Recent Developments in Statistical Theory and Methods Based on Distributed Computing (18w5089)

Sujit Ghosh (North Carolina State University),
Xiaoming Huo (Georgia Institute of Technology),
Hua Zhou (University of California, Los Angeles),
Mu Zhu (University of Waterloo)

May 20-25, 2018

1 Overview of the Field

In the era of "big data," in many applications, it is impossible to store data in a single device or central location and thus it has become a major challenge to make inference by developing methodologies that can be performed on data distributed across many devices or locations. For example, a search engine company may have data coming from a large number of locations, and each location collects Tera-bytes of data per day; building a data center to store all existing data can be very costly or unsecured. On a different setting, high volume of data (like images or videos) have to be stored in relatively smaller parts, instead of on a centralized server, and aggregated inference are needed. This workshop enabled participants to explore algorithmic and theoretical problems in distributed inference and computation.

2 Recent Developments and Open Problems

The workshop addressed a contemporary issue in mathematical sciences with significant impact in industrial applications by bring together researchers from both statistics and applied mathematics who are working on distributed inference and distributed computing. The applied focus of the workshop was on the explorations of the methodology for the computing and modeling architecture for distributed data, including data ingestion and staging platform, enterprise data warehouse and analytics platform. The workshop brought academic and industrial researchers together for the exploration and scientific discussions on recent challenges faced by practitioners and related cutting-edge theories and proven best practices in both academia and industries on distributed data analytics.

3 Presentation Highlights

One feature of this workshop is the diversity among its presentations and participants. Some give overview talks on methodology. Some describe various applications. There is one talk that is focused on the optimization technique, which is unique from the rest of the presentations, however is enlightening. Three student speakers (Emily Hector, Bochao Jia, Shanshan Cao) and one postdoc (Luc Villandré) give excellent presentations.

3.1 Methodological presentations

In the workshop, six presentations are focused on the methodological aspect of the distributed inference. These six talks were given by Xiaoming Huo, Stanislav Minsker, Peter Song, Ding-Xuan Zhou, Rajarshi Guhaniyogi, and Min-ge Xie. They represent methodology innovation from the frequentist, Bayesian, fiducial inference, and machine learning perspectives.

In his talk titled "Computationally and Statistically Efficient distributed Inference with Theoretical Guarantees," Huo points out that in many contemporary data-analysis settings, it is expensive and/or infeasible to assume that the entire data set is available at a central location. In recent works of computational mathematics and machine learning, great strides have been made in distributed optimization and distributed learning (i.e., machine learning). On the other hand, classical statistical methodology, theory, and computation are typically based on the assumption that the entire data are available at a central location - this is a significant shortcoming in classical statistical knowledge. The statistical methodology and theory for distributed inference have been actively developed. Huo discusses one distributed statistical method that is computationally efficient, requiring minimal communication, and has comparable statistical properties. Theoretical guarantees of this distributed statistical estimator are presented.

In the talk titled "Distributed Statistical Estimation and Rates of Convergence in Normal Approximation," Minsker presents algorithms for distributed statistical estimation that can take advantage of the divide-and-conquer approach. He shows that one of the key benefits attained by an appropriate divide-and-conquer strategy is robustness, an important characteristic of large distributed systems. Moreover, they introduce a class of algorithms that are based on the properties of the spatial median, establish connections between performance of these distributed algorithms and rates of convergence in normal approximation, and provide tight deviations guarantees for resulting estimators in the form of exponential concentration inequalities. Techniques are illustrated with several examples; in particular, they have obtained new results for the median-of-means estimator, as well as provide performance guarantees for robust distributed maximum likelihood estimation. This talk is based on his joint work with Nate Strawn.

Song, in his talk titled "Meta Estimation of Normal Mean Parameter: Seven Perspectives of Data Integration," describes a synergic treatment on the estimation of mean parameter of a normal distribution from seven different schools of statistics, which sheds light on the future development of data integration analytics. They include best linear unbiased estimation (BLUE), maximum likelihood estimation (MLE), Bayesian estimation, empirical Bayesian estimation (EBE), Fisher's fiducial estimation, generalized methods of moments (GMM) estimation, and empirical likelihood estimation (ELE). Their properties of scalability and distributed inference are discussed and compared analytically and numerically. This talk presents a nice overall architecture.

Zhou gave a talk titled "Theory of Deep Convolutional Neural Networks and Distributed Learning." Deep learning has been widely applied and brought breakthroughs in speech recognition, computer vision, and many other domains. The involved deep neural network architectures and computational issues have been well studied in machine learning. But there lacks a theoretical foundation for understanding the approximation or generalization ability of deep learning methods with network architectures such as deep convolutional neural networks with convolutional structures. This talk describes a mathematical theory of deep convolutional neural networks (CNNs). In particular, they show the universality of a deep CNN, meaning that it can be used to approximate any continuous function to an arbitrary accuracy when the depth of the neural network is large enough. Their quantitative estimate, given tightly in terms of the number of free parameters to be computed, verifies the efficiency of deep CNNs in dealing with large dimensional data. Some related distributed learning algorithms are discussed.

Guhaniyogi's talk, titled "DISK: Divide and Conquer Spatial Kriging for Massive Sea Surface Database", presents a divide-and-conquer Bayesian approach to large-scale kriging for geostatistical data. His talk is unique in several aspects: it is the only Bayesian methodology talk, and it tackles correlated data. Guhanijogi and his collaborators propose a three-step divide-and-conquer strategy within the Bayesian paradigm to achieve massive scalability for any spatial process model. They partition the data into a large number of subsets, apply a readily available Bayesian spatial process model on every subset in parallel, and optimally combine the posterior distributions estimated across all the subsets into a pseudo posterior distribution that conditions on the entire data. The combined pseudo posterior distribution is used for predicting the responses at arbitrary locations and for performing posterior inference on the model parameters and the residual spatial

surface. Under the standard theoretical setup, they show that if the number of subsets is not too large, then the Bayes L2-risk of estimating the true residual spatial surface using the DISK posterior distribution decays to zero at a nearly optimal rate. A variety of simulations and a geostatistical analysis of the Pacific Ocean sea surface temperature data validate our theoretical results.

The last methodological talk is given by Xie with the title "On Combination of Inferences After Split-and-Conquer." Xie gives an overview of scientific methods that are used for combining estimations or inferences from different data (or subsets of data). He discusses pros and cons of different methodologies that are commonly used in scientific research. He stresses the use of 'distribution estimators' to combining inferences and provide a 'unified' angle to view both Bayesian and frequentist approaches. He provides examples to illustrate some pitfalls of some well-known approaches.

The above five talks collectively present a nice overview of the current research front.

3.2 Applications

Five speakers presented interesting applications.

In her talk titled "Variance Component Testing and Selection for a Longitudinal Microbiome Study," Jin Zhou points out that high-throughput sequencing technology has enabled population-based studies of the role of the human microbiome in disease etiology and exposure response. Due to the high cost of sequencing technology such studies usually have limited sample sizes. They study the association of microbiome composition and clinical phenotypes by testing the nullity of variance components. When the null model has more than one variance parameters and sample sizes are limited, such as in longitudinal metagenomics studies, testing zero variance components remains an open challenge. In her talk, she first introduce a series of efficient exact tests (score test, likelihood ratio test, and restricted likelihood ratio test) of testing zero variance components in presence of multiple variance components. Their approach does not rely on the asymptotic theory thus significantly boosts the power in small samples. Furthermore, to further conquer limited sample size and high dimensional features of metagenomics data, they introduce a variance component selection scheme with lasso penalization. They propose an minorization-maximization (MM) algorithm for the difficult optimization problem. Extensive simulations demonstrate the superiority of our methods vs existing methods. Finally, they apply their method to a longitudinal microbiome study of HIV infected patients.

Martin Lysy presents a talk with title "Applications of a Distributed Computational Method for Microparticle Tracking in Biological Fluids." State-of-the-art techniques in passive particle-tracking microscopy provide high-resolution path trajectories of diverse foreign particles in biological fluids. In order to analyze experiments often tracking thousands of particles at once, scientists must account for many sources of unwanted variability, such as heterogeneity of the fluid environment and measurement error. Lysy presents a versatile family of hierarchical stochastic process models, along with a scalable split-merge distributed computing strategy for parameter inference. Also presented are several applications to quantifying subdiffusive mobility of tracer particles in human lung mucus.

Sharmistha Guha's talk is titled "Bayesian Regression with Undirected Network Predictors with an Application to Brain Connectome Data." She proposes a Bayesian approach to regression with a continuous scalar response and an undirected network predictