Distributed linear regression by averaging

Edgar Dobriban[*] and Yue Sheng[†]

October 2, 2018

Abstract

Modern massive datasets pose an enormous computational burden to practitioners. Distributed computation has emerged as a universal approach to ease the burden: Datasets are partitioned over machines, which compute locally, and communicate short messages. Distributed data also arises due to privacy reasons, such as in medicine. It is important to study how to do statistical inference and machine learning in a distributed setting.

In this paper, we study *one-step parameter averaging* in statistical *linear models* under *data parallelism*. We do linear regression on each machine, and take a weighted average of the parameters. How much do we lose compared to doing linear regression on the full data? Here we study the performance loss in *estimation error*, *test error*, and *confidence interval length* in high dimensions, where the number of parameters is comparable to the training data size.

We discover several key phenomena. First, averaging is *not optimal*, and we find the exact performance loss. Our results are simple to use in practice. Second, different problems are affected differently by the distributed framework. Estimation error and confidence interval length increases a lot, while prediction error increases much less. These results match simulations and a data analysis example. We rely on recent results from random matrix theory, where we develop a new calculus of deterministic equivalents as a tool of broader interest.

1 Introduction

Datasets are constantly increasing in size and complexity. This leads to important challenges for practitioners. Statistical inference and machine learning, which used to be computationally convenient on small datasets, now bring an enormous computational burden.

Distributed computation is a universal approach to deal with large datasets. Datasets are partitioned across several machines (or workers). The machines perform computations locally and communicate only small bits of information with each other. They coordinate to compute the desired quantity. This is the standard approach taken at large technology companies, which routinely deal with huge datasets spread over computer clusters. What are the best ways to divide up and coordinate the work?

The same problem arises when the data is distributed due to privacy, security, or ethical concerns. For instance, medical and healthcare data is typically distributed across hospitals or medical units. The parties agree that they want to aggregate the results. At the same time, they are concerned with privacy, and do not want other parties to look at their data. How can they compute the desired aggregates, without sharing the entire data?

[*] Wharton Statistics Department, University of Pennsylvania. E-mail: dobriban@wharton.upenn.edu.

[†] Graduate Group in Applied Mathematics and Computational Science, Department of Mathematics, University of Pennsylvania. E-mail: yuesheng@sas.upenn.edu.

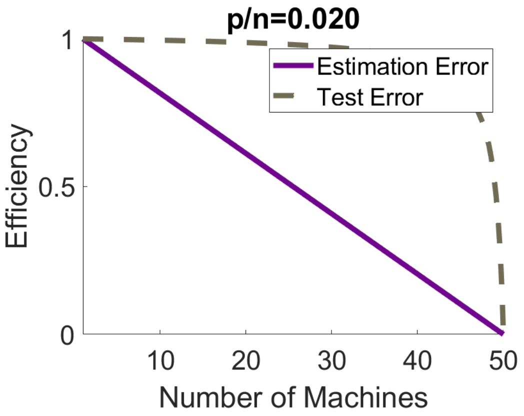

Figure 1: How much accuracy do we lose in distributed regression? The plots show the relative efficiency, i.e., the ratio of errors, of the global least squares (OLS) estimator, compared to the distributed estimator averaging the local least squares estimators. This efficiency is at most unity, because the global estimator is more accurate. If the efficiency is close to unity, then one-step averaging is accurate. We show the behavior of estimation and test error, as a function of number of machines. We see that estimation error is *much more affected* than test error. The specific formulas are given in Table 1.

In both cases, the key question is how to do statistical estimation and machine learning in a distributed setting. And what performance can the best methods achieve? This is an important question of broad interest, and it is expected that the area of distributed estimation and computation will grow even more in the future.

In this paper, we develop precise theoretical answers to fundamental questions in distributed estimation. We study *one-step parameter averaging* in statistical *linear models* under *data parallelism*. Specifically, suppose we do linear regression (Ordinary Least Squares, OLS) on each subset of a dataset distributed over k machines, and take an optimal weighted average of the regression coefficients. How do the statistical and predictive properties of this estimator compare to doing OLS on the full data?

We study the behavior of several learning and inference problems, such as *estimation error*, *test error* (out-of-sample *prediction error*), and *confidence intervals*. We also consider a high-dimensional (or proportional-limit) setting where the number of parameters is of the same order as the number of total samples (i.e., the size of the training data). We discover the following key phenomena, some of which are surprising in the context of existing work:

1. **Sub-optimality.** One-step averaging is not optimal, meaning that it leads to a performance decay. In contrast to some recent work (see the related work section), we find that there is a clear performance loss due to one-step averaging. This loss is due to the *essential high-dimensional* nature of our problem. However, we can quantify this loss precisely. This paves the way to developing improved methods for distributed estimation.
2. **Strong problem-dependence.** Different learning and inference problems are affected differently by the distributed framework. Specifically, *estimation error and the length of confidence intervals increases a lot, while prediction error increases less.* This phenomenon was apparently not noticed before.
3. **Simple form and universality.** We discover that the asymptotic efficiencies have simple

Table 1: Estimation, Confidence Interval, and test efficiency as a function of number of machines k, the sample size n, and the dimension p. This is how much smaller the error of the global estimator is compared to the distributed estimator. These functions are plotted and described in Figure 1.

Quantity	Relative efficiency (n, p, k)
Estimation & CIs	$\frac{n-kp}{n-p}$
Test error	$\frac{1}{1+\frac{p^2(k-1)}{n(n-kp)}}$

forms that are often *universal.* Specifically, they do not depend on the covariance matrix of the data, or on the sample sizes on the local machines. For instance, the estimation efficiency *decreases linearly in the number of machines k* (see Figure 1 and Table 1).

While there is already a lot of work in this direction (see Section 4) our results are new and complementary. The key elements of novelty are: (1) The sample size and the dimension are comparable, and we do not assume sparsity. Hence the problems we study are *essentially high-dimensional.* There is very little work in this direction. (2) We have a new mathematical approach, using recent results from asymptotic random matrix theory. Our approach also develops a novel theoretical tool, the *calculus of deterministic equivalents*, which may be useful in other problems as well.

1.1 Summary of our results

Our contributions, and the structure of our paper, are as follows (see Section 4 for some related work):

1. We start with studying estimation error in linear models. We find an explicit expression for the *finite sample relative efficiency* of one-step averaging compared to OLS. We show directly that it is less than or equal to unity, by showing that the function $1/\operatorname{tr}(X^{-1})$ is concave on positive definite matrices.
2. We then consider asymptotics, first under Marchenko-Pastur models where the data is iid from a distribution with a general covariance matrix. We find a simple expression for the limit of the relative efficiency, called the *asymptotic relative efficiency.* We give a *multi-response regression* characterization that gives the correct "degrees of freedom" for distributed regression.
3. Next, we study more general *elliptical models*, where the different samples have different scales. We find an expression for the ARE, albeit less explicit than the previous one. We show that the ARE is monotone and convex. We also perform a detailed *worst-case analysis*, giving examples of elliptical scale distributions for which the split across two machines leads to a huge increase in estimation error.
4. We then develop a more general framework for evaluating the relative efficiency of predicting *linear functionals of the regression coefficients.* We show that *test error* (out-of-sample prediction), *training error* (in-sample prediction), *confidence intervals*, and *regression function estimation* all fall into this general framework. We find the optimal relative efficiency, and show that it depends on the traces of certain functionals of X. We also generalize the concavity property obtained above.

5. To evaluate the needed trace functionals, we develop a *calculus of deterministic equivalents* in random matrix theory. Such deterministic equivalents have appeared in prior work, but we develop a more general approach. Specifically, we define two sequences of matrices to be *asymptotically equivalent*, if they have the same limits of inner products with any other fixed sequences of matrices. In terms of this definition, we present a general Marchenko-Pastur theorem in elliptical models, which is a streamlined version of previously obtained results. We also give several *rules of the calculus*, including rules for sums, products, traces and Stieltjes transforms.
6. As an application of the calculus of deterministic equivalents, we find the limits of the relative efficiencies for the four functionals introduced above, test and training error, confidence intervals, and regression function estimation, in a distributed setting. We show that the efficiency loss *depends strongly on the learning problem.* See Figure 1 and Table 1. For instance, estimation error and CI length is *much more affected* than test error.
7. We show that our theoretical results are very accurate in numerical simulations throughout the paper, and also in Section 7. We also illustrate that our results are accurate in an empirical data example using the NYC Flights dataset.

The code for our paper is available at `github.com/dobriban/dist`.

2 Estimation

2.1 Finite sample relative efficiency

We start by studying estimation error in linear models. Consider the standard linear model

$$Y = X\beta + \varepsilon.$$

Here we have an outcome variable y along with some p covariates $x = (x^1, \ldots, x^p)^\top$, and want to understand their relationship. We observe n such data points, arranging their outcomes into the $n \times 1$ vector Y, and their covariates into the $n \times p$ matrix X. We assume that Y depends linearly on X, via some unknown $p \times 1$ parameter vector β.

We consider the case where there are more samples than training data points, i.e., $n > p$, while p can also be large. In that case, a gold standard is the usual least squares estimator (ordinary least squares or OLS)

$$\hat\beta = (X^\top X)^{-1} X^\top Y.$$

We also assume that the coordinates of the noise ε are uncorrelated and have variance σ^2. Then is well known that the mean squared error (MSE) of OLS equals $\mathbb{E}\|\hat\beta - \beta\|^2 = \sigma^2 \operatorname{tr}[(X^\top X)^{-1}]$.

Suppose now that the samples are distributed across k machines (these can be real machines, but they can also be—say—sites or hospitals in medical applications). The i-th machine has the $n_i \times p$ matrix X_i, containing n_i samples, and also the $n_i \times 1$ vector Y_i of the corresponding outcomes for those samples. Thus, the i-th worker has access to only a subset of training n_i data points out of the total of n training data points. For instance, if the data points denote n users, then they may be partitioned into k sets based on country of residence, and we may have n_1 samples from the United States on one server, n_2 samples from Canada on another server, etc. The broad question is: How can we estimate the unknown regression parameter β if we need to do most of the computations locally?

Let us write the partitioned data as

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

We also assume that each *local* OLS estimator $\hat{\beta}_i = (X_i^\top X_i)^{-1} X_i^\top Y_i$ is well defined, which requires that the number of local training data points n_i must be at least p on each machine. We consider combining the local OLS estimators at a parameter server via one-step weighted averaging. Since they are uncorrelated, unbiased for β, and have MSE $M_i = \sigma^2 \operatorname{tr}[(X_i^\top X_i)^{-1}]$, we can find the optimal unbiased weighted estimator

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

with $\sum_{i=1}^k w_i = 1$, and its mean squared error. As a consequence, we can find an explicit expression for the finite sample *relative efficiency* for estimation.

Lemma 2.1 (Relative efficiency in OLS)**.** *Consider the distributed linear regression problem described above. The optimal unbiased weighted estimator $\hat{\beta}_{dist} = \hat{\beta}_{dist}(w^*) = \sum_{i=1}^k w_i^* \hat{\beta}_i$ has weights proportional to $w_i^* \propto 1/\operatorname{tr}[(X_i^\top X_i)^{-1}]$. Its mean squared error equals*

$$MSE^*_{dist} = \mathbb{E}\|\hat{\beta}_{dist} - \beta\|^2 = \sigma^2 \frac{1}{\sum_{i=1}^k \frac{1}{\operatorname{tr}[(X_i^\top X_i)^{-1}]}}.$$

Therefore, the relative efficiency *of the distributed estimator with respect to the full estimator equals*

$$RE(X_1, \ldots, X_k) = \frac{\mathbb{E}\|\hat{\beta} - \beta\|^2}{\mathbb{E}\|\hat{\beta}_{dist} - \beta\|^2} = \operatorname{tr}[(X^\top X)^{-1}] \cdot \left[\sum_{i=1}^k \frac{1}{\operatorname{tr}[(X_i^\top X_i)^{-1}]}\right].$$

Recall here that $X^\top X = \sum_{i=1}^k X_i^\top X_i$.

See Section 9.1 for the proof.

The relative efficiency is a fundamental quantity, because it quantifies the loss of efficiency from distributed estimation. It is of great interest to understand the behavior of this quantity. For instance, for what matrices does it equal nearly one? In a simple special case, when each local problem is orthonormal, in the sense that $X_i^\top X_i = I_p$, we see that the relative efficiency is unity, so that there is no loss of efficiency in distributed OLS. More generally, the same holds when the Gram matrices $X_i^\top X_i$ are proportional to each other.

To achieve a deeper understanding of the relative efficiency, we will first prove directly that it is at most unity. This turns out to require the convexity of the matrix functional $1/\operatorname{tr}(X^{-1})$. See Section 9.2 for the proof.

Proposition 2.2 (Concavity for relative efficiency)**.** *The function*

$$f(X) = 1/\operatorname{tr}(X^{-1})$$

is concave on positive definite matrices. As a consequence, the relative efficiency for distributed estimation is at most unity for any matrices X_i:

$$RE(X_1, \ldots, X_k) \le 1.$$

The concavity claim seems to be distantly related to the so-called matrix ϕ-entropies (Chen and Tropp, 2014), however there does not seem to be a direct connection.

If not all sample sizes n_i are larger than p, and the OLS estimator is not well-defined on each machine, then we can still take the average of the estimators that are well-defined. It is easy to see that this corresponds to reducing the sample size from n to the *effective sample size* $n^* = \sum_{i \in S} n_i$, where S is the set of machines such that OLS is well-defined. Therefore, our results can still be used, with n replaced by n^*.

However, the form of the finite sample relative efficiency does not show clearly how the performance of averaging depends on n, p, k. Therefore, we will consider an asymptotic setting below.

2.2 Asymptotics

To get more insightful results about the efficiency, we will take an asymptotic approach. In general, we notice that the RE only depends on the eigenvalues of the Gram matrices $X_i^\top X_i$ and $X^\top X$, and therefore it makes sense to study models where the eigenvalues of those matrices are precisely characterized. Indeed, we will adopt such models from asymptotic random matrix theory.

Specifically, recall that the empirical spectral distribution (e.s.d.) of a symmetric matrix M is simply the CDF of its eigenvalues (which are all real-valued). More formally, it is the discrete distribution F_p that places equal mass on all eigenvalues of M. There are many models in random matrix theory where the dimensions of the matrices grow, while with probability one, the e.s.d. F_p converges weakly to some limiting spectral distributions (l.s.d.) F_γ, i.e., $F_p \Rightarrow F_\gamma$. In that case, it follows that for suitable test functions f of the eigenvalues, the trace functional

$$\frac{1}{p}\operatorname{tr} f(M) = \frac{\sum_{i=1}^{p} f(\lambda_i(M))}{p}$$

has a well-defined limit in terms of F_γ, namely $\mathbb{E}_{F_\gamma} f(T)$. This explains how we can use results from random matrix theory to analyze the relative efficiency.

We will consider models such that $n_i^{-1} X_i^\top X_i$ have almost sure limiting spectral distributions F_{γ_i}. Then, we should have the limits

$$\operatorname{tr}[(X_i^\top X_i)^{-1}] \to \gamma_i \cdot \mathbb{E}_{F_{\gamma_i}} T^{-1},$$

assuming that the limits are finite. We will make assumptions to ensure this holds. Similarly, assuming that $n^{-1}X^\top X$ has almost sure limiting spectral distribution F_γ, and that the limit exists, we should have

$$\operatorname{tr}[(X^\top X)^{-1}] \to \gamma \cdot \mathbb{E}_{F_\gamma} T^{-1}.$$

Hence, the RE should converge to an asymptotic relative efficiency (ARE) of the form

$$RE(X_1, \ldots, X_k) \to \gamma \cdot \mathbb{E}_{F_\gamma} T^{-1} \cdot \sum_{i=1}^{k} \frac{1}{\gamma_i \cdot \mathbb{E}_{F_{\gamma_i}} T^{-1}}.$$

Next, we will consider some specific models for X_i, under which the ARE has a more explicit form. First, we will consider "Marchenko-Pastur" type sample covariance matrices, which are fundamental in multivariate statistics. See e.g., Bai and Silverstein (2009); Anderson (2003); Paul and Aue (2014); Yao et al. (2015) for reference.

In this basic model, the rows of X are iid p-dimensional observations x_i, for $i = 1, \ldots, n$. The samples are drawn from a population with covariance matrix Σ. The classical model is that the data points have the form $x_i = \Sigma^{1/2} z_i$, for some vector z_i with iid entries. Arranging the data points x_i as the rows of the $n \times p$ data matrix X, this has the form

$$X = Z\Sigma^{1/2},$$

where the $n \times p$ matrix Z has iid standardized entries, and Σ is a $p \times p$ deterministic positive semi-definite matrix. Let H_p be the empirical spectral distribution of Σ.

In this model, the *Marchenko-Pastur distribution* describes the weak limit of the spectral distribution F_p of $\widehat{\Sigma}$. Suppose the entries of Z come from an infinite array of iid variables with mean zero and variance 1, and we take a sequence of such problems with both the dimension and the sample size growing, $n, p \to \infty$, with asymptotically fixed aspect ratio $p/n \to \gamma < 1$. If the e.s.d. of Σ converges to some limit distribution, i.e., $H_p \Rightarrow H$ weakly, then with probability 1, the e.s.d. of $\widehat{\Sigma}$ also converges, so that $F_p \Rightarrow F_\gamma$ for a probability measure $F_\gamma = F_\gamma(H)$

(Marchenko and Pastur, 1967; Bai and Silverstein, 2009). We assume moreover that H is compactly supported away from the origin, in which case the same is true for F_γ.

In this model, we obtain the following surprisingly simple expression for the ARE.

Theorem 2.3 (ARE for Marchenko-Pastur models)**.** *Consider the above high-dimensional asymptotic limit, where the data matrix is random, and its rows are iid from a population with some covariance matrix Σ. Specifically, the data has the form $X = Z\Sigma^{1/2}$, where X is $n \times p$, and $n, p \to \infty$ such that $p/n \to \gamma > 0$. Suppose the data is distributed over k machines with sample sizes $n_i > p$, and the sample sizes are all proportional to the dimension, with $p/n_i \to \gamma_i > 0$. Then, the ARE of the distributed one-step averaging OLS estimator with respect to the full OLS estimator equals*

$$ARE = \frac{1 - k\gamma}{1 - \gamma}. \tag{1}$$

Moreover, for any finite sample size n, dimension p, and number of machines k, we can approximate the ARE as

$$ARE \approx \frac{n - kp}{n - p}.$$

See Section 9.3 for the proof.

We find this to be a surprisingly simple formula, which can moreover be easily computed in practice. Moreover, the formula has several more interesting properties:

1. The ARE *decreases linearly* with the number of machines k. This holds as long as $ARE \geq 0$. At the threshold case $ARE = 0$, there is a phase transition. The reason is that there is a singularity, and the OLS estimator is not well defined anymore for at least one machine.

