How to deal with big data? Understanding large-scale distributed regression

Edgar Dobriban and Yue Sheng

University of Pennsylvania

April 1, 2019
Overview

Background

Setup

General framework

Proof ideas and more general models

Distributed ridge regression
  Algorithm

Summary
Overview

Background

Setup

General framework

Proof ideas and more general models

Distributed ridge regression

Summary
The Age of Data

- We live in the *Age of Data*
- An enormous variety of digital data is generated and recorded every day
- Examples: Web (pages, links, ads, reviews), Science (health records, genetics, high-energy physics), ...
Processing massive datasets

**Figure:** A datacenter
Implications

▶ Creates new problems: how to process, analyze, learn from...
▶ Example:
  ▶ In 2014, Facebook reported storing 300 Petabytes (PB) of data. i.e., 300,000 Terabytes
  ▶ Typical hard drive can store 1Tb...
  ▶ So the data must be distributed over many computers
  ▶ Compute locally, and communicate to get final answer
▶ New area: distributed computation and statistical learning
State of the field

- Active area of research at large tech companies (Google, Facebook,...) and in the AI/ML community
- Grand challenges: resource-adaptive, easy to use (tuning-free), reliable and resilient, verifiable guarantees...
- Standard frameworks: MPI, MapReduce, Spark, GraphLab
MapReduce (Dean, Ghemawat, 2004)
Typical problem: minimize loss over $W$:

$$\sum_{i=1}^{n} L(W, x_i, y_i)$$

$(x_i, y_i)$ are the training datapoints (features and labels)

$W$ are the parameters

$L$ is the loss, e.g. $L(W, x_i, y_i) = (Wx_i - y_i)^2$
Current approach to stats/ML problems

- Data parallelism: Distribute training data over machines. The loss is a sum over training examples. Do iterative calculation (e.g., gradient descent), where compute gradient by summing over machines.

- Made efficient and reliable by e.g. Spark (Zaharia et al 2010)
Statistics and ML research

- Increasing volume of research in the last few years
- Typical research results: one-shot learning does not lose efficiency if the number of machines is not too large
- Low-dimensional mean estimation, kernel ridge regression (Y Zhang, J Duchi, M Wainwright, M Jordan, ...)
- High-dimensional sparse regression, PCA (Y Sun, JD Lee, J Taylor, H Battey, J Fan, H Liu, Z Zhu, ...)
- Other problems: proportional limit asymptotics (J Rosenblatt, B Nadler), nonparametrics (J Lafferty, Y Zhu)
Overview

Background

Setup

General framework

Proof ideas and more general models

Distributed ridge regression

Summary
Our work

- Linear regression
- Data parallelism
- One step of communication
- Parameter averaging
- High (moderate) dimension $n \propto p$
Setup

- Standard linear model $Y = X\beta + \varepsilon$, where
  1. $Y$ is $n \times 1$ outcome, $X$ is $n \times p$ feature matrix.
  2. $\beta$ is $p$-dim parameter

- Samples distributed across $k$ machines. The $i$-th machine has matrix $X_i$ ($n_i \times p$) and outcomes $Y_i$.

\[ X = \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix}, \quad Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_k \end{bmatrix} \]

- Global least squares - infeasible

\[ \hat{\beta} = (X^T X)^{-1} X^T Y \]

- Local least squares estimator $\hat{\beta}_i = (X_i^T X_i)^{-1} X_i^T Y_i$ (assume $n_i > p$)

- Send to parameter server, average

- How does this compare to OLS on full data?
Discoveries

1. **Sub-optimality.** One-shot learning is not optimal, leads to a clear performance decay. (In contrast to recent work)

2. **Strong problem-dependence.** Different learning problems are affected differently by the distributed framework. *Estimation error and the length of confidence intervals increases a lot, while prediction error (test error) increases less.*
Parameter estimation

- Weighted distributed estimator, $\sum_{i=1}^{k} w_i = 1$

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

- Mean squared error (MSE) of OLS

$$\mathbb{E}\|\hat{\beta} - \beta\|^2 = \sigma^2 \text{tr}[(X^\top X)^{-1}]$$

MSE on i-th machine is $\mathbb{E}\|\hat{\beta}_i - \beta\|^2 = \sigma^2 \text{tr}[(X_i^\top X_i)^{-1}]$

- Optimal ”inverse variance weighting” : $w_i^* \propto 1/\left[\sigma^2 \text{tr}[(X_i^\top X_i)^{-1}]\right]$ 

- Relative efficiency

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

How does this depend on $n, p, k$?
Discoveries under asymptotics

- Surprising discovery: Under reasonable conditions, the RE has a simple approximation \((n\ \text{samples}, \ p\ \text{dimensions}, \ k\ \text{machines})\)

\[
\frac{\mathbb{E}\|\hat{\beta} - \beta\|^2}{\mathbb{E}\|\hat{\beta}_{\text{dist}} - \beta\|^2} \approx \frac{n - kp}{n - p}
\]

- Notes
  1. Can be computed conveniently in practice. e.g., \(n = 10^9, \ p = 10^6, \ k = 100\), then \(RE \approx 10/11 \approx 0.91\), so we keep 90% efficiency
  2. Decreases \textit{linearly} in \(k\), the number of machines - effective dimension is \(kp\) (sample size calculation)
  3. \textit{Does not depend} on the sample sizes \(n_i\), or the data
  4. Accurate in simulations and in a data analysis example
Asymptotics

- Recall relative efficiency. Only depends on the eigenvalue spectra of $X^\top X$, $X_i^\top X_i$. 

  \[ RE = \text{tr}[(X^\top X)^{-1}] \left[ \sum_{i=1}^{k} \frac{1}{\text{tr}[(X_i^\top X_i)^{-1}]} \right] \]

- $\text{tr}[(X^\top X)^{-1}] = \sum_{j=1}^{p} 1/\lambda_j(X^\top X)$

- So it makes sense to study models that describe and characterize these spectra

- We will leverage models from asymptotic random matrix theory
Empirical spectral distribution (esd) $F_p$ of symmetric matrix $M$: the cdf of its eigenvalues, $F_p = p^{-1} \sum_i \delta_{\lambda_i(M)}$. If $T \sim F_p$, for any function $f$

$$E_{F_p} f(T) = \frac{\sum_{i=1}^{p} f(\lambda_i(M))}{p} = \frac{1}{p} \text{tr } f(M)$$

Let $n, n_i, p \to \infty$ such that $p/n \to \gamma$, $p/n_i \to \gamma_i$

Limiting spectral distribution (lsd): weak limit of esd
Asymptotics: random matrix theory

- Each $p$-dimensional datapoint $x_i$ is sampled iid from a population with covariance matrix $\Sigma$
- Moreover, it has the form $x_i = \Sigma^{1/2}z_i$, where $z_i$ has iid entries
- e.g., if $z_i$ are normal rvs, then $x_i \sim \mathcal{N}(0, \Sigma)$
- It turns out that the lsd of $\hat{\Sigma} = n^{-1}X^\top X$ is well characterized as $n, p \to \infty$, $p/n \to \gamma$ (Marchenko, Pastur 1967, Bai, Silverstein, 1990s, Tao, Vu, Erdos, Yau 2010s ...)
Asymptotics

By leveraging the Marchenko-Pastur law, we find the following:

The ARE has the simple form \((n, p \to \infty, p/n \to \gamma, k\) is number of machines\)

\[
\frac{\mathbb{E}\|\hat{\beta} - \beta\|^2}{\mathbb{E}\|\hat{\beta}_{dist} - \beta\|^2} \to_{a.s.} \frac{1 - k\gamma}{1 - \gamma} \approx \frac{n - kp}{n - p},
\]
Figure: The loss of efficiency is much worse for estimation $\left( \frac{\mathbb{E}\|\hat{\beta} - \beta\|^2}{\mathbb{E}\|\hat{\beta}_{dist} - \beta\|^2} \right)$ than for test error $\left( \frac{\mathbb{E}(x_t^T \hat{\beta} - y_t)^2}{\mathbb{E}(x_t^T \hat{\beta}_{dist} - y_t)^2} \right)$. 
Figure: Test error relative efficiency on NYC flights data (from nycflights R package). On the data we do not make *any* assumptions, just compare our formulas with the prediction error.
Overview

