A Review of Distributed Statistical Inference

Gao Yuan

School of Statistics
East China Normal University

December 21, 2020

Abstract

The rapid emergence of massive datasets in various fields poses a serious challenge to traditional statistical methods. Meanwhile, it provides opportunities for researchers to develop novel algorithms. Inspired by the idea of divide-and-conquer, various distributed frameworks for statistical estimation and inference have been proposed. They were developed to deal with large-scale statistical optimization problems. This report aims to provide a comprehensive review for related literature. It includes parametric models, nonparametric models, and other frequently used models. Their key ideas and theoretical properties are summarized. The trade-off between communication cost and estimate precision together with other concerns are discussed.

Outline

1. Introduction
2. Parametric Models
   - One-Shot Approach
   - Iterative Approach
   - Shrinkage Methods
   - Non-Smooth Loss Based Models
3. Nonparametric Models
   - Local Smoothing
   - RKHS Methods
4. Other Related Works
   - Principal Component Analysis
   - Feature Screening
   - Bootstrap
5. Summary and Future Study

1. Introduction
2. Parametric Models
   - One-Shot Approach
   - Iterative Approach
   - Shrinkage Methods
   - Non-Smooth Loss Based Models
3. Nonparametric Models
   - Local Smoothing
   - RKHS Methods
4. Other Related Works
   - Principal Component Analysis
   - Feature Screening
   - Bootstrap
5. Summary and Future Study

Introduction

- Massive datasets in practice:
  - Transaction data in e-commerce
  - Gene expression data in Bioinformatics
  - Text, image, voice, video data on the Internet
  - ...
- Difficult to process whole data on one central machine:
  - Insufficient computing power and memory
  - Network bandwidth
  - Privacy or security considerations

Introduction

- Divide-and-conquer (DC):
  - Divide a large task into many small pieces
  - Tackle them simultaneously on multiple CPUs or machines
- Traditional parallel computing system:
  - All the CPUs share the same memory
- Distributed computing system:
  - Different machines are physically separated and connected by a network
  - Inter-machine communication cost should be considered

Introduction

- Target of distributed statistical inference:
  - Design novel distributed algorithms for statistical problems
  - Balance the communication cost, computing time and estimation precision
  - Study the statistical properties of the resulting estimators

1. Introduction
2. Parametric Models
   - One-Shot Approach
   - Iterative Approach
   - Shrinkage Methods
   - Non-Smooth Loss Based Models
3. Nonparametric Models
   - Local Smoothing
   - RKHS Methods
4. Other Related Works
   - Principal Component Analysis
   - Feature Screening
   - Bootstrap
5. Summary and Future Study

Notations

- N observations: $Z_i = (X_i^\top, Y_i)^\top \in \mathbb{R}^{p+1},\ 1 \le i \le N$. Z_i's are *i.i.d.* with the distribution $\mathbb{P}_{\boldsymbol{\theta}^*}$
- True parameter: $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_p^*)^\top \in \mathbb{R}^p$
- Covariate vector: $X_i \in \mathbb{R}^p$
- Scalar response: $Y_i \in \mathbb{R}$
- K local machines: $\mathcal{M}_k,\ 1 \le k \le K$
- Central machine: $\mathcal{M}_{\text{center}}$, connected with all local machines
- Whole sample: $\mathbb{S} = \{1, \ldots, N\}$
- Local sample on $\mathcal{M}_k$: $\mathcal{S}_k$

Notations

- Local sample size: $|\mathcal{S}_k| = n$, then $N = nK$
- Loss function: $\mathcal{L} : \Theta \times \mathbb{R}^{p+1} \mapsto \mathbb{R}$
  - Assume $\boldsymbol{\theta}^*$ minimizes the population risk $\mathcal{L}^*(\boldsymbol{\theta}) = \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}; Z)]$
- Local loss function on $\mathcal{M}_k$: $\mathcal{L}_k(\boldsymbol{\theta}) = n^{-1} \sum_{i \in \mathcal{S}_k} \mathcal{L}(\boldsymbol{\theta}; Z_i)$
  - Assume $\hat{\boldsymbol{\theta}}_k = \arg\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_k(\boldsymbol{\theta})$
- Global loss function: $\mathcal{L}(\boldsymbol{\theta}) = N^{-1} \sum_{i \in \mathbb{S}} \mathcal{L}(\boldsymbol{\theta}; Z_i) = K^{-1} \sum_{k=1}^{K} \mathcal{L}_k(\boldsymbol{\theta})$
  - Assume $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}(\boldsymbol{\theta})$
  - If N is too large, the whole sample estimator $\hat{\boldsymbol{\theta}}$ is hard to compute

One-Shot Approach

- Basic idea:
  - Calculate relevant statistics on each local machine
  - Assemble these statistics into the final estimator on central machine
- Simple averaging estimator:
  - Compute $\hat{\boldsymbol{\theta}}_k = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_k(\boldsymbol{\theta})$ on each $\mathcal{M}_k$
  - Obtain averaging estimator as $\bar{\boldsymbol{\theta}} = K^{-1}\sum_{k=1}^{K} \hat{\boldsymbol{\theta}}_k$ on $\mathcal{M}_{\text{center}}$
- Advantages:
  - Simple to apply
  - Communication cost is low: $O(Kp)$

One-Shot Approach

- Mean Squared Error (MSE):

$$\mathbb{E}\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_2^2 \leq \frac{C_1}{N}+\frac{C_2}{n^2}+O\left(\frac{1}{Nn}+\frac{1}{n^3}\right), \tag{1}$$

  where $C_1,\ C_2>0$ are some constants
  - If $n \gg N^{1/2}$, then (1) is of the order $O(N^{-1})$
- Rosenblatt and Nadler (2016); Huang and Huo (2015) showed that
  - $\bar{\boldsymbol{\theta}}$ is first order equivalent to $\hat{\boldsymbol{\theta}}$
  - The second-order error terms of $\bar{\boldsymbol{\theta}}$ can be non-negligible for nonlinear models
- Problem: Some local machines might suffer from data of poor quality

One-Shot Approach

- Robust aggregation strategy (Minsker et al., 2019):

$$\hat{\boldsymbol{\theta}}_{\text{robust}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \sum_{k=1}^{K} \rho(|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_k|)$$

where $\rho(\cdot)$ is a robust loss function

  - If $\rho(u) = u$ and $p = 1$, $\hat{\boldsymbol{\theta}}_{\text{robust}}$ is the median of $\hat{\boldsymbol{\theta}}_k$'s
  - Under some regularity conditions, $\hat{\boldsymbol{\theta}}_{\text{robust}}$ achieves the global convergence rate provided $K = O(\sqrt{N})$

Iterative Approach

- Limitations of one-shot approach:
    - Local machines need to have sufficient amount of data (e.g., $n \gg \sqrt{N}$) to achieve the global convergence rate (Wang et al., 2017; Jordan et al., 2019)
    - Simple averaging estimator is often poor in performance for nonlinear models (Rosenblatt and Nadler, 2016; Huang and Huo, 2015; Jordan et al., 2019)
    - When p is diverging with N, the situation could be even worse (Rosenblatt and Nadler, 2016; Lee et al., 2017)

Iterative Approach

- One-step estimator (Huang and Huo, 2015):
  - $\mathcal{M}_{\text{center}}$ broadcasts the averaging estimator $\bar{\boldsymbol{\theta}}$ to each local machine
  - $\mathcal{M}_k$ computes local gradient $\nabla \mathcal{L}_k(\bar{\boldsymbol{\theta}})$ and local Hessian $\nabla^2 \mathcal{L}_k(\bar{\boldsymbol{\theta}})$
  - $\mathcal{M}_{\text{center}}$ computes the one-step updated estimator as

$$\hat{\boldsymbol{\theta}}_{\text{one-step}} = \bar{\boldsymbol{\theta}} - [\nabla^2 \mathcal{L}(\bar{\boldsymbol{\theta}})]^{-1} \nabla \mathcal{L}(\bar{\boldsymbol{\theta}}) \tag{2}$$