   However, we should be somewhat cautious about interpreting the linear decrease. In some cases, it may make more sense to study the root mean squared error (RMSE). That quantity has a different loss of efficiency, namely the square root of the ARE presented above.

2. The ARE has two important *universality* properties.

   (a) First, it *does not depend* on how the samples are distributed across the different machines, i.e., it is independent of the specific sample sizes n_i.

   (b) Second, it *does not depend* on the covariance matrix Σ of the samples. This is in contrast to the estimation error of OLS, which does in fact depend on the covariance structure. Therefore, we think that the cancellation of Σ in the ARE is noteworthy.

The ARE is also very accurate in simulations. See Figure 2 for an example. Here we report the results of a simulation where we generate an $n \times p$ random matrix X such that the rows are distributed independently as $x_i \sim \mathcal{N}(0, \Sigma)$. We take Σ to be diagonal with entries chosen uniformly at random between 1 and 2. We choose $n > p$, and for each value of k such that $k < n/p$, we split the data into k groups of a random size n_i. To ensure that each group has a size $n_i \geq p$, we first let $n_i^0 = p$, and then distribute the remaining samples uniformly at random. We then show the results of the expression for the RE from Lemma 2.1 compared to the theoretical ARE. We observe that the two agree closely.

It is also of interest to understand the performance of one-step averaging if we use suboptimal weights w_i. How much do we lose compared to the optimal performance if we do not use the right weights? In practice, it may seem reasonable to take a simple average of all estimators. We have performed that analysis in Section 9.4, and we found that the loss can be viewed in terms of an inequality between the arithmetic and harmonic means.

There are several more remarkable properties to note. We will discuss them in turn. We have studied the monotonicity properties and interpretation of the relative efficiency, see Section

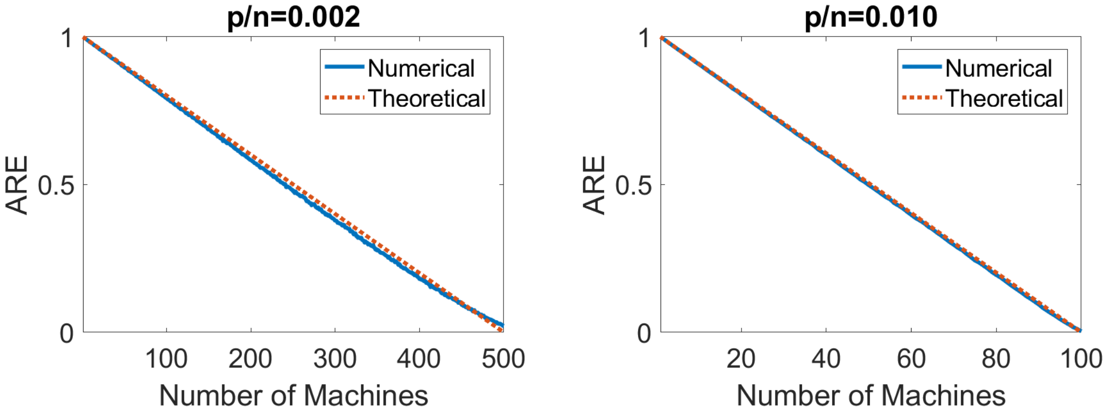

Figure 2: Comparison of empirical and theoretical ARE for standard sample covariance matrices. Left: $n = 10,000$, $p = 20$. Right: $n = 10,000$, $p = 100$.

9.5. Our results also show that the distributed regression estimator is minimax rate-optimal as long as the number of machines is not too large (Section 9.6).

Next we give a multi-response regression characterization that heuristically gives the correct "*degrees of freedom*" for distributed regression. This will be helpful to understand the asymptotic formulas derived above.

We re-parametrize $Y = X\beta + \varepsilon$, treating the samples on each machine as a different outcome. We write the $n \times k$ multi-response outcome matrix $\underline{Y}$, the $n \times pk$ feature matrix $\underline{X}$, and the corresponding noise $\underline{\varepsilon}$ as

$$\underline{Y} = \begin{bmatrix} Y_1 & 0 & \dots & 0 \\ 0 & Y_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Y_k \end{bmatrix}, \quad \underline{X} = \begin{bmatrix} X_1 & 0 & \dots & 0 \\ 0 & X_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & X_k \end{bmatrix}, \quad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 & 0 & \dots & 0 \\ 0 & \varepsilon_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \varepsilon_k \end{bmatrix}.$$

We also introduce $\underline{\beta}$, the $pk \times k$ parameter matrix, which shares parameters across the k outcomes:

$$\underline{\beta} = \begin{bmatrix} \beta & 0 & \dots & 0 \\ 0 & \beta & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \beta \end{bmatrix} = I_k \otimes \beta$$

Note that $Y = X\beta + \varepsilon$ is equivalent to $\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon}$. The OLS estimator of $\underline{\beta}$ is $\hat{\underline{\beta}} = (\underline{X}^\top \underline{X})^{-1}\underline{X}^\top \underline{Y}$. This can be calculated as

$$\hat{\underline{\beta}} = \begin{bmatrix} \hat\beta_1 & 0 & \dots & 0 \\ 0 & \hat\beta_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat\beta_k \end{bmatrix} = \begin{bmatrix} (X_1^\top X_1)^{-1}X_1^\top Y_1 & 0 & \dots & 0 \\ 0 & (X_2^\top X_2)^{-1}X_2^\top Y_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & (X_k^\top X_k)^{-1}X_k^\top Y_k \end{bmatrix}$$

Notice that the estimators of the coefficients of different outcomes are the familiar distributed OLS estimators. Now, we can find a plug-in estimator of β, based on $\underline{\beta}$. Given the form of $\underline{\beta}$

above, for any vector w such that $\sum_{i=1}^{k} w_i = 1$, we have that β can be expressed in terms of the tensorized parameter $\underline{\beta}$ as a weighted combination

$$\beta = (1_p^\top \otimes I_k)\underline{\beta}w.$$

Therefore, for any unbiased estimator of $\underline{\beta}$, the corresponding weighted combination estimators given below are unbiased for β:

$$\hat{\beta}(w) = (1_p^\top \otimes I_k)\underline{\hat{\beta}}w.$$

In our case, given the zeros in the estimator, this simply reduces to the weighted sum $\hat{\beta}(w) = \sum_{i=1}^{k} w_i \hat{\beta}_i$.

This explains how our problem can be understood in the framework of multi-response regression. Also, the number of parameters in that problem is kp, so the "degrees of freedom" is $n - kp$. Indeed, the residual effective degrees of freedom $\hat{r} = y - Hy$ is usually defined as $\mathrm{tr}(I - H)$. Let H_i be the hat matrix on the i-th machine, so that $H_i = X_i(X_i^\top X_i)^{-1}X_i^\top$. Then it is easy to see that $\mathrm{tr}(I - H_i) = n_i - p$, for all i. Since H_{dist} is simply the block diagonal matrix with H_i as blocks, we see that $\mathrm{tr}(I - H_{dist}) = n - pk$, as required.

This provides a simple explanation for why the residual "degrees of freedom" of a distributed estimation problem is $n - kp$, and also for why the relative efficiency is approximately $(n - kp)/(n - p)$.

2.3 Elliptical models

Second, we study the more general setting of elliptical data. In this model the data samples may have different scalings, having the form $x_i = g_i^{1/2}\Sigma^{1/2}z_i$, for some vector z_i with iid entries, and for datapoint-specific *scale parameters* g_i. Arranging the data as the rows of the matrix X, that takes the form

$$X = \Gamma^{1/2}Z\Sigma^{1/2},$$

where Z and Γ are as before: Z has iid standardized entries, while Σ is the covariance matrix of the features. Now Γ is the diagonal *scaling matrix* containing the scales g_i of the samples. This model has a long history in multivariate statistical analysis (e.g., Mardia et al., 1979).

In this model, it turns out that the ARE can be expressed in a simple way via the η-transform (Tulino and Verdú, 2004). The η-transform of a distribution G is

$$\eta(x) = \mathbb{E}_G \frac{1}{1 + xT},$$

for all x for which this expectation is well-defined. We will see that the ARE can be expressed in terms of the functional inverse f of the η-transform evaluated at the specific value $1 - \gamma$:

$$f(\gamma, G) = \eta_G^{-1}(1 - \gamma). \tag{2}$$

For some insight on the behavior of η and f, consider first the case when G is a point mass at unity, $G = \delta_1$. In this case, all scales are equal, so this is just the usual Marchenko-Pastur model. Then, we have $\eta(x) = 1/(1 + x)$, while $f(\gamma, G) = \gamma/(1 - \gamma)$. See Figure 3 for the plots. The key points to notice are that η is a decreasing function of x, with $\eta(0) = 1$, and $\lim_{x\to\infty} \eta(x) = 0$. Moreover, f is an increasing function on $[0, 1]$ with $f(0) = 0$, $\lim_{\eta\to 1} f(\eta) = +\infty$. The same qualitative properties hold in general for compactly supported distributions G bounded away from 0.

In the elliptical model, we find the following expression for the ARE.

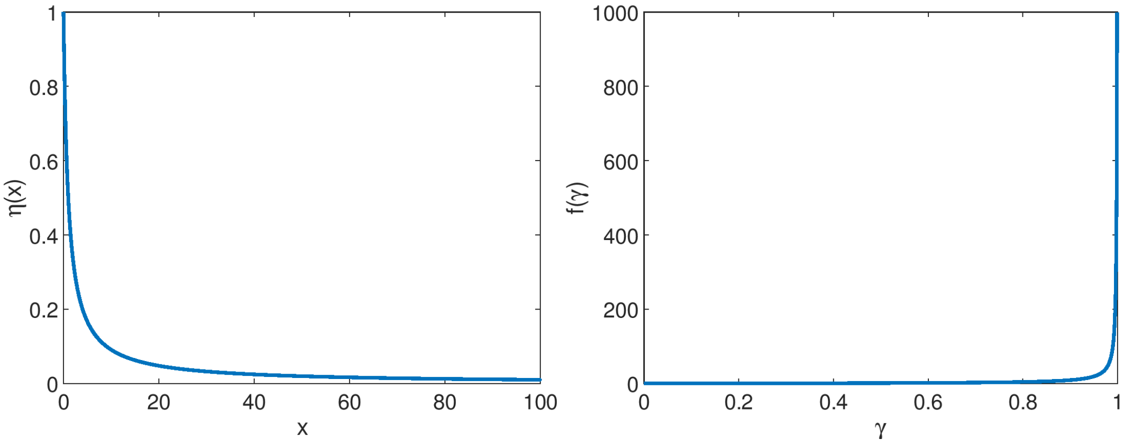

Figure 3: Plots of η and f for G being the point mass at unity.

Theorem 2.4 (ARE for elliptical models)**.** *Consider the above high-dimensional asymptotic limit, where the data matrix is random, and the samples have the form $X = \Gamma^{1/2} Z \Sigma^{1/2}$. Suppose that, as $n_i \to \infty$ with $p/n_i \to \gamma_i > 0$, the e.s.d. of Γ converges weakly to G, the e.s.d. of each Γ_i converges weakly to some G_i, and that the e.s.d. of Σ converges weakly to H. Suppose that H is compactly supported away from the origin, G is also compactly supported and does not have point mass at the origin. Then, the ARE has the form*

$$ARE = f(\gamma, G) \cdot \sum_{i=1}^{k} \frac{1}{f(\gamma_i, G_i)}.$$

See Section 9.7 for the proof.

There are two implicit relations in the above formula. First, $\sum 1/\gamma_i = 1/\gamma$, because $\sum n_i/p = n/p$. Second, $n \cdot G = \sum_{i=1}^{k} n_i \cdot G_i$, or equivalently $G/\gamma = \sum_{i=1}^{k} G_i/\gamma_i$, because Γ contains all entries of each Γ_i.

For the special case when all aspect ratios γ_i are equal, and all scale distributions G_i are equal to G, we can say more about the ARE. We have the following theorem.

Theorem 2.5 (Properties of ARE for elliptical models)**.** *Consider the behavior of distributed regression in elliptical models under the conditions of Theorem 2.4. Suppose that the data sizes n_i on all machines are equal, so that $\gamma_i = \gamma_j = k\gamma$ for all i, j. Suppose moreover that the scale distributions G_i on all machines are also equal. Then, the ARE has the following properties*

1. *It can be expressed equivalently as*

$$ARE(k) = \frac{k \cdot \eta_G^{-1}(1-\gamma)}{\eta_G^{-1}(1-k\gamma)} = \frac{k \cdot f(\gamma, G)}{f(k\gamma, G)} = \frac{e(\gamma, G)}{e(k\gamma, G)}.$$

Here η_G is the η-transform of G, f is defined above, while e is the unique positive solution of the equation

$$\int \frac{se}{1+\gamma se} dG(s) = 1.$$

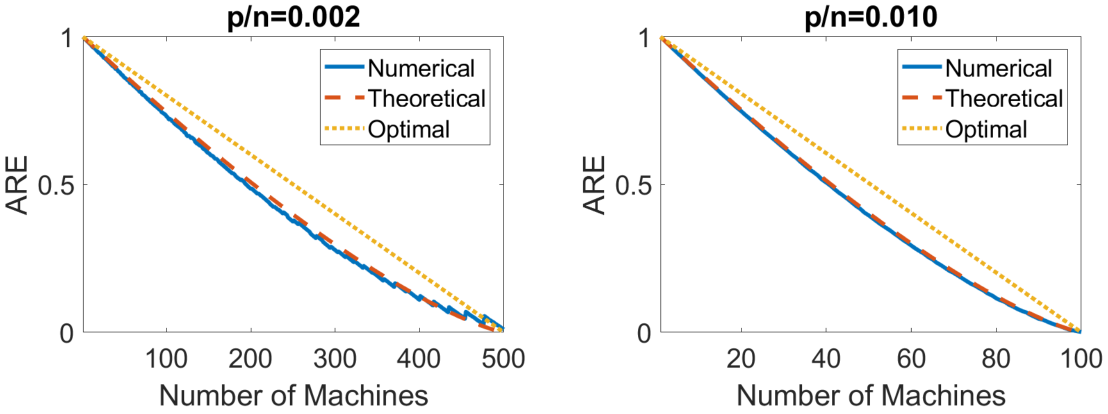

Figure 4: Comparison of empirical and theoretical ARE for elliptical distributions. Left: $n = 10,000$, $p = 20$. Right: $n = 10,000$, $p = 100$.

2. *Suppose also that G does not have a point mass at the origin. Then, the ARE is a strictly decreasing smooth convex function for $k \in [1, 1/\gamma]$. Here k is viewed as a continuous variable. Moreover $ARE(1) = 1$, and*

$$\lim_{k \to 1/\gamma} ARE(k) = 0.$$

See Section 9.8 for the proof. These theoretical formulas again match simulation results well, see Figure 4. On that figure, we use the same simulation setup as for Figure 2, and in addition we choose the scale distribution to be uniform on $[0, 1]$.

The ARE for a constant scale distribution is a straight line in k, $\text{ARE}(k) = (1-k\gamma)/(1-\gamma)$. For a general scale distribution, the graph of ARE is a curve below that straight line. The interpretation is that for elliptical distributions, there is a larger efficiency loss in one-step averaging. Intuitively, the problem becomes ”more non-orthogonal” due to the additional variability from sample to sample.

2.3.1 Worst-case analysis

It is natural to ask which elliptical distributions are difficult for distributed estimation. That is, for what scale distributions G does the distributed setting have a strong effect on the learning accuracy? Intuitively, if some of the scales are much larger than others, then they ”dominate” the problem, and may effectively reduce the sample size.

Here we show that this intuition is correct. We find a sequence of scale distributions G_τ such that distributed estimation is ”arbitrarily bad”, so that the ARE decreases very rapidly, and approaches zero even for two machines.

Proposition 2.6 (Worst-case example)**.** *Consider elliptical models with scale distributions that are a mixture of small variances τ, and larger variances $1/\gamma$, with weights γ and $1/\gamma$, i.e., $G_\tau = (1-\gamma)\delta_\tau + \gamma\delta_{1/\gamma}$. Then, as $\tau \to 0$, we have $ARE(2) = O(\tau^{1/2}) \to 0$. Therefore, the relative efficiency for any $k \geq 2$ tends to zero.*

See Section 9.9 for the proof.

Next, we consider more general scale distributions that are a mixture of small scales τ, and larger scales $\alpha\tau$, with arbitrary weights $1-c$ and c:

Table 2: ARE as a function of number of machines $k \geq 1$ and the fraction of large scales $c = M\gamma$, as the ratio α tends to infinity. Note that for any fixed M, we need $k/M \geq 1/M$ for the result to be well-defined. We also need k to be an integer. In the table below, we mark by a $\diagup$ the cases where this is not satisfied.

ARE(k) M $\frac{k}{M}$	$0 < M < 1$	$M = 1$	$1 < M < 2$	$M = 2$	$M > 2$
$\frac{k}{M} < 1$	$\diagup$	$\diagup$	1	1	$\frac{c-k\gamma}{c-\gamma}$
$\frac{k}{M} = 1$	$\diagup$	1	$\diagup$	$O(\alpha^{-1/2})$	$O(\alpha^{-1/2})$
$\frac{k}{M} > 1$	$\frac{k(\gamma-c)(1-k\gamma)}{(1-\gamma)(k\gamma-c)}$	$O(\alpha^{-1/2})$	$O(\alpha^{-1})$	$O(\alpha^{-1})$	$O(\alpha^{-1})$

$$G = (1-c)\delta_\tau + c\delta_{\alpha\tau}, \tag{3}$$

where $c \in [0,1], \tau > 0$, and $\alpha > 1$. To gain some intuition about the setting where $\alpha \to \infty$, we notice that only the large scales contribute a non-negligible amount to the sample covariance matrix $X^\top X$. Therefore, the sample size is reduced by a factor equal to the fraction of large scales, which equals c. More specifically, we have

$$n^{-1}X^\top X = n^{-1}\sum_{j=1}^{n} x_j x_j^\top = \alpha^2 n^{-1}\sum_{j\leq cn} z_j z_j^\top + n^{-1}\sum_{j>cn} z_j z_j^\top \approx \alpha^2 n^{-1}\sum_{j\leq cn} z_j z_j^\top.$$

The last approximation follows because the sample covariance matrix has p eigenvalues, out of which we expect $\min(cn,p)$ to be large, of the order of $\alpha^2 \gg 1$. The remaining $\max(p-cn,0)$ are smaller, of unit order. Thus, heuristically, this matrix is well approximated by a scaled sample covariance matrix of cn vectors. Therefore, the sample size is reduced to the number of large scales. If $cn < p$, the matrix is nearly singular, while if $cn > p$, it is well-conditioned. This should provide some intuition for the results to follow.

Theorem 2.7 (More general worst-case example)**.** *Consider elliptical models with scale distribution*

$$G = (1-c)\delta_\tau + c\delta_{\alpha\tau},$$

as in (3)*, where* $c \in [0,1]$*, and* $\tau > 0$*. When* α *tends to infinity, the ARE will depend on* c *and* γ *as summarized in Table 2.*

See Section 9.10 for the proof.

As an example of special interest, if $c = M\gamma$ for some $M > 2$, then

$$\lim_{\alpha\to+\infty} ARE(k) = \begin{cases} \frac{c-k\gamma}{c-\gamma}, & k < M, \\ O(\alpha^{-1/2}), & k = M, \\ O(\alpha^{-1}), & k > M. \end{cases}$$

As before, this result can be understood in terms of reducing the effective sample size to cn. When $k < (cn)/p$, i.e., $k < M$, each local problem can be well-conditioned. However, when $k > (cn)/p$, i.e., $k > M$, all local OLS problems are ill-conditioned, so the ARE is small.

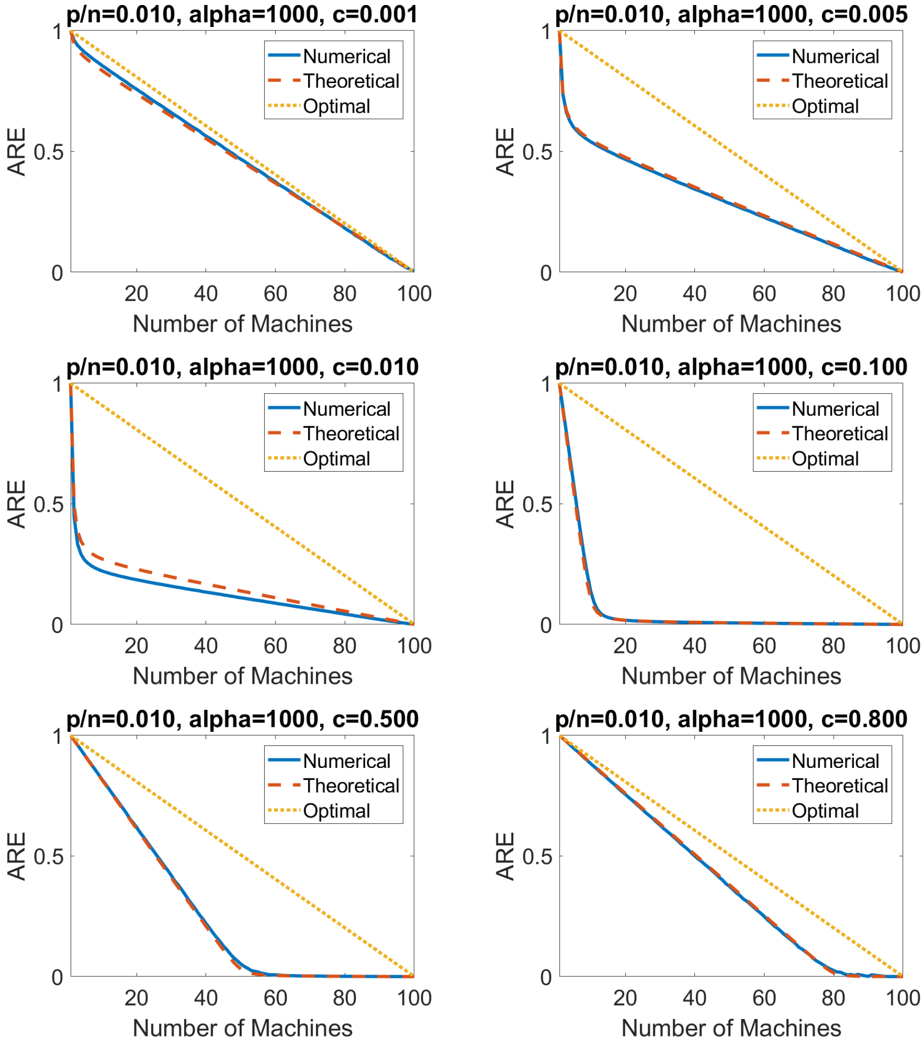

Figure 5: Comparison of empirical and theoretical ARE for worst case elliptical distributions. We fix $n = 10,000$, $p = 100$, so that $\gamma = 0.01$. We also fix $\alpha = 1000$. We vary c.

3 A general framework

After our study of estimation, we now introduce a more general framework, which allows us to study the behavior of many learning tasks in a unified way, including prediction error, confidence intervals (i.e., statistical inference), and regression function estimation. In the general framework, we predict *linear functionals* of the regression coefficients β of the form

$$L_A = A\beta + Z.$$

Table 3: A general framework for finite-sample efficiency calculations. The rows show the various statistical problems studied in our work, namely estimation, confidence interval formation, in-sample prediction, out-of-sample prediction and regression function estimation. The elements of the row show how these tasks fall in the framework of linear functional prediction described in the main body.

Statistical learning problem	L_A	$\hat{L}_A$	A	h	N
Estimation	β	$\hat{\beta}$	I_p	0	0
Regression function estimation	$X\beta$	$X\hat{\beta}$	X	0	0
Confidence interval	β_j	$\hat{\beta}_j$	$E_j^\top$	0	0
Test error	$x_t^\top \beta + \varepsilon_t$	$x_t^\top \hat{\beta}$	$x_t^\top$	1	0
Training error	$X\beta + \varepsilon$	$X\hat{\beta}$	X	1	$\sigma^2 I_n$

Here A is a fixed $d \times p$ matrix, and Z is a zero-mean Gaussian noise vector of dimension d, with covariance matrix $\text{Cov}[Z] = h\sigma^2 I_d$, for some scalar parameter $h \geq 0$. We denote the covariance matrix between ε and Z by N, so that $\text{Cov}[\varepsilon, Z] = N$. If $h = 0$, we say that there is no noise. In that case, we necessarily have $N = 0$.

We predict the linear functional L_A via plug-in based on some estimator $\hat{\beta}_0$ (typically OLS or distributed OLS)

$$\hat{L}_A(\hat{\beta}_0) = A\hat{\beta}_0.$$

We measure the quality of estimation by the mean squared error

$$M(\hat{\beta}_0) = \mathbb{E}\|L_A - \hat{L}_A(\hat{\beta}_0)\|^2.$$

We compute the relative efficiency of OLS $\hat{\beta}$ compared to a weighted distributed estimator $\hat{\beta}_{dist} = \hat{\beta}_{dist}(w)$:

$$E(A, d; X_1, \ldots, X_k) := \frac{M(\hat{\beta})}{M(\hat{\beta}_{dist})}.$$

3.1 Examples

We now show how several learning and inference problems fall into the general framework. See Table 3 for a concise summary. In addition to parameter estimation, we will discuss out-of-sample prediction (test error), in-sample prediction (training error), and confidence intervals.

- **Estimation**. In parameter estimation, we want to estimate the regression coefficient vector β using $\hat{\beta}$. This is an example of the general framework where the transform matrix is $A = I_p$, and there is no noise (so that $h = 0$).
- **Regression function estimation**. We can use $X\hat{\beta}$ to estimate the regression function $\mathbb{E}(Y|X) = X\beta$. In this case, the transform matrix is $A = X$, the linear functional is $L_A = X\beta$, the predictor is $\hat{L}_A = X\hat{\beta}$, and there is no noise.
- **Out-of-sample prediction (Test error)**. For out-of-sample prediction, or test error, we consider a test data point (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, and only x_t is observable. We want to use $x_t^\top \hat{\beta}$ to predict y_t.

This corresponds to predicting the linear functional $L_{x_t} = x_t^\top \beta + \varepsilon_t$, so that $A = x_t^\top$, and the noise is $Z = \varepsilon_t$, which is uncorrelated with the noise ε in the original problem.

- **In-sample prediction (Training error)**. For in-sample prediction, or training error, we consider predicting the response vector Y, using the model fit $X\hat{\beta}$. Therefore, the functional L_A is $L_A = Y = X\beta + \varepsilon$. This agrees with regression function estimation, except for the noise $Z = \varepsilon$ which is identical to the original noise. Hence, the noise scale is $h = 1$, and $N = \mathrm{Cov}\left[\varepsilon, Z\right] = \sigma^2 I_n$.
- **Confidence intervals**. To construct confidence intervals for individual coordinates, we consider the normal model $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$. Assuming σ^2 is known, a confidence interval with coverage $1 - \alpha$ for a given coordinate β_j is

$$\hat{\beta}_j \pm \sigma z_{\alpha/2} V_j^{1/2},$$

where $z_\alpha = \Phi^{-1}(\alpha)$ is the inverse normal CDF, and V_j is the j-th diagonal entry of $(X^\top X)^{-1}$.

Therefore, we can measure the difficulty of the problem by V_j. The larger V_j is, the longer the confidence interval. This corresponds to measuring the difficulty of estimating the coordinate $L_A = \beta_j$. This can be fit in our general framework by choosing $A = E_j^\top$, the $1 \times p$ vector of zeros, with only a one in the j-th coordinate. This problem is noiseless.

3.2 Finite sample results

We now show how to calculate the efficiency explicitly in the general framework. We start with the simpler case where $h = 0$. We then have for the OLS estimator

$$M(\hat{\beta}) = \sigma^2 \cdot \left(\mathrm{tr}\left[(X^\top X)^{-1} A^\top A\right]\right).$$

For the distributed estimator with weights w_i summing to one, given by $\hat{\beta}_{dist}(w) = \sum_i w_i \hat{\beta}_i$, we have

$$M(\hat{\beta}_{dist}) = \sigma^2 \cdot \left(\sum_{i=1}^{k} w_i^2 \cdot \mathrm{tr}\left[(X_i^\top X_i)^{-1} A^\top A\right]\right).$$

Following the same steps as before, we find that the optimal efficiency is

$$E(A; X_1, \ldots, X_k) = \frac{\mathrm{tr}\left[(X^\top X)^{-1} A^\top A\right]}{\frac{1}{\sum_{i=1}^{k} \frac{1}{\mathrm{tr}\left[(X_i^\top X_i)^{-1} A^\top A\right]}}} = \mathrm{tr}\left[(X^\top X)^{-1} A^\top A\right] \cdot \sum_{i=1}^{k} \frac{1}{\mathrm{tr}\left[(X_i^\top X_i)^{-1} A^\top A\right]}. \tag{4}$$

This shows that the key to understanding the efficiency are the traces $\mathrm{tr}\left[(X_i^\top X_i)^{-1} A^\top A\right]$.

By following proposition 2.2 with minor modification, we can prove that the efficiency is at most unity.

Proposition 3.1 (Concavity for general efficiency)**.** *The function $f(X) = 1/\mathrm{tr}(X^{-1} A^\top A)$ is a concave function defined on positive definite matrices. As a consequence, the general relative efficiency for distributed estimation is at most unity for any matrices X_i:*

$$E(A; X_1, \ldots, X_k) \leq 1.$$

See Section 9.11 for the proof.

For the more general case when $h \neq 0$, we can also find the OLS MSE as

$$M(\hat{\beta}) = \sigma^2 \cdot \left[\operatorname{tr}\left((X^\top X)^{-1} A^\top A\right) - 2\operatorname{tr}\left(A(X^\top X)^{-1} X^\top N\right) + hd\right].$$

For the distributed estimator, we can find, denoting $N_i := \operatorname{Cov}[\varepsilon_i, Z]$,

$$M(\hat{\beta}_{dist}) = \sigma^2 \cdot \left(\sum_{i=1}^{k} w_i^2 \cdot \operatorname{tr}\left[(X_i^\top X_i)^{-1} A^\top A\right] - 2w_i \cdot \operatorname{tr}\left(A(X_i^\top X_i)^{-1} X_i^\top N_i\right)\right) + \sigma^2 hd.$$

Let $a_i = \operatorname{tr}\left[(X_i^\top X_i)^{-1} A^\top A\right]$, and $b_i = \operatorname{tr}\left(A(X_i^\top X_i)^{-1} X_i^\top N_i\right)$. The optimal weights can be found from a quadratic optimization problem:

$$w_i = \frac{\lambda^* + b_i}{a_i},\ \lambda^* := \frac{1 - \sum_{i=1}^k \frac{b_i}{a_i}}{\sum_{i=1}^k \frac{1}{a_i}}.$$

The resulting formula for the optimal weights, and for the global optimum, can be calculated explicitly. However, we have not found the result particularly insightful, so we not report it here. The details can be found in Section 9.12.

4 Some related work

In this section we discuss some related work. There is a great deal of work in computer science and optimization on parallel and distributed computation (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). These areas are quite mature, with well-developed algorithmic frameworks, such as the PRAM (or parallel random access memory) model, which allows precise analyses of the computational resources involved in distributed computation.

In addition, there are several popular examples of distributed computing environments, some of them specifically designed for machine learning and statistics: for instance MapReduce (Dean and Ghemawat, 2008), Hadoop (White, 2012), and Spark (Zaharia et al., 2010).

In contrast, there is less work on understanding the statistical properties, and the inherent computation-statistics tradeoffs, in distributed computation environments. This area has attracted increasing attention only in recent years, see for instance Mcdonald et al. (2009); Zhang et al. (2012, 2013b,a); Duchi et al. (2014); Zhang et al. (2015); Braverman et al. (2016); Jordan et al. (2016); Rosenblatt and Nadler (2016); Smith et al. (2016); Fan et al. (2017); Lin et al. (2017); Lee et al. (2017); Battey et al. (2018); Zhu and Lafferty (2018), and the references therein. See (Huo and Cao) for a review. We can only discuss the most closely related papers due to space limitations.

Mcdonald et al. (2009) show finite-sample bounds on the expected error of averaged estimators in multinomial regression. Zinkevich et al. (2010) propose parallel stochastic gradient descent algorithms, deriving the speed of convergence of parameter distributions to their asymptotic limits.

Zhang et al. (2013b) bound the leading order term for MSE of averaged estimation in empirical risk minimization. Their bounds do not explicitly take dimension into account. However, their empirical data example clearly has large p, considering a logistic regression with $n = 2.4 \cdot 10^8$, and $p = 740,000$, so that $1/\gamma \approx 340$. In their experiments, they distribute the data over up to 128 machines. So, our regime, where k is of the same order as $1/\gamma$ matches well their simulation setup. In addition, their concern is on regularized estimators, where they propose to estimate and reduce bias by subsampling.

Liu and Ihler (2014) study distributed estimation in statistical exponential families, showing that the efficiency loss compared to the global setting relates to how much the underlying distributions deviate from full exponential families. They also propose nonlinear KL-divergence-based combination methods, which can be more efficient than linear averaging.

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. Due to their generality, their results are more involved, and also their dimension is fixed. Lin et al. (2017) improve those results. Duchi et al. (2014) derive minimax bounds on distributed estimation where the number of bits communicated is controlled.

Rosenblatt and Nadler (2016) consider the distributed learning problem in three different settings. The first two settings are fixed dimensional. The third setting is high-dimensional M-estimation, where they study the first order behavior of estimators using prior results from Donoho and Montanari (2013); El Karoui et al. (2013). This is possibly the most closely related work to ours in the literature. They use the following representation, derived in the previous works mentioned above: a high-dimensional M-estimator can be written as $\hat\beta = \beta + r(\gamma)\Sigma^{-1/2}\xi(1+o_P(1))$, where $\xi \sim \mathcal{N}(0, I_p/p)$, and $r(\gamma)$ is a constant depending on the loss function, whose expression can be found in Donoho and Montanari (2013); El Karoui et al. (2013).

They derive a relative efficiency formula in this setting, which for OLS takes the form

$$\frac{\mathbb{E}\|\hat\beta_{dist} - \beta\|^2}{\mathbb{E}\|\hat\beta - \beta\|^2} = 1 + \gamma(1 - 1/k) + O(\gamma^2).$$

In contrast, our result for this case is equal to

$$\frac{1-\gamma}{1-k\gamma} = 1 + \gamma\frac{k-1}{1-k\gamma}.$$

Thus, our result is much more precise, and in fact exact, while of course being limited to the special case of linear regression.

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. Battey et al. (2018) study a similar problem, also including hypothesis testing under more general sparse models.

5 Calculus of deterministic equivalents

5.1 A calculus of deterministic equivalents in RMT

In the previous section we saw that the relative efficiency depends on the trace functionals $\operatorname{tr}[(X^\top X)^{-1}\ A^\top A]$, for specific matrices A. To find the limits of these functionals, we will use the technique of *deterministic equivalents* from random matrix theory. This is a method to find the almost sure limits of random quantities. See for example Hachem et al. (2007); Couillet et al. (2011) and the related work section below.

For instance, the Marchenko-Pastur (MP) law itself states that the eigenvalue distribution of certain random matrices is asymptotically deterministic. More generally, one of the best ways to understand the MP law is that *resolvents are asymptotically deterministic*. Indeed, let $\widehat\Sigma = n^{-1}X^\top X$, where $X = Z\Sigma^{1/2}$ and Z has iid entries. Then the MP law means that for any z with positive imaginary part, we have the equivalence

$$(\widehat\Sigma - zI)^{-1} \asymp (x_p\Sigma - zI)^{-1},$$

for a certain scalar $x_p = x(\Sigma, n, p, z)$. At this stage we can think of the equivalence entry-wise, but we will make this precise next. The above formulation has appeared in some early works by VI Serdobolskii, see e.g., Serdobolskii (1983), and Theorem 1 on page 15 of Serdobolskii (2007) for a very clear statement.

The consequence is that simple linear functionals of the random matrix $(\widehat{\Sigma} - zI)^{-1}$ have a deterministic equivalent based on $(x_p\Sigma - zI)^{-1}$. In particular, we may be able to express the trace functionals we need in a simpler deterministic way. In order to do this, we will take a principled approach and define some appropriate notions for a *calculus of deterministic equivalents*, which allows us to do calculations in a simple and effective way.

First, we make more precise the notion of equivalence. We say that the (deterministic or random) symmetric matrix sequences A_n, B_n of growing dimensions are *equivalent*, and write

$$A_n \asymp B_n$$

if

$$\lim_{n\to\infty} |\mathrm{tr}\,[C_n(A_n - B_n)]| = 0$$

almost surely, for any sequence C_n of symmetric deterministic matrices with bounded trace norm, i.e., such that

$$\limsup \|C_n\|_{tr} < \infty.$$

We call such a sequence C_n a *standard sequence*. Recall here that the trace norm (or nuclear norm) is defined by $\|M\|_{tr} = \mathrm{tr}((M^\top M)^{1/2}) = \sum_i \sigma_i$, where σ_i are the singular values of M. Since the matrices considered here are real symmetric, we also have that $\|M\|_{tr} = \sum_i |\lambda_i|$, where λ_i are the eigenvalues of M.

5.2 General MP theorem

To find the limits of the efficiencies, the most important deterministic equivalent will be the following general Marchenko-Pastur theorem. See Section 9.13 for the proof, which relies on the generalized Marchenko-Pastur theorem of Rubio and Mestre (2011).

Theorem 5.1 (Deterministic equivalent in elliptical models, consequence of Rubio and Mestre (2011))**.** *Let the $n \times p$ data matrix X follow the elliptical model*

$$X = \Gamma^{1/2} Z \Sigma^{1/2},$$

where Γ is an $n \times n$ diagonal matrix with non-negative entries representing the scales of the n observations, and Σ is a $p \times p$ positive definite matrix representing the covariance matrix of the p features. Assume the following:

1. *The entries of Z are iid random variables with mean zero, unit variance, and finite $8+c$-th moment, for some $c > 0$.*
2. *The eigenvalues of Σ, and the entries of Γ, are uniformly bounded away from zero and infinity.*
3. *We have $n, p \to \infty$, with $\gamma_p = p/n$ bounded away from zero and infinity.*

Let $\widehat{\Sigma} = n^{-1}X^\top X$ be the sample covariance matrix. Then $\widehat{\Sigma}$ is equivalent to a scaled version of the population covariance

$$\widehat{\Sigma}^{-1} \asymp \Sigma^{-1} \cdot e_p.$$

Here $e_p = e_p(n, p, \Gamma) > 0$ is the unique solution of the fixed-point equation

$$1 = \frac{1}{n} \mathrm{tr}\left[e_p\Gamma(I + \gamma_p e_p \Gamma)^{-1}\right].$$

Thus, the inverse sample covariance has a deterministic equivalent in terms of a scaled version of the inverse population covariance. This result does not require the convergence of the aspect ratio p/n, or of the e.s.d. of Σ, Γ. However, if the empirical spectral distribution of Γ tends to G, the above equation has a limit which agrees with the previous equation, namely

$$\int \frac{se}{1+\gamma se} dG(s) = 1.$$

The usual MP theorem is a special case of the above result where $\Gamma = I_n$. As a result, we obtain the following corollary:

Corollary 5.2 (Deterministic equivalent in MP models). *Let the $n \times p$ data matrix X follow the model $X = Z\Sigma^{1/2}$, where Σ is a $p \times p$ positive definite matrix representing the covariance matrix of the p features. Assume the same conditions on Σ from Theorem 5.1. Then $\widehat{\Sigma}$ is equivalent to a scaled version of the population covariance*

$$\widehat{\Sigma}^{-1} \asymp \frac{1}{1-\gamma_p} \cdot \Sigma^{-1}.$$

The proof is immediate, by checking that $e_p = 1/(1-\gamma_p)$ in this case.

5.2.1 Related work on deterministic equivalents

There are several works in random matrix theory on deterministic equivalents. One of the early works is Serdobolskii (1983), see Serdobolskii (2007) for a modern summary. The name ”deterministic equivalents” and technique was more recently introduced and re-popularized by Hachem et al. (2007) for signal-plus-noise matrices. Later Couillet et al. (2011) developed deterministic equivalents for matrix models of the type $\sum_{k=1}^{B} R_k^{1/2} X_k T_k X_k^\top R_k^{1/2}$, motivated by wireless communications. See the book Couillet and Debbah (2011) for a summary of related work. See also Müller and Debbah (2016) for a tutorial. However, many of these results are stated only for some fixed functional of the resolvent, such as the Stieltjes transform. One of our points here is that there is a much more general picture.

Rubio and Mestre (2011) is one of the few works that explicitly states more general convergence of arbitrary trace functionals of the resolvent. Our results are essentially a consequence of theirs.

However, we think that it is valuable to define a set of rules, a ”calculus” for working with deterministic equivalents, and we use those techniques in our paper. Similar ideas for operations on deterministic equivalents have appeared in Peacock et al. (2008), for the specific case of a matrix product. Our approach is more general, and allows many more matrix operations, see below.

5.3 Rules of calculus

The calculus of deterministic equivalents has several properties that simplify calculations. We think these justify the name of *calculus*. Below, we will denote by A_n, B_n, C_n etc, sequences of deterministic or random matrices. See Section 9.14 for the proof.

Theorem 5.3 (Rules of calculus). *The calculus of deterministic equivalents has the following properties.*

1. **Equivalence.** *The $\asymp$ relation is indeed an equivalence relation.*
2. **Sum.** *If $A_n \asymp B_n$ and $C_n \asymp D_n$, then $A_n + C_n \asymp B_n + D_n$.*
3. **Product.** *If A_n is a sequence of matrices with bounded operator norms i.e., $\|A_n\|_{op} < \infty$, and $B_n \asymp C_n$, then $A_nB_n \asymp A_nC_n$.*

4. **Trace.** *If $A_n \asymp B_n$, then $\operatorname{tr}\{n^{-1}A_n\} - \operatorname{tr}\{n^{-1}B_n\} \to 0$ almost surely.*
5. **Stieltjes transforms.** *As a consequence, if $(A_n - zI_n)^{-1} \asymp (B_n - zI_n)^{-1}$, then $m_{A_n}(z) - m_{B_n}(z) \to 0$ almost surely. Here $m_{X_n}(z) = n^{-1}\operatorname{tr}(X_n - zI_n)^{-1}$ is the Stieltjes transform of the empirical spectral distribution of X_n.*

In addition, we note that the calculus of deterministic equivalents has additional properties, such as continuous mapping theorems, differentiability, etc. However, we do not need those explicitly in the current paper, so we leave them for future work.

6 Examples

We now show how to use the calculus of deterministic equivalents to find the limits of the trace functionals in our general framework. We study each problem in turn.

6.1 Regression function estimation

For estimating the regression function, we have $\mathbb{E}\|X(\beta - \hat\beta)\|^2 = \sigma^2 p$. We then find via equation (4) the prediction efficiency

$$FE(X_1, \ldots, X_k) = \sum_{i=1}^{k} \frac{p}{\operatorname{tr}((X_i^\top X_i)^{-1} X^\top X))}.$$

For asymptotics, we consider as before elliptical models.

Theorem 6.1 (FE for elliptical and MP models)**.** *Under the conditions of Theorems 5.1 and 2.4, the FE has the almost sure limit*

$$FE(X_1, \ldots, X_k) \to_{a.s.} \sum_{i=1}^{k} \frac{1}{1 + \left(\frac{1}{\gamma}\mathbb{E}_G T - \frac{1}{\gamma_i}\mathbb{E}_{G_i} T\right) f(\gamma_i, G_i)}.$$

Under the conditions of Corollary 5.2, the FE has the almost sure limit $(1/\gamma - k)/(1/\gamma - 1)$. So for Marchenko-Pastur models, the limit is the same as for parameter estimation from Theorem 2.3.

See Section 9.15 for the proof. This efficiency is more complex than that for estimation error. In particular, it does not depend on a simple linear way on k, but rather via a ratio of two linear functions of k. However, it can be checked that many of the properties (e.g., monotonicity) for ARE still hold here.

6.2 In-sample prediction (Training error)

For in-sample prediction, we start with the well known formula

$$\mathbb{E}||X(\beta - \hat\beta) + \varepsilon||^2 = \sigma^2[n - \operatorname{tr}((X^\top X)^{-1} X^\top X] = \sigma^2(n - p).$$

As we saw, to fit in-sample prediction in the general framework, we need to take the transform matrix $A = X$, the noise $Z = \varepsilon$, and the covariance matrices $N_i = \operatorname{Cov}[\varepsilon_i, Z] = \operatorname{Cov}[\varepsilon_i, \varepsilon]$. Then, in the formula for optimal weights we need to take $a_i = \operatorname{tr}[(X_i^\top X_i)^{-1} X^\top X]$ and $b_i = \operatorname{tr}(X(X_i^\top X_i)^{-1} X_i^\top N_i) = \operatorname{tr}[(X_i^\top X_i)^{-1} X_i^\top N_i X] = \operatorname{tr}[(X_i^\top X_i)^{-1} X_i^\top X_i] = p$. Therefore, the optimal error for distributed regression is achieved by the weights

$$w_i = \frac{\lambda - b_i}{a_i} = \frac{\lambda - p}{a_i}, \quad \lambda = \frac{1 - \sum_{i=1}^{k} \frac{b_i}{a_i}}{\sum_{i=1}^{k} \frac{1}{a_i}} = \frac{1}{\sum_{i=1}^{k} \frac{1}{a_i}} - p.$$

Plugging these into $M(\hat\beta_{dist})$ given in the general framework, we find

$$M(\hat\beta_{dist}) = \sigma^2\left(n - 2p + \frac{1}{\sum_{i=1}^k \frac{1}{a_i}}\right), \quad a_i = \operatorname{tr}((X_i^\top X_i)^{-1} X^\top X).$$

Thus, the optimal in-sample prediction efficiency is

$$IE(X_1, \ldots, X_k) = \frac{n-p}{n - 2p + \frac{1}{\sum_{i=1}^k \frac{1}{\operatorname{tr}((X_i^\top X_i)^{-1} X^\top X)}}}.$$

For asymptotics in elliptical models, we find:

Theorem 6.2 (IE for elliptical and MP models)**.** *Under the conditions of Theorems 5.1 and 2.4, the IE has the almost sure limit*

$$IE(X_1, \ldots, X_k) \to_{a.s.} \frac{1-\gamma}{1 - 2\gamma + \frac{1}{\sum_{i=1}^k \psi(\gamma_i, G_i)}},$$

where ψ is the following functional of the distributions G_i and G, depending on the inverse of the η-transform f defined in equation (2):

$$\psi(\gamma_i, G_i) = \frac{1}{\gamma + (\mathbb{E}_G T - \frac{\gamma}{\gamma_i}\mathbb{E}_{G_i} T) f(\gamma_i, G_i)}.$$

Under the conditions of Corollary 5.2, the IE has the almost sure limit

$$IE(X_1, \ldots, X_k) \to_{a.s.} \frac{1-\gamma}{1 - 2\gamma + \frac{\gamma(1-\gamma)}{1-k\gamma}} = \frac{1}{1 + \frac{(k-1)\gamma^2}{(1-k\gamma)(1-\gamma)}}.$$

See Section 9.16 for the proof.

6.3 Out-of-sample prediction (Test error)

In out-of-sample prediction, 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, and only x_t is observable. We want to use $x_t^\top \hat\beta$ to predict y_t. We compare the prediction error of two estimators:

$$OE(x_t; X_1, \ldots, X_k) := \frac{\mathbb{E}\left[(y_t - x_t^\top \hat\beta)^2\right]}{\mathbb{E}\left[(y_t - x_t^\top \hat\beta_{dist})^2\right]}.$$

In our general framework, we saw that this corresponds to predicting the linear functional $x_t^\top \beta + \varepsilon_t$. Based on equation (4), the optimal out-of-sample prediction efficiency is

$$OE(x_t; X_1, \ldots, X_k) = \frac{1 + x_t^\top (X^\top X)^{-1} x_t}{1 + \frac{1}{\sum_{i=1}^k \frac{1}{x_t^\top (X_i^\top X_i)^{-1} x_t}}}.$$

For asymptotics in elliptical models, we find the following result. Since the samples have the form $x_i = g_i^{1/2} \Sigma^{1/2} z_i$, the test sample depends on a scale parameter g_t.

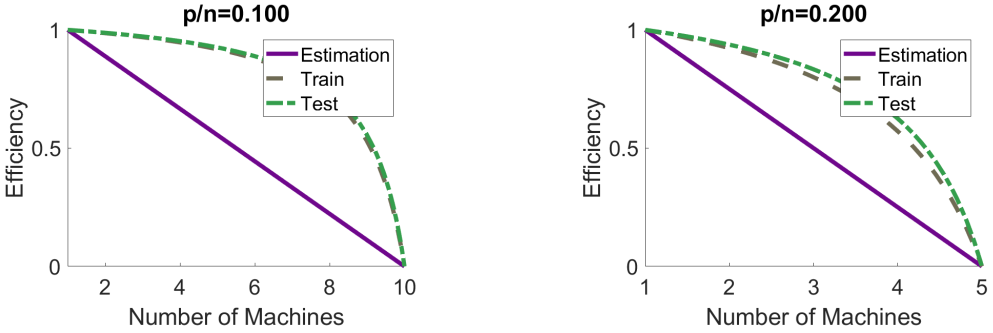

Figure 6: Relative efficiency for Marchenko-Pastur model.

Theorem 6.3 (OE for elliptical and MP models)**.** *Under the conditions of Theorems 5.1 and 2.4, the OE has the almost sure limit, conditional on* g_t

$$OE(x_t; X_1, \ldots, X_k) \to_{a.s.} \frac{1 + g_t \cdot f(\gamma, G)}{1 + \frac{g_t}{\sum_{i=1}^k \frac{1}{f(\gamma_i, G_i)}}}.$$

For Marchenko-Pastur models under the conditions of Corollary 5.2, the OE has the almost sure limit

$$\frac{\frac{1}{1-\gamma}}{1 + \frac{\gamma}{1-k\gamma}} = \frac{1}{1 + \frac{(k-1)\gamma^2}{1-k\gamma}}.$$

See Section 9.17 for the proof. If the scale parameter g_t is random, then the OE typically does not have an almost sure limit, and converges in distribution to a random variable instead. We mention that Theorem 6.3 holds under even weaker conditions, if we are only given the $4{+}c$-th moment of z_1 instead of $8{+}c$-th one. The argument is slightly different, and is presented in the location referenced above.

One can check that that $OE \geq RE$. Thus, out-of-sample prediction incurs a smaller efficiency loss than estimation. The intuition is that the out-of-sample prediction always involves a fixed loss due to the *irreducible noise* in the test sample, which ”amortizes” the problem. Moreover,

$$OE \geq IE \geq RE.$$

The intuition here is that IE incurs a smaller fixed loss than OE, because the noise in the training set is effectively reduced, as it is already partly fit by our estimation process. So the graph of IE will be in between the other two criteria. See Figure 6. We also see that the IE is typically very close to OE.

In addition, we also mention that if we care about the increase of the reducible part of the error, then this is the same as for the estimation error. That is, the prediction error has two components: the irreducible noise, and the reducible error. The reducible error has the same behavior as for estimation, and thus on figure 6 it would have the same plot as the curve for estimation.

6.4 Confidence intervals

To form confidence intervals, we consider the normal model $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$. Recall that in this model the OLS estimator has distribution $\hat\beta \sim \mathcal{N}(\beta, \sigma^2 (X^\top X)^{-1})$. Assuming σ^2 is known, an exact level $1-\alpha$ confidence interval for a given coordinate β_j can be formed as

$$\hat{\beta}_j \pm \sigma z_{\alpha/2} V_j^{1/2},$$

where $z_\alpha = \Phi^{-1}(\alpha)$ is the inverse normal CDF, and V_j is the j-th diagonal entry of $(X^\top X)^{-1}$. We follow the same program as before, comparing the length of the confidence intervals formed based on our two estimators. However, for technical reasons it is more convenient to work with squared length.

Thus we consider the criterion

$$CE(j; X_1, \ldots, X_k) := \frac{V_j}{V_{j,dist}}.$$

Here $V_{j,dist}$ is the variance of the j-th entry of an optimally weighted distributed estimator. As we saw in our framework, this is equivalent to estimating the j-th coordinate of β. Hence the optimal confidence interval efficiency is

$$CE(j; X_1, \ldots, X_k) = [(X^\top X)^{-1}]_{jj} \cdot \sum_{i=1}^{k} \frac{1}{[(X_i^\top X_i)^{-1}]_{jj}}. \tag{5}$$

For asymptotics, we find:

Theorem 6.4 (CE for elliptical and MP models)**.** *Under the conditions of Theorems 5.1 and 2.4, the CE has the same limit as the ARE from Theorem 2.4. Therefore, for Marchenko-Pastur models, the CE also has the form before, $CE(j) = (1/\gamma - k)/(1/\gamma - 1)$.*

See Section 9.18 for the proof.

6.5 Understanding and comparing the efficiencies

We give two perspectives for understanding and comparing the efficiencies. The key qualitative insight is that estimation and CIs are much more affected than prediction by distributed computation. We make this precise below.

Criticality of k. We ask: What is the largest number of machines we can use such that the asymptotic efficiency is at least 1/2? Let us call this the *critical* number of machines. It is easy to check that for estimation and CIs, $k_R = (\gamma+1)/(2\gamma)$. For training error, $k_{Tr} = (\gamma^2-\gamma+1)/\gamma$, while for test error, $k_{Te} = (\gamma^2+1)/(\gamma^2+\gamma)$.

We also have the following asymptotics as $\gamma \to 0$:

$$k_R \asymp 1/(2\gamma),$$

while

$$k_{Tr} \asymp k_{Te} \asymp 1/\gamma.$$

So the number of machines that can be used is nearly maximal (i.e., n/p) for training and test error, while it is about *half that* for estimation error and CIs. This shows quantitatively that estimation and CIs are much more affected than prediction by distributed computation.

Edge efficiency. The maximum number of machines that we can use is approximately $k^* = 1/\gamma - 1$, for small γ. Let us define the *edge efficiency* e^* as the relative efficiency achieved at this edge case. For estimation and CIs, we have $e_R^* = \gamma/(1-\gamma)$. For training error, $e_{Tr}^* = (1-\gamma)/(2-3\gamma)$, and for test error, $e_{Te}^* = 1/[2(1-\gamma)]$.

We also have the following asymptotic values as $\gamma \to 0$:

$$e_R^* \asymp \gamma,$$

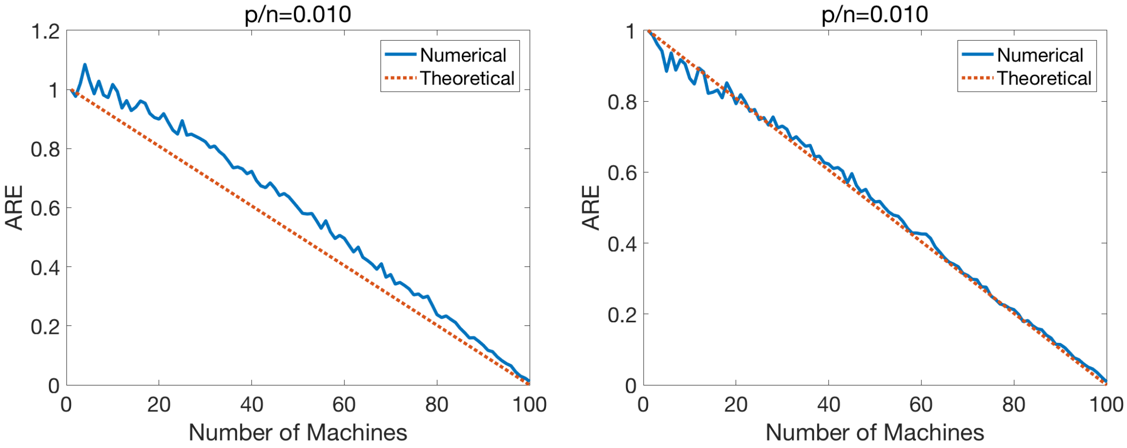

Figure 7: Relative efficiency in regression.

while

$$e^*_{Tr} \asymp \frac{1}{2} + \frac{\gamma}{4}, \text{ and } e^*_{Te} \asymp \frac{1}{2} + \frac{\gamma}{2}.$$

This shows that for $n \gg p$ the edge efficiency is vanishing for estimation and CIs, while it is approximately $1/2$ for training and test error. Thus, even for the maximal number of machines, prediction error is not greatly increased.

7 Numerical simulations

We present a few numerical results to complement our theory, in addition to the numerical results already shown in the paper.

7.1 Relative efficiency for regression

Figure 7 shows a comparison of our theoretical formulas for ARE and realized relative efficiency in a regression simulation. Here we consider regression problems with $Y = X\beta + \varepsilon$, where X is $n \times p$ with iid standard Gaussian entries, $\beta = 0$, and ε has iid standard Gaussian entries. We choose $n > p$, and for each value of k such that $k < n/p$, we split the data into k equal groups. We then 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 take $n = 10,000$ and $p = 100$.

We observe that the two agree closely. However, there is more sampling variation than in the previously reported simulations, where we only compared the expected values of the relative efficiency to its asymptotic limit. In particular, the realized relative efficiency can be greater than unity. This is not a contradiction as our theoretical results only concern the expected values. However, we find that the simulations still match the theoretical results quite well.

7.2 Relative efficiencies for the elliptical model

For the elliptical model, we can also study the relation between different asymptotic efficiencies and show a plot similar to Figure 1 for the Marchenko-Pastur model. Intuitively, we cannot

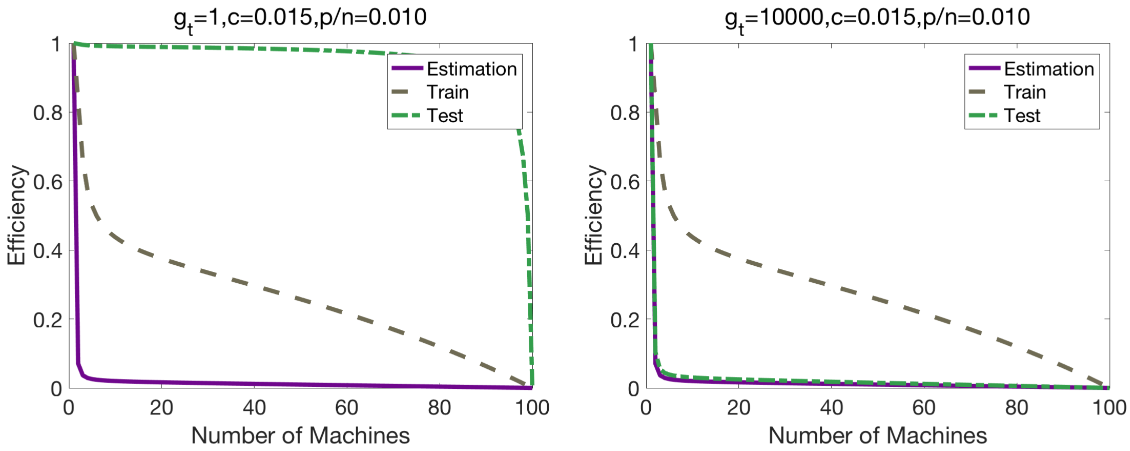

Figure 8: Relative efficiency for elliptical model.

expect a universal phenomenon in this situation since all efficiencies depend on the distribution G. Let us consider the worst-case example from Theorem 2.7. Figure 8 shows that the asymptotic relative efficiency for out-of-sample prediction could be either very good or as bad as the ARE.

In the first plot, we take $p = 100, n = 10000$, while $\alpha = 10000$, and $c = 0.015$. The test datapoint has magnitude $g_t = 1$. In the second plot, we choose the same parameters but change the magnitude of the test datapoint to $g_t = 10000$. Intuitively, when g_t is large, the irreducible noise is negligible. Otherwise, the irreducible noise will make the problem easier. This is precisely what we observe in our figure, where in the first case, test error increases only a little (the efficiency is nearly unity), while in the second case test error increases a lot. We also expect this from our formula in Theorem 6.3.

8 Empirical data analysis

In this section we present an empirical data example to assess the accuracy of our theoretical results. Specifically, figure 9 shows a comparison of our theoretical formulas for OE and actual out-of-sample prediction error (test error) on the NYC flights dataset (Wickham, 2018). We observe a quite good match.

Specifically, we performed the following steps in our data analysis. We downloaded the flights data as included in the nycflights13 R package (Wickham, 2018). We joined the separate datasets (weather, planes, and airlines). We omitted data points with missing entries. We removed one out of each pair of variables with absolute correlation higher than 0.8. This left a total of $N = 60,448$ samples and $p = 17$ variables. For each n in the range $n = 500, 1000, 3000, 10000$, we randomly sampled a training set of size n, and a non-overlapping test set of size also equal to n. The test set size does not have equal the training set size, and we only followed this protocol for simplicity.

We then fit linear regression estimators to this data in a global and distributed way. For the distributed version, we split the train data as equally as possible into k subsets, for each $k \leq n/p$. We then fit a linear regression to each subset, and took a weighted average with the optimal weights. We computed the test error of both the global and the distributed estimators

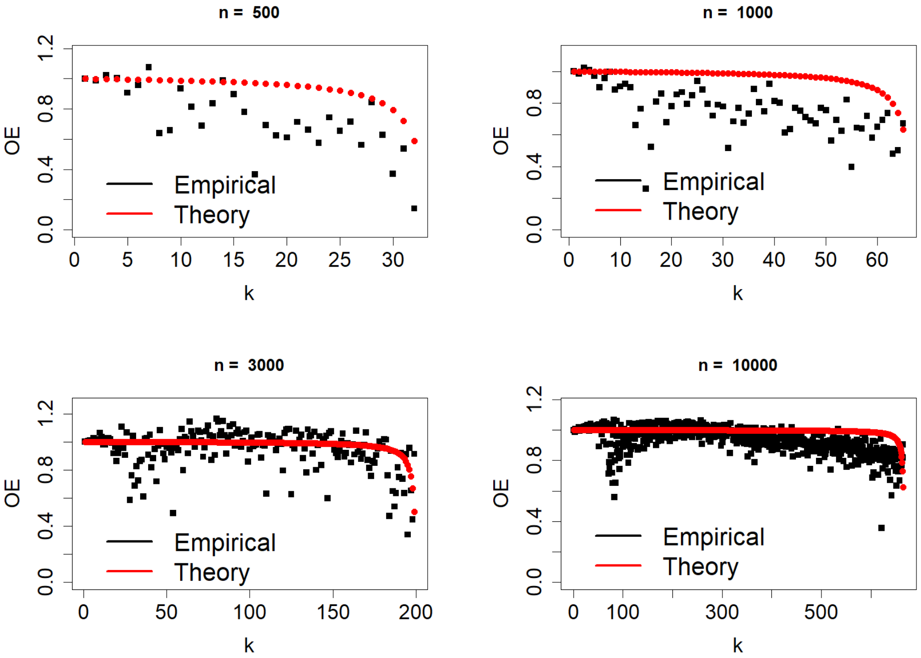

Figure 9: NYC flights data.

over the test sample, and defined their ratio to be the empirical OE. We compared this to our theoretical formula for the OE, see figure 9.

We observe a quite good match between the theoretical and empirical results. However, the empirical estimate of OE can be larger than unity. This is because of sampling noise. Our results show that $OE \leq 1$, but only for the theoretical quantity where we have taken expectations. To get estimators with reduced variance, one could average over multiple Monte Carlo trials. However, those are beyond the our scope.

Acknowledgements

The authors thank Jason D. Lee, Philip Gressman, Andreas Haeberleen, Boaz Nadler, Balasubramanian Narashiman and Ziwei Zhu for helpful discussions. The authors are grateful to Sifan Liu for providing an initial script for processing the empirical data. ED was partially supported by NSF BIGDATA grant IIS 1837992.

9 Proofs

9.1 Proof of Lemma 2.1

Each local estimator is unbiased, and has MSE $M_i = \sigma^2 \operatorname{tr}[(X_i^\top X_i)^{-1}]$. If we restrict to $\sum_{i=1}^k w_i = 1$, then the weighted estimator is unbiased and its MSE equals

$$MSE(w) = \sum_i w_i^2 \cdot MSE(\hat{\beta}_i) = \sum_i w_i^2 \cdot M_i.$$

Clearly, to minimize this subject to $\sum_i w_i = 1$, by the Cauchy-Schwarz inequality we should take $w_i^* = M_i^{-1}/(\sum_j M_j^{-1})$, and the minimum is $1/(\sum_j M_j^{-1})$. This finishes the proof.

9.2 Proof of Proposition 2.2

Let us define $g(t) = f(X + tV)$, where $X \succ 0$ is a positive definite matrix and V is any symmetric matrix such that $X + tV \succ 0$ is still positive definite. Then $f(X)$ is concave iff $g(t)$ is concave on its domain for any X and V.
Now we have

$$\begin{aligned} g(t) = \frac{1}{\operatorname{tr}[(X+tV)^{-1}]} &= \frac{1}{\operatorname{tr}[X^{-1}(I + tX^{-1/2}VX^{-1/2})^{-1}]} \\ &= \frac{1}{\operatorname{tr}[X^{-1}Q(I+t\Lambda)^{-1}Q^\top]} \\ &= \frac{1}{\operatorname{tr}[Q^\top X^{-1}Q(I+t\Lambda)^{-1}]} \\ &= \left(\sum_{i=1}^n \frac{(Q^\top X^{-1}Q)_{ii}}{1+t\lambda_i}\right)^{-1}, \end{aligned}$$

where λ_i-s are eigenvalues of $X^{-1/2}VX^{-1/2}$. From the assumption, we always have $1 + t\lambda_i > 0$. Since $Q^\top X^{-1}Q$ is a positive definite matrix, its diagonal elements are all positive. We may use the notation $\alpha_i = (Q^\top X^{-1}Q)_{ii}$. Then, let us compute $g'(t)$ and $g''(t)$. First we have

$$g'(t) = \left(\sum_{i=1}^n \frac{\alpha_i}{1+\lambda_i t}\right)^{-2} \cdot \left(\sum_{i=1}^n \frac{\alpha_i\lambda_i}{(1+\lambda_i t)^2}\right),$$

Next, we find

$$g''(t) = 2\left(\sum_{i=1}^n \frac{\alpha_i}{1+\lambda_i t}\right)^{-3} \cdot \left[\left(\sum_{i=1}^n \frac{\alpha_i\lambda_i}{(1+\lambda_i t)^2}\right)^2 - \left(\sum_{i=1}^n \frac{\alpha_i}{1+\lambda_i t}\right)\left(\sum_{i=1}^n \frac{\alpha_i\lambda_i^2}{(1+\lambda_i t)^3}\right)\right]$$

Multiplying by $-2\left(\sum_{i=1}^n \frac{\alpha_i}{1+\lambda_i t}\right)^3$, we get the expression

$$\begin{aligned} &\sum_{1\le i<j\le n} \frac{\alpha_i\alpha_j\lambda_j^2}{(1+\lambda_i t)(1+\lambda_j t)^3} + \frac{\alpha_j\alpha_i\lambda_i^2}{(1+\lambda_j t)(1+\lambda_i t)^3} - \frac{2\alpha_i\alpha_j\lambda_i\lambda_j}{(1+\lambda_i t)^2(1+\lambda_j t)^2} \\ &= \sum_{1\le i<j\le n} \frac{\alpha_i\alpha_j}{(1+\lambda_i t)^3(1+\lambda_j t)^3}[\lambda_j^2(1+\lambda_i t)^2 + \lambda_i^2(1+\lambda_j t)^2 - 2\lambda_i\lambda_j(1+\lambda_i t)(1+\lambda_j t)] \\ &= \sum_{1\le i<j\le n} \frac{\alpha_i\alpha_j(\lambda_i-\lambda_j)^2}{(1+\lambda_i t)^3(1+\lambda_j t)^3} \ge 0. \end{aligned}$$

Hence $g(t)$ concave, and so is $f(X)$.

We can use the convexity directly to check RE is less than or equal to unity. Indeed, f is affine, in the sense that $f(cX) = cf(X)$ for any $c > 0$. The concavity result that we proved implies that, with $A_i = X_i^\top X_i$,

$$\sum_{i=1}^k f(A_i)/k \le f\left(\sum_{i=1}^k A_i/k\right).$$

By the affine nature of f, this result implies that f is sub-additive. This can be checked to be equivalent to $RE \le 1$, finishing the proof.

9.3 Proof of Theorem 2.3

It follows from known results on the Marchenko-Pastur distribution (Marchenko and Pastur, 1967; Bai and Silverstein, 2009) that, with $F = F_\gamma$ the l.s.d. of $\widehat{\Sigma}$

$$\mathbb{E}_F T^{-1} = \frac{\mathbb{E}_H T^{-1}}{1-\gamma}.$$

Indeed, recall that the Stieltjes transform of a signed measure μ on $[0, \infty)$ is defined as the map $m : \mathbb{C} \setminus [0, \infty) \to \mathbb{C}$, $m(z) = \int (x-z)^{-1} d\mu(x)$. Let $m(z) = m_\gamma(z; H)$ be the Stieltjes transform of the limiting e.s.d. F. This satisfies the Marchenko-Pastur equation (Marchenko and Pastur, 1967)

$$m(z) = \int \frac{1}{t[1-\gamma-\gamma z m(z)] - z} dH(t).$$

where H is the *l.s.d.* of the *e.s.d.* of Σ. By taking $z \to 0$, we have

$$m(0) = \int \frac{1}{t(1-\gamma)} dH(t).$$

The rigorous argument for this claim is provided in the proof of Theorem 2.4. Since $m(0) = \mathbb{E}_F T^{-1}$ and the left-hand side equals $\mathbb{E}_H T^{-1}/(1-\gamma)$, this proves the required claim.

Thus, the ARE equals

$$ARE = \frac{\gamma \cdot \mathbb{E}_H T^{-1}}{1-\gamma} \cdot \sum_{i=1}^k \frac{1-\gamma_i}{\gamma_i \cdot \mathbb{E}_H T^{-1}} = \frac{\sum_{i=1}^k (1/\gamma_i - 1)}{1/\gamma - 1} = \frac{1/\gamma - k}{1/\gamma - 1}.$$

In the last line, we used that $\sum 1/\gamma_i = 1/\gamma$, because $\sum n_i/p = n/p$. Moreover, we clearly have the finite sample approximation $(n-kp)/(n-p)$ to this expression. This finishes the proof.

9.4 Suboptimal weights

If we take all weights w_i to be equal, i.e. $w_i = 1/k$, then the MSE is

$$MSE_{subopt} = \frac{\sigma^2}{k^2} \sum_{i=1}^k \operatorname{tr}(X_i^\top X_i)^{-1} \to \frac{\sigma^2}{k^2} \sum_{i=1}^k \frac{\gamma_i \cdot \mathbb{E}_H T^{-1}}{1-\gamma_i}.$$

Thus the ARE of the equally weighted estimator becomes (with the notation AE denoting asymptotic MSE)

$$ARE_{subopt} = \frac{AE(\hat{\beta}_{dist}(1/k, \ldots, 1/k))}{AE_{subopt}} = \frac{k^2 \frac{\gamma}{1-\gamma}}{\sum_{i=1}^k \frac{\gamma_i}{1-\gamma_i}}.$$

Now, ARE_{subopt} can be viewed as a harmonic mean of the numbers

$$k\frac{\gamma}{1-\gamma}\frac{1-\gamma_i}{\gamma_i},$$

while the optimal ARE is the corresponding arithmetic mean. Therefore, we have $ARE_{subopt} \leq ARE$.

9.5 Properties and interpretation of the relative efficiency.

Let $f(n,p,k)$ be the relative efficiency for estimation, $(n-kp)/(n-p)$. If $kp > n$, that expression is negative, but in that case it is more proper to define the relative efficiency as 0. So, we consider

$$f(n,p,k) = \max\left(\frac{n-kp}{n-p}, 0\right).$$

This has the following properties. Each of these has a statistical interpretation.

1. **Well-definedness**. f is well-defined for all n,p,k such that $n > p$
2. **Range**. $0 \leq f \leq 1$ for all n,p,k. Clearly the efficiency should be between zero and unity. Also, f is zero for $k \geq n/p$. In this case, some machine has an OLS estimator that is not well-defined.
   Moreover, $f = 1$ when $k = 1$ or when $p = 0$. When we have one machine, the efficiency is unity by definition. When $p = 0$, the problem is not well-defined, as there are no parameters to estimate.
3. **Monotonicity**.
   (a) **f is monotone decreasing in k.** This property is easy to interpret. The distributed regression problem gets harder as k increases.
   (b) **f is monotone increasing in n.** The linear regression problem should get easier as n grows. However, it turns out that more is true. The distributed problem gets relatively easier compared to the "shared" problem.
   (c) **f is monotone decreasing in p.** Similarly, a typical linear regression problem should get harder as p grows. However, the relative difficulty of solving the distributed problem also gets larger.
4. **Limits and singularity**.
   (a) $n \to \infty$. When $n \to \infty$ with fixed k,p, then f tends to unity. When $n \to \infty$, the distributed estimator becomes asymptotically efficient.
   (b) $p = n$. The function is singular when $p = n$, because the OLS estimator itself is singular when $p = n$.

Note that these properties are not enough to characterize the relative efficiency. In fact, for any monotone increasing transform such that $g(0) = 0$ and $g(1) = 1$, $g(f(n,p,k))$ has the same properties.

9.6 Minimax optimality

Our results show that the distributed regression estimator is minimax rate-optimal as long as the number of machines is not too large. Indeed, it is well known that the minimax estimation error in linear regression is $\sigma^2 \operatorname{tr}[(X^\top X)^{-1}]$ (Lehmann and Casella, 1998). Asymptotically, our results show that the estimation error of OLS is less than that of one-step averaging by a factor ARE given in (1). Thus, as long as $ARE > c > 0$ for some universal constant $c > 0$, we can say that one-step averaging in distributed linear regression is asymptotically minimax rate-optimal.

Indeed, in finite samples minimax rate optimality of a sequence of estimators $\hat{\theta}_n$ with respect to the risk functions R_n is defined as $R_n(\hat{\theta}_n) \le CR_n^*$, where R_n^* is the minimax risk with n samples, and where C is any universal constant independent of n. Asymptotically, this is equivalent to $\liminf_n R_n^*/R_n(\hat{\theta}_n) > 0$. For our problem, it can be checked that this holds precisely if

$$\limsup_{n,p\to\infty,\, k\in\mathbb{N}} \frac{kp}{n} < 1.$$

This gives a precise condition under which one-step averaging is rate-optimal. However, we note that our results are much stronger than that, because we find the *exact limit* of the risk, and not just up to unspecified constants.

9.7 Proof of Theorem 2.4

Consider the matrix

$$\widehat{\Sigma} = \frac{1}{n} X^\top X = \frac{1}{n} \Sigma^{\frac{1}{2}} Z^\top \Gamma Z \Sigma^{\frac{1}{2}}.$$

Recall that the *e.s.d.* of Γ converges to G, and that the *e.s.d.* of Σ converges to H. According to Paul and Silverstein (2009), with probability 1, the *e.s.d.* of $\widehat{\Sigma}$ converges to a distribution F, whose Stieltjes transform $m(z), z \in \mathbb{C}^+$ is given by

$$m(z) = \int \frac{1}{t \int \frac{s}{1+\gamma s e} dG(s) - z} dH(t),$$

where $e = e(z)$ is the unique solution in $\mathbb{C}^+$ of the equation

$$e = \int \frac{t}{t \int \frac{s}{1+\gamma s e} dG(s) - z} dH(t).$$

Since $\mathrm{tr}[(X^\top X)^{-1}] \to \gamma m(0)$, we only need to solve for $m(0)$. As we will show below, we can take $z \to 0$, and obtain that

$$m(0) = \frac{\int \frac{1}{t} dH(t)}{\int \frac{s}{1+\gamma s e} dG(s)}, \quad e(0) = \frac{1}{\int \frac{s}{1+\gamma s e} dG(s)},$$

hence $e := e(0)$ can be checked to be the unique positive solution to the equation

$$\int \frac{se}{1+\gamma s e} dG(s) = 1.$$

To make this rigorous, we need to use some results from Couillet and Hachem (2014). Let μ_F, μ_G, μ_H be the probability measures corresponding to the distributions F, G, H. Our goal is to show that, when μ_H is compactly supported away from the origin, μ_G is compactly supported and does not have a point mass at the origin, then μ_F is also compactly supported away from the origin and the solutions $m(z)$, $e(z)$ to the above equations can be extended to the origin.

First, for any $z \in \mathbb{C}^+$, Couillet and Hachem (2014) showed that the system of equations

$$\delta = \int \frac{\gamma t}{-z(1+\tilde{\delta} t)} d\mu_H(t), \ \tilde{\delta} = \int \frac{t}{-z(1+\delta t)} d\mu_G(t)$$

admits a unique solution $(\delta, \tilde{\delta}) \in (\mathbb{C}^+)^2$. Let $\delta(z)$ and $\tilde{\delta}(z)$ be these solutions. Notice that $\delta(z)$ and $e(z)$ have the following relation: $\delta(z) = \gamma e(z)$. Therefore, we can equivalently study $\delta(z)$ instead of $e(z)$. The function $m(z)$, which is the Stieltjes transform of μ_F, can be expressed as:

$$m(z) = \int \frac{1}{-z(1+\tilde{\delta}(z) t)} d\mu_H(t), \quad z \in \mathbb{C}^+.$$

We will use this expression later.

A important and useful proposition from Couillet and Hachem (2014) is that the functions $\delta(z), \tilde{\delta}(z)$ admit the representations

$$\delta(z) = \int_0^\infty \frac{1}{t-z} d\rho(t), \quad \tilde{\delta}(z) = \int_0^\infty \frac{1}{t-z} d\tilde{\rho}(t),$$

where ρ and $\tilde{\rho}$ are two Radon positive measures on $\mathbb{R}^+$ such that

$$0 < \int_0^\infty \frac{1}{1+t} d\rho(t) < \infty, \quad 0 < \int_0^\infty \frac{1}{1+t} d\tilde{\rho}(t) < \infty.$$

Thus, $\delta(z)$ and $\tilde{\delta}(z)$ can be analytically extended to $\mathbb{C}\backslash\text{supp}(\rho)$ and $\mathbb{C}\backslash\text{supp}(\tilde{\rho})$ respectively.

For the support of measures $\mu_F, \mu_G, \mu_H, \rho$ and $\tilde{\rho}$, we have the following relations from Couillet and Hachem (2014):

1. $$\mu_F(\{0\}) = 1 - \min[1 - \mu_H(\{0\}), \frac{1 - \mu_G(\{0\})}{\gamma}].$$
   So under our assumption that each of H, G have zero point mass at the origin, and that $\gamma < 1$, we have $\mu_F(\{0\}) = 0$.
2. Let $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$, then $\text{supp}(\rho) \cap \mathbb{R}^* = \text{supp}(\tilde{\rho}) \cap \mathbb{R}^* = \text{supp}(\mu_F) \cap \mathbb{R}^*$.
3. Suppose $\inf(\text{supp}(\mu_H) \cap \mathbb{R}^*) > 0$, i.e. the support of μ_H is away from the origin, then $\inf(\text{supp}(\mu_F) \cap \mathbb{R}^*) > 0$, the support of μ_F is also away from the origin.
4. $\text{supp}(\mu_F)$ is compact if and only if $\text{supp}(\mu_G)$ and $\text{supp}(\mu_H)$ are both compact.
5. Under our assumption, $\tilde{\rho}(\{0\}) = \lim_{y\downarrow 0}(-iy\tilde{\delta}(iy)) > 0$. Since $\tilde{\delta}(z) \to \infty$ as $z \to 0$ and μ_H is compactly supported away from the origin, by the dominated convergence theorem (DCT),
   $$\rho(\{0\}) = \lim_{y\downarrow 0}(-iy\delta(iy)) = \lim_{y\downarrow 0} \int \frac{\gamma t}{1 + \tilde{\delta}(iy)t} d\mu_H(t) = 0.$$

Given these, the picture is now clear. That is, under our assumption, $\text{supp}(\mu_F) = \text{supp}(\rho) = K$, $\text{supp}(\tilde{\rho}) = \{0\} \cup K$, where K is some compact set on $\mathbb{R}^+$ away from the origin. Thus, $m(z)$ and $\delta(z)$ can be analytically extended to $\mathbb{C} \setminus K$. And $\tilde{\delta}(z)$ can be extended to a meromorphic function on $\mathbb{C} \setminus K$, with a simple pole at $z = 0$.

Let us rewrite the system of equations as

$$\delta(z) = \int \frac{\gamma t}{-z(1 + \tilde{\delta}(z)t)} d\mu_H(t), \; -z\tilde{\delta}(z) = \int \frac{t}{1 + \delta(z)t} d\mu_G(t),$$

where $z \in \mathbb{C}^+$. Now, using the integral representations of $\delta, \tilde{\delta}$ given above,

$$\delta(0) = \int_0^\infty \frac{1}{t} d\rho(t) > 0, \quad \lim_{z\to 0} z\tilde{\delta}(z) = -\tilde{\rho}(\{0\}) < 0,$$

it is easy to check that the right-hand sides of both equations above are analytic at least in a small neighborhood U of the origin. By the uniqueness property of analytic functions, the above system of equations will hold for all $z \in U$. This means that we can evaluate the equation at $z = 0$. For the equation

$$m(z) = \int \frac{1}{-z(1 + \tilde{\delta}(z)t)} d\mu_H(t), \quad z \in \mathbb{C}^+,$$

we find by a similar argument that we can also evaluate $m(z)$ at $z = 0$. This finishes the proof for the expressions of $m(0), e(0)$ given at the beginning of the proof of the main theorem.

Moreover, the Stieltjes transform we are looking for has the form $m(0) = e(0) \cdot \mathbb{E}_H T^{-1}$. Let us write $e(\gamma, G)$ for $e(0)$ showing the dependence on γ, G explicitly. Then,

$$\operatorname{tr}\left[(X^\top X)^{-1}\right] \to \gamma e(\gamma, G) \cdot \mathbb{E}_H T^{-1}.$$

Similarly, since X_i has the same elliptical form $X_i = \Gamma_i^{1/2} Z_i \Sigma^{1/2}$, and by assumption the e.s.d. of Γ_i converges to G_i, we obtain that

$$\operatorname{tr}\left[(X_i^\top X_i)^{-1}\right] \to \gamma_i e(\gamma_i, G_i) \cdot \mathbb{E}_H T^{-1}.$$

Thus, the ARE equals

$$ARE = \gamma e(\gamma, G) \cdot \mathbb{E}_H T^{-1} \cdot \sum_{i=1}^{k} \frac{1}{\gamma_i e(\gamma_i, G_i) \cdot \mathbb{E}_H T^{-1}} = \gamma e(\gamma, G) \cdot \sum_{i=1}^{k} \frac{1}{\gamma_i e(\gamma_i, G_i)}.$$

Notice now that $f(\gamma, G) = \gamma e(\gamma, G)$ so the ARE also equals $f(\gamma, G) \cdot \sum_{i=1}^k \frac{1}{f(\gamma_i, G_i)}$. This finishes the proof.

9.8 Proof of Theorem 2.5

Under the assumptions of the theorem, we have:

$$ARE(k) = \frac{kf(\gamma, G)}{f(k\gamma, G)} = \frac{k \cdot \eta_G^{-1}(1-\gamma)}{\eta_G^{-1}(1-k\gamma)} = \frac{e(\gamma, G)}{e(k\gamma, G)}.$$

The second form given in the theorem follows directly from the definition of e.

Next, we assume G does not have a point mass at the origin. From the definition of η-transform, we have the following observation. For any G, $\eta_G(x)$ is a smooth decreasing function on $[0, +\infty)$ with $\eta_G(0) = 1$ and $\lim_{x\to+\infty} \eta_G(x) = 0$. So $\eta_G^{-1}(x)$ defined on $(0, 1]$ is also smooth and decreasing with $\eta_G^{-1}(1) = 0$ and $\lim_{x\to 0+} \eta_G^{-1}(x) = +\infty$. This means that $ARE(k)$ is indeed a well-defined function for $k \in [1, 1/\gamma]$.

Next we show that ARE(k) is a decreasing convex function. Convexity is equivalent to saying that $1/e(k\gamma, G)$ is decreasing and convex in k. Let $\psi(k) = 1/e(k\gamma, G)$. Then $\psi(k)$ is the unique positive solution to the equation

$$\int \frac{t}{\psi(k) + k\gamma t} dG(t) = 1.$$

We can differentiate with respect to k on both sides to get

$$\psi'(k) = -\frac{\int \frac{\gamma t^2}{(\psi+k\gamma t)^2} dG(t)}{\int \frac{t}{(\psi+k\gamma t)^2} dG(t)} \le 0.$$

Similarly, we differentiate it twice to get

$$\psi''(k) = \frac{\int \frac{2t(\psi'+\gamma t)^2}{(\psi+k\gamma t)^3} dG(t)}{\int \frac{t}{(\psi+k\gamma t)^2} dG(t)} \ge 0.$$

This proves that ψ is decreasing and convex, and finishes the proof.

9.9 Proof of Proposition 2.6

Recall that $G_\tau = (1-\gamma)\delta_\tau + \gamma\delta_{1/\gamma}$. Since ARE($k$) is always decreasing, we only need to show that $\lim_{\tau\to 0}$ARE(2) $= 0$. Now,

$$ARE(2) = \frac{2 \cdot \eta_G^{-1}(1-\gamma)}{\eta_G^{-1}(1-2\gamma)}.$$

Recall also that $\eta_G(x) = \mathbb{E}_G 1/(1+xT)$, and so for $G_\tau = (1-\gamma)\delta_\tau + \gamma\delta_{1/\gamma}$,

$$\eta_{G_\tau}(x) = \frac{1-\gamma}{1+\tau x} + \frac{\gamma}{1+\frac{x}{\gamma}}.$$

To find $x_1 = \eta_{G_\tau}^{-1}(1-\gamma)$, $x_2 = \eta_{G_\tau}^{-1}(1-2\gamma)$, it is sufficient to solve the following quadratic equations:

$$\frac{1-\gamma}{1+\tau x_1} + \frac{\gamma}{1+\frac{x_1}{\gamma}} = 1-\gamma, \quad \frac{1-\gamma}{1+\tau x_2} + \frac{\gamma}{1+\frac{x_2}{\gamma}} = 1-2\gamma,$$

We can rewrite this as

$$(1-\gamma)\tau x_1^2 + \tau(1-2\gamma)\gamma x_1 - \gamma^2 = 0,$$
$$(1-2\gamma)\tau x_2^2 + \gamma(\tau - 1 - 3\gamma\tau)x_2 - 2\gamma^2 = 0.$$

Since we are looking for a positive solution, we find that

$$\begin{aligned}
x_1 &= \frac{-\tau\gamma(1-2\gamma) + \sqrt{\tau^2\gamma^2(1-2\gamma)^2 + 4\tau\gamma^2(1-\gamma)}}{2(1-\gamma)\tau} \\
&= \frac{4\tau\gamma^2(1-\gamma)}{2(1-\gamma)\tau} \cdot \frac{1}{\tau\gamma(1-2\gamma) + \sqrt{\tau^2\gamma^2(1-2\gamma)^2 + 4\tau\gamma^2(1-\gamma)}} \\
&= \frac{2\gamma^2}{\tau\gamma(1-2\gamma) + \sqrt{\tau^2\gamma^2(1-2\gamma)^2 + 4\tau\gamma^2(1-\gamma)}} \\
&= O(\tau^{-1/2}),
\end{aligned}$$

$$x_2 = \frac{\gamma(1+3\gamma\tau-\tau) + \sqrt{\gamma^2(1+3\gamma\tau-\tau)^2 + 8\tau\gamma^2(1-2\gamma)}}{2(1-2\gamma)\tau} \sim \frac{2\gamma}{2(1-2\gamma)\tau} = O(\tau^{-1}).$$

The order of magnitude calculations follow as $\tau \to 0$. Specifically, for the first case, one can check that the numerator is of order $\tau^{1/2}$ by using the formula for conjugate square roots. For the second case, we only need to notice that the numerator tends to 2γ as $\tau \to 0$. So ARE(2) $= 2x_1/x_2 = O(\tau^{1/2}) \to 0$.

9.10 Proof of Theorem 2.7

As before, to find $x_1 = \eta_G^{-1}(1-\gamma)$, $x_k = \eta_G^{-1}(1-k\gamma)$, we solve the quadratic equations:

$$\frac{1-c}{1+\tau x_1} + \frac{c}{1+\alpha\tau x_1} = 1-\gamma, \quad \frac{1-c}{1+\tau x_k} + \frac{c}{1+\alpha\tau x_k} = 1-k\gamma,$$

and choose the positive solutions:

$$x_1 = \frac{(\gamma-c)\alpha + c + \gamma - 1 + \sqrt{((\gamma-c)\alpha + c + \gamma - 1)^2 + 4\gamma(1-\gamma)\alpha}}{2(1-\gamma)\alpha\tau},$$

$$x_k = \frac{(k\gamma-c)\alpha + c + k\gamma - 1 + \sqrt{((k\gamma-c)\alpha + c + k\gamma - 1)^2 + 4k\gamma(1-k\gamma)\alpha}}{2(1-k\gamma)\alpha\tau}.$$

So

$$ARE(k) = \frac{kx_1}{x_k} = \frac{k(1-k\gamma)}{1-\gamma} \cdot \frac{(\gamma-c)\alpha + c + \gamma - 1 + \sqrt{((\gamma-c)\alpha + c + \gamma - 1)^2 + 4\gamma(1-\gamma)\alpha}}{(k\gamma-c)\alpha + c + k\gamma - 1 + \sqrt{((k\gamma-c)\alpha + c + k\gamma - 1)^2 + 4k\gamma(1-k\gamma)\alpha}}$$

is independent of τ, which intuitively makes sense. Indeed, the problem is scale invariant. We can rescale our data by any constant and the relative efficiency does not change.

Observe that, when we take α to tend to infinity, the limit will depend on the choice of c and γ. There are five sub-cases:

1. $0 < c < \gamma$.

$$\begin{aligned} ARE(2) &= \frac{2x_1}{x_2} = \frac{2(1-2\gamma)}{1-\gamma} \cdot \frac{(\gamma-c)\alpha + c + \gamma - 1 + \sqrt{((\gamma-c)\alpha + c + \gamma - 1)^2 + 4\gamma(1-\gamma)\alpha}}{(2\gamma-c)\alpha + c + 2\gamma - 1 + \sqrt{((2\gamma-c)\alpha + c + 2\gamma - 1)^2 + 8\gamma(1-2\gamma)\alpha}} \\ &= O(1), \quad \alpha \to +\infty. \end{aligned}$$

In this case, since the limit of $2x_1/x_2$ is $O(1)$, we should look at the limit of kx_1/x_k for $k > 2$, which turns out to be also $O(1)$:

$$\begin{aligned} \lim_{\alpha\to+\infty} ARE(k) &= \lim_{\alpha\to+\infty} \frac{kx_1}{x_k} \\ &= \lim_{\alpha\to+\infty} \frac{k(1-k\gamma)}{1-\gamma} \cdot \frac{(\gamma-c)\alpha + c + \gamma - 1 + \sqrt{((\gamma-c)\alpha + c + \gamma - 1)^2 + 4\gamma(1-\gamma)\alpha}}{(k\gamma-c)\alpha + c + k\gamma - 1 + \sqrt{((k\gamma-c)\alpha + c + k\gamma - 1)^2 + 4k\gamma(1-k\gamma)\alpha}} \\ &= \frac{k(\gamma-c)(1-k\gamma)}{(1-\gamma)(k\gamma-c)}. \end{aligned}$$

2. $c = \gamma$.

The previous example is a sub-case of this case, where the ratio between the large and small variances equals $\alpha := 1/(\gamma\tau) \to \infty$.

$$ARE(2) = \frac{2x_1}{x_2} = \frac{2(1-2\gamma)}{1-\gamma} \cdot \frac{2\gamma - 1 + \sqrt{(2\gamma-1)^2 + 4\gamma(1-\gamma)\alpha}}{\gamma\alpha + 3\gamma - 1 + \sqrt{(\gamma\alpha + 3\gamma - 1)^2 + 8\gamma(1-2\gamma)\alpha}} = O(\alpha^{-1/2}), \;\; \alpha \to +\infty.$$

3. $\gamma < c < 2\gamma$.

$$\begin{aligned} ARE(2) &= \frac{2x_1}{x_2} = \frac{2(1-2\gamma)}{1-\gamma} \cdot \frac{(\gamma-c)\alpha + c + \gamma - 1 + \sqrt{((\gamma-c)\alpha + c + \gamma - 1)^2 + 4\gamma(1-\gamma)\alpha}}{(2\gamma-c)\alpha + c + 2\gamma - 1 + \sqrt{((2\gamma-c)\alpha + c + 2\gamma - 1)^2 + 8\gamma(1-2\gamma)\alpha}} \\ &= O(\alpha^{-1}), \quad \alpha \to +\infty. \end{aligned}$$

4. $c = 2\gamma$.

$$\begin{aligned} ARE(2) &= \frac{2x_1}{x_2} = \frac{2(1-2\gamma)}{1-\gamma} \cdot \frac{-\gamma\alpha + 3\gamma - 1 + \sqrt{(-\gamma\alpha + 3\gamma - 1)^2 + 4\gamma(1-\gamma)\alpha}}{4\gamma - 1 + \sqrt{(4\gamma-1)^2 + 8\gamma(1-2\gamma)\alpha}} \\ &= O(\alpha^{-1/2}), \quad \alpha \to +\infty. \end{aligned}$$

5. $c > 2\gamma$.

Here, we can easily find $x_1/x_2 = O(1)$ as $\alpha \to +\infty$, but now the exact value of c matters. That is, suppose $c = M\gamma$ for some $M > 2$. Then we will find that

$$\lim_{\alpha \to +\infty} ARE(k) = \lim_{\alpha \to +\infty} \frac{kx_1}{x_k} = \begin{cases} \frac{c-k\gamma}{c-\gamma}, & k < M, \\ O(\alpha^{-1/2}), & k = M, \\ O(\alpha^{-1}), & k > M. \end{cases}$$

9.11 Proof of Proposition 3.1

Notice that, it is sufficient to show that, for any given A, the function $f(X) = 1/\operatorname{tr}(X^{-1}A^\top A)$ is concave on positive definite matrices. Similar to the proof of proposition 2.2., we can define $g(t) = f(X + tV)$. The constraints on $X, V, X + tV$ are the same as before. Now, we have

$$\begin{aligned} g(t) &= \frac{1}{\operatorname{tr}((X+tV)^{-1}A^\top A)} = \frac{1}{\operatorname{tr}((I + tX^{-1/2}VX^{-1/2})^{-1}X^{-1/2}A^\top AX^{-1/2})} \\ &= \frac{1}{\operatorname{tr}((I+t\Lambda)^{-1}Q^\top X^{-1/2}A^\top AX^{-1/2}Q)} \\ &= \left(\sum_{i=1}^n \frac{(Q^\top X^{-1/2}A^\top AX^{-1/2}Q)_{ii}}{1+t\lambda_i}\right)^{-1}. \end{aligned}$$

Since $Q^\top X^{-1/2}A^\top AX^{-1/2}Q$ is always nonnegative definite, we can get the desired result by following the proof of proposition 2.2.

9.12 Computing optimal weights in the general framework, Section 3.1

Recall that we have

$$\begin{aligned} M(\hat\beta_0) &= \mathbb{E}\|L_A - \hat L_A(\hat\beta_0)\|^2 = \mathbb{E}\|A(\beta - \hat\beta_0) + Z\|^2 = \operatorname{tr}\left(\operatorname{Cov}\left[A\hat\beta_0 - Z\right]\right) \\ &= \operatorname{tr}\left(\operatorname{Cov}\left[A\hat\beta_0\right]\right) - 2\operatorname{tr}\left(\operatorname{Cov}\left[A\hat\beta_0, Z\right]\right) + \operatorname{tr}\left(\operatorname{Cov}\left[Z, Z\right]\right) \\ &= \operatorname{tr}\left(\operatorname{Cov}\left[\hat\beta_0\right]A^\top A\right) - 2\operatorname{tr}\left(A\operatorname{Cov}\left[\hat\beta_0, Z\right]\right) + hd\sigma^2 \end{aligned}$$

For OLS, we can calculate, recalling $N = \operatorname{Cov}[\varepsilon, Z]$, $\operatorname{Cov}\left[\hat\beta, Z\right] = (X^\top X)^{-1}X^\top N$. Hence

$$M(\hat\beta) = \sigma^2 \cdot \left[\operatorname{tr}\left[(X^\top X)^{-1}A^\top A\right] - 2\operatorname{tr}\left[A(X^\top X)^{-1}X^\top N\right] + hd\right].$$

For the distributed estimator $\hat\beta_{dist}(w) = \sum_i w_i\hat\beta_i$, we have

$$\begin{aligned} \operatorname{Cov}\left[\hat\beta_{dist}, Z\right] &= \operatorname{Cov}\left[\sum_i w_i(X_i^\top X_i)^{-1}X_i^\top \varepsilon_i, Z\right] \\ &= \sum_i w_i(X_i^\top X_i)^{-1}X_i^\top \operatorname{Cov}\left[\varepsilon_i, Z\right] = \sum_i w_i(X_i^\top X_i)^{-1}X_i^\top N_i \end{aligned}$$

Above, we denoted $N_i := \operatorname{Cov}[\varepsilon_i, Z]$. Therefore,

$$M(\hat\beta_{dist}) = \sigma^2 \cdot \left(\sum_{i=1}^k w_i^2 \cdot \operatorname{tr}\left[(X_i^\top X_i)^{-1}A^\top A\right] - 2w_i \cdot \operatorname{tr}\left[A(X_i^\top X_i)^{-1}X_i^\top N_i\right]\right) + \sigma^2 hd.$$

To find the optimal weights, we consider more generally the quadratic optimization problem

$$\min_{w\in\mathbb{R}^k} \sum_{i=1}^{k} \frac{a_i}{2} w_i^2 - b_i w_i$$

subject to $\sum_{i=1}^k w_i = 1$. We assume that $a_i > 0$. In that case, the problem is convex, and we can use a simple Lagrangian reformulation to solve it. Note that we do not impose the constraint $w_i \geq 0$, because in principle one could allow negative weights, and because it is usually satisfied without imposing the constraint.

Denoting by $\Psi(w)$ the objective, we consider the problem of minimizing the Lagrangian $\Psi_\lambda(w) = \Psi(w) - \lambda(\sum_i w_i - 1)$. It is easy to check that the condition $\frac{\partial \Psi_\lambda}{\partial w_i} = 0$ reduces to

$$w_i = \frac{\lambda + b_i}{a_i}.$$

In order for the constraint $\sum_{i=1}^k w_i = 1$ to be satisfied, we need that

$$\lambda = \lambda^* := \frac{1 - \sum_{i=1}^k \frac{b_i}{a_i}}{\sum_{i=1}^k \frac{1}{a_i}}.$$

Plugging back this value of λ, we obtain the optimal value or the weights w_i^*. To apply this result to our problem, we choose $a_i = \operatorname{tr}\left[(X_i^\top X_i)^{-1} A^\top A\right]$, and $b_i = \operatorname{tr}\left(A (X_i^\top X_i)^{-1} X_i^\top N_i\right)$. This finishes the proof.

9.13 Proof of Theorem 5.1

We want to show

$$\widehat{\Sigma}^{-1} \asymp \Sigma^{-1} \cdot e_p.$$

As mentioned, the proof of this result relies on the generalized Marchenko-Pastur theorem of Rubio and Mestre (2011). From that result, we have under the stated assumptions

$$(\widehat{\Sigma} - zI)^{-1} \asymp (x_p \Sigma - zI)^{-1},$$

where $x_p = x_p(z), e_p = e_p(z)$ are the unique solutions of the system

$$e_p = \frac{1}{p} \operatorname{tr}\left[\Sigma (x_p \Sigma - zI)^{-1}\right], \; x_p = \frac{1}{n} \operatorname{tr}\left[\Gamma (I + \gamma_p e_p \Gamma)^{-1}\right].$$

From section 2.2 of Paul and Silverstein (2009), when the e.s.d of Σ converges to H and the e.s.d of Γ converges to G, x_p and e_p will converge to x and e respectively, where $x = x(z)$ and $e = e(z)$ are the unique solutions of the system

$$e = \int \frac{t}{tx - z} dH(t), \; x = \int \frac{t}{1 + \gamma t e} dG(t).$$

Recall that, in Section 9.7, we have the system of equations

$$\delta = \int \frac{\gamma t}{-z(1 + \tilde{\delta} t)} d\mu_H(t), \; \tilde{\delta} = \int \frac{t}{-z(1 + \delta t)} d\mu_G(t).$$

Then, it's easy to check that $\delta = \gamma e$ and $x = -z\tilde{\delta}$. We will use these relations later.

Now, we want to show that we can take $z = 0$, i.e.

$$\widehat{\Sigma}^{-1} \asymp (x_p(0)\Sigma)^{-1} = \Sigma^{-1} \cdot e_p(0).$$

So for a given sequence of matrices C_p with bounded trace norm we need to bound

$$\begin{aligned}\Delta_p :&= \operatorname{tr}[C_p(\widehat{\Sigma}^{-1} - (x_p(0)\Sigma)^{-1})]\\ &= \operatorname{tr}[C_p(\widehat{\Sigma}^{-1} - (\widehat{\Sigma} - zI)^{-1})] + \operatorname{tr}[C_p((\widehat{\Sigma} - zI)^{-1} - (x_p\Sigma - zI)^{-1}))]\\ &+ \operatorname{tr}[C_p((x_p\Sigma - zI)^{-1} - (x_p\Sigma)^{-1})] + \operatorname{tr}[C_p((x_p\Sigma)^{-1} - (x_p(0)\Sigma)^{-1})]\\ &= \Delta_p^1 + \Delta_p^2 + \Delta_p^3 + \Delta_p^4.\end{aligned}$$

We can bound the four error terms in turn:

1. Bounding Δ_p^1:
   We have
   $$D_1(z) = (\widehat{\Sigma} - zI)^{-1} - \widehat{\Sigma}^{-1} = z(\widehat{\Sigma} - zI)^{-1}\widehat{\Sigma}^{-1}.$$
   Hence, the operator norm of $D(z)$ can be bounded as
   $$\|D_1(z)\|_{op} \leq \frac{2|z|}{\lambda_{\min}(\widehat{\Sigma})^2}$$
   for sufficiently small z.
   Recall that $X = \Gamma^{1/2}Z\Sigma^{1/2}$, where Γ is a diagonal matrix with positive entries and Σ is a symmetric positive definite matrix. Let us consider the least singular value of X. By assumption, the entries of Γ and the eigenvalues of Σ are uniformly bounded below by some constant K. So we can bound $\sigma_{\min}(X)$ as follows:
   $$\sigma_{\min}(X) = \sigma_{\min}(\Gamma^{1/2}Z\Sigma^{1/2}) \geq \sigma_{\min}(\Gamma^{1/2})\sigma_{\min}(Z)\sigma_{\min}(\Sigma^{1/2}) \geq K \cdot \sigma_{\min}(Z).$$
   By using the bound above, we have
   $$\begin{aligned}\lambda_{\min}(\hat{\Sigma}) = \lambda_{\min}(\frac{X^\top X}{n}) &= \frac{(\sigma_{\min}(X))^2}{n} \geq \frac{K^2 \cdot (\sigma_{\min}(Z))^2}{n}\\ &= K^2 \cdot \lambda_{\min}(\frac{Z^\top Z}{n}) \to_{a.s.} K^2(1-\sqrt{\gamma})^2,\end{aligned}$$
   where the final step comes from the well-known Bai-Yin law (Bai and Silverstein, 2009). Thus, we conclude that
   $$\begin{aligned}\lim_{p\to+\infty} |\Delta_p^1| &= \lim_{p\to+\infty} |\operatorname{tr}[C_p(\widehat{\Sigma}^{-1} - (\widehat{\Sigma} - zI)^{-1})]|\\ &\leq \lim_{p\to+\infty} \|C_p\|_{tr} \cdot \|D_1(z)\|_{op} \leq \lim_{p\to+\infty} \|C_p\|_{tr} \cdot \frac{2|z|}{(K^2 \cdot \lambda_{\min}(Z^\top Z/n))^2} \leq C'|z|.\end{aligned}$$
   This holds almost surely, for some fixed constant $C' > 0$.

2. Bounding Δ_p^2:
   This just follows Theorem 1 of Rubio and Mestre (2011):
   $$|\Delta_p^2| = \operatorname{tr}[C_p((\widehat{\Sigma} - zI)^{-1} - (x_p\Sigma - zI)^{-1}))] \to_{a.s.} 0.$$

3. Bounding Δ_p^3:

   By a similar logic, we can obtain a bound on the operator norm of

$$D_2(z) = (x_p\Sigma - zI)^{-1} - (x_p\Sigma)^{-1}$$

   for sufficiently small z, of the form

$$\|D_2(z)\|_{op} \leq \frac{2|z|}{|x_p(z)|^2 \cdot \lambda_{\min}(\Sigma)^2}$$

   for sufficiently small z. Again, we have assumed that the smallest eigenvalues of Σ are always bounded away from zero, so that $\lambda_{\min}(\Sigma) > c > 0$ for some fixed constant $c > 0$. Since $x_p(z) \to x(z) = -z\tilde{\delta}(z)$ as $p \to +\infty$, and we know that $-z\tilde{\delta}(z)$ is analytic in a neighborhood of the origin with $x(0) = \lim_{z\to 0}[-z\tilde{\delta}(z)] = \tilde{\rho}(\{0\}) > 0$. We can argue that $|x(z)|$ is bounded below in a neighborhood of the origin.

   So we conclude that

$$\begin{aligned}\lim_{p\to+\infty} |\Delta_p^3| &= \lim_{p\to+\infty} |\operatorname{tr}[C_p((x_p\Sigma - zI)^{-1} - (x_p\Sigma)^{-1})]| \\ &\leq \lim_{p\to+\infty} \|C_p\|_{tr} \cdot \|D_2(z)\|_{op} \\ &\leq \limsup \|C_p\|_{tr} \cdot \frac{2|z|}{|x(z)|^2 \cdot \lambda_{\min}(\Sigma)^2} \leq C''|z|.\end{aligned}$$

   This holds almost surely, for some fixed constant $C'' > 0$.

4. Bounding Δ_p^4:

$$\begin{aligned}\lim_{p\to+\infty} |\Delta_p^4| &= \lim_{p\to+\infty} |\operatorname{tr}[C_p((x_p\Sigma)^{-1} - (x_p(0)\Sigma)^{-1})] \\ &\leq \lim_{p\to+\infty} \|C_p\|_{tr} \cdot \frac{|x_p(z)^{-1} - x_p(0)^{-1}|}{\lambda_{\min}(\Sigma)} \\ &\leq \limsup \|C_p\|_{tr} \cdot \frac{|x(z)^{-1} - x(0)^{-1}|}{\lambda_{\min}(\Sigma)} \leq C'''|z|.\end{aligned}$$

   This holds almost surely for some fixed constant $C''' > 0$, since $x(z)$ is analytic near the origin with $x(0) > 0$.

Combining these, we have

$$\lim_{p\to+\infty} |\Delta_p| = \lim_{p\to+\infty} |\Delta_p^1 + \Delta_p^2 + \Delta_p^3 + \Delta_p^4| \leq (C' + C'' + C''')|z|.$$

Since $|z|$ can be arbitrarily small, we conclude that, almost surely

$$\lim_{p\to+\infty} |\Delta_p| = \lim_{p\to+\infty} \operatorname{tr}[C_p(\widehat{\Sigma}^{-1} - (x_p(0)\Sigma)^{-1})] = 0.$$

This finishes the proof.

9.14 Proof of Theorem 5.3

Recall that we defined $A_n \asymp B_n$ if $\lim_{n\to\infty} |\mathrm{tr}\,[E_n(A_n - B_n)]| = 0$ a.s., for any standard sequence E_n (of symmetric deterministic matrices with bounded trace norm). Below, E_n will always denote such a sequence.

1. The three required properties are that the $\asymp$ relation is reflexive, symmetric and transitive. The reflexivity and symmetry are obvious. To verify transitivity, we suppose $A_n \asymp B_n$ and $B_n \asymp C_n$. Then, for any standard sequence E_n, by the triangle inequality,

$$|\mathrm{tr}\,[E_n(A_n - C_n)]| \le |\mathrm{tr}\,[E_n(A_n - B_n)]| + |\mathrm{tr}\,[E_n(B_n - C_n)]|\,.$$

   Since the two sequences on the right hand side converge to zero almost surely, the conclusion follows.

2. Let $D_n^1 = A_n - B_n$ and $D_n^2 = C_n - D_n$. Then we can bound by the triangle inequality

$$\left|\mathrm{tr}\left[E_n(D_n^1 + D_n^2)\right]\right| \le \left|\mathrm{tr}\left[E_n D_n^1\right]\right| + \left|\mathrm{tr}\left[E_n D_n^2\right]\right|.$$

   As before, the two sequences on the right hand side converge to zero almost surely, so the conclusion follows.

3. We need to show that $A_n B_n \asymp A_n C_n$. Let E_n be any standard sequence. For this it is enough to show that $A_n E_n$ is still a standard sequence. However, this is clear, because

$$\limsup \|A_n E_n\|_{tr} \le \limsup \|A_n\|_{op}\|E_n\|_{tr} \le \limsup \|A_n\|_{op} \limsup \|E_n\|_{tr} < \infty.$$

4. We know that $\lim_{n\to\infty} |\mathrm{tr}\,[E_n(A_n - B_n)]| = 0$ a.s., for any standard sequence E_n. Consider $E_n = n^{-1} I_n$. Then $\|E_n\|_{tr} = 1$, so that E_n is a standard sequence. Therefore, $\lim_{n\to\infty} |\mathrm{tr}\,[A_n - B_n]| = 0$ a.s., as desired.

5. This is a direct consequence of the trace property.

9.15 Proof of Theorem 6.1

In the proof of Theorem 6.2, we derive the limit for

$$\frac{\mathrm{tr}((X_i^\top X_i)^{-1} X^\top X)}{p} = \frac{p + \sum_{j\ne i} \mathrm{tr}((X_i^\top X_i)^{-1} X_j^\top X_j)}{p},$$

which is

$$1 + (\frac{1}{\gamma}\mathbb{E}_G T - \frac{1}{\gamma_i}\mathbb{E}_{G_i} T) f(\gamma_i, G_i).$$

Then the desired result follows.

9.16 Proof of Theorem 6.2

We consider the Elliptical type sample covariance matrices first. Recall that we have $X = \Gamma^{1/2} Z \Sigma^{1/2}$, where Z is an $n \times p$ matrix with standardized entries, Γ is an $n \times n$ diagonal matrix with positive entries and Σ is a $p \times p$ nonnegative-definite matrix. Our goal is to understand the limit of

$$\mathrm{tr}[(X_i^\top X_i)^{-1} X^\top X] = \mathrm{tr}[(X_i^\top X_i)^{-1} X_i^\top X_i] + \sum_{j\ne i} \mathrm{tr}[(X_i^\top X_i)^{-1} X_j^\top X_j] = p + \sum_{j\ne i} \mathrm{tr}[(X_i^\top X_i)^{-1} X_j^\top X_j].$$

If we delete all rows of X_i from X and denote the remaining matrix by $\tilde{X}_i$, then this can be written as

$$p + \mathrm{tr}[\tilde{X}_i (X_i^\top X_i)^{-1} \tilde{X}_i^\top].$$

Since $X_i = \Gamma_i^{1/2} Z_i \Sigma^{1/2}, \tilde{X}_i = \tilde{\Gamma}_i^{1/2} \tilde{Z}_i \Sigma^{1/2}$, where the $n_i \times p$ matrix Z_i and the $(n - n_i) \times p$ matrix $\tilde{Z}_i$ both have i.i.d. standardized entries, the $(n - n_i) \times (n - n_i)$ diagonal matrix $\tilde{\Gamma}_i$ is the remaining matrix after deleting all the entries of Γ_i from Γ. Then we find that the population covariance Σ will cancel out:

$$\begin{aligned}
&\mathrm{tr}[\tilde{X}_i (X_i^\top X_i)^{-1} \tilde{X}_i^\top] = \mathrm{tr}[\tilde{\Gamma}_i^{1/2} \tilde{Z}_i \Sigma^{1/2} (\Sigma^{1/2} Z_i^\top \Gamma_i Z_i \Sigma^{1/2})^{-1} \Sigma^{1/2} \tilde{Z}_i^\top \tilde{\Gamma}_i^{1/2}] \\
&= \mathrm{tr}[\tilde{\Gamma}_i^{1/2} \tilde{Z}_i (Z_i^\top \Gamma_i Z_i)^{-1} \tilde{Z}_i^\top \tilde{\Gamma}_i^{1/2}].
\end{aligned}$$

To evaluate the limit, we will use the following lemma from Rubio and Mestre (2011).