Background

Setup

General framework

Proof ideas and more general models

Distributed ridge regression

Summary
A general framework

• Important to study not only estimation, but also prediction/test error, residual error, confidence intervals etc

• Predict the linear functional

\[ L_A = A\beta + Z \]

• Using the plug-in estimator

\[ \hat{L}_A(\hat{\beta}_0) = A\hat{\beta}_0 \]

• A - fixed \( d \times p \) matrix; mean and covariance has the structure:

\[ Z \sim (0, h\sigma^2 I_d), \ h \geq 0 \]

• The noise can be correlated with \( \varepsilon \): \( \text{Cov} [\varepsilon, Z] = N \) (e.g., to study residuals)
Examples: Predict $L_A = A\beta + Z$ by $\hat{L}_A(\hat{\beta}_0) = A\hat{\beta}_0$, $Z \sim (0, h\sigma^2 I_d)$

- **Parameter estimation.** Estimate the regression parameter $\beta$ using $\hat{\beta}$. Here $A = I_p$ and $h = 0$.

- **Regression function estimation.** Estimate the regression function $E(Y|X) = X\beta$ using $X\hat{\beta}$. Here $A = X$ and $h = 0$.

- **Out-of-sample prediction (Test error).** Observe test datapoint $x_t$, predict $\hat{y}_t = x_t^\top \hat{\beta}$. Assume $y_t = x_t^\top \beta + \varepsilon_t$. So $A = x_t^\top$, $Z = \varepsilon_t$. 
Examples: Predict $L_A = A\beta + Z$ by $\hat{L}_A(\hat{\beta}_0) = A\hat{\beta}_0$

<table>
<thead>
<tr>
<th>Statistical learning problem</th>
<th>$L_A$</th>
<th>$\hat{L}_A$</th>
<th>$A$</th>
<th>$h$</th>
<th>$N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter estimation</td>
<td>$\beta$</td>
<td>$\hat{\beta}$</td>
<td>$I_p$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Regression function estimation</td>
<td>$X\beta$</td>
<td>$X\hat{\beta}$</td>
<td>$X$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Confidence interval</td>
<td>$\beta_j$</td>
<td>$\hat{\beta}_j$</td>
<td>$E_j^T$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Test error</td>
<td>$x_t^T\beta + \epsilon_t$</td>
<td>$x_t^T\hat{\beta}$</td>
<td>$x_t^T$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Training error/Residual</td>
<td>$X\beta + \epsilon$</td>
<td>$X\hat{\beta}$</td>
<td>$X$</td>
<td>1</td>
<td>$\sigma^2 I_n$</td>
</tr>
</tbody>
</table>
Our results in the general framework

- Predict $L_A = A\beta + Z$ by $\hat{L}_A(\hat{\beta}_0) = A\hat{\beta}_0$
- Relative efficiency:
  
  $$E(A; X_1, \ldots, X_k) := \frac{\mathbb{E}\|L_A - \hat{L}_A(\hat{\beta})\|^2}{\mathbb{E}\|L_A - \hat{L}_A(\hat{\beta}_{\text{dist}})\|^2}.$$  

- For each problem, we find its limits of under both Marchenko-Pastur models and elliptical models (samples have different scales).
Overview

Background

Setup

General framework

Proof ideas and more general models

Distributed ridge regression

Summary
Finite sample results

- When $h = 0$ (no noise), the MSE of estimating $L_A = A\beta$ by OLS $\hat{L}_A = A\hat{\beta} = A(X^\top X)^{-1}X^\top Y$ is
  \[ M(\hat{\beta}) = \sigma^2 \cdot \text{tr} \left[ (X^\top X)^{-1}A^\top A \right]. \]

- For the distributed estimator $\hat{\beta}_{\text{dist}}(w) = \sum_i w_i \hat{\beta}_i$, $\sum_i w_i = 1$
  \[ M(\hat{\beta}_{\text{dist}}) = \sigma^2 \cdot \sum_{i=1}^{k} w_i^2 \cdot \text{tr} \left[ (X_i^\top X_i)^{-1}A^\top A \right]. \]

- So optimal efficiency is
  \[ E(A; X_1, \ldots, X_k) = \text{tr} \left[ (X^\top X)^{-1}A^\top A \right] \cdot \sum_{i=1}^{k} \frac{1}{\text{tr} \left[ (X_i^\top X_i)^{-1}A^\top A \right]}. \]