- MSE of one-step estimator:

$$\mathbb{E}\left\|\hat{\boldsymbol{\theta}}_{\text{one-step}} - \boldsymbol{\theta}^*\right\|_2^2 \leq \frac{C_1}{N} + O\left(\frac{1}{n^4} + \frac{1}{N^2}\right),$$

where $C_1 > 0$ is some constant

Iterative Approach

- Extension of one-step estimator: Allow the iteration (2) to be executed many times; the communication cost is about $O(K(p^2 + p))$
- Drawbacks: Communication cost is heavy when p is very large
- Remedy:
  - Replace the global Hessian $\nabla^2 \mathcal{L}(\bar{\boldsymbol{\theta}})$ in (2) by a local Hessian computed on some machine (e.g., $\mathcal{M}_{\text{center}}$) (Shamir et al., 2014; Jordan et al., 2019)
  - Fan et al. (2019a) relaxed the heavy dependence on the good choice of the local machine to update Hessian

Shrinkage Methods

- General form:

$$\min_{\boldsymbol{\theta}\in\Theta}\{\mathcal{L}(\boldsymbol{\theta})+\sum_{j=1}^{p}\rho_\lambda(|\theta_j|)\},$$

where $\rho_\lambda(\cdot)$ is a penalty function with a regularization parameter $\lambda > 0$

  - Popular choices: LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001) and others discussed in Zhang et al. (2012)

- LASSO (whole sample estimator):

$$\hat{\boldsymbol{\theta}}_\lambda = \arg\min_{\boldsymbol{\theta}\in\Theta}\Big\{\frac{1}{N}\sum_{i\in\mathbb{S}}(Y_i - X_i^\top\boldsymbol{\theta})^2 + \lambda\sum_{j=1}^{p}|\theta_j|\Big\}.$$

  - Difficulty: LASSO estimator is basically **biased**

Shrinkage Methods

- One-shot averaging estimator (Lee et al., 2017):
  - Compute **debiased** LASSO estimator $\hat{\boldsymbol{\theta}}_{k,\lambda}$ (Javanmard and Montanari, 2014) on each $\mathcal{M}_k$
  - Obtain averaging estimator as $\bar{\boldsymbol{\theta}}_\lambda = K^{-1}\sum_{k=1}^{K}\hat{\boldsymbol{\theta}}_{k,\lambda}$ on $\mathcal{M}_{\text{center}}$
- Problems:
  - Sparsity level can be seriously degraded: by hard threshold
  - Debiasing step is computationally expensive: an improved algorithm

Shrinkage Methods

- Hypothesis testing: Battey et al. (2018)
- Majority voting method: Chen and Xie (2014), for GLMs
- Adaptive-LASSO type method: Zhu et al. (2019), for problems with $p \ll n$ and smooth loss function
- Iterative algorithm: Wang et al. (2017); Jordan et al. (2019), by using local Hessian to iterate
- Global convergence rate of resulting estimators were studied in their works

Non-Smooth Loss Based Models

- Methods described above typically require the loss function to be sufficiently smooth
- Useful methods with non-smooth loss:
    - Quantile regression (QR)
    - Support vector machine (SVM)
- How to deal with these problems?

Non-Smooth Loss Based Models

- QR model:

$$Y_i = X_i^\top \boldsymbol{\theta}^* + \varepsilon_i,\ i \in \mathbb{S}$$

where ε_i is the random noise satisfying $\mathbb{P}(\epsilon_i \le 0|X_i) = \tau \in (0,1)$

- Whole sample estimator:

$$\arg\min_{\boldsymbol{\theta}\in\Theta} N^{-1} \sum_{i\in\mathbb{S}} \rho_\tau(Y_i - X_i^\top \boldsymbol{\theta})$$

where $\rho_\tau(u) = u(\tau - \mathbf{1}\{u \le 0\}) = u(\mathbf{1}\{u > 0\} + \tau - 1)$ is the non-differentiable check-loss function

Non-Smooth Loss Based Models

- Difficulties of distributed estimation:
  - The non-smooth loss function makes the one-shot averaging type estimator perform not well
- Remedies:
  - Approximate $\mathbf{1}\{u > 0\}$ by a smooth function $H(u/h)$ (Chen et al., 2019)
  - A Bahadur representation based method, where the unknown parameters can be replaced by a consistent pilot estimator (Pan et al., 2020)
- The smoothing technique can be applied to SVM (Wang et al., 2019)

1 Introduction

2 Parametric Models
- One-Shot Approach
- Iterative Approach
- Shrinkage Methods
- Non-Smooth Loss Based Models

3 Nonparametric Models
- Local Smoothing
- RKHS Methods

4 Other Related Works
- Principal Component Analysis
- Feature Screening
- Bootstrap

5 Summary and Future Study

Nonparametric Models

- Nonparametric regression:

$$Y_i = f^*(X_i) + \varepsilon_i,\ i \in \mathbb{S}$$

  where
  - $f^*(\cdot)$ is an unknown but sufficiently smooth function
  - ε_i is the random noise with zero mean
- Target: Estimate f^* in a given nonparametric class $\mathcal{F}$
- Difficulty: Hard to obtain unbiased estimators for nonparametric models

Local Smoothing

- Whole sample estimator:

$$\hat{f}_h(x) = \sum_{i \in \mathbb{S}} W_{h,X_i}(x) Y_i,$$

where

  - $W_{h,X_i}(x) \geq 0$ is the local weight at $X = x$
  - $h > 0$ is the bandwidth

- Nadaraya-Watson kernel estimator:

$$W_{h,X_i}(x) = \frac{K\big((X_i - x)/h\big)}{\sum_{i' \in \mathbb{S}} K\big((X_{i'} - x)/h\big)}$$

where $K(\cdot)$ is a kernel function (e.g., $K(u) = \mathbf{1}_{\|u\|_2 \leq 1}$)

Local Smoothing

- One-shot averaging estimator (Chang et al., 2017a):
  - Compute the local estimator $\hat{f}_{k,h}(x)$ on each $\mathcal{M}_k$
  - Obtain averaging estimator as $\bar{f}_h(x) = K^{-1} \sum_{k=1}^{K} \hat{f}_{k,h}(x)$ on $\mathcal{M}_{\text{center}}$
- Problem: For some k, $\|X_i - x\|_2 > h, \ \forall i \in \mathcal{S}_k$; i.e. there may be no sample unit around x on $\mathcal{M}_k$
- Remedies:
  - A data dependent bandwidth
  - An extra qualification step

Local Smoothing

- Nearest neighbors classification: Qiao et al. (2019)
- Density estimation: Li et al. (2013)
- Optimal bandwidth: The bandwidth (or local smoothing parameter) should be adjusted according to the **whole sample size** N (Kaplan, 2019)

RKHS Methods

- Reproducing kernel Hilbert space (RKHS):
  - A special **Hilbert** space
  - Can be induced by a continuous, symmetric and positive semi-definite kernel function $K(\cdot,\cdot)$
  - $\|\cdot\|_{\mathcal{H}}$ is the associated norm
- Kernel ridge regression (KRR):

$$\hat{f}_\lambda = \arg\min_{f\in\mathcal{H}} \left\{ \frac{1}{N} \sum_{i\in\mathbb{S}} (Y_i - f(X_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}, \tag{3}$$

- Representer theorem (Wahba, 1990):

$$\hat{f}_\lambda(x) = \sum_{i\in\mathcal{S}} \alpha_i K(X_i, x)$$

RKHS Methods

- One-shot averaging estimator (Zhang et al., 2015):
  - Compute the local estimator $\hat{f}_{k,\lambda}$ on each $\mathcal{M}_k$
  - Obtain averaging estimator as $\bar{f}_\lambda = K^{-1} \sum_{k=1}^{K} \hat{f}_{k,\lambda}$ on $\mathcal{M}_{\text{center}}$
- Lin et al. (2017) studied the same problem and derived some improved theoretical results
- Xu et al. (2016) extended the loss function in (3) to a further general form