Lemma 9.1 (Concentration of average of quadratic forms, Lemma 4 in Rubio and Mestre (2011))**.** *Let $\mathcal{U} = \{\xi_k \in \mathbb{C}^M, 1 \le k \le N\}$ denote a collection of i.i.d. random vectors with i.i.d entries that have mean 0, variance 1 and finite $4 + \delta$ moment, $\delta > 0$. Furthermore, consider a collection of random matrices $\{\mathbf{C}_{(k)} \in \mathbb{C}^{M \times M}, 1 \le k \le N\}$ such that, for each k, $\mathbf{C}_{(k)}$ may depend on all the elements of $\mathcal{U}$ except for ξ_k, and the trace norm of $\mathbf{C}_{(k)}$, $||\mathbf{C}_{(k)}||_{\mathrm{tr}}$ is almost surely uniformly bounded for all M. Then, almost surely as $N \to \infty$,*

$$\left| \frac{1}{N} \sum_{k=1}^{N} \left(\xi_k^H \mathbf{C}_{(k)} \xi_k - \mathrm{tr}\, \mathbf{C}_{(k)} \right) \right| \to 0.$$

For our purpose, we can take the number of summands to be $N = n - n_i$, the dimension $M = p$, and the inner matrices to be $\mathbf{C}_{(k)} = \frac{n_i}{p} (Z_i^\top \Gamma_i Z_i)^{-1} \cdot (\tilde{\Gamma}_i)_{kk}$. Also, we let $\xi_k^\top$ to be the k-th row of $\tilde{Z}_i$. By using the well-known result on spectrum separation(see e.g., Bai and Silverstein, 2009), almost surely, the smallest eigenvalue of ${n_i}^{-1} Z_i^\top Z_i$ is uniformly bounded below by some constant. Since $\lambda_{\min}({n_i}^{-1} Z_i^\top \Gamma_i Z_i) \ge \lambda_{\min}(\Gamma_i) \cdot \lambda_{\min}({n_i}^{-1} Z_i^\top Z_i)$, $\lambda_{\min}({n_i}^{-1} Z_i^\top \Gamma_i Z_i)$ is also uniformly bounded below almost surely. So under the assumption of Theorem 6.2, we can check that the trace norm of $\frac{n_i}{p} (Z_i^\top \Gamma_i Z_i)^{-1} \cdot (\tilde{\Gamma}_i)_{kk}$ is indeed uniformly bounded. Then by Lemma 9.1, we will have as $n \to \infty$

$$\left| \frac{1}{n - n_i} \sum_{k=1}^{n - n_i} \left[\frac{n_i}{p} (\tilde{\Gamma}_i)_{kk} \cdot \xi_k^\top (Z_i^\top \Gamma_i Z_i)^{-1} \xi_k - \mathrm{tr} \left(\frac{n_i}{p} (Z_i^\top \Gamma_i Z_i)^{-1} \cdot (\tilde{\Gamma}_i)_{kk} \right) \right] \right| \to_{a.s.} 0.$$

This implies

$$\frac{1}{n - n_i} \mathrm{tr}[\tilde{\Gamma}_i^{1/2} \tilde{Z}_i (Z_i^\top \Gamma_i Z_i)^{-1} \tilde{Z}_i^\top \tilde{\Gamma}_i^{1/2}] \to_{a.s.} f(\gamma_i, G_i) \left(\frac{\gamma_i}{\gamma_i - \gamma} \mathbb{E}_G T - \frac{\gamma}{\gamma_i - \gamma} \mathbb{E}_{G_i} T \right),$$

since $\mathrm{tr}(Z_i^\top \Gamma_i Z_i)^{-1} \to_{a.s.} f(\gamma_i, G_i)$ and

$$\frac{\sum_k (\tilde{\Gamma}_i)_{kk}}{n - n_i} = \frac{n}{n - n_i} \cdot \frac{\mathrm{tr}(\Gamma)}{n} - \frac{n_i}{n - n_i} \cdot \frac{\mathrm{tr}(\Gamma_i)}{n_i} \to_{a.s.} \left(\frac{\gamma_i}{\gamma_i - \gamma} \mathbb{E}_G T - \frac{\gamma}{\gamma_i - \gamma} \mathbb{E}_{G_i} T \right).$$

This holds for all i. Thus, for the elliptical model, we have

$$IE(X_1, \ldots, X_k) = \frac{1 - \frac{p}{n}}{1 - \frac{2p}{n} + \frac{1}{\sum_{i=1}^{k} \frac{n}{\mathrm{tr}[(X_i^\top X_i)^{-1} X^\top X]}}} \to_{a.s.} \frac{1 - \gamma}{1 - 2\gamma + \frac{1}{\sum_{i=1}^{k} \psi(\gamma_i, G_i)}},$$

where

$$\psi(\gamma_i, G_i) = \frac{1}{\gamma + (\mathbb{E}_G T - \frac{\gamma}{\gamma_i}\mathbb{E}_{G_i} T) f(\gamma_i, G_i)}.$$

Now, for the MP model, we can simply take Γ to be identity matrix, then the above result reduces to

$$IE(X_1, \ldots, X_k) = \frac{1 - \frac{p}{n}}{1 - \frac{2p}{n} + \frac{1}{\sum_{i=1}^k \frac{n}{\operatorname{tr}[(X_i^\top X_i)^{-1} X^\top X]}}} \rightarrow_{a.s.} \frac{1-\gamma}{1 - 2\gamma + \frac{\gamma(1-\gamma)}{1-k\gamma}},$$

which finishes the proof.

9.17 Proof of Theorem 6.3

We first provide the proof for the MP model. Since Σ is positive definite, we have

$$x_t^\top (X_i^\top X_i)^{-1} x_t = z_t^\top \Sigma^{1/2} (\Sigma^{1/2} Z_i^\top Z_i \Sigma^{1/2})^{-1} \Sigma^{1/2} z_t = z_t^\top (Z_i^\top Z_i)^{-1} z_t.$$

This cancellation shows that the test error does not depend on the covariance matrix.

For the null case, we will show below that we have, almost surely

$$z_t^\top (Z_i^\top Z_i)^{-1} z_t \to \frac{\gamma_i}{1-\gamma_i}.$$

Hence, we obtain that

$$OE(x_t; X_1, \ldots, X_k) \rightarrow_{a.s.} \frac{1 + \frac{\gamma}{1-\gamma}}{1 + \frac{1}{\sum_{i=1}^k \frac{1-\gamma_i}{\gamma_i}}} = \frac{\frac{1}{1-\gamma}}{1 + \frac{1}{\frac{1}{\gamma} - k}}.$$

Under the elliptical model, we have $x_t^\top (X_i^\top X_i)^{-1} x_t = g_t z_t^\top (Z_i^\top \Gamma_i Z_i)^{-1} z_t$. While Σ still cancels out, the scale parameters do not cancel out anymore. Therefore, we must take them into account when taking the limits. However, similarly to the proof of Theorem 6.1, we find that, almost surely

$$z_t^\top (Z_i^\top \Gamma_i Z_i)^{-1} z_t \to f(\gamma_i, G_i).$$

Putting these together finishes the proof.

To see the reason for the convergence of quadratic forms, we present a slightly different argument. In fact, Theorem 6.3 will still hold if we are only given the $4+c$-th moment of z_1 instead of $8+c$-th one. This follows by the concentration of quadratic forms $x^\top A x - p^{-1} \operatorname{tr} A \to 0$ for matrices A whose spectral distribution converges, and for vectors x with iid entries. Specifically, we will use the following well-known statement about concentration of quadratic forms. To use this result, we simply choose $x = z_t/\sqrt{p}$, and $A_p = (Z_i^\top \Gamma_i Z_i/p)^{-1}$, and the desired claim follows.

Lemma 9.2 (Concentration of quadratic forms, consequence of Lemma B.26 in Bai and Silverstein (2009))**.** *Let $x \in \mathbb{R}^p$ be a random vector with i.i.d. entries and $\mathbb{E}[x] = 0$, for which $\mathbb{E}\left[(\sqrt{p}x_i)^2\right] = \sigma^2$ and $\sup_i \mathbb{E}\left[(\sqrt{p}x_i)^{4+\eta}\right] < C$ for some $\eta > 0$ and $C < \infty$. Moreover, let A_p be a sequence of random $p \times p$ symmetric matrices independent of x, with uniformly bounded eigenvalues. Then the quadratic forms $x^\top A_p x$ concentrate around their means at the following rate*

$$P(|x^\top A_p x - p^{-1}\sigma^2 \operatorname{tr} A_p|^{2+\eta/2} > C) \le C p^{-(1+\eta/4)}.$$

To prove this, we will use the following Trace Lemma quoted from Bai and Silverstein (2009), see also Dobriban et al. (2017) for a similar argument.

Lemma 9.3 ([Trace Lemma, Lemma B.26 of Bai and Silverstein (2009))**.** *Let y be a p-dimensional random vector of i.i.d. elements with mean 0. Suppose that $\mathbb{E}\left[y_i^2\right] = 1$, and let A_p be a fixed $p \times p$ matrix. Then*

$$\mathbb{E}\left[|y^\top A_p y - \operatorname{tr} A_p|^q\right] \le C_q \left\{ \left(\mathbb{E}\left[y_1^4\right] \operatorname{tr}[A_p A_p^\top]\right)^{q/2} + \mathbb{E}\left[y_1^{2q}\right] \operatorname{tr}[(A_p A_p^\top)^{q/2}] \right\},$$

for some constant C_q that only depends on q.

Proof. Under the conditions of Lemma 9.2, the operator norms $\|A_p\|_2$ are almost surely uniformly bounded by a constant C, thus $\operatorname{tr}[(A_p A_p^\top)^{q/2}] \le pC^q$ and $\operatorname{tr}[A_p A_p^\top] \le pC^2$. Consider now a random vector x with the properties assumed in the present lemma. For $y = \sqrt{p}x/\sigma$ and $q = 2 + \eta/2$, using that $\mathbb{E}\left[y_i^{2q}\right] \le C$ and the other the conditions in Lemma 9.2, Lemma 9.3 thus yields

$$\frac{p^q}{\sigma^{2q}} \mathbb{E}\left[|x^\top A_p x - \frac{\sigma^2}{p} \operatorname{tr} A_p|^q\right] \le C \left\{ \left(pC^2\right)^{q/2} + pC^q \right\},$$

or equivalently $\mathbb{E}\left[|x^\top A_p x - \frac{\sigma^2}{p} \operatorname{tr} A_p|^{2+\eta/2}\right] \le Cp^{-(1+\eta/4)}$.

By Markov's inequality applied to the $2+\frac{\eta}{2}$-th moment of $\varepsilon_p = x^\top A_p x - \frac{\sigma^2}{p} \operatorname{tr} A_p$, we obtain as required

$$P(|\varepsilon_p|^{2+\eta/2} > C) \le Cp^{-(1+\eta/4)}.$$

□

9.18 Proof of Theorem 6.4

From Theorem 5.1, it follows that the inverse sample covariance matrix $\widehat{\Sigma}$ is equivalent to a scaled version of the population covariance

$$\widehat{\Sigma}^{-1} \asymp \Sigma^{-1} \cdot e_p,$$

for some scalar sequence $e_p > 0$. By taking in Theorem 5.1 the matrix $C_p = E_j E_j^\top$, the $p \times p$ matrix with a 1 in the (j, j)-th entry, and zeros otherwise, we find that almost surely,

$$[\widehat{\Sigma}^{-1}]_{jj} - [\Sigma^{-1}]_{jj} \cdot e_p \to 0,$$

We can apply this to each sub-matrix X_i to find

$$n_i \cdot [(X_i^\top X_i)^{-1}]_{jj} - [\Sigma^{-1}]_{jj} \cdot e_p(i) \to 0.$$

Here $e_p(i)$ is the solution to the fixed point equation

$$1 = \frac{1}{n_i} \operatorname{tr}\left[e_p(i)\Gamma_i(I_{n_i} + \gamma_{p,i} \cdot e_p(i)\Gamma_i)^{-1}\right].$$

Moreover, $\gamma_{p,i} = p/n_i$ and Γ_i is the $n_i \times n_i$ sub-matrix of Γ corresponding to the i-th machine. It follows that the CE has a deterministic equivalent equal to

$$\frac{[\Sigma^{-1}]_{jj} \cdot e_p}{n} \cdot \sum_{i=1}^k \frac{n_i}{[\Sigma^{-1}]_{jj} \cdot e_p(i)} =$$

$$= \frac{p \cdot e_p}{n} \cdot \sum_{i=1}^k \frac{n_i}{p \cdot e_p(i)} \to \gamma \cdot e(\gamma, G) \cdot \sum_{i=1}^k \frac{1}{\gamma_i \cdot e(\gamma_i, G_i)}.$$

Here $e(\gamma_i, G_i)$ are the quantities encountered before, discussed after Theorem 5.1. The convergence follows from the discussion after Theorem 5.1. Also, from the definition of $f(\gamma, G)$ it follows that $f(\gamma, G) = \gamma e(\gamma, G)$, so that we get the desired result. This finishes the proof.

References

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley New York, 2003.

Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices.* Springer Series in Statistics. Springer, 2009.

H. Battey, J. Fan, H. Liu, J. Lu, and Z. Zhu. Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46(3):1352–1382, 2018.

D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and distributed computation: numerical methods*, volume 23. Prentice hall Englewood Cliffs, NJ, 1989.

G. E. Blelloch and B. M. Maggs. Parallel algorithms. In *Algorithms and theory of computation handbook*, pages 25–25. Chapman & Hall/CRC, 2010.

S. Boyd, N. Parikh, E. Chu, B. Peleato, J. 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.

M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. 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.

R. Y. Chen and J. A. Tropp. Subadditivity of matrix ϕ-entropy and concentration of random matrices. *Electron. J. Probab*, 19(27):1–30, 2014.

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

R. Couillet and W. Hachem. Analysis of the limiting spectral measure of large random matrices of the separable covariance type. *Random Matrices: Theory and Applications*, 3(04):1450016, 2014.

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

J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

E. Dobriban, W. Leeb, and A. Singer. Optimal prediction in the linearly transformed spiked model. *arXiv preprint arXiv:1709.03393*, 2017.

D. L. Donoho and A. Montanari. High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *arXiv preprint arXiv:1310.7320*, 2013.

J. C. Duchi, M. I. Jordan, M. J. Wainwright, and Y. Zhang. Optimality guarantees for distributed statistical estimation. *arXiv preprint arXiv:1405.0782*, 2014.

N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu. On robust regression with high-dimensional predictors. *Proc. Natl. Acad. Sci. USA*, 110(36):14557–14562, 2013.

J. Fan, D. Wang, K. Wang, and Z. Zhu. Distributed estimation of principal eigenspaces. *arXiv preprint arXiv:1702.06488*, 2017.

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

X. Huo and S. Cao. Aggregated inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, page e1451.

M. I. Jordan, J. D. Lee, and Y. Yang. Communication-efficient distributed statistical inference. *arXiv preprint arXiv:1605.07689*, 2016.

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

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

E. Lehmann and G. Casella. Theory of point estimation. *Springer Texts in Statistics*, 1998.

S.-B. Lin, X. Guo, and D.-X. Zhou. Distributed learning with regularized least squares. *The Journal of Machine Learning Research*, 18(1):3202–3232, 2017.

Q. Liu and A. T. Ihler. Distributed estimation, information loss and exponential families. In *Advances in Neural Information Processing Systems*, pages 1098–1106, 2014.

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

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

K. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate analysis.* Academic Press, 1979.

R. Mcdonald, M. Mohri, N. Silberman, D. Walker, and G. S. Mann. Efficient large-scale distributed training of conditional maximum entropy models. In *Advances in Neural Information Processing Systems*, pages 1231–1239, 2009.

A. Müller and M. Debbah. Random matrix theory tutorial–introduction to deterministic equivalents. *TRAITEMENT DU SIGNAL*, 33(2-3):223–248, 2016.

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

D. Paul and J. W. Silverstein. No eigenvalues outside the support of the limiting empirical spectral distribution of a separable covariance matrix. *Journal of Multivariate Analysis*, 100 (1):37–57, 2009.

M. J. Peacock, I. B. Collings, and M. L. Honig. Eigenvalue distributions of sums and products of large random matrices via incremental matrix expansions. *IEEE Transactions on Information Theory*, 54(5):2123–2138, 2008.

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

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

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

V. I. Serdobolskii. On minimum error probability in discriminant analysis. In *Dokl. Akad. Nauk SSSR*, volume 27, pages 720–725, 1983.

V. I. Serdobolskii. *Multiparametric Statistics.* Elsevier, 2007.

V. Smith, S. Forte, C. Ma, M. Takác, M. I. Jordan, and M. Jaggi. Cocoa: A general framework for communication-efficient distributed optimization. *arXiv preprint arXiv:1611.02189*, 2016.

A. M. Tulino and S. Verdú. Random matrix theory and wireless communications. *Communications and Information theory*, 1(1):1–182, 2004.

T. White. *Hadoop: The definitive guide.* " O'Reilly Media, Inc.", 2012.

H. Wickham. *nycflights13: Flights that Departed NYC in 2013*, 2018. URL `https://CRAN.R-project.org/package=nycflights13`. R package version 1.0.0.

J. Yao, Z. Bai, and S. Zheng. *Large Sample Covariance Matrices and High-Dimensional Data Analysis.* Cambridge University Press, 2015.

M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: Cluster computing with working sets. *HotCloud*, 10(10-10):95, 2010.

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

Y. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression. In *Conference on Learning Theory*, pages 592–617, 2013a.

Y. Zhang, J. C. Duchi, and M. J. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363, 2013b.

Y. Zhang, J. Duchi, and M. 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.

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

M. Zinkevich, M. Weimer, L. Li, and A. J. Smola. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, pages 2595–2603, 2010.