Key: the traces $\text{tr} \left[ (X_i^\top X_i)^{-1}A^\top A \right]$. 
Calculus of deterministic equivalents

- Deterministic equivalents are a powerful tool in random matrix theory (Serdobolskii 1980s, Hachem et al 2007, etc). Here we develop a systematic approach.
- We have sequences of symmetric $k_n \times k_n$ random matrices $A_n$ and deterministic matrices $B_n$ of growing dimensions
- Definition: $A_n, B_n$ are equivalent,

$$A_n \sim B_n$$

if

$$\lim_{n \to \infty} |\text{tr}(C_nA_n) - \text{tr}(C_nB_n)| = 0$$

almost surely, for any $k_n \times k_n$ sequence $C_n$ of deterministic matrices with bounded trace norm, i.e.,

$$\lim \sup \|C_n\|_{tr} = \lim \sup \sum_i |\lambda_i(C_n)| < \infty.$$
Calculus of deterministic equivalents

- $\text{tr}(C_n A_n)$ is a linear combination of entries of $A_n$
- $A_n \preceq B_n$ if each entry, and each linear combination of entries, of $A_n$ can be approximated by $B_n$
  - Same for traces
- Typically $A_n$ random and complicated, $B_n$ deterministic and simple
General MP theorem (Rubio, Mestre 2011)

- Suppose $x_i = \Sigma^{1/2}z_i \in \mathbb{R}^p$ for $i = 1, \ldots, n$, and $n, p \to \infty$, with $\gamma = p/n$.
- Then with $\hat{\Sigma} = n^{-1}X^\top X$,
  $$\hat{\Sigma}^{-1} \approx \frac{1}{1-\gamma} \cdot \Sigma^{-1}.$$  
- More generally:
  $$(\hat{\Sigma} - zI_p)^{-1} \approx x_p(\gamma, \Sigma, z) \cdot (\Sigma - zI_p)^{-1}.$$  
- This is the simplest way I know how to think of a broad class of results in random matrix theory.
- From this, we can derive
  $$\text{tr} \left[ (X^\top X)^{-1} A^\top A \right] = n^{-1} \text{tr} \left[ \hat{\Sigma}^{-1} A^\top A \right] \approx \frac{1}{1-\gamma} \text{tr} \left[ \Sigma^{-1} \cdot n^{-1} A^\top A \right].$$
The calculus of deterministic equivalents has the following properties.

1. **Sum.** If $A_n \asymp B_n$ and $C_n \asymp D_n$, then $A_n + C_n \asymp B_n + D_n$.
2. **Product.** If $\|A_n\|_{op} < \infty$, and $B_n \asymp C_n$, then $A_n B_n \asymp A_n C_n$.
3. **Trace.** If $A_n \asymp B_n$, then $\text{tr}\{n^{-1}A_n\} - \text{tr}\{n^{-1}B_n\} \to 0$ almost surely.
Elliptical models

- The datapoints can have different scalings: $x_i = g_i^{1/2} \Sigma^{1/2} z_i$ or

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


- We can still do everything, and discover new phenomena

- $\eta$-transform of a distribution $G$ is (Tulino & Verdu, 2004)

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

Inverse $f$ of the $\eta$-transform

$$f(\gamma, G) = \eta^{-1}_G(1 - \gamma).$$

- If the esd of $\Gamma$ and each $\Gamma_i$ converges to $G$, then

$$RE \rightarrow f(\gamma, G) \sum_{i=1}^{k} \frac{1}{f(\gamma_i, G)}.$$  

[Does not depend on $H$, but depends on $n_i$ (or $\gamma_i$)]
Elliptical models

- Can find formulas for all efficiencies
- Elliptical always harder than uniform (convexity)
- There are arbitrarily difficult examples (split across two - lose all)
Overview

Background

Setup

General framework

Proof ideas and more general models

Distributed ridge regression
  Algorithm

Summary
Distributed ridge regression

- Global ridge estimator \( \hat{\beta}(\lambda) = (X^\top X + n\lambda I_p)^{-1}X^\top Y \)
- Local ridge estimator \( \hat{\beta}_i(\lambda_i) = (X_i^\top X_i + n_i\lambda_i I_p)^{-1}X_i^\top Y_i \)
- One-shot weighted estimator

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

- Key point: no constraints on the weights \( w \) because ridge estimator is biased!
- In fact, this leads to some surprising consequences, e.g. optimal weights do not sum to unity
- Also, do not require \( n_i > p \) anymore
Finite sample optimal weights and MSE

- Goal: find optimal weights $w$ to minimize $\mathbb{E}\|\hat{\beta}_{dist}(w) - \beta\|^2$
- Notations: $\hat{\Sigma} = X^T X / n$, $\hat{\Sigma}_i = X_i^T X_i / n_i$ and $Q_i = (\hat{\Sigma}_i + \lambda_i I_p)^{-1}\hat{\Sigma}_i$
- Optimal weights $w^* = (A + R)^{-1}v$, where

$$v_i = \beta^T Q_i \beta, \quad A_{ij} = \beta^T Q_i Q_j \beta, \quad R_{ii} = \frac{\sigma^2}{n_i} \text{tr}[(\hat{\Sigma}_i + \lambda_i I_p)^{-2}\hat{\Sigma}_i]$$

- Corresponding MSE

$$M_k = \mathbb{E}\|\hat{\beta}_{dist}(w^*) - \beta\|^2 = \|\beta\|^2 - v^T (A + R)^{-1}v$$

- A sanity check: when the number of machines is $k = 1$, the optimal weight is $w^* = 1$
A random-effects model

- Ridge estimator is biased, unlike OLS
- MSE of ridge estimator involves $\beta$, so we need some assumptions on $\beta$
- Based on $Y = X\beta + \varepsilon$, further assume $\beta$ is random and independent of $\varepsilon$
- Mean and variance: $E\varepsilon_i = 0, E\varepsilon_i^2 = \sigma^2, E\beta_i = 0, E\beta_i^2 = \sigma^2 \alpha^2 / p$
- $\sigma^2$ is the noise level, $\alpha^2$ is the signal-to-noise ratio
- Widely used model and standard parametrization
- Concentration of quadratic forms:
  \[
  \beta^\top M \beta - \frac{\alpha^2 \sigma^2}{p} \cdot \text{tr}(M) \to_{a.s.} 0
  \]
- Need to know the limits of
  \[
  \text{tr } Q_i = \text{tr}[(\hat{\Sigma}_i + \lambda_i I_p)^{-1}\hat{\Sigma}_i], \quad \text{tr } Q_i Q_j, \quad R_{ii}
  \]
Some asymptotic random matrix theory

- **Distributional assumptions on data $X$:**
  - $X = Z \Sigma^{1/2}$ for an $n \times p$ matrix $Z$ with i.i.d. entries, satisfying $\mathbb{E} Z_{ij} = 0$ and $\mathbb{E} Z_{ij}^2 = 1$, and $\Sigma$ is a $p \times p$ population covariance matrix
  - $n, p \to \infty, p/n \to \gamma \in (0, \infty)$
  - The spectral distribution $F_\Sigma$ of $\Sigma$ converges weakly to a limiting distribution $H$ supported on $[0, \infty)$

- Then the spectral distribution $\hat{F}_\Sigma$ of the sample covariance matrix $\hat{\Sigma}$ also converges to a limiting distribution $F_\gamma$ supported on $[0, \infty)$

- Example: When $\Sigma = I$, $H = \delta_1$, $F_\gamma$ is the Marchenko-Pastur distribution
Marchenko-Pastur distribution
Stieltjes transform

- How to find the limit of \( \text{tr}[(\hat{\Sigma} + \lambda I_p)^{-1}] \)? Stieltjes transform!
- For any probability distribution \( G \), the Stieltjes transform of \( G \) is defined as
  \[
  m_G(z) := \mathbb{E}_G \frac{1}{X - z} = \int \frac{1}{x - z} dG(x), \quad z \in \mathbb{C} \setminus \text{supp}G
  \]
- With this definition, we have
  \[
  m_{F_{\hat{\Sigma}}}(z) = \int \frac{1}{x - z} dF_{\hat{\Sigma}}(x) = \frac{\text{tr}[(\hat{\Sigma} - zI_p)^{-1}]}{p}
  \]
- Since \( F_{\hat{\Sigma}} \to F_\gamma \), we have \( m_{F_{\hat{\Sigma}}}(z) \to m_{F_\gamma}(z) \)
- Then
  \[
  \frac{\text{tr} Q_i}{p} = \frac{\text{tr}[(\hat{\Sigma}_i + \lambda_i I_p)^{-1}\hat{\Sigma}_i]}{p} = 1 - \lambda_i \frac{\text{tr}[(\hat{\Sigma}_i + \lambda_i I_p)^{-1}]}{p} \to 1 - \lambda_i m_{F_\gamma i}(-\lambda_i)
  \]
Asymptotic MSE

- The limiting MSE $M_k = \mathbb{E}\|\hat{\beta}_{\text{dist}}(w^*) - \beta\|^2 = \|\beta\|^2 - \nu^\top (A + R)\nu$ is
  \[ M_k = \sigma^2 \alpha^2 - V^\top (A + R)^{-1} V \]

- The limiting optimal weights $w^* = (A + R)^{-1} \nu$ is
  \[ W_k = (A + R)^{-1} V \]

- Now $M_k$ and $W_k$ depend on the following quantities:
  - The noise level $\sigma^2$ and the SNR $\alpha^2$
  - The number of machines $k$
  - The local aspect ratios $\gamma_i$
  - The local tuning parameters $\lambda_i$
  - The distribution $F_{\gamma_i}$ which depends the population covariance matrix $\Sigma$ and $\gamma_i$
How to choose $\lambda_i$?

- An important issue is: how to choose the tuning parameters $\lambda_i$?
- We want to choose $\lambda_1, \ldots, \lambda_k$ that minimize $M_k$
- An easier question: how to choose the parameter $\lambda$ when $k = 1$?
- It turns out that, asymptotically, the optimal $\lambda = \gamma/\alpha^2$ (Dobriban and Wager, 2018 AoS) which does not depend on the structure of $\Sigma$
- Intuitively, from the Bayesian interpretation of ridge regression, the optimal parameter is $\approx p/(n\alpha^2)$
- $\lambda_i = \gamma_i/\alpha^2$ is locally optimal
- Heuristics from distributed kernel ridge regression (Y Zhang, J Duchi, M Wainwright): the variance, but not the bias, will be reduced when we take the average. So we should choose smaller local parameters
How to choose $\lambda_i$?

- For general $\Sigma$, it is hard to choose $\lambda_i$ to minimize the MSE $M_k$
- Main difficulty: limit of $\text{tr} \ Q_i Q_j$ is complicated, cannot invert $A + R$
- If $p^{-1} \text{tr} \ Q_i Q_j \approx p^{-1} \text{tr} \ Q_i \cdot p^{-1} \text{tr} \ Q_j$, then $A + R$ can be written as a rank-one perturbation of a diagonal matrix
- Under what condition?
- Free probability theory: $\Sigma = I$ is ”almost” the only choice
- For example, when $X^\top X$ is Wishart, we need $X^\top X$ to be orthogonally-invariant: $X^\top X \overset{d}{=} U^\top X^\top X U$
Identity population covariance

▶ When $\Sigma = I$, the limiting Stieltjes transform $m_{F\gamma} := m_{\gamma}$ of $\hat{\Sigma}$ has the explicit form

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

▶ The limiting MSE $M_k$ has the form

$$M_k = \frac{\sigma^2 \alpha^2}{1 + \sum_{i=1}^{k} \frac{V_i^2}{\sigma^2 \alpha^2 (R_{ii} + A_{ii}) - V_i^2}},$$

where $V_i = \sigma^2 \alpha^2 [1 - \lambda_i m_{\gamma i}(-\lambda_i)]$

$A_{ii} = \sigma^2 \alpha^2 [1 - 2\lambda_i m_{\gamma i}(-\lambda_i) + \lambda_i^2 m'_{\gamma i}(-\lambda_i)]$ and

$R_{ii} = \sigma^2 \gamma_i [m_{\gamma i}(-\lambda_i) - \lambda_i m'_{\lambda_i}(-\lambda_i)]$

▶ MSE decouples over $k$ machines, which means locally optimal $\lambda_i = \gamma_i/\alpha^2$ is also globally optimal!
Properties of the asymptotic relative efficiency (ARE)

- Asymptotic relative efficiency (ARE):
  \[
  \lim_{n,p \to \infty} \frac{\mathbb{E} \| \hat{\beta} - \beta \|^2}{\mathbb{E} \| \hat{\beta}_{dist} - \beta \|^2} = \frac{M_1}{M_k}
  \]

- ARE has the explicit form
  \[
  \frac{M_1}{M_k} = \frac{\gamma m_{\gamma}(-\gamma/\alpha^2)}{\alpha^2} \left[ 1 + \sum_{i=1}^{k} \left( \frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)} - 1 \right) \right]
  \]

- Worst case is equally distributed data: The ARE attains its minimum when \( \gamma_1 = \gamma_2 = \cdots = \gamma_k = k\gamma \)
  \[
  \min_{\gamma_1, \ldots, \gamma_k} \text{ARE} = \psi(k, \gamma, \alpha^2) := \frac{\gamma m_{\gamma}(-\gamma/\alpha^2)}{\alpha^2} \left( 1 - k + \frac{\alpha^2}{\gamma m_{k\gamma}(-k\gamma/\alpha^2)} \right)
  \]
Properties of the asymptotic relative efficiency (ARE)

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

  $$\lim_{k \to \infty} \psi(k, \gamma, \alpha^2) = h(\alpha^2, \gamma) > 0.$$ 

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

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

- **Unlike distributed OLS where we lose all the efficiency when $k$ is large**
A plot of $h(\alpha^2, \gamma)$

This suggests that one-shot learning is practical and has good performance

In the "low dimension and high SNR" region, perhaps should use other methods, e.g. iterative methods
Properties of the optimal weights

- Recall the limit of optimal weights $\mathcal{W}_k = (\mathcal{A} + \mathcal{R})^{-1} V$
- The $i$-th coordinate of $\mathcal{W}_k$ is:

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

- We always have $\sum_{i=1}^{k} \mathcal{W}_{k,i} \geq 1$ ($\sum_{i=1}^{k} \mathcal{W}_{k,i} > 1$ when $k \geq 2$)
- When the samples are equally distributed $\gamma_i = k\gamma$, then all $\mathcal{W}_{k,i}$ equal to

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

- $k=2, \alpha=0.50$
- $k=2, \alpha=8.00$
Why are the weights large?

- The short intuitive answer is that ridge estimator is *downward* biased, and so we should counter the effect of bias by upweighting.
- This also can be understood as a way of *debiasing*.
- Similar technique (debiased lasso) has been used in the distributed sparse regression setting (Y Sun, J Lee, J Taylor, H Battey, J Fan, H Liu, Z Zhu, ...).
Algorithm

- Data matrices \((n_i \times p)\) and outcomes \((n_i \times 1)\), \((X_i, Y_i)\) distributed across \(k\) sites
- Compute the MLE \(\hat{\theta}_i = (\hat{\sigma}_i^2, \hat{\alpha}_i^2)\) locally on \(i\)-th machine
- Set local aspect ratio \(\gamma_i = p/n_i\) and set regularization parameter \(\lambda_i = \gamma_i/\hat{\alpha}_i^2\)
- Compute the local ridge estimator \(\hat{\beta}_i(\lambda_i) = (X_i^\top X_i + n_i \lambda_i I_p)^{-1} X_i^\top Y_i\)
- Send \(\hat{\theta}_i, \gamma_i\) and \(\hat{\beta}_i\) to the global data center
- At the data center, combine \(\hat{\theta}_i\) to get a global estimator \(\hat{\theta} = (\hat{\sigma}^2, \hat{\alpha}^2)\), by \(\hat{\theta} = k^{-1} \sum_{i=1}^k \hat{\theta}_i\)
Algorithm

Compute the optimal weights $\omega$, where the $i$-{th} coordinate of $\omega$ is

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

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

Background

Setup

General framework

Proof ideas and more general models

Distributed ridge regression

Summary
Summary

- Broad area of distributed statistical learning
- Interesting discoveries for averaging in distributed linear regression
- Many important problems in this area