RKHS Methods

- Problem: Performance of One-shot methods depends heavily on the local sample size n
- Remedies:
    - Semi-supervised learning framework: Chang et al. (2017b)
    - More communication: Lin et al. (2020)

RKHS Methods

- Semi-parametric model:
  - Zhao et al. (2016) constructed a special RKHS to estimate the nonparametric part of a partially linear model
  - Lv and Lian (2017) used a debiasing technique for the high-dimensional sparse partial linear models
- How decide the regularized parameter λ?
  - The optimal order of λ should be chosen according to the **whole sample size** N
  - Xu et al. (2018) proposed a *distributed generalized cross-validation* (dGCV) to select an asymptotically optimal λ

1 Introduction

2 Parametric Models
- One-Shot Approach
- Iterative Approach
- Shrinkage Methods
- Non-Smooth Loss Based Models

3 Nonparametric Models
- Local Smoothing
- RKHS Methods

4 Other Related Works
- Principal Component Analysis
- Feature Screening
- Bootstrap

5 Summary and Future Study

Principal Component Analysis

- Principal component analysis (PCA) is a common procedure to reduce the dimension of the data
- Procedures on one machine:
  - Compute the sample covariance matrix $\hat{\Sigma} = N^{-1} \sum_{i \in \mathbb{S}} X_i X_i^\top$
  - Standard SVD gives $\hat{\Sigma} = \hat{V} \hat{D} \hat{V}^\top$, then the columns of $\hat{V}$ are the principal component directions that we need
- Problem: Simple average of the eigenvectors estimated locally cannot give a valid result

Principal Component Analysis

- A one-shot distributed algorithm (Fan et al., 2019b):
  1. Each $\mathcal{M}_k$ computes d leading eigenvectors of the local sample covariance matrix $\hat{\Sigma}_k = n^{-1} \sum_{i \in \mathcal{S}_k} X_i X_i^\top$, denoted by $\hat{v}_{1,k}, \cdots, \hat{v}_{d,k} \in \mathbb{R}^p$. Next, $\hat{V}_k = (\hat{v}_{1,k}, \cdots, \hat{v}_{d,k}) \in \mathbb{R}^{p \times d}$ is sent to $\mathcal{M}_{\text{center}}$
  2. $\mathcal{M}_{\text{center}}$ averages K local projection matrices to obtain $\tilde{\Sigma} = K^{-1} \sum_{k=1}^{K} \hat{V}_k \hat{V}_k^\top$. Then it computes d leading eigenvectors of $\tilde{\Sigma}$, denoted by $\tilde{v}_1, \cdots, \tilde{v}_d \in \mathbb{R}^p$, which are the principal component directions that we need

- The communication cost is of the order $O(Kdp)$, where d is usually very small

- Conditions to achieve the global convergence rate were studied in their work

Feature Screening

- Standard linear model:

$$Y_i = X_i^\top \boldsymbol{\theta}^* + \epsilon_i,\ i \in \mathbb{S}$$

  Suppose $\mathcal{A}^* = \{1 \leq j \leq p : \theta_j^* \neq 0\}$ is the true sparse model

- Target: Screen out the irrelevant features not in $\mathcal{A}^*$

- Sure independence screening (SIS) (Fan and Lv, 2008):

$$\hat{\mathcal{A}}_\gamma = \{1 \leq j \leq p : |\hat{\omega}_j| > \gamma\}$$

  where

  - γ is a prespecified threshold
  - $\hat{\omega}_j$ is the whole sample estimator of ω_j, the Pearson correlation between jth feature and Y

Feature Screening

- Difficulty: $\hat{\omega}_j$ is usually biased for many correlation measures
- Distributed feature screening (Li et al., 2020):
  - Express the correlation measure as $\omega_j = g(\nu_1, \ldots, \nu_s)$
  - Use U-statistic to estimate ν_q's on each $\mathcal{M}_k$
  - Obtain the one-shot averaging estimators $\bar{\nu}_q$'s on $\mathcal{M}_{\text{center}}$
  - Use the distributed estimator $\tilde{\omega}_j = g(\bar{\nu}_1, \ldots, \bar{\nu}_s)$ to select features as

$$\tilde{\mathcal{A}}_\gamma = \{1 \le j \le p : |\tilde{\omega}_j| > \gamma\}$$

- The sure screening property was shown as

$$\mathbb{P}(\mathcal{A}^* \subset \tilde{\mathcal{A}}_\gamma) \to 1 \quad \text{as } N \to \infty$$

Bootstrap

- Target: Assess the accuracy of some estimator $\hat{\boldsymbol{\theta}}$ (e.g., variance)
- General procedures:
  - Draw r samples of size N from $\mathcal{S}$ with replacement
  - Compute r estimates of $\boldsymbol{\theta}$ based on the r resamples
  - Calculate the variance of above r estimators
- Problem:
  - Bootstrap is computationally expensive, especially when N is very large
  - Variants of classic Bootstrap need an additional correction step

Bootstrap

- Bag of little bootstraps (BLB) (Kleiner et al., 2014):
  1. Each $\mathcal{M}_k$ draws r samples of size N (instead of n) from $\mathcal{S}_k$ with replacement.
  2. Computes r estimates of $\boldsymbol{\theta}$ based on the r resamples drawn above
  3. Each $\mathcal{M}_k$ computes some accuracy measure, denoted by $\hat{\xi}_k$, by the r estimates above
  4. Average these $\hat{\xi}_k$'s as $\bar{\xi} = K^{-1} \sum_{k=1}^{K} \hat{\xi}_k$ on $\mathcal{M}_{\text{center}}$
- Generating some certain weight vectors of length n suffices to approximate the resampling process

1 Introduction

2 Parametric Models
- One-Shot Approach
- Iterative Approach
- Shrinkage Methods
- Non-Smooth Loss Based Models

3 Nonparametric Models
- Local Smoothing
- RKHS Methods

4 Other Related Works
- Principal Component Analysis
- Feature Screening
- Bootstrap

5 Summary and Future Study

Summary and Future Study

- How to analyze middle-sized data?
  - Can be easily stored in a hard drive
  - Cannot be read into the memory
  - Not large enough to justify an expensive distributed system
- How to analyze heterogeneous and unbalanced local data?
  - Meta analysis may be applicable (Zhou and Song, 2017)

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

Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

Bickel, P. J., Götze, F., and van Zwet, W. R. (2012). Resampling fewer than n observations: gains, losses, and remedies for losses. In *Selected works of Willem van Zwet*, pages 267–297. Springer.

Chang, X., Lin, S.-B., Wang, Y., et al. (2017a). Divide and conquer local average regression. *Electronic Journal of Statistics*, 11(1):1326–1350.

Chang, X., Lin, S.-B., and Zhou, D.-X. (2017b). Distributed semi-supervised learning with kernel ridge regression. *The Journal of Machine Learning Research*, 18(1):1493–1514.

Chen, X., Liu, W., Zhang, Y., et al. (2019). Quantile regression under memory constraint. *The Annals of Statistics*, 47(6):3244–3273.

Chen, X. and Xie, M.-g. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, pages 1655–1684.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press.

Fan, J., Guo, Y., and Wang, K. (2019a). Communication-efficient accurate statistical estimation. *arXiv preprint arXiv:1906.04870*.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, J., Wang, D., Wang, K., Zhu, Z., et al. (2019b). Distributed estimation of principal eigenspaces. *The Annals of Statistics*, 47(6):3009–3031.

Huang, C. and Huo, X. (2015). A distributed one-step estimator. *arXiv preprint arXiv:1511.01443*.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909.

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

Kaplan, D. M. (2019). Optimal smoothing in divide-and-conquer for big data. Technical report, working paper available at https://faculty. missouri. edu/˜ kaplandm.

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816.

Koenker (2005). *Quantile Regression (Econometric Society Monographs; No. 38)*. Cambridge University Press.

Lee, J. D., Liu, Q., Sun, Y., and Taylor, J. E. (2017). Communication-efficient sparse regression. *The Journal of Machine Learning Research*, 18(1):115–144.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Li, R., Lin, D. K., and Li, B. (2013). Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry*, 29(5):399–409.

Li, X., Li, R., Xia, Z., and Xu, C. (2020). Distributed feature screening via componentwise debiasing. *Journal of Machine Learning Research*, 21(24):1–32.

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

Lin, S.-B., Wang, D., and Zhou, D.-X. (2020). Distributed kernel ridge regression with communications. *arXiv preprint arXiv:2003.12210*.

Liu, D., Liu, R. Y., and Xie, M. (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. *Journal of the American Statistical Association*, 110(509):326–340.

Lv, S. and Lian, H. (2017). A debiased distributed estimation for sparse partially linear models in diverging dimensions. *arXiv preprint arXiv:1708.05487*.

Minsker, S. et al. (2019). Distributed statistical estimation and rates of convergence in normal approximation. *Electronic Journal of Statistics*, 13(2):5213–5252.

Pan, R., Ren, T., Guo, B., Li, G., and Wang, H. (2020). A note on distributed quantile regression by pilot sampling and one-step updating. Technical report, working paper to be available at arXiv.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer Science & Business Media.

Qiao, X., Duan, J., and Cheng, G. (2019). Rates of convergence for large-scale nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 10768–10779.

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

Shamir, O., Srebro, N., and Zhang, T. (2014). Communication-efficient distributed optimization using an approximate newton-type method. In *International Conference on Machine Learning*, pages 1000–1008.

Steinwart, I., Hush, D. R., Scovel, C., et al. (2009). Optimal rates for regularized least squares regression. In *COLT*, pages 79–93.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Wahba, G. (1990). *Spline models for observational data*, volume 59. Siam.

Wang, F., Huang, D., Zhu, Y., and Wang, H. (2020). Efficient estimation for generalized linear models on a distributed system with nonrandomly distributed data. *arXiv preprint arXiv:2004.02414*.

Wang, J., Kolar, M., Srebro, N., and Zhang, T. (2017). Efficient distributed learning with sparsity. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3636–3645. JMLR. org.

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

Xu, C., Zhang, Y., Li, R., and Wu, X. (2016). On the feasibility of distributed kernel regression for big data. *IEEE Transactions on Knowledge and Data Engineering*, 28(11):3041–3052.

Xu, G., Shang, Z., and Cheng, G. (2018). Optimal tuning for divide-and-conquer kernel ridge regression with massive data. *Proceedings of Machine Learning Research*.

Xu, M. and Shao, J. (2020). Meta-analysis of independent datasets using constrained generalised method of moments. *Statistical Theory and Related Fields*, 4(1):109–116.

Zhang, C.-H., Zhang, T., et al. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593.

Zhang, T. (2005). Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098.

Zhang, Y., Duchi, J., and Wainwright, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research*, 16(1):3299–3340.

Zhang, Y., Duchi, J. C., and Wainwright, M. J. (2013). Communication-efficient algorithms for statistical optimization. *The Journal of Machine Learning Research*, 14(1):3321–3363.

Zhao, T., Cheng, G., and Liu, H. (2016). A partially linear framework for massive heterogeneous data. *Annals of Statistics*, 44(4):1400.

Zhou, L. and Song, P. X.-K. (2017). Scalable and efficient statistical inference with estimating functions in the mapreduce paradigm for big data. *arXiv preprint arXiv:1709.04389*.

Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496):1464–1475.

Zhu, X., Li, F., and Wang, H. (2019). Least squares approximation for a distributed system. *arXiv preprint arXiv:1908.04904*.

Thanks!