha)} = \mathbb{E}_\pi(\beta|y, X)$. So the corresponding Bayes risk is $r_B(\alpha) = \mathbb{E}_\pi ||\hat{\beta}_{\mathbb{S}^{p-1}(\alpha)} - \beta||^2$. Then, the Bayes estimator also minimizes the original minimax risk and $r(\alpha) = r_B(\alpha)$ (Dicker, 2016).

Recall that the ridge estimator with optimally tuned regularization parameter is

$$\hat{\beta}_r(\alpha) = (X^\top X + \frac{p}{\alpha^2} I_p)^{-1} X^\top Y,$$

which can be interpreted as the posterior mean of β under the normal prior assumption $\beta \sim \mathcal{N}(0, \alpha^2/p I_p)$. When p is very large, the normal distribution $\mathcal{N}(0, \alpha^2/p I_p)$ is very close to the uniform distribution on $\mathbb{S}^{p-1}(\alpha)$, so we would expect that $\hat{\beta}_{\mathbb{S}^{p-1}(\alpha)} \approx \hat{\beta}_r(\alpha)$. With this intuition, Dicker (2016) further showed that, as $p, n \to \infty, p/n \to \gamma \in (0, \infty)$, for any $\beta \in \mathbb{S}^{p-1}(\alpha)$

$$\lim_{n,p \to \infty} \left[R(\hat{\beta}_{\mathbb{S}^{p-1}(\alpha)}, \beta) - R(\hat{\beta}_r(\alpha), \beta) \right] = 0.$$

So the global ridge estimator is asymptotically exact minimax.

We call an estimator is *asymptotically rate-minimax* if asymptotically its risk is at most a constant times the minimax risk. For our distributed problem, we have the following result:

Theorem 10 (Minimax optimality) *For fixed signal strength α^2, the optimally weighted distributed ridge estimator is asymptotically rate minimax. Specifically, its risk $\mathcal{M}_k$ is less than the risk $\mathcal{M}_1$ of the global ridge estimator multiplied by a constant $C = 1 + \alpha^2$ which only depends on the signal strength α^2, and not on the aspect ratio $\gamma = \lim p/n$ and number of machines k. Specifically*

$$\mathcal{M}_k \leq (1 + \alpha^2)\mathcal{M}_1.$$

Moreover, for fixed aspect ratio $\gamma > 1$, the distributed risk $\mathcal{M}_k$ is less than the global risk $\mathcal{M}_1$ times a constant $C' = \gamma^2/(\gamma^2 - 1)$ which is independent of α^2 and k, i.e.

$$\mathcal{M}_k \leq \frac{\gamma^2}{\gamma^2 - 1}\mathcal{M}_1.$$

Therefore, in either case, the optimally weighted distributed ridge estimator is asymptotically rate minimax.

See Section F.10 for the proof. The minimax optimality result is nontrivial, and does not hold for some simpler estimators. For instance, for the null estimator $\hat{\beta}_{null} = 0$, the corresponding ARE can be written in terms of the optimal risk function $\phi(\gamma)$ as

$$\lim_{n,p\to\infty} \frac{R(\hat{\beta}_r(\alpha), \beta)}{R(\hat{\beta}_{null}, \beta)} = \frac{\phi(\gamma)}{\alpha^2} = \frac{\gamma m_\gamma(-\gamma/\alpha^2)}{\alpha^2}.$$

When $\gamma \to \infty$, we know that $\gamma/\alpha^2 m_\gamma(-\gamma/\alpha^2) \to 1$, so that even *the null estimator is asymptotically exact minimax*. In this regime, exact minimaxity is a weak result. When $\gamma \to 0$ however, we have $\gamma/\alpha^2 m_\gamma(-\gamma/\alpha^2) \to 0$ for any α, and so the null estimator does not perform well (has zero efficiency). However, the distributed estimator is still asymptotically rate-minimax.

5. WONDER: Algorithms for Weighted One-shot Distributed Ridge Regression

So far, most of our results on distributed ridge regression are purely theoretical. In practice, it would be very helpful to have an implementable algorithm. In fact, our theory for distributed ridge regression allows us to develop an efficient algorithm which works for designs X with arbitrary covariance structures Σ.

Recall that we have n samples distributed across k machines. For simplicity, let us assume the samples are equally distributed. On the i-th machine, we compute a local ridge estimator $\hat{\beta}_i$, local estimators $\hat{\sigma}_i^2$, $\hat{\alpha}_i^2$ of the signal-to-noise ratio and the noise level. From Theorem 3, we know that the other quantities needed to find the optimal weights are m, m' and λ. For m and m', by the definition of the Stieltjes transform, they can be approximated by

$$\frac{\mathrm{tr}[(\widehat{\Sigma}_i + \lambda I)^{-1}]}{p} \approx m(-\lambda) \quad \text{and} \quad \frac{\mathrm{tr}[(\widehat{\Sigma}_i + \lambda I)^{-2}]}{p} \approx m'(-\lambda).$$

Algorithm 1: WONDER: Weighted ONe-shot DistributEd Ridge regression algorithm, general design

Input : Data matrices $(n_i \times p)$ and outcomes $(n_i \times 1)$, (X_i, Y_i) distributed across k sites

Output: Distributed ridge estimator $\hat{\beta}_{dist}$ of regression coefficients β

for $i \leftarrow 1$ **to** k **do**
  Compute the MLE $\hat{\theta}_i = (\hat{\sigma}_i^2, \hat{\alpha}_i^2)$ locally on i-th machine;
  Send $\hat{\theta}_i$ to the global data center;
end
At the data center, combine $\hat{\theta}_i$ to get a global estimator $\hat{\theta} = (\hat{\sigma}^2, \hat{\alpha}^2) = k^{-1} \sum_{i=1}^k \hat{\theta}_i$ and send it back to the local machines;
Choose a set of tuning parameters $\mathcal{S}$ around the initial guess $\lambda_0 = kp/(n\hat{\alpha}^2)$;
for $\lambda \in \mathcal{S}$ **do**
  for $i \leftarrow 1$ **to** k **do**
    Compute the local ridge estimator $\hat{\beta}_i(\lambda) = (X_i^\top X_i + n_i \lambda I_p)^{-1} X_i^\top Y_i$;
    Compute the weight ω_i for the i-th local estimator by using the formulas from Theorem 3:

$$\omega_i(\lambda) = \frac{\hat{\sigma}^2 \hat{\alpha}^2 (1 - \lambda m)}{\mathcal{F} + k\mathcal{G}}$$

    where we use $\mathrm{tr}[(X_i^\top X_i/n_i + \lambda I)^{-1}]/p$ to approximate m, and use $\mathrm{tr}[(X_i^\top X_i/n_i + \lambda I)^{-2}]/p$ to approximate m';
    Send $\hat{\beta}_i(\lambda)$ and $\omega_i(\lambda)$ to the global data center;
  end
  Evaluate the performance of the distributed ridge estimator $\hat{\beta}_{dist}(\lambda) = \sum_{i=1}^k \omega_i(\lambda) \hat{\beta}_i(\lambda)$ on validation sets;
end
Select the best tuning parameter λ^* and output the corresponding distributed ridge estimator $\hat{\beta}_{dist}(\lambda^*) = \sum_{i=1}^k \omega_i(\lambda^*) \hat{\beta}_i(\lambda^*)$.

Here we only need to use local data. The remaining question is: how do we choose the tuning parameter λ? One way may be grid search. From the theory for the isotropic design, a proper initial guess would be $\lambda = kp/(n\alpha^2)$. Then we can search around this initial guess to find a good parameter with small prediction error.

We assume the data are already mean-centered, which can be performed exactly in one additional round of communication, or approximately by centering the individual data sets.

Now we have all the quantities we need for our Weighted ONe-shot DistributEd Ridge regression algorithm (WONDER). We send them to a global machine or data center, and aggregate them to compute a weighted ridge estimator. See Algorithm 1 for more details. WONDER is communication efficient as the local machines only need to send the local ridge estimator $\hat{\beta}_i$ and some scalars to the global datacenter.

For identity covariance, our results lead to a much simpler WONDER algorithm which requires even less communication and computation. See Algorithm 2.

Algorithm 2: WONDER: Weighted ONe-shot DistributEd Ridge regression algorithm, isotropic design

Input : Data matrices $(n_i \times p)$ and outcomes $(n_i \times 1)$, (X_i, Y_i) distributed across k sites

Output: Distributed ridge estimator $\hat{\beta}_{dist}$ of regression coefficients β

for $i \leftarrow 1$ **to** k **do**

  Compute the MLE $\hat{\theta}_i = (\hat{\sigma}_i^2, \hat{\alpha}_i^2)$ locally on i-th machine;

  Set local aspect ratio $\gamma_i = p/n_i$;

  Set regularization parameter $\lambda_i = \gamma_i/\hat{\alpha}_i^2$;

  Compute the local ridge estimator $\hat{\beta}_i(\lambda_i) = (X_i^\top X_i + n_i \lambda_i I_p)^{-1} X_i^\top Y_i$;

  Send $\hat{\theta}_i, \gamma_i$ and $\hat{\beta}_i$ to the global data center.

end

At the data center, combine $\hat{\theta}_i$ to get a global estimator $\hat{\theta} = (\hat{\sigma}^2, \hat{\alpha}^2)$, by $\hat{\theta} = k^{-1} \sum_{i=1}^k \hat{\theta}_i$;

Evaluate the optimal risk functions for $i = 1, 2, \ldots, k$

$$\phi(\gamma_i) = \gamma_i m_{\gamma_i}(-\gamma_i/\hat{\alpha}^2) = \frac{-\gamma_i/\hat{\alpha}^2 + \gamma_i - 1 + \sqrt{(-\gamma_i/\hat{\alpha}^2 + \gamma_i - 1)^2 + 4\gamma_i^2/\hat{\alpha}^2}}{2\gamma_i/\hat{\alpha}^2};$$

Compute the optimal weights ω, where the i-th coordinate of ω is

$$\omega_i = \left(\frac{\hat{\alpha}^2}{\phi(\gamma_i)}\right) \cdot \left(\frac{1}{1 + \sum_{i=1}^k \left[\frac{\hat{\alpha}^2}{\phi(\gamma_i)} - 1\right]}\right);$$

Output the distributed ridge estimator $\hat{\beta}_{dist} = \sum_{i=1}^k \omega_i \hat{\beta}_i$.

In the above WONDER algorithms, we combine the local estimators of the noise level and signal strength $\hat{\theta}_i$ to find a global estimator $\hat{\theta}$. A simple method is to take the average: $\hat{\theta} = k^{-1} \sum_{i=1}^k \hat{\theta}_i$. Another option is to use inverse-variance weighting, based on the asymptotic variance of the MLE (which then of course has to be estimated).

Based on the results so far, it follows that our WONDER algorithm can consistently estimate the limiting optimal weights, and moreover it has asymptotically optimal mean squared error among all weighted distributed ridge estimators, at least for the identity covariance case. We omit the details.

6. Experimental Results

We present some numerical results in addition to the ones already shown in the paper.

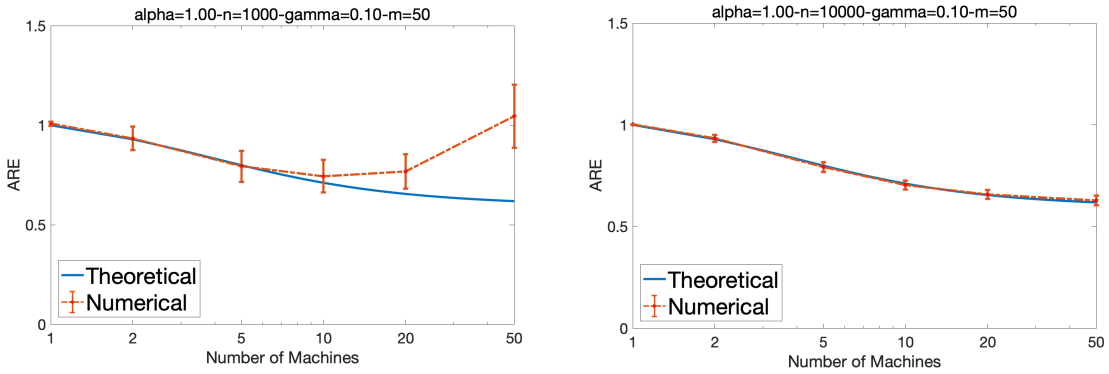


Figure 8: Realized relative efficiency in a regression simulation.

