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

Edgar Dobriban and Yue Sheng

University of Pennsylvania

November 16, 2018

Overview

Overview

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

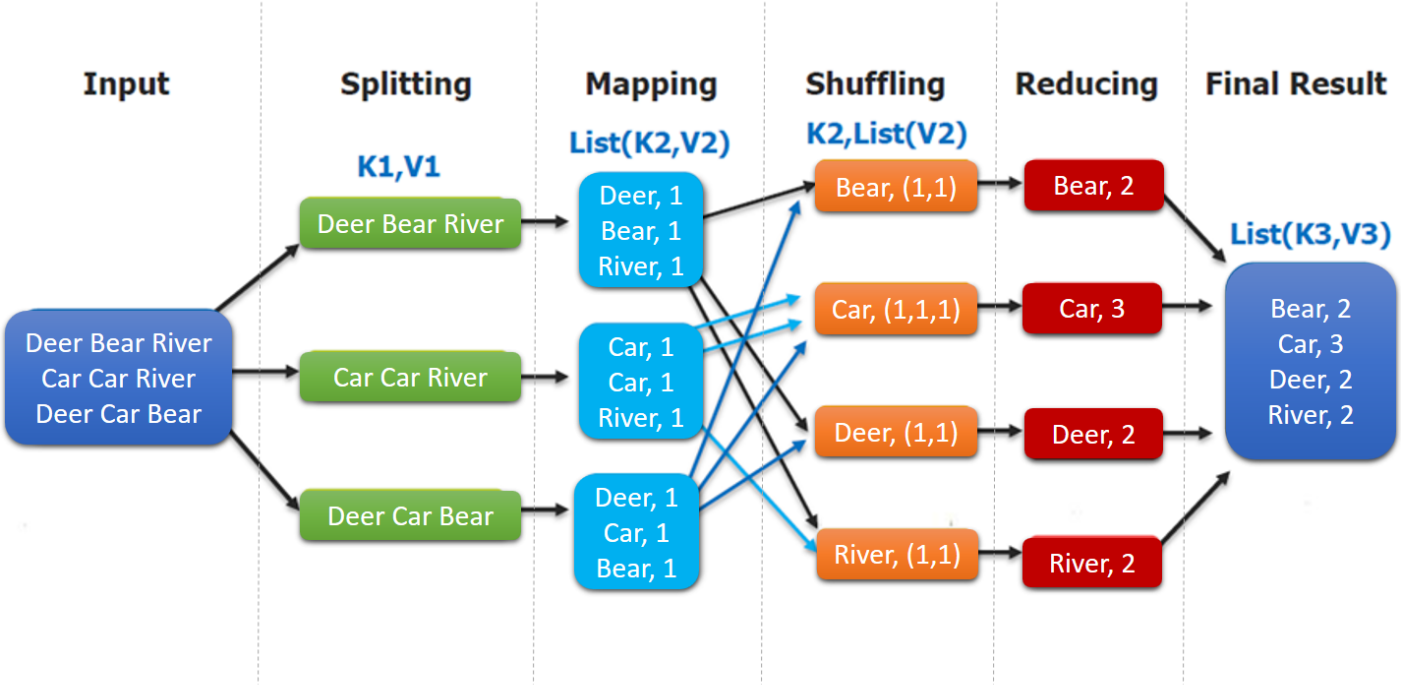
Implications

- 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)



MapReduce Word Count Process

Current approach to stats/ML problems

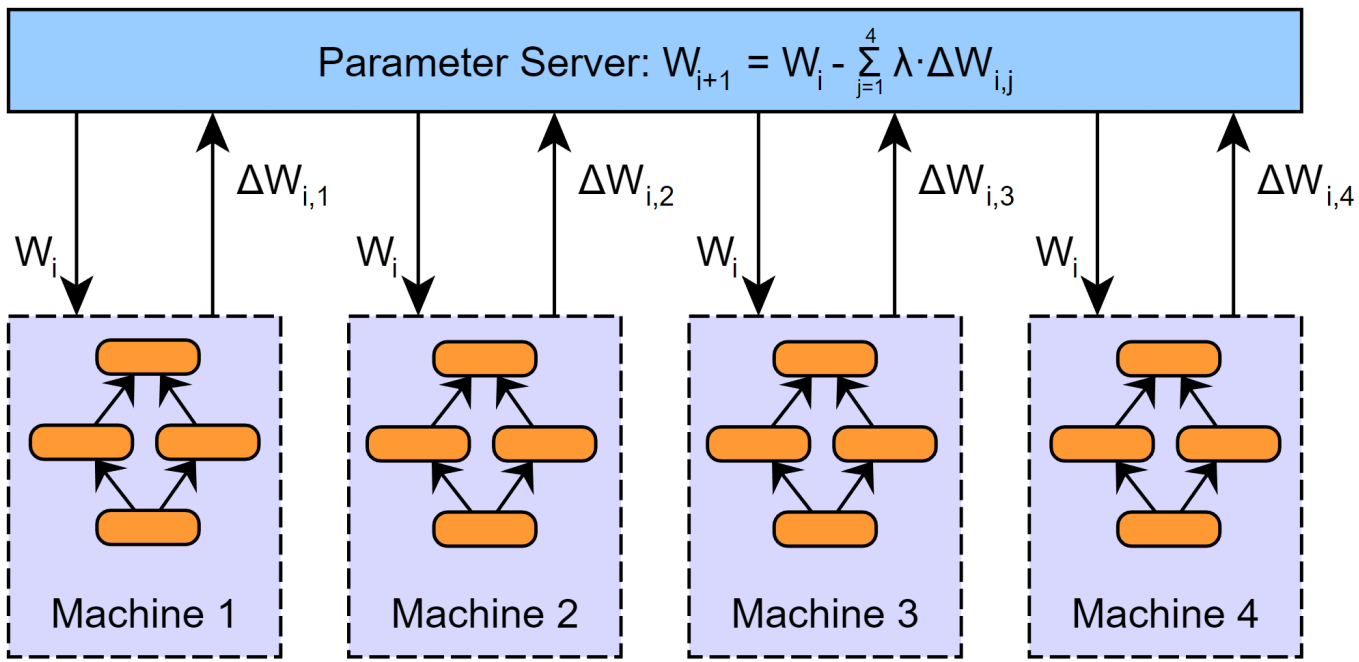
- Typical problem: minimize loss over W:

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

- (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-step averaging 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, ...)

Overview

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. β 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.

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

- Global least squares - infeasible

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

- *Local* least squares estimator $\hat{\beta}_i = (X_i^\top X_i)^{-1} X_i^\top Y_i$ (assume $n_i > p$)
- Send to parameter server, average
- How does this compare to OLS on full data?

Discoveries

1. **Sub-optimality.** One-step averaging 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 errors) 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

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

  MSE on i-th machine is $\mathbb{E}\|\hat{\beta}_i - \beta\|^2 = \sigma^2 \operatorname{tr}[(X_i^\top X_i)^{-1}]$
- Optimal "inverse variance weighting": $w_i^* \propto 1/[\sigma^2 \operatorname{tr}[(X_i^\top X_i)^{-1}]]$
- *Relative efficiency*

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

  How does this depend on n, p, k?

Discoveries under asymptotics

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

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

- 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 *linearly* in k, the number of machines
  3. *does not depend* on the sample sizes n_i, or the data
  4. Accurate in simulations and in data analysis examples

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

Asymptotics

- 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 $T \sim F_p$

$$\mathbb{E}_{F_p} f(T) = \frac{\sum_{i=1}^{p} f(\lambda_i(M))}{p} = \frac{1}{p} \operatorname{tr} f(M)$$

- 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 Σ
- 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 $\widehat{\Sigma} = n^{-1} X^\top X$ 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)

$$\frac{\mathbb{E}\|\hat{\beta} - \beta\|^2}{\mathbb{E}\|\hat{\beta}_{dist} - \beta\|^2} \to_{a.s.} \frac{1 - k\gamma}{1 - \gamma} \approx \frac{n - kp}{n - p},$$

Plot efficiencies

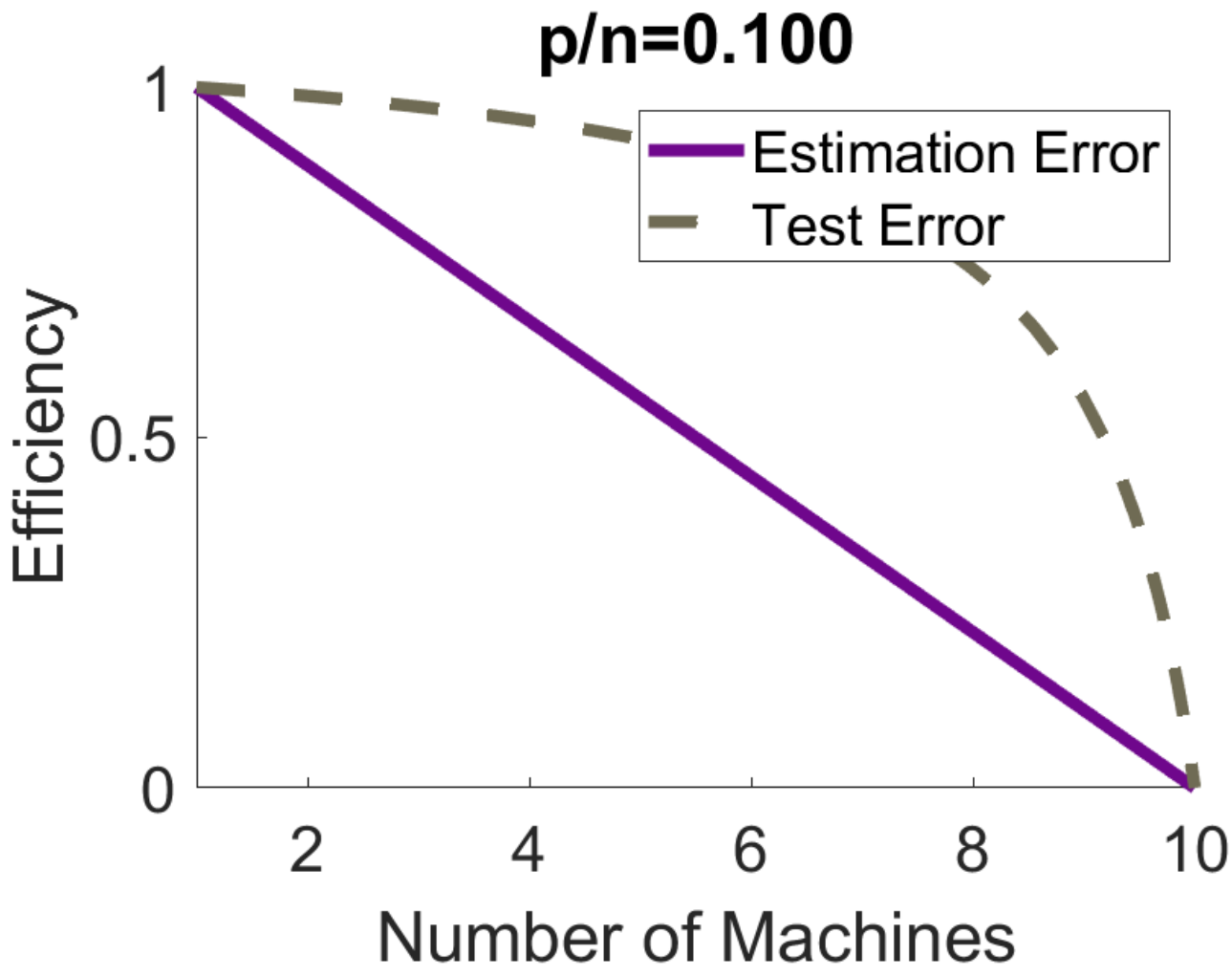


Figure: The loss of efficiency is much worse for estimation ($\frac{\mathbb{E}\|\hat{\beta}-\beta\|^2}{\mathbb{E}\|\hat{\beta}_{dist}-\beta\|^2}$) than for test error ($\frac{\mathbb{E}(x_t^\top \hat{\beta}-y_t)^2}{\mathbb{E}(x_t^\top \hat{\beta}_{dist}-y_t)^2}$).

Empirical data example

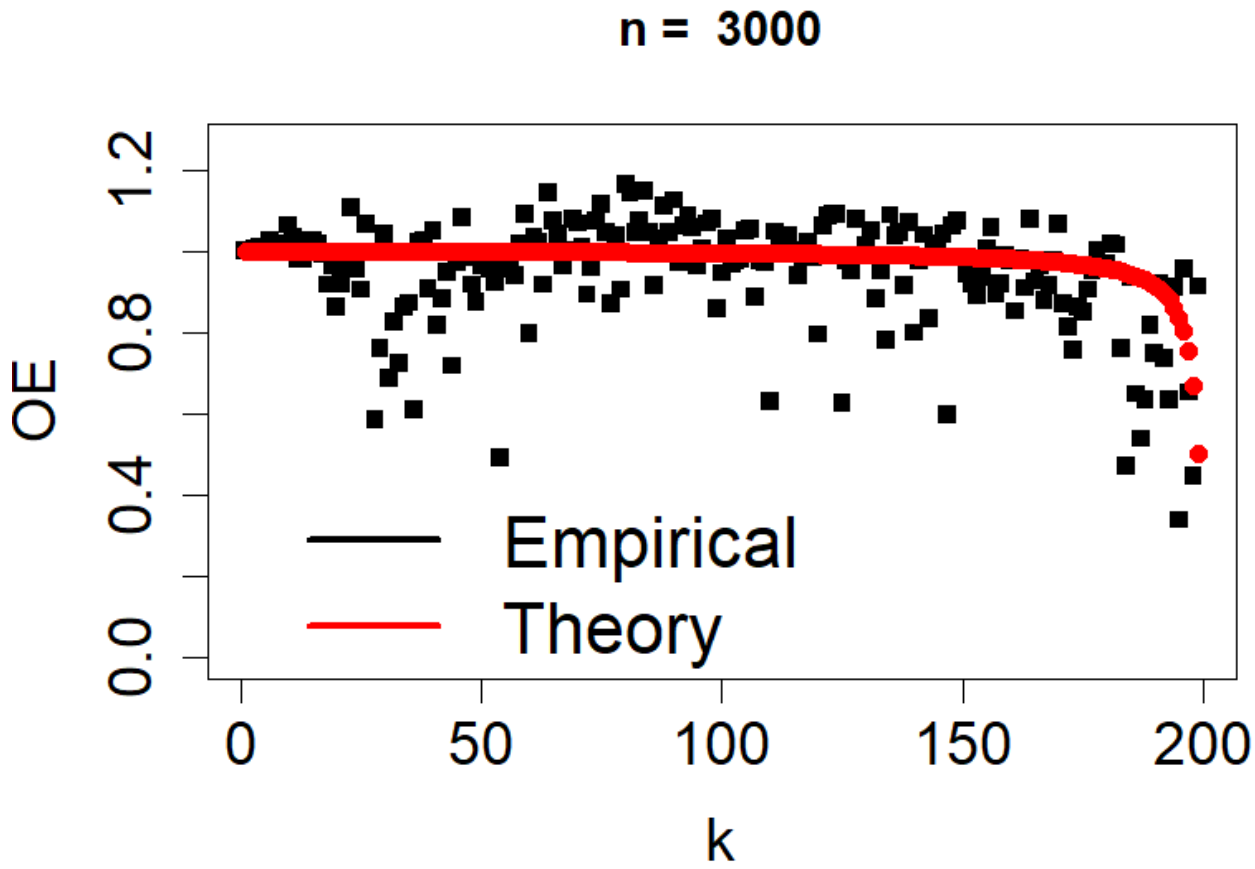


Figure: Test error relative efficiency on NYC flights data (from nycflights R package). On the data we do not make *any* assumptions, just compare our formulas with the prediction error.

Overview

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; $Z \sim (0, h\sigma^2 I_d)$, $h \geq 0$
- The noise can be correlated with ε: $\mathrm{Cov}\,[\varepsilon, Z] = N$ (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 β using $\hat{\beta}$. Here $A = I_p$ and $h = 0$.
- **Regression function estimation**. Estimate the regression function $\mathbb{E}(Y|X) = X\beta$ 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 $\hat{L}_A(\hat{\beta}_0) = A\hat{\beta}_0$

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/Residual	$X\beta + \varepsilon$	$X\hat{\beta}$	X	1	$\sigma^2 I_n$

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:

$$E(A; X_1, \ldots, X_k) := \frac{\mathbb{E}\|L_A - \hat{L}_A(\hat{\beta})\|^2}{\mathbb{E}\|L_A - \hat{L}_A(\hat{\beta}_{dist})\|^2}.$$

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

Overview

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

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

- For the distributed estimator $\hat{\beta}_{dist}(w) = \sum_i w_i \hat{\beta}_i$, $\sum_i w_i = 1$

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

- So optimal efficiency is

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

Key: the traces $\mathrm{tr}\left[(X_i^\top X_i)^{-1}A^\top A\right]$.

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 random matrices A_n and deterministic matrices B_n of growing dimensions
- Definition: A_n, B_n are *equivalent*,

$$A_n \asymp B_n$$

if

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

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

$$\limsup \|C_n\|_{tr} = \limsup \sum_i |\lambda_i(C_n)| < \infty.$$

Calculus of deterministic equivalents

- $\mathrm{tr}(C_n A_n)$ is a linear combination of entries of A_n
- $A_n \asymp B_n$ if each entry, and each linear combination of entries, of A_n can be approximated by B_n

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 $\widehat{\Sigma} = n^{-1} X^\top X$,

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

- From this, we can derive

$$\operatorname{tr}\left[(X^\top X)^{-1} A^\top A\right] = n^{-1} \operatorname{tr}\left[\widehat{\Sigma}^{-1} A^\top A\right] \asymp \frac{1}{1-\gamma} \operatorname{tr}\left[\Sigma^{-1} \cdot n^{-1} A^\top A\right].$$

Rules of calculus

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 $\operatorname{tr}\{n^{-1} A_n\} - \operatorname{tr}\{n^{-1} B_n\} \to 0$ almost surely.

Elliptical models

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

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

  Long history in multivariate statistics (Anderson, 1958).
- Can do everything, and discover new phenomena
- η-transform of a distribution G is (Tulino & Verdu, 2004)

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

  Inverse f of the η-transform

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

- If the esd of Γ and each Γ_i converges to G, then

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

  [Does not depend on H, but depends on n_i (or γ_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

Summary

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