6.1. Finite-sample comparison of relative efficiency for isotropic covariance

Figure 8 shows a comparison of the theoretical formulas for ARE and realized relative efficiency in a regression simulation. Here the regression model is $Y = X\beta + \varepsilon$, where X is $n \times p$ with i.i.d. standard normal entries, β is a p-dimensional random vector with i.i.d. mean 0, variance α^2/p normal entries, and ε also has i.i.d standard normal entries. For each $k = 1, 2, 5, 10, 20, 50$, we split the data equally into k groups and perform ridge regression on each group. For each group, we choose the same tuning parameter $\lambda_i = p/(n_i\alpha^2)$. For the global regression on the entire data set, we choose the tuning parameter $\lambda = p/(n\alpha^2)$ optimally.

We show the results of the expression for the realized relative efficiency $\|\hat{\beta} - \beta\|^2/\|\hat{\beta}_{dist} - \beta\|^2$ compared to the theoretical ARE. We generate 100 independent copies of ε, perform regression, recording the realized relative efficiency $||\hat{\beta} - \beta||^2/||\hat{\beta}_{dist} - \beta||^2$, as well as its overall Monte Carlo mean. For the first plot, we take $n = 1000$, $p = 100$, and $\alpha = \sigma = 1$.

As we can see in the plot, the theoretical formula is accurate only for a small number of machines. It turns out that this is due to finite-sample effects. In the second plot, we set $n = 10000$, $p = 1000$ and $\alpha = \sigma = 1$ such that the aspect ratio $\gamma = p/n$ is the same as before. In that case the theoretical formula becomes very accurate.

6.2. Choosing the regularization for general covariance

How can we choose the optimal regularization parameters when the predictors have a general covariance structure Σ? In this case, our theoretical results do not give an explicit expression for the optimal regularization parameters. In practice, one can use techniques like cross-validation to do selections. Here we present simulation results to shed light on the important question of how to choose them.

We use a similar simulation setup as in the previous sections, except we generate the datapoints independently from an autoregressive model of order one (AR-1), i.e., each datapoint x_i is generated as $x_i \sim \mathcal{N}(0, \Sigma)$, where $\Sigma_{ij} = \rho^{|i-j|}$, and ρ is the autocorrelation parameter. We choose $\rho = 0.9$. We also choose $n = 3000$, $p = 500$, and report the results of a simulation where we average over $n_{mc} = 20$ independent realizations of β. Figure 9 shows the optimal distributed risk $M^*(k)$ as a function of the local regularization parameter λ. We

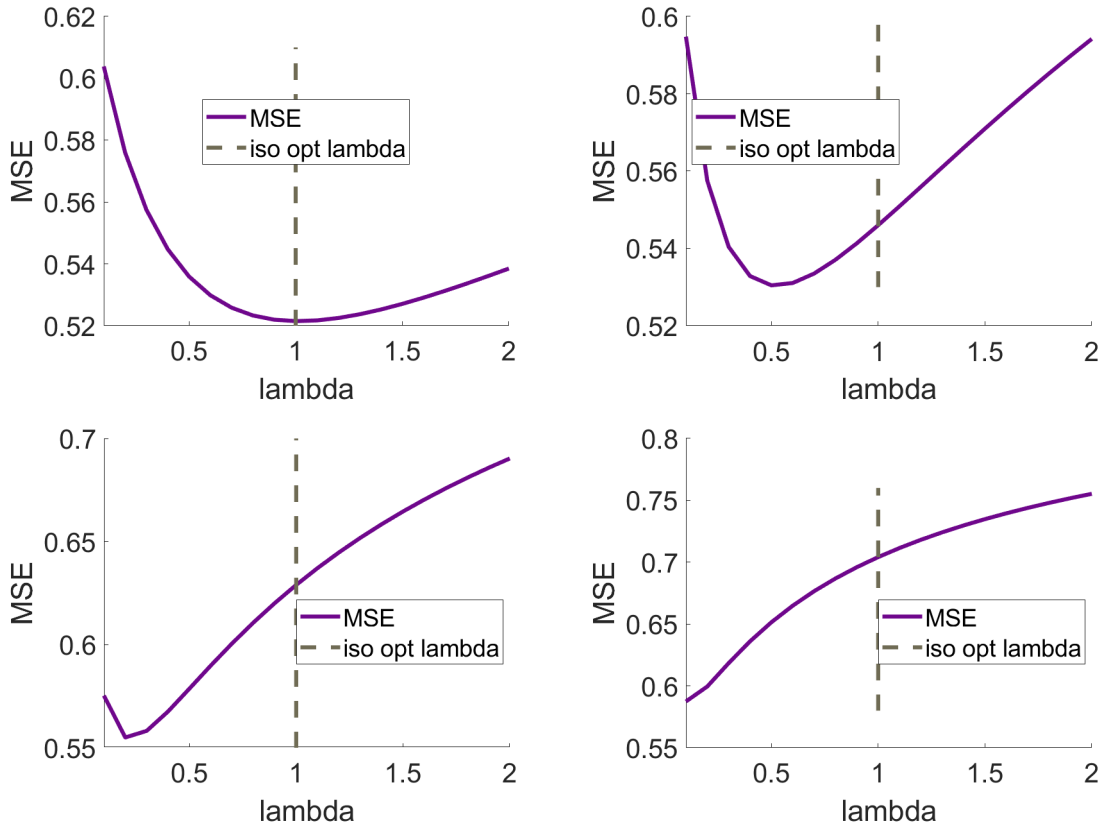


Figure 9: Distributed risk as a function of the regularization parameter. We plot the risk of the optimally weighted distributed estimator for an AR-1 covariance structure. We set $\alpha = 1$, $\gamma = 0.17$ and $k = 1, 2, 5, 10$.

set all local regularization parameters to equal values, which is reasonable, since the local problems are exchangeable. We also parametrize the regularization parameters as multiples of the optimal parameter for the isotropic case (which equals $k\gamma/\alpha^2$).

We observe that for $k = 1$, the optimal parameter is the same as in the isotropic case. This makes sense, because the optimal regularization parameter for one machine is always the same, regardless of the structure of the design. However for $k > 1$, we observe that the regularization parameters are *smaller* than the isotropic ones. This is an insight that has apparently not been available before. It is an interesting topic of future work to develop an intuitive understanding.

6.3. Experiments on empirical data

In this section, we present an empirical data example to examine the accuracy of our theoretical results. It is reasonable to compare the performance of different estimators in terms of the prediction (test) error. Figure 10 shows a comparison of three estimators including our optimal weighted estimator on the Million Song Year Prediction Dataset (MSD) (Bertin-Mahieux et al., 2011).

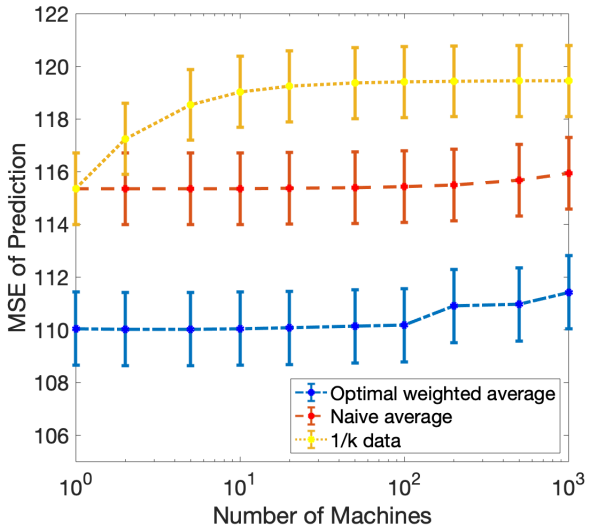


Figure 10: Million Song Year Prediction Dataset (MSD). Optimal weighted average (WONDER), Naive average, and regression on $1/k$ fraction of data.

Specifically, we perform the following steps in our data analysis. We download the data set from the UC Irvine Machine Learning Repository. The original data set has $N = 515,345$ samples and $p = 91$ features. The data set has already been divided into a training set and a test set. The training set consists of the first $463,715$ samples and the test set contains the rest. We attempt to predict the release year of a song. Before doing distributed regression, we first center and normalize both the design matrix X and the outcome Y. Now we are ready to do ridge regression under the distributed setting.

For each experiment, we randomly choose $n_{train} = 10,000$ samples from the training set and $n_{test} = 1,000$ samples from the test set. We construct the estimators on the training samples. Then we perform ridge regression in a distributed way to obtain our optimal weighted WONDER estimator as described in Algorithm 1. We measure its performance on the test data by computing its MSE for prediction. We choose the number of machines to be $k = 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000$, and we distribute the data evenly across the k machines. Here we try different tuning parameters λ around $kp/(n_{train} \cdot \hat{\alpha}^2)$, and use $\lambda = 3kp/(n_{train} \cdot \hat{\alpha}^2)$ as our final parameter. (In practice, one may try a 1-D grid search to find the right scale.)

For comparison, we also consider two other estimators:

1. The distributed estimator where we take the naive average (weight for each local estimator is simply $1/k$) and choose the local tuning parameter $\lambda = p/(n_{train} \cdot \hat{\alpha}^2)$. This formally agrees with the divide-and-conquer type estimator proposed in Zhang et al. (2015).

2. The estimator using only a fraction $1/k$ of the data, which is just one of the local estimators. For this estimator, we choose the tuning parameter $\lambda = kp/(n_{train} \cdot \hat{\alpha}^2)$.

We repeat the experiment for $T = 100$ times, and report the average and 1/4 standard deviation over all experiments on Figure 10. Each time we randomly collect new training and test sets.

From Figure 10, we observe the following:

1. The WONDER estimator has smaller MSE than both the local estimator and the naive averaged estimator, which means optimal weighting can indeed help.

2. It seems that data splitting does not have huge impact on the performance of the WONDER estimator. This phenomenon is compatible with our theory. Since the signal-to-noise ratio α^2 is about 1.2 for this data set, we are in a low SNR scenario. From Proposition 9 and Figure 7, we see that the performance of the distributed estimator is close to the global estimator in terms of the prediction error.

To conclude, in terms of computation-statistics tradeoff, this example suggests a very positive outlook on using distributed ridge regression via WONDER: The accuracy is affected very little even though the data is split up into 100 parts. Thus we save at least 100x in computation time, while we have nearly no loss in performance.

Finally, we mention that in Figure 4 of Zhang et al. (2015), the authors also compare the performance of the distributed estimator to the local estimator on the same Million Song data set. We notice that the MSE of prediction in their experiments is usually between 80 and 90, and variance is typically very small. In our experiments, both the MSE and variance are larger. The reason for this seems to be that they consider more general kernel ridge regression.

Acknowledgments

The authors are very grateful to the Action Editor and the referees for their work in editing and reviewing the paper. The authors thank Yuekai Sun for discussions motivating our study, as well as John Duchi, Jason D. Lee, Xinran Li, Jonathan Rosenblatt, Feng Ruan, and Linjun Zhang for helpful discussions. They are grateful to Sifan Liu for thorough comments on an earlier version of the manuscript. They are also grateful to the associate editor and referees for valuable suggestions. ED was partially supported by NSF BIGDATA grant IIS 1837992.

Appendix A. Adding a Constant to the Regression

We show below the details of the derivation of optimal weights for ridge regression when we also add a constant to the (biased) local estimators. In our calculation from Theorem 1, we need to change some details as follows:

We need to define a new matrix $\widehat{B} = [\hat{\beta}_1, \ldots, \hat{\beta}_k, p^{-1/2}1_p]$ and new weights $w = [w; w_{k+1}]$. Clearly, we still have that

$$B = [\mathbb{E}\hat{\beta}_1, \ldots, \mathbb{E}\hat{\beta}_k, p^{-1/2}1_p] = [Q_1\beta; \ldots; Q_k\beta, p^{-1/2}1_p].$$

The new matrix R is now diagonal with all entries as before, and the lower right corner entry is $R_{k+1} = 0$.

We consider the same regression problem as before, except we add an intercept into the matrix B as above. The same algebraic form of the optimal weights and risk holds, with the new definitions above. The optimal risk is now

$$M^*(k) = \|\beta\|^2 - v^\top (A+R)^{-1} v$$

where

$$\begin{aligned} v &= B^\top \beta = [vec[\beta^\top Q_i \beta]; p^{-1/2} 1_p^\top \beta] \\ A &= \begin{bmatrix} mx[\beta^\top Q_i Q_j \beta] & vec[p^{-1/2} 1_p^\top Q_i \beta] \\ vec[p^{-1/2} 1_p^\top Q_i \beta] & 1 \end{bmatrix} \\ R &= \text{diag}\left[n_i^{-1} \,\text{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-2} \widehat{\Sigma}_i]; 0 \right] \\ Q_i &= (\widehat{\Sigma}_i + \lambda_i I_p)^{-1} \widehat{\Sigma}_i \end{aligned}$$

In simulation studies, we have observed that this approach typically does not lead to a significant decrease in MSE.

Appendix B. Differentiation Rule for Calculus of Deterministic Equivalents

Theorem 11 (Differentiation rule) *Suppose $T = T_n$ and $S = S_n$ are two (deterministic or random) matrix sequences of growing dimensions such that $f(z, T_n) \asymp g(z, S_n)$, where the entries of f and g are analytic functions in $z \in D$ and D is an open connected subset of $\mathbb{C}$. Suppose that for any sequence C_n of deterministic matrices with bounded trace norm we have*

$$|\,\text{tr}\,[C_n(f(z,T_n) - g(z,S_n))]\,| \le M$$

for every n and $z \in D$. Then we have $f'(z, T_n) \asymp g'(z, S_n)$ for $z \in D$, where the derivatives are entry-wise with respect to z.

To prove this theorem, we need to introduce a lemma from complex analysis which is a consequence of the dominated convergence theorem and Cauchy's integral formula.

Lemma 12 (see Lemma 2.14 in Bai and Silverstein (2010)) *Let $f_1, f_2, \ldots$ be analytic on the domain D, satisfying $|f_n(z)| \le M$ for every n and $z \in D$. Suppose that there is an analytic function on D such that $f_n(z) \to f(z)$ for all $z \in D$. Then it also holds that $f_n'(z) \to f'(z)$ for all $z \in D$.*

The proof of theorem 11 is clear. Since $\text{tr}\,[C_n(f(z,T_n) - g(z,S_n))]$ is a sequence of analytic functions on D with uniform bound, then from the definition of the deterministic equivalence, we have $\text{tr}\,[C_n(f(z,T_n) - g(z,S_n))] \to 0$. By lemma 12, the derivative also converges to 0 for all $z \in D$, which finishes the proof.

Appendix C. Gaussian MLE for Signal and Noise Components

Recall that our model is $Y = X\beta + \varepsilon$ where β and ε are independent. Let $\theta = (\sigma^2, \alpha^2)$ and define the Gaussian log-likelihood,

$$\ell(\theta) = -\frac{1}{2}\log(\sigma^2) - \frac{1}{2n}\log\det\left(\frac{\alpha^2}{p}XX^\top + I\right) - \frac{1}{2\sigma^2 n}Y^\top\left(\frac{\alpha^2}{p}XX^\top + I\right)^{-1}Y.$$

Note that $\ell(\theta)$ is the log-likelihood for θ under the Gaussian assumption of $\beta \sim \mathcal{N}(0, (\sigma^2\alpha^2/p)I)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. For the MLE

$$\hat{\theta} = (\hat{\sigma}^2, \hat{\alpha}^2) = \underset{\sigma^2,\alpha^2\geq 0}{\operatorname{argmax}}\ \ell(\theta),$$

we have the following result from Dicker and Erdogdu (2017).

Theorem 13 (Consistency and asymptotic normality, Dicker and Erdogdu (2017)) *Suppose $\theta = (\sigma^2, \alpha^2)$ are the true parameters, then $\hat{\theta} \to \theta$ in probability as $p/n \to \gamma$. Furthermore, define the Fisher information matrix for θ under the Gaussian assumption model*

$$\mathcal{I}_n(\theta) = \begin{bmatrix} I_2(\theta) & I_3(\theta) \\ I_3(\theta) & I_4(\theta) \end{bmatrix},$$

where

$$I_k(\theta) = \frac{1}{2n\sigma^{8-2k}}\operatorname{tr}\left[\left(\frac{1}{p}XX^\top\right)^{k-2}\left(\frac{\alpha^2}{p}XX^\top + I\right)^{2-k}\right], \quad k = 2, 3, 4.$$

Then $n^{1/2}\mathcal{I}_n(\theta)^{1/2}(\hat{\theta} - \theta) \to \mathcal{N}(0, I_2)$ in distribution as $p/n \to \gamma$.

In addition, if we put some assumptions on X as we did in Theorem 1 and denote the limiting spectral distribution of $p^{-1}XX^\top$ by F_γ, then the entries of the Fisher information matrix $\mathcal{I}_n(\theta)$ have limits

$$I_k(\theta) \to_{a.s.} \mathcal{J}_k(\theta) = \frac{1}{2\sigma^{8-2k}}\int\left(\frac{s}{\alpha^2 s + 1}\right)^{k-2} dF_\gamma(s), \quad k = 2, 3, 4.$$

Thus $\mathcal{I}_n(\theta)$ converges almost surely to a limiting information matrix $\mathcal{I}(\theta)$ which characterizes the asymptotic variance of the MLE $\hat{\theta}$.

Appendix D. Explaining Decoupling via Free Probability Theory

In this section, we provide an explanation via free probability theory for why the limiting distributed risk decouples. Specifically, we explain why the limit of the quantities $\beta^\top Q_i\beta \cdot \beta^\top Q_j\beta$ for $i \neq j$ becomes a product of terms depending on i, j.

We will use some basic notions from free probability theory (Voiculescu et al., 1992; Hiai and Petz, 2006; Nica and Speicher, 2006; Anderson et al., 2009; Couillet and Debbah, 2011). Let us define our non-commutative probability space as

$$\left(\mathcal{A} = (L^{\infty-} \otimes M_p(\mathbb{R})), \tau = \frac{1}{p}\operatorname{tr}\right),$$

where $L^{\infty-}$ denotes the collection of random variables with all moments finite and $M_p(\mathbb{R})$ is the space of $p\times p$ real matrices. Recall that, a sequence of random variables $\{a_{1,p}, a_{2,p}, \ldots, a_{k,p}\} \subset \mathcal{A}$ is said to be asymptotically free almost surely if

$$\tau[\prod_{j=1}^{m} P_j(a_{i_j,p} - \tau(P_j(a_{i_j,p})))] \to_{a.s.} 0,$$

for any positive integer m, any polynomials $P_1, \ldots, P_m$ and any indices $i_1, \ldots, i_m \in [k]$ with no two adjacent i_j equal. Suppose A_p, B_p are two sequences of independent random matrices and at least one of them is orthogonally invariant, then it is well-known that $\{A_p, B_p\} \subset \mathcal{A}$ is asymptotically free almost surely.

Now, let us assume that $X^\top X$ is orthogonally invariant, which is the case when $X^\top X$ follows the white Wishart distribution. Then clearly $X_i^\top X_i$ and $X_j^\top X_j$ are asymptotically free almost surely. It follows that Q_i and Q_j are also asymptotically free almost surely. By using the definition of asymptotic freeness, we have for $i \neq j$

$$\tau[(Q_i - \frac{1}{p}\operatorname{tr}(Q_i)I)(Q_j - \frac{1}{p}\operatorname{tr}(Q_j)I)] \to_{a.s.} 0,$$

which is equivalent to

$$\frac{1}{p}\operatorname{tr}(Q_iQ_j) - \frac{1}{p}\operatorname{tr}(Q_i)\frac{1}{p}\operatorname{tr}(Q_j) \to_{a.s.} 0.$$

Hence, under the random-effects assumption for β, the limit of $\beta^\top\beta \cdot \beta^\top Q_iQ_j\beta$ $(i \neq j)$ will decouple and is the same as the limit of $\beta^\top Q_i\beta \cdot \beta^\top Q_j\beta$.

Appendix E. Intuitive Explanation for the Need to Use Weights Summing to Greater than Unity

Consider a much more simplified problem, where we are estimating a scalar parameter θ. We have an estimator $\hat{\theta}$, which is generally biased, and we would like to find the scale multiple $c \cdot \hat{\theta}$ that minimizes the mean squared error. A calculation reveals that

$$M(c) = \mathbb{E}(c \cdot \hat{\theta} - \theta)^2 = c^2\mathbb{E}(\hat{\theta}^2) - 2c \cdot \mathbb{E}\hat{\theta} \cdot \theta + \theta^2$$

Hence the optimal scale factor is $c = \mathbb{E}\hat{\theta} \cdot \theta / \mathbb{E}(\hat{\theta}^2)$.

We can achieve a better understanding of this optimal scale if we consider the bias-variance decomposition of $\hat{\theta}$. Let us define the bias and the variance as

$$B = \mathbb{E}\hat{\theta} - \theta$$
$$V = \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2$$

We then see that the optimal scale factor is

$$c = \frac{B + \theta}{V + (B + \theta)^2}\theta = 1 - \frac{V + B(B + \theta)}{V + (B + \theta)^2}.$$

This quantity is an "inflation factor", i.e., greater than or equal to unity, if $V+B(B+\theta) \leq 0$. This can be written as

$$V + B^2 \leq -B\theta$$

Hence, this condition can only hold if the bias B has opposite sign with θ. This would be the case for a *shrinkage estimator* θ. In that case, the condition could hold if the parameter θ has a large magnitude.

Returning to our main problem, ridge regression is a shrinkage estimator, and averages of ridge regression estimators are still shrinkage estimators. Therefore, it makes sense that their weighted average should be inflated to minimize mean squared error. This provides an intuitive explanation for why the weights sum to greater than one.

Appendix F. Proofs

F.1. Proof of Theorem 1

We can calculate the MSE of the weighted sum as

$$\begin{aligned} M(w) = \mathbb{E}\left\|\sum w_i\hat{\beta}_i - \beta\right\|^2 &= \mathbb{E}(\sum w_i\hat{\beta}_i - \beta)^\top(\sum w_j\hat{\beta}_j - \beta) \\ &= \sum_{ij} w_i w_j \cdot \mathbb{E}\hat{\beta}_i^\top\hat{\beta}_j - 2\sum_i w_i\mathbb{E}\hat{\beta}_i^\top\beta + \|\beta\|^2. \end{aligned}$$

Let $\widehat{B}$ be the $p \times k$ matrix defined as $\widehat{B} = [\hat{\beta}_1, \ldots, \hat{\beta}_k]$. Then we can write the above MSE as

$$M(w) = w^\top\mathbb{E}\widehat{B}^\top\widehat{B}w - 2\mathbb{E}\beta^\top\widehat{B}w + \|\beta\|^2.$$

Let also

$$B = \mathbb{E}\widehat{B} = [\mathbb{E}\hat{\beta}_1, \ldots, \mathbb{E}\hat{\beta}_k].$$

Since the local estimators are independent, we can write

$$M(w) = w^\top(B^\top B + R)w - 2\beta^\top Bw + \|\beta\|^2,$$

where R is a diagonal matrix with entries

$$R_i = \mathbb{E}\|\hat{\beta}_i\|^2 - \|\mathbb{E}\hat{\beta}_i\|^2 = \mathbb{E}\|\hat{\beta}_i - \mathbb{E}\hat{\beta}_i\|^2.$$

The objective function $M(w)$ can be viewed as corresponding to a k-parameter linear regression problem, with unknown parameters w_i, design matrix B and outcome vector β. Specifically, we regress β on $\mathbb{E}\widehat{B} = \mathbb{E}[\hat{\beta}_1, \ldots, \hat{\beta}_k]$. Therefore, the optimal weights are

$$w^* = (B^\top B + R)^{-1}B^\top\beta,$$

and the optimal risk equals

$$M^* = M(w^*) = \beta^\top\left[I_p - B(B^\top B + R)^{-1}B^\top\right]\beta.$$

Now, to find $B = \mathbb{E}\widehat{B}$, we need $\mathbb{E}\hat{\beta}_i$. The expectation of the ridge regression estimator for the full data set is

$$\mathbb{E}\hat{\beta}(\lambda) = \mathbb{E}(X^\top X + n\lambda I_p)^{-1}X^\top Y = (X^\top X + n\lambda I_p)^{-1}X^\top X\beta.$$

Letting $\widehat{\Sigma} = n^{-1}X^\top X$, this equals $\mathbb{E}\hat{\beta}(\lambda) = (\widehat{\Sigma} + \lambda I_p)^{-1}\widehat{\Sigma}\beta$. Similarly,

$$\mathbb{E}\hat{\beta}_i(\lambda_i) = (X_i^\top X_i + n_i\lambda_i I_p)^{-1}X_i^\top X_i\beta.$$

Let $Q_i = Q_i(\lambda_i) = (X_i^\top X_i + n_i\lambda_i I_p)^{-1}X_i^\top X_i$ be those matrices and let $\widehat{\Sigma}_i = n^{-1}X_i^\top X_i$. Then the above equals $Q_i = (\widehat{\Sigma}_i + \lambda_i I_p)^{-1}\widehat{\Sigma}_i$, and

$$B = [Q_1\beta; \ldots; Q_k\beta].$$

Therefore, $B^\top B$ has entries $\beta^\top Q_i Q_j \beta$, while $B^\top\beta$ has entries $\beta^\top Q_i\beta$. Moreover,

$$R_i = \mathbb{E}\|\hat{\beta}_i - \mathbb{E}\hat{\beta}_i\|^2 = \mathbb{E}\|(X_i^\top X_i + n_i\lambda_i I_p)^{-1}X_i^\top\varepsilon_i\|^2 = \sigma^2\,\mathrm{tr}[(X_i^\top X_i + n_i\lambda_i I_p)^{-2}X_i^\top X_i].$$

We can also write this as $R_i = n_i^{-1}\sigma^2\,\mathrm{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-2}\widehat{\Sigma}_i]$. To conclude the optimal risk, we have

$$M^*(k) = \|\beta\|^2 - v^\top(A + R)^{-1}v,$$

where

$$\begin{aligned} v &= B^\top\beta = vec[\beta^\top Q_i\beta], \\ A &= mat[\beta^\top Q_i Q_j\beta], \\ R &= \mathrm{diag}\left[n_i^{-1}\sigma^2\,\mathrm{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-2}\widehat{\Sigma}_i]\right], \\ Q_i &= (\widehat{\Sigma}_i + \lambda_i I_p)^{-1}\widehat{\Sigma}_i. \end{aligned}$$

Here we used the vectorization and to-matrix operators vec, mat. For the global MSE, we only need to consider the special case where $k = 1$, which gives us

$$\mathbb{E}||\hat{\beta} - \beta||^2 = M^*(1) = \|\beta\|^2 - \frac{(\beta^\top Q\beta)^2}{\beta^\top Q^2\beta + \sigma^2\,\mathrm{tr}[(X^\top X + n\lambda I_p)^{-2}X^\top X]},$$

where $Q = (\widehat{\Sigma} + \lambda I_p)^{-1}\widehat{\Sigma}$. This finishes the argument.

F.2. Proof of Theorem 2

The first step is to use the well-known concentration of quadratic forms to reduce to trace functionals (See e.g. Lemma C.3 of Dobriban and Wager (2018) which is based on Lemma B.26 of Bai and Silverstein (2010)). Since β is independent of the data X with mean zero and finite variance, under the moment assumptions imposed in the theorem, we have

$$\begin{aligned} \beta^\top Q_i\beta - \sigma^2\alpha^2/p \cdot \mathrm{tr}\, Q_i &\to_{a.s.} 0, \\ \beta^\top Q_i Q_j\beta - \sigma^2\alpha^2/p \cdot \mathrm{tr}\, Q_i Q_j &\to_{a.s.} 0, \\ \beta^\top Q_i^2\beta - \sigma^2\alpha^2/p \cdot \mathrm{tr}\, Q_i^2 &\to_{a.s.} 0. \end{aligned}$$

Let us compute the limits of v, A and R respectively.

1. Limit of v: First of all, we have already known that

$$\beta^\top Q_i \beta - \sigma^2\alpha^2/p \cdot \operatorname{tr} Q_i \to_{a.s.} 0,$$

so it is sufficient to consider the limit of $\operatorname{tr} Q_i/p$. Since

$$\operatorname{tr} Q_i/p = 1 - \lambda_i \operatorname{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-1}]/p.$$

assuming that the spectral distribution of $\widehat{\Sigma}_i$ converges almost surely to F_{γ_i}, we thus have

$$\operatorname{tr} Q_i/p \to_{a.s.} 1 - \lambda_i \mathbb{E}_{F_{\gamma_i}}(T + \lambda_i)^{-1} = 1 - \lambda_i m_{F_{\gamma_i}}(-\lambda_i).$$

Above we have introduced the Stieltjes transform $m_{F_{\gamma_i}}$ as a limiting object. So,

$$\beta^\top Q_i \beta \to_{a.s.} \sigma^2\alpha^2[1 - \lambda_i m_{F_{\gamma_i}}(-\lambda_i)].$$

For the form in terms of the population spectral distribution H, if $p/n \to \gamma$ and the spectral distribution of Σ converges to H, we have by the general Marchenko-Pastur (MP) theorem of Rubio and Mestre (Rubio and Mestre, 2011), that

$$(\widehat{\Sigma} + \lambda I)^{-1} \asymp (x_p \Sigma + \lambda I)^{-1},$$

where x_p is the unique positive solution of the fixed point equation

$$1 - x_p = \frac{x_p}{n} \operatorname{tr}\left[\Sigma(x_p\Sigma + \lambda I)^{-1}\right].$$

When $n, p \to \infty$, $x_p \to x$ and x satisfies the equation

$$1 - x = \gamma\left[1 - \lambda \int_0^\infty \frac{dH(t)}{xt + \lambda}\right].$$

We remark that the assumptions made in the theorem suffice for using the Rubio-Mestre result. Moreover, we only use a special case of their result, similar to Dobriban and Sheng (2018). Hence from the calculus of deterministic equivalents (Dobriban and Sheng, 2018), we can take the traces of the matrices in question to obtain

$$\operatorname{tr} Q_i/p = 1 - \lambda_i \operatorname{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-1}]/p \asymp 1 - \lambda_i \operatorname{tr}[(x_i\Sigma + \lambda_i I)^{-1}]/p \to_{a.s.} \mathbb{E}_H \frac{x_i T}{x_i T + \lambda_i},$$

where $x_i = x(H, \gamma_i, -\lambda_i)$ is the unique solution of

$$1 - x_i = \gamma_i\left[1 - \lambda_i \int_0^\infty \frac{dH(t)}{x_i t + \lambda_i}\right].$$

2. Limit of A: Let us consider the cases $i \neq j$ and $i = j$ separately.

(a) $i \neq j$: We begin by

$$\beta^\top Q_i Q_j \beta - \sigma^2 \alpha^2 / p \cdot \operatorname{tr} Q_i Q_j \to_{a.s.} 0.$$

Based on the above expression for Q_i, we have

$$Q_i Q_j = I_p - \lambda_i (\widehat{\Sigma}_i + \lambda_i I_p)^{-1} - \lambda_j (\widehat{\Sigma}_j + \lambda_j I_p)^{-1} + \lambda_i \lambda_j (\widehat{\Sigma}_i + \lambda_i I_p)^{-1} (\widehat{\Sigma}_j + \lambda_j I_p)^{-1}.$$

So the key will be to find the limit of

$$E_{ij} = p^{-1} \operatorname{tr}\{(\widehat{\Sigma}_i + \lambda_i I_p)^{-1} (\widehat{\Sigma}_j + \lambda_j I_p)^{-1}\}.$$

From the general MP theorem, since $p/n_i \to \gamma_i$, we have for all i,

$$(\widehat{\Sigma}_i + \lambda_i I_p)^{-1} \asymp (x_{ip} \Sigma + \lambda_i I_p)^{-1}.$$

Here x_{ip} is the unique positive solution of the fixed point equation

$$1 - x_{ip} = \frac{x_{ip}}{n_i} \operatorname{tr}\left[\Sigma (x_{ip} \Sigma + \lambda_i I)^{-1}\right],$$

and $x_{ip} \to x_i$ as $n_i, p \to \infty$. By the product rule of the calculus of deterministic equivalents, we have for $i \neq j$

$$(\widehat{\Sigma}_i + \lambda_i I_p)^{-1} (\widehat{\Sigma}_j + \lambda_j I_p)^{-1} \asymp (x_{ip} \Sigma + \lambda_i I_p)^{-1} (x_{jp} \Sigma + \lambda_j I_p)^{-1}.$$

Hence by the trace rule of deterministic equivalents,

$$E_{ij} \asymp p^{-1} \operatorname{tr}[(x_{ip} \Sigma + \lambda_i I_p)^{-1} (x_{jp} \Sigma + \lambda_j I_p)^{-1}]$$

Moreover, since the spectral distribution of Σ converges to H, we find for $i \neq j$

$$E_{ij} \to \mathbb{E}_H \frac{1}{(x_i T + \lambda_i)(x_j T + \lambda_j)}.$$

Putting it together,

$$Q_i Q_j \asymp I_p - \lambda_i (x_{ip} \Sigma + \lambda_i I_p)^{-1} - \lambda_j (x_{jp} \Sigma + \lambda_j I_p)^{-1} + \lambda_i \lambda_j (x_{ip} \Sigma + \lambda_i I_p)^{-1} (x_{jp} \Sigma + \lambda_j I_p)^{-1}.$$

So, again by the trace rule of deterministic equivalents, we have

$$\begin{aligned} p^{-1} \operatorname{tr}\{Q_i Q_j\} &\to_{a.s.} 1 - \mathbb{E}_H \frac{\lambda_i}{x_i T + \lambda_i} - \mathbb{E}_H \frac{\lambda_j}{x_j T + \lambda_j} + \mathbb{E}_H \frac{\lambda_i \lambda_j}{(x_i T + \lambda_i)(x_j T + \lambda_j)} \\ &= x_i x_j \mathbb{E}_H \frac{T^2}{(x_i T + \lambda_i)(x_j T + \lambda_j)}. \end{aligned}$$

Therefore, for $i \neq j$

$$A_{ij} \to \sigma^2 \alpha^2 \left[x_i x_j \mathbb{E}_H \frac{T^2}{(x_i T + \lambda_i)(x_j T + \lambda_j)} \right].$$

(b) $i = j$: In this case,

$$\beta^\top Q_i^2 \beta - \sigma^2\alpha^2/p \cdot \operatorname{tr} Q_i^2 \to 0,$$

where $Q_i^2 = I_p - 2\lambda_i(\widehat{\Sigma}_i + \lambda_i I_p)^{-1} + \lambda_i^2(\widehat{\Sigma}_i + \lambda_i I_p)^{-2}$. We can easily find the limit of $\operatorname{tr} Q_i^2/p$ in terms of empirical quantities, based on our knowledge of the convergence of Stieltjes transforms and its derivatives:

$$\operatorname{tr} Q_i^2/p \to 1 - 2\lambda_i m_{F_{\gamma_i}}(-\lambda_i) + \lambda_i^2 m'_{F_{\gamma_i}}(-\lambda_i).$$

Therefore, for $i = j$

$$A_{ii} \to \sigma^2\alpha^2[1 - 2\lambda_i m_{F_{\gamma_i}}(-\lambda_i) + \lambda_i^2 m'_{F_{\gamma_i}}(-\lambda_i)].$$

We can also express the limit of A_{ii} in terms of the population spectral distribution H by using Theorem 11. For our purpose, let $T = \Sigma$, $S = \widehat{\Sigma}$, while

$$f(z, T) = (x_p T - zI)^{-1},$$
$$g(z, S) = (S - zI_p)^{-1}.$$

From Rubio and Mestre (2011), we know that for each $z \in D := \mathbb{C}\backslash\mathbb{R}^+$, $f(z, \Sigma) \asymp g(z, \widehat{\Sigma})$. x_p is defined as

$$x_p = \frac{1}{n}\operatorname{tr}[(I + \frac{p}{n}e_p I)^{-1}] = \frac{1}{1 + (p/n)e_p} = \frac{1}{1 + \gamma_p e_p},$$

and $e_p = e_p(z)$ is the Stieltjes transform of a certain positive measure on $\mathbb{R}^+$, obtained as the unique solution of the equation

$$e_p = \frac{1}{p}\operatorname{tr}[\Sigma(x_p\Sigma - zI_p)^{-1}].$$

It is well-known that $x_p(z), e_p(z)$ are both analytic functions on D. Then we can check that the conditions of theorem 11 hold in this case. First of all, for an invertible matrix A, $A^{-1} = (\det A)^{-1}A^*$, where A^* is the adjugate matrix of A. Since x_p is analytic, it is easy to verify that $\det(x_p\Sigma - zI_p), \det(\widehat{\Sigma} - zI_p)$ and all entries of $(x_p\Sigma - zI_p)^*, (\widehat{\Sigma} - zI_p)^*$ are analytic functions of z. So the entries of $f(z, \Sigma)$ and $g(z, \widehat{\Sigma})$ are analytic in D.
Next, we want to bound

$$\operatorname{tr}[C_n((x_p\Sigma - zI_p)^{-1} - (\widehat{\Sigma} - zI_p)^{-1})] \le ||C_n||_{\operatorname{tr}} \cdot ||(x_p\Sigma - zI_p)^{-1} - (\widehat{\Sigma} - zI_p)^{-1}||_2.$$

For a fixed $\delta > 0$, let us define a domain $D_\delta := \{z \in D : \operatorname{Re} z < -\delta\} \cup \{z \in D : |\operatorname{Im} z| > \delta\}$. Then, it is sufficient to find a uniform bound for $||(x_p\Sigma - zI_p)^{-1} - (\widehat{\Sigma} - zI_p)^{-1}||_2$ on D_δ. In fact, we can bound $||(x_p\Sigma - zI_p)^{-1}||_2$ and $||(\widehat{\Sigma} - zI_p)^{-1}||_2$ separately.

i. Bounding $||(\widehat{\Sigma} - zI_p)^{-1}||_2$:

$$||(\widehat{\Sigma} - zI_p)^{-1}||_2 = \sigma_{\max}((\widehat{\Sigma} - zI_p)^{-1}) = \max_i \frac{1}{|\hat{l}_i - z|},$$

where $\hat{l}_i$ is the i-th eigenvalue of $\widehat{\Sigma}$. Since $\hat{l}_i$ is always non-negative, we have

$$\frac{1}{|\hat{l}_i - z|} = \frac{1}{|\hat{l}_i - \mathrm{Re}z - i\mathrm{Im}z|} = \frac{1}{\sqrt{(\hat{l}_i - \mathrm{Re}z)^2 + (\mathrm{Im}z)^2}} \leq \frac{1}{\delta}.$$

ii. Bounding $||(x_p\Sigma - zI_p)^{-1}||_2$:
In this case, we need to use the properties of e_p and x_p. Recall that e_p is the Stieltjes transform of a certain measure on $\mathbb{R}^+$, i.e.

$$\begin{aligned} e_p(z) &= \int_0^\infty \frac{1}{t-z} d\mu(t) = \int_0^\infty \frac{1}{t - \mathrm{Re}z - i\mathrm{Im}z} d\mu(t) \\ &= \int_0^\infty \frac{t - \mathrm{Re}z}{(t-\mathrm{Re}z)^2 + (\mathrm{Im}z)^2} d\mu(t) + i\int_0^\infty \frac{\mathrm{Im}z}{(t-\mathrm{Re}z)^2 + (\mathrm{Im}z)^2} d\mu(t). \end{aligned}$$

So

$$\begin{aligned} x_p &= \frac{1}{1+\gamma_p e_p} = \frac{1}{1 + \gamma_p \mathrm{Re}(e_p) + i\gamma_p \mathrm{Im}(e_p)} \\ &= \frac{1 + \gamma_p \mathrm{Re}(e_p)}{(1+\gamma_p \mathrm{Re}(e_p))^2 + (\gamma_p \mathrm{Im}(e_p))^2} - i\frac{\gamma_p \mathrm{Im}(e_p)}{(1+\gamma_p \mathrm{Re}(e_p))^2 + (\gamma_p \mathrm{Im}(e_p))^2}. \end{aligned}$$

When $z \in D_\delta$, we can check that $\mathrm{Re}(x_p) > 0$. Meanwhile, $\mathrm{Im}(x_p)$ and $\mathrm{Im}(z)$ have opposite signs.
Now, let us consider

$$||(x_p\Sigma - zI_p)^{-1}||_2 = \sigma_{\max}((x_p\Sigma - zI_p)^{-1}) = \max_k \frac{1}{|x_p l_k - z|},$$

where l_k is the k-th eigenvalue of Σ. Since l_k is non-negative, we have

$$\begin{aligned} \frac{1}{|x_p l_k - z|} &= \frac{1}{|l_k \mathrm{Re}(x_p) + i l_k \mathrm{Im}(x_p) - \mathrm{Re}z - i\mathrm{Im}z|} \\ &= \frac{1}{\sqrt{(l_k \mathrm{Re}(x_p) - \mathrm{Re}z)^2 + (l_k \mathrm{Im}(x_p) - \mathrm{Im}z)^2}} \\ &\leq \frac{1}{\delta}. \end{aligned}$$

Finally, since δ is arbitrary, we can conclude that $f'(z,\Sigma) \asymp g'(z,\widehat{\Sigma})$ for all $z \in D$. Then let us compute the derivatives. For invertible $A = A(z)$, we have

$$\frac{d(A^{-1})}{dz} = -A^{-1}\frac{dA}{dz}A^{-1},$$

where the derivative is entry-wise. Thus

$$\begin{aligned} &f'(z,T) = -(x_pT - zI)^{-1}(x_p'T - I)(x_pT - zI_p)^{-1} = -(x_pT - zI_p)^{-2}(x_p'T - I), \\ &g'(z,S) = (S - zI_p)^{-2}. \end{aligned}$$

Next, we need to calculate $x' = dx/dz$, where $x(z)$ is the limit of $x_p(z)$. In fact, by looking at the expression of $x_p(z)$, it is not hard to find that $x_p(z)$ is uniformly bounded on D. By using a similar argument, we have $x_p' \to x'$ on D. To find x', let us start from the following fixed-point equation

$$1 - x = \gamma \left[1 + z\mathbb{E}_H \frac{1}{xT - z} \right].$$

Take derivatives on both sides to get

$$\begin{aligned} -x' &= \gamma \left[z\mathbb{E}_H \frac{1}{xT - z} \right]' \\ -x' &= \gamma \left[\mathbb{E}_H \frac{1}{xT - z} + \mathbb{E}_H \frac{z - zTx'}{(xT - z)^2} \right] \\ x' \left[-1 + \gamma z \mathbb{E}_H \frac{T}{(xT - z)^2} \right] &= \gamma \mathbb{E}_H \frac{xT}{(xT - z)^2} \\ x' &= \frac{\gamma \mathbb{E}_H \frac{xT}{(xT-z)^2}}{-1 + \gamma z \mathbb{E}_H \frac{T}{(xT-z)^2}}. \end{aligned}$$

Therefore we obtain

$$\begin{aligned} (\widehat{\Sigma} - zI)^{-2} &\asymp (x_p\Sigma - zI_p)^{-2}(I - x_p'\Sigma) \\ p^{-1} \operatorname{tr}(\widehat{\Sigma} - zI)^{-2} &\asymp -x_p' p^{-1} \operatorname{tr}[\Sigma(x_p\Sigma - zI)^{-2}] + p^{-1} \operatorname{tr}[(x_p\Sigma - zI_p)^{-2}] \\ &\to \frac{\gamma \mathbb{E}_H \frac{xT}{(xT-z)^2}}{1 - \gamma z \mathbb{E}_H \frac{T}{(xT-z)^2}} \mathbb{E}_H \frac{T}{(xT - z)^2} + \mathbb{E}_H \frac{1}{(xT - z)^2} \\ &= \frac{\gamma x \left(\mathbb{E}_H \frac{T}{(xT-z)^2} \right)^2}{1 - \gamma z \mathbb{E}_H \frac{T}{(xT-z)^2}} + \mathbb{E}_H \frac{1}{(xT - z)^2}. \end{aligned}$$

Now, let $z = -\lambda$ and then we will have

$$\begin{aligned} (\widehat{\Sigma} + \lambda I)^{-2} &\asymp (x_p\Sigma + \lambda I)^{-2}(I - x_p'\Sigma) \\ p^{-1} \operatorname{tr}(\widehat{\Sigma} + \lambda I)^{-2} &\to \frac{\gamma x \left(\mathbb{E}_H \frac{T}{(xT+\lambda)^2} \right)^2}{1 + \gamma \lambda \mathbb{E}_H \frac{T}{(xT+\lambda)^2}} + \mathbb{E}_H \frac{1}{(xT + \lambda)^2}. \end{aligned}$$

Finally, we can simply replace $\widehat{\Sigma}, \lambda, \gamma, x$ by $\widehat{\Sigma}_i, \lambda_i, \gamma_i, x_i$ to get the desired results.

3. Limit of R: Recall that $R_i = n_i^{-1}\sigma^2 \operatorname{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-2}\widehat{\Sigma}_i]$. We note $p^{-1} \operatorname{tr}(\widehat{\Sigma} + \lambda I)^{-2} \to m_{F_\gamma}'(-\lambda)$ and $\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-2} = (\widehat{\Sigma} + \lambda I)^{-1} - \lambda(\widehat{\Sigma} + \lambda I)^{-2}$, so

$$\frac{\operatorname{tr}[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-2}]}{n} \to \gamma[m_{F_\gamma}(-\lambda) - \lambda m_{F_\gamma}'(-\lambda)].$$

Hence

$$R_{ii} \to \sigma^2 \left[\gamma_i [m_{F_{\gamma_i}}(-\lambda_i) - \lambda m_{F_{\gamma_i}}'(-\lambda_i)] \right].$$

Next, we find a limit in terms of population parameters

$$\begin{aligned}
\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-2} &= (\widehat{\Sigma}+\lambda I)^{-1} - \lambda(\widehat{\Sigma}+\lambda I)^{-2} \\
&\asymp (x_p\Sigma+\lambda I)^{-1} - \lambda(x_p\Sigma+\lambda I)^{-2}(I - x_p'\Sigma) \\
p^{-1}\operatorname{tr}\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-2} &\asymp p^{-1}\operatorname{tr}(x_p\Sigma+\lambda I)^{-1} - \lambda p^{-1}\operatorname{tr}\left[(I-x_p'\Sigma)(x_p\Sigma+\lambda I)^{-2}\right] \\
&\to \mathbb{E}_H\frac{1}{xT+\lambda} - \lambda\frac{\gamma x\left(\mathbb{E}_H\frac{T}{(xT+\lambda)^2}\right)^2}{1+\gamma\lambda\mathbb{E}_H\frac{T}{(xT+\lambda)^2}} - \mathbb{E}_H\frac{\lambda}{(xT+\lambda)^2} \\
&= \mathbb{E}_H\frac{xT}{(xT+\lambda)^2} - \lambda\frac{\gamma x\left(\mathbb{E}_H\frac{T}{(xT+\lambda)^2}\right)^2}{1+\gamma\lambda\mathbb{E}_H\frac{T}{(xT+\lambda)^2}} \\
&= \frac{x\mathbb{E}_H\frac{T}{(xT+\lambda)^2}}{1+\lambda\gamma\mathbb{E}_H\frac{T}{(xT+\lambda)^2}},
\end{aligned}$$

where we used the differentiation rule of the calculus of deterministic equivalents. Hence we finally find the limit

$$R_{ii} \to \sigma^2\left[\frac{x_i\mathbb{E}_H\frac{T}{(x_iT+\lambda_i)^2}}{1+\lambda_i\gamma_i\mathbb{E}_H\frac{T}{(x_iT+\lambda_i)^2}}\right].$$

F.3. Proof of Theorem 3

Notice that, when the samples are equally distributed and we use the same tuning parameter λ for all the local estimators, a direct consequence is that $x_i = x_j = x$ for all i, j, where x is the unique solution of the following fixed point equation

$$1-x = k\gamma\left[1-\lambda\int_0^\infty\frac{dH(t)}{xt+\lambda}\right] = k\gamma\left[1-\mathbb{E}_H\frac{\lambda}{xT+\lambda}\right] = k\gamma(1-\lambda m_{F_{k\gamma}}(-\lambda)) = k\gamma(1-\lambda m).$$

In this case, we can express $\mathcal{A}_{ij}$ as

$$\mathcal{A}_{ij} = \sigma^2\alpha^2\mathbb{E}_H\frac{(xT)^2}{(xT+\lambda)^2} = \sigma^2\alpha^2\int\frac{(xt)^2}{(xt+\lambda)^2}dH(t).$$

In order to express $\mathcal{A}_{ij}$ in terms of the sample quantities, we can start from the following equality

$$\int\frac{1}{xt+\lambda}dH(t) = m.$$

Take derivatives with respect to λ, we have

$$\int\frac{x't+1}{(xt+\lambda)^2}dH(t) = m'.$$

Rearrange terms, we have

$$\int\frac{x't+1}{(xt+\lambda)^2}dH(t) = \int\frac{(xt+\lambda-\lambda)\cdot\frac{x'}{x}+1}{(xt+\lambda)^2}dH(t) = \frac{x'}{x}m + \left(1-\frac{\lambda x'}{x}\right)\int\frac{1}{(xt+\lambda)^2}dH(t) = m'.$$

On the other hand, take derivatives with respect to λ on the fixed point equation for x gives us

$$x' = k\gamma(m - \lambda m').$$

So

$$\int \frac{1}{(xt+\lambda)^2} dH(t) = \frac{xm' - x'm}{x - \lambda x'} = \frac{(1-k\gamma)m' + 2k\gamma\lambda mm' - k\gamma m^2}{1 - k\gamma + k\gamma\lambda^2 m'}.$$

Then we have

$$\begin{aligned}\int \frac{(xt)^2}{(xt+\lambda)^2} dH(t) &= \int \frac{(xt+\lambda-\lambda)^2}{(xt+\lambda)^2} dH(t) \\ &= 1 - 2\lambda m + \lambda^2 \int \frac{1}{(xt+\lambda)^2} dH(t) \\ &= 1 - 2\lambda m + \lambda^2 m' - \frac{k\gamma\lambda^2(m-\lambda m')^2}{1-k\gamma+k\gamma\lambda m'}.\end{aligned}$$

Now we the expressions for $V, \mathcal{A}, \mathcal{R}$, we also know $\mathcal{W}_k^* = (\mathcal{A}+\mathcal{R})^{-1}V$ and $\mathcal{M}_k = \sigma^2\alpha^2 - V^\top(\mathcal{A}+\mathcal{R})^{-1}V$. By using the auxiliary functions $\mathcal{F}, \mathcal{G}$ defined in the theorem, we have

$$\mathcal{M}_k = \sigma^2\alpha^2 - \frac{1}{\mathcal{G}} V^\top (\mathbf{1}\cdot\mathbf{1}^\top + \mathrm{diag}(\mathcal{F}/\mathcal{G}))^{-1} V = \sigma^2\alpha^2 - \frac{(\sigma^2\alpha^2(1-\lambda m))^2}{\mathcal{G}} \mathbf{1}^\top (\mathbf{1}\cdot\mathbf{1}^\top + \mathrm{diag}(\mathcal{F}/\mathcal{G}))^{-1}\mathbf{1},$$

where $\mathbf{1} = (1, 1, \cdots, 1)^\top$ is the all-one vector. Then similar to the proof of Theorem 2, we can use the Sherman-Morrison formula to simply the expression, this leads to

$$\begin{aligned}\mathcal{M}_k &= \sigma^2\alpha^2 - \frac{(\sigma^2\alpha^2(1-\lambda m))^2}{\mathcal{G}} \mathbf{1}^\top \left(\mathrm{diag}(\mathcal{F}/\mathcal{G})^{-1} - \frac{\mathrm{diag}(\mathcal{F}/\mathcal{G})^{-1}\mathbf{1}\cdot\mathbf{1}^\top \mathrm{diag}(\mathcal{F}/\mathcal{G})^{-1}}{1 + \mathbf{1}^\top \mathrm{diag}(\mathcal{F}/\mathcal{G})^{-1}\mathbf{1}} \right) \mathbf{1} \\ &= \sigma^2\alpha^2 - \frac{(\sigma^2\alpha^2(1-\lambda m))^2}{\mathcal{G}} \left(\frac{k\mathcal{G}}{\mathcal{F}} - \frac{k^2\mathcal{G}^2/\mathcal{F}^2}{1 + k\mathcal{G}/\mathcal{F}} \right) \\ &= \sigma^2\alpha^2 \left(1 - \frac{\sigma^2\alpha^2(1-\lambda m)^2 k}{\mathcal{F} + k\mathcal{G}} \right).\end{aligned}$$

Similarly, we can express the optimal weights $\mathcal{W}_k^*$ as

$$\begin{aligned}\mathcal{W}_k^* &= \frac{1}{\mathcal{G}} (\mathbf{1}\cdot\mathbf{1}^\top + \mathrm{diag}(\mathcal{F}/\mathcal{G}))^{-1} V \\ &= \frac{\sigma^2\alpha^2(1-\lambda m)}{\mathcal{G}} (\mathbf{1}\cdot\mathbf{1}^\top + \mathrm{diag}(\mathcal{F}/\mathcal{G}))^{-1}\mathbf{1} \\ &= \frac{\sigma^2\alpha^2(1-\lambda m)}{\mathcal{G}} \left(\mathrm{diag}(\mathcal{F}/\mathcal{G})^{-1} - \frac{\mathcal{G}^2/\mathcal{F}^2}{1+k\mathcal{G}/\mathcal{F}} \mathbf{1}\cdot\mathbf{1}^\top \right) \mathbf{1} \\ &= \frac{\sigma^2\alpha^2(1-\lambda m)}{\mathcal{F}+k\mathcal{G}} \mathbf{1}.\end{aligned}$$

Why do we need to assume the samples are uniformly distributed across machines? This is a technical assumption. The key difficulty in analyzing MSE and optimal weights comes from the off-diagonal entries of $\mathcal{A}$:

$$\mathcal{A}_{ij} = \sigma^2\alpha^2 \mathbb{E}_H \frac{x_i x_j T^2}{(x_i T + \lambda_i)(x_j T + \lambda_j)}.$$

Here, x_i is uniquely determined by the aspect ratio $\gamma_i = p/n_i$, the tuning parameter λ_i, and the population distribution H. The terms we can handle are of form

$$\mathbb{E}_H \frac{(x_i T)^k}{(x_i T + \lambda_i)^l}, \quad k, l \in \mathbb{N}^*.$$

We can calculate these by taking derivatives of the Stieltjes transform $\mathbb{E}_H \frac{1}{x_i T + \lambda_i}$ and doing further calculations.

We may decompose $\mathcal{A}_{ij}$ into

$$\mathcal{A}_{ij} = \sigma^2 \alpha^2 \left(1 - \lambda_i \mathbb{E}_H \frac{1}{x_i T + \lambda_i} - \lambda_j \mathbb{E}_H \frac{1}{x_j T + \lambda_j} + \lambda_i \lambda_j \mathbb{E}_H \frac{1}{(x_i T + \lambda_i)(x_j T + \lambda_j)}\right).$$

In general, it is hard to further simplify the last term of the above decomposition. But if we assume the subsample sizes are all equal, then the optimal tuning parameters λ_i should also be equal by symmetry. Thus, all x_i are the same and we can rewrite the last term into

$$\lambda_i^2 \mathbb{E}_H \frac{1}{(x_i T + \lambda_i)^2},$$

which is something we can deal with as mentioned above. Besides, from a practical perspective, if the samples are not uniformly distributed, we may need to do much more rounds of communication (compare to Algorithm 1) to find the optimal tuning parameters λ_i and the optimal weights ω_i.

F.4. Proof of Theorem 4

The proof for v and R is clear by Theorem 2. For the limit of A, the diagonal case is also direct. When $i \neq j$, recall that

$$E_{ij} = p^{-1} \operatorname{tr}\{(\widehat{\Sigma}_i + \lambda_i I_p)^{-1} (\widehat{\Sigma}_j + \lambda_j I_p)^{-1}\} \to \mathbb{E}_H \frac{1}{(x_i T + \lambda_i)(x_j T + \lambda_j)}.$$

For $H = \delta_1$, the expectation decouples, we find

$$E_{ij} \to \frac{1}{x_i + \lambda_i} \cdot \frac{1}{x_i + \lambda_j} = m_{\gamma_i}(-\lambda_i) m_{\gamma_j}(-\lambda_j).$$

Therefore,

$$A_{ij} \to \sigma^2 \alpha^2 [1 - \lambda_i m_{\gamma_i}(-\lambda_i)] \cdot [1 - \lambda_j m_{\gamma_j}(-\lambda_j)].$$

Now let us put everything together. Recall that the optimal risk has the form $\mathrm{MSE}^*_{dist} = \|\beta\|^2 - v^\top (A + R)^{-1} v$. Based on the above discussion, we have

$$\sigma^2 \alpha^2 (A + R) \to \sigma^2 \alpha^2 (\mathcal{A} + \mathcal{R}) = V V^\top + D,$$

where D is a diagonal matrix with i-th diagonal entry $\sigma^2 \alpha^2 (\mathcal{R}_{ii} + \mathcal{A}_{ii}) - V_i^2$. Then, by using the Sherman−Morrison formula, we have

$$V^\top (V V^\top + D)^{-1} V = \frac{V^\top D^{-1} V}{1 + V^\top D^{-1} V}.$$

So the limiting distributed risk is

$$\mathcal{M}_k = \sigma^2\alpha^2 - \sigma^2\alpha^2\frac{V^\top D^{-1}V}{1+V^\top D^{-1}V} = \frac{\sigma^2\alpha^2}{1+V^\top D^{-1}V} = \frac{\sigma^2\alpha^2}{1+\sum_{i=1}^k \frac{V_i^2}{D_i}},$$

which finishes the proof.

F.5. Proof of Proposition 5

Recall that

$$\begin{aligned}\frac{V_i^2}{D_i} &= \frac{\sigma^4\alpha^4(1-\lambda_i m_{\gamma_i}(-\lambda_i))^2}{\sigma^4\alpha^4\lambda_i^2[m'_{\gamma_i}(-\lambda_i) - m^2_{\gamma_i}(-\lambda_i)] + \sigma^4\alpha^2\gamma_i[m_{\gamma_i}(-\lambda_i) - \lambda_i m'_{\gamma_i}(-\lambda_i)]} \\ &= \frac{\alpha^2(1-\lambda_i m_{\gamma_i}(-\lambda_i))^2}{\alpha^2\lambda_i^2[m'_{\gamma_i}(-\lambda_i) - m^2_{\gamma_i}(-\lambda_i)] + \gamma_i[m_{\gamma_i}(-\lambda_i) - \lambda_i m'_{\gamma_i}(-\lambda_i)]},\end{aligned}$$

and our goal is to find λ_i that maximizes V_i^2/D_i. Luckily, from Dobriban and Wager (2018) it follows that for $k=1$, i.e. when there is only one machine, the optimal choice of the tuning parameter λ is γ/α^2. This means that the maximizer of V^2/D is $\lambda = \gamma/\alpha^2$. Now, due to the decoupled structure of $\mathcal{M}_k$, the optimal tuning parameters are $\lambda_i = \gamma_i/\alpha^2$. Plugging in the parameters, we have

$$\frac{V_i^2}{D_i} = \frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)} - 1.$$

Then the optimal risk can be simplified to

$$\mathcal{M}_k = \frac{\sigma^2\alpha^2}{1+\sum_{i=1}^k\left[\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)} - 1\right]}.$$

When $k=1$, this equals to $\sigma^2\gamma m_\gamma(-\gamma/\alpha^2)$ which matches the known result from Dobriban and Wager (2018).

F.6. Proof of Proposition 6

The explicit form is easy to derive by plugging $z = -\gamma/\alpha^2$ into the formula of $m_\gamma(z)$. Next, we can check monotonicity by computing $\phi'(\gamma)$:

$$\phi'(\gamma) = \frac{\alpha^2}{2\gamma^2}\left(1 + \frac{(1-1/\alpha^2)\gamma - 1}{\sqrt{[(1-1/\alpha^2)\gamma-1]^2 + 4\gamma^2/\alpha^2}}\right) > 0.$$

Finally, for the convexity, let us consider the two cases separately.

1. $\alpha \le 1$: With some effort, we find the second derivative of ϕ

$$\phi''(\gamma) = \frac{\alpha^2\left(\frac{2\gamma^2}{\alpha^2} - \left(((1-\frac{1}{\alpha^2})\gamma-1)^2 + \frac{4\gamma^2}{\alpha^2}\right)\left((1-\frac{1}{\alpha^2})\gamma-1\right) - \left(((1-\frac{1}{\alpha^2})\gamma-1)^2 + \frac{4\gamma^2}{\alpha^2}\right)^{3/2}\right)}{\gamma^3[((1-1/\alpha^2)\gamma-1)^2 + 4\gamma^2/\alpha^2]^{3/2}}.$$

To analyze the second derivative, it is helpful to denote $1-(1-\frac{1}{\alpha^2})\gamma$ by Δ. Clearly, in this case, $\Delta \geq 1$. Then we can rewrite ϕ'' as

$$\begin{aligned}
\phi''(\gamma) &= \frac{\alpha^2}{\gamma^3[\Delta^2+4\gamma^2/\alpha^2]^{3/2}}\left(\frac{2\gamma^2}{\alpha^2}+(\Delta^2+\frac{4\gamma^2}{\alpha^2})\Delta-(\Delta^2+\frac{4\gamma^2}{\alpha^2})^{3/2}\right)\\
&= \frac{\alpha^2}{\gamma^3[\Delta^2+4\gamma^2/\alpha^2]^{3/2}}\left(\frac{2\gamma^2}{\alpha^2}+(\Delta^2+\frac{4\gamma^2}{\alpha^2})\left(\Delta-\sqrt{\Delta^2+\frac{4\gamma^2}{\alpha^2}}\right)\right)\\
&= \frac{\alpha^2}{\gamma^3[\Delta^2+4\gamma^2/\alpha^2]^{3/2}}\left(\frac{2\gamma^2}{\alpha^2}-\frac{4\gamma^2}{\alpha^2}\cdot\frac{\Delta^2+4\gamma^2/\alpha^2}{\Delta+\sqrt{\Delta^2+\frac{4\gamma^2}{\alpha^2}}}\right)\\
&= \frac{\alpha^2}{\gamma^3[\Delta^2+4\gamma^2/\alpha^2]^{3/2}}\left(\frac{2\gamma^2}{\alpha^2}-\frac{4\gamma^2}{\alpha^2}\cdot\frac{\sqrt{\Delta^2+4\gamma^2/\alpha^2}}{\Delta+\sqrt{\Delta^2+\frac{4\gamma^2}{\alpha^2}}}\right)\\
&\leq \frac{\alpha^2}{\gamma^3[\Delta^2+4\gamma^2/\alpha^2]^{3/2}}\left(\frac{2\gamma^2}{\alpha^2}-\frac{4\gamma^2}{\alpha^2}\cdot\frac{1}{2}\right)=0.
\end{aligned}$$

Thus, $\phi(\gamma)$ is always concave in this case.

2. $\alpha > 1$: Here we can consider the Taylor expansion of ϕ'' near the origin. We can check that $\phi''(\gamma) = 2(1-1/\alpha^2)\gamma^3 + o(\gamma^3)$ as $\gamma \to 0$, which means $\phi''(\gamma) > 0$ for small γ. When γ is very large, we can immediately see that $\phi''(\gamma) < 0$, since the leading order in the numerator of $\phi''(\gamma)$ is $-\gamma^3$. Then the desired result follows.

F.7. Proof of Theorem 7

For the first property, minimizing the ARE is equivalent to maximizing the following quantity

$$\sum_{i=1}^{k}\frac{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)}{\alpha^2}=\sum_{i=1}^{k}\frac{\phi(\gamma_i)}{\alpha^2}.$$

It is helpful to introduce $r(t)=\phi(\gamma)$, where $t=1/\gamma$. We can easily compute that

$$r'(t)=\frac{\alpha^2}{2}\left(-1+\frac{t-(1-1/\alpha^2)}{\sqrt{(t-(1-1/\alpha^2))^2+4/\alpha^2}}\right)<0\ \ ,r''(t)=\frac{2}{[(t-(1-1/\alpha^2))^2+4/\alpha^2]^{3/2}}>0.$$

Thus, $r(t)$ is a decreasing and convex function. We can show the ARE achieves minimum when the samples are equally distributed by considering the following optimization problem

$$\begin{aligned}
\max_{t_i} \quad & \sum_{i=1}^{k}\frac{r(t_i)}{\alpha^2}\\
\text{subject to} \quad & \sum_{i=1}^{k}t_i=\frac{1}{\gamma},\\
& t_i\geq 0, i=1,2,\ldots,k.
\end{aligned}$$

We denote the objective by $R(t_1, \ldots, t_k)$, and the corresponding Lagrangian by $R_\xi = R - \xi(\sum_i t_i - 1/\gamma)$. Then it is easy to check that the condition $\frac{\partial R_\xi}{\partial t_i} = 0$ reduces to

$$\frac{r'(t_i)}{\alpha^2} - \xi = 0, \ i = 1, 2, \ldots, k.$$

Since $r'(t)$ is also monotone, the unique solution to the stationary condition is $t_1 = t_2 = \cdots = t_k = 1/(k\gamma)$. If some t_i equals to 0, then it reduces to a problem with $k-1$ machines. So it remains to check the boundary case where only one t_i is non-zero and equals to $1/\gamma$. Obviously, this is the trivial case where the ARE is 1. Therefore, we have shown that the ARE attains its minimum when the samples are equally distributed across k machines.

Next, for fixed α^2 and γ, we can check

$$\frac{\partial \psi}{\partial k} = \frac{\gamma m_\gamma(-\gamma/\alpha^2)}{\alpha^2}\left(\frac{\alpha^2}{2\gamma} \cdot \frac{\left(\gamma/\alpha^2+\gamma\right)^2 k + \gamma/\alpha^2 - \gamma}{\sqrt{\left(\gamma/\alpha^2+\gamma\right)^2 k^2 + 2\left(\gamma/\alpha^2 - \gamma\right)k + 1}} - \frac{1+\alpha^2}{2}\right) \leq 0.$$

Moreover, the limit of ψ is

$$\begin{aligned} h(\alpha^2, \gamma) = \lim_{k\to\infty} \psi(k, \gamma, \alpha^2) &= \frac{\gamma m_\gamma(-\gamma/\alpha^2)}{\alpha^2}\left(1 + \frac{\alpha^2}{\gamma(1+\alpha^2)}\right) \\ &= \frac{-\gamma/\alpha^2 + \gamma - 1 + \sqrt{(-\gamma/\alpha^2+\gamma-1)^2 + 4\gamma^2/\alpha^2}}{2\gamma}\left(1 + \frac{\alpha^2}{\gamma(1+\alpha^2)}\right). \end{aligned}$$

Then for fixed α^2, we can differentiate $h(\alpha^2, \gamma)$ with respect to γ:

$$\begin{aligned} \frac{\partial h}{\partial \gamma} = &-\frac{\alpha^2}{\gamma^2(1+\alpha^2)} \cdot \frac{-\gamma/\alpha^2 + \gamma - 1 + \sqrt{(-\gamma/\alpha^2+\gamma-1)^2 + 4\gamma^2/\alpha^2}}{2\gamma} \\ &+ \left(1 + \frac{\alpha^2}{\gamma(1+\alpha^2)}\right) \cdot \frac{1 - 1/\alpha^2 + \frac{(-\gamma/\alpha^2+\gamma-1)(1-1/\alpha^2)+4\gamma/\alpha^2}{\sqrt{(-\gamma/\alpha^2+\gamma-1)^2+4\gamma^2/\alpha^2}}}{2\gamma} \\ &- \left(1 + \frac{\alpha^2}{\gamma(1+\alpha^2)}\right) \cdot \frac{-\gamma/\alpha^2 + \gamma - 1 + \sqrt{(-\gamma/\alpha^2+\gamma-1)^2 + 4\gamma^2/\alpha^2}}{2\gamma^2}. \end{aligned}$$

After tedious calculation, we find $\frac{\partial h}{\partial \gamma} \geq 0$. Finally, we can evaluate the limit of h as $\gamma \to 0$ and $\gamma \to \infty$

$$\lim_{\gamma\to 0} h(\alpha^2, \gamma) = \frac{1}{1+\alpha^2}, \quad \lim_{\gamma\to\infty} h(\alpha^2, \gamma) = 1.$$

On the other hand, for fixed γ, we can check that h is a decreasing function of α^2 and

$$\lim_{\alpha^2\to 0} h(\alpha^2, \gamma) = 1, \quad \lim_{\alpha^2\to\infty} h(\alpha^2, \gamma) = \begin{cases} 1 - \frac{1}{\gamma^2}, & \gamma > 1, \\ 0, & 0 < \gamma \leq 1. \end{cases}$$

F.8. Proof of Theorem 8

Recall that the optimal weights are $w^* = (A+R)^{-1}v$ and $\sigma^2\alpha^2(A+R) \to VV^\top + D$. Denote the limit of the optimal weights by W, so that we have

$$W = \sigma^2\alpha^2(VV^\top + D)^{-1}V = \frac{\sigma^2\alpha^2 D^{-1}V}{1 + V^\top D^{-1}V}.$$

When we choose $\lambda_i = \gamma_i/\alpha^2$ for each i, we can write the limiting optimal weights as

$$W = \mathcal{M}_k \cdot D^{-1} V.$$

So, it follows from the formulas of $\mathcal{M}_k, D$ and V that

$$W_i = \left(\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)}\right) \cdot \left(\frac{1}{1+\sum_{i=1}^k \left[\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)} - 1\right]}\right).$$

For the sum of the coordinates, we have

$$1^\top W = \frac{\sum_{i=1}^k \left(\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)}\right)}{1+\sum_{i=1}^k \left[\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)} - 1\right]} = \frac{\sum_{i=1}^k \left(\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)}\right)}{1-k+\sum_{i=1}^k \left(\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)}\right)} \geq 1.$$

In the special case where all γ_i are equal, i.e., $\gamma_i = k\gamma$, we have all W_i equal to

$$W_i = \frac{\frac{\alpha^2}{k\gamma \cdot m_{k\gamma}(-k\gamma/\alpha^2)}}{1-k+\frac{\alpha^2}{\gamma \cdot m_{k\gamma}(-k\gamma/\alpha^2)}} = \frac{1}{k+(1-k)\cdot k\gamma/\alpha^2 \cdot m_{k\gamma}(-k\gamma/\alpha^2)}.$$

In terms of the optimal risk function $\phi(\gamma) = \phi(\gamma, \alpha) = \gamma m_\gamma(-\gamma/\alpha^2)$ defined before, this can also be written as the following optimal weight function

$$\mathcal{W}(k, \gamma, \alpha) = \frac{1}{k-(k-1)\cdot \phi(k\gamma)/\alpha^2}.$$

The monotonicity and the limits of $\mathcal{W}$ can be checked directly.

F.9. Proof of Proposition 9

Recall the definition of the out-of-sample prediction error is $\mathbb{E}\|y_t - x_t^\top \hat{\beta}\|^2$. So for any estimator $\hat{\beta}$, under the assumption $\Sigma = I$, we have

$$\begin{aligned}
\mathbb{E}\|y_t - x_t^\top \hat{\beta}\|^2 &= \mathbb{E}\|x_t^\top(\hat{\beta}-\beta) + \varepsilon_t\|^2 = \mathbb{E}\|x_t^\top(\hat{\beta}-\beta)\|^2 + \sigma^2 \\
&= \mathbb{E}[(\hat{\beta}-\beta)^\top x_t \cdot x_t^\top(\hat{\beta}-\beta)] + \sigma^2 \\
&= \mathbb{E}[(\hat{\beta}-\beta)^\top \Sigma(\hat{\beta}-\beta)] + \sigma^2 \\
&= \mathbb{E}\|\hat{\beta}-\beta\|^2 + \sigma^2.
\end{aligned}$$

When we consider the distributed estimator and take the limit, we obtain

$$\mathcal{O}_k = \sigma^2 + \mathcal{M}_k,$$

and the formula for OE. For the inequality between OE and ARE, it is sufficient to notice that $ARE \leq 1$. Finally, the explicit formulas follow easily from previous results.

F.10. Proof of Theorem 10

It is equivalent to show that the ARE is always greater than or equal to $1/(1+\alpha^2)$. To do this, we need to use Theorem 7. From the first property, we have ARE$\geq \psi(k,\gamma,\alpha^2)$. Then, since ψ is a decreasing function of k, it is lower bounded by its limit at infinity, which is $h(\alpha^2,\gamma)$. Finally, $h(\alpha^2,\gamma)$ is an increasing function of γ, so it is lower bounded by the limit at 0, which is $1/(1+\alpha^2)$. When $\gamma > 1$, $h(\alpha^2,\gamma)$ is a decreasing function of α^2, so it is lower bounded by the limit at infinity, which is $1-1/\gamma^2$. The desired result follows.

References

Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2009.

Theodore W Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley New York, 2003.

Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, 2010.

Moulinath Banerjee, Cecile Durot, and Bodhisattva Sen. Divide and conquer in non-standard problems and the super-efficiency phenomenon. *The Annals of Statistics*, 47 (2):720–757, 2019.

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46(3): 1352–1382, 2018.

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Inc., 1989.

Guy E Blelloch and Bruce M Maggs. Parallel algorithms. In *Algorithms and theory of computation handbook*. Chapman & Hall/CRC, 2010.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Mark Braverman, Ankit Garg, Tengyu Ma, Huy L Nguyen, and David P Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1011–1020. ACM, 2016.

T Tony Cai and Hongji Wei. Distributed gaussian mean estimation under communication constraints: Optimal rates and communication-efficient algorithms. *arXiv preprint arXiv:2001.08877*, 2020.

Xi Chen, Weidong Liu, and Yichen Zhang. First-order newton-type estimator for distributed estimation and inference. *arXiv preprint arxiv:1811.11368*, 2018.

Xi Chen, Weidong Liu, and Yichen Zhang. Quantile regression under memory constraint. *The Annals of Statistics*, 47(6):3244–3273, 2019.

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

Romain Couillet and Merouane Debbah. *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.

Romain Couillet, Mérouane Debbah, and Jack W Silverstein. A deterministic equivalent for the analysis of correlated mimo multiple access channels. *IEEE Trans. Inform. Theory*, 57(6):3493–3514, 2011.

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

Lee H Dicker. Variance estimation in high-dimensional linear models. *Biometrika*, 101(2): 269–284, 2014.

Lee H Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1–37, 2016.

Lee H. Dicker and Murat A. Erdogdu. Maximum likelihood for variance estimation in high-dimensional linear models. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 159–167. PMLR, 2016.

Edgar Dobriban and Yue Sheng. Distributed linear regression by averaging. *arXiv preprint arxiv:1810.00412*, 2018.

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

Jiexin Duan, Xingye Qiao, and Guang Cheng. Distributed nearest neighbor classification. *arXiv preprint arXiv:1812.05005*, 2018.

Jianqing Fan, Dong Wang, Kaizheng Wang, and Ziwei Zhu. Distributed estimation of principal eigenspaces. *The Annals of Statistics*, 47(6):3009–3031, 2019.

Zheng-Chu Guo, Lei Shi, and Qiang Wu. Learning theory of distributed regression with bias corrected regularization kernel network. *The Journal of Machine Learning Research*, 18(1):4237–4261, 2017.

Walid Hachem, Philippe Loubaton, and Jamal Najim. Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875–930, 2007.

Fumio Hiai and Dénes Petz. *The semicircle law, free random variables and entropy*, volume 77 of *Mathematical Surveys and Monographs*. American Mathematical Soc., 2006.

Xiaoming Huo and Shanshan Cao. Aggregated inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, page e1451, 2018.

Jiming Jiang. Reml estimation: asymptotic behavior and related topics. *The Annals of Statistics*, 24(1):255–286, 1996.

Jiming Jiang, Cong Li, Debashis Paul, Can Yang, and Hongyu Zhao. On high-dimensional misspecified mixed model analysis in genome-wide association study. *The Annals of Statistics*, 44(5):2127–2160, 2016.

Michael I. Jordan, Jason D. Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.

Paraschos Koutris, Semih Salihoglu, Dan Suciu, et al. Algorithmic aspects of parallel data processing. *Foundations and Trends® in Databases*, 8(4):239–370, 2018.

Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18(5):1–30, 2017.

Runze Li, Dennis KJ Lin, and Bing Li. Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry*, 29(5):399–409, 2013.

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.

Meimei Liu, Zuofeng Shang, and Guang Cheng. How many machines can we use in parallel computing for kernel ridge regression? *arXiv preprint arXiv:1805.09948*, 2018.

Nancy A Lynch. *Distributed algorithms*. Elsevier, 1996.

Lester W Mackey, Michael I Jordan, and Ameet Talwalkar. Divide-and-conquer matrix factorization. In *Advances in neural information processing systems*, pages 1134–1142, 2011.

Vladimir A Marchenko and Leonid A Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sb.*, 114(4):507–536, 1967.

Ryan Mcdonald, Mehryar Mohri, Nathan Silberman, Dan Walker, and Gideon S Mann. Efficient large-scale distributed training of conditional maximum entropy models. In *Advances in Neural Information Processing Systems*, pages 1231–1239, 2009.

Alexandru Nica and Roland Speicher. *Lectures on the combinatorics of free probability*, volume 13. Cambridge University Press, 2006.

Debashis Paul and Alexander Aue. Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29, 2014.

Armin Pourshafeie, Carlos D Bustamante, and Snehit Prabhu. Caring without sharing: Meta-analysis 2.0 for massive genome-wide association studies. *bioRxiv*, page 436766, 2018.

Thomas Rauber and Gudula Rünger. *Parallel programming: For multicore and cluster systems*. Springer Science & Business Media, 2013.

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.

Francisco Rubio and Xavier Mestre. Spectral convergence for a general class of random matrices. *Statistics & Probability Letters*, 81(5):592–602, 2011.

Shayle R Searle, George Casella, and Charles E McCulloch. *Variance components*, volume 391. John Wiley & Sons, 2009.

Vadim Ivanovich Serdobolskii. *Multiparametric Statistics*. Elsevier, 2007.

Zuofeng Shang and Guang Cheng. Computational limits of a distributed algorithm for smoothing spline. *The Journal of Machine Learning Research*, 18(1):3809–3845, 2017.

Chengchun Shi, Wenbin Lu, and Rui Song. A massive data framework for m-estimators with cubic-rate. *Journal of the American Statistical Association*, pages 1–12, 2018.

Virginia Smith, Simone Forte, Chenxin Ma, Martin Takác, Michael I Jordan, and Martin Jaggi. Cocoa: A general framework for communication-efficient distributed optimization. *Journal of Machine Learning Research*, 18(230):1–49, 2018.

Alexander Smola. Course notes on scalable machine learning, 2012.

Dan V Voiculescu, Ken J Dykema, and Alexandru Nica. *Free random variables*, volume 1 of *CRM Monograph Series*. American Mathematical Soc., 1992.

Stanislav Volgushev, Shih-Kang Chao, and Guang Cheng. Distributed inference for quantile regression processes. *The Annals of Statistics*, 47(3):1634–1662, 2019.

Xiaozhou Wang, Zhuoyi Yang, Xi Chen, and Weidong Liu. Distributed inference for linear support vector machine. *Journal of Machine Learning Research*, 20(113):1–41, 2019.

Ganggang Xu, Zuofeng Shang, and Guang Cheng. Optimal tuning for divide-and-conquer kernel ridge regression with massive data. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5483–5491, 2018.

Jianfeng Yao, Zhidong Bai, and Shurong Zheng. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press, 2015.

Yuchen Zhang, Martin J Wainwright, and John C Duchi. Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems*, pages 1502–1510, 2012.

Yuchen Zhang, John Duchi, Michael I Jordan, and Martin J Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems 26*, pages 2328–2336. Curran Associates, Inc., 2013a.

Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression. In *Conference on Learning Theory*, pages 592–617, 2013b.

Yuchen Zhang, John C Duchi, and Martin J Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363, 2013c.

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.

Tianqi Zhao, Guang Cheng, and Han Liu. A partially linear framework for massive heterogeneous data. *Annals of statistics*, 44(4):1400, 2016.

Yuancheng Zhu and John Lafferty. Distributed nonparametric regression under communication constraints. *arXiv preprint arXiv:1803.01302*, 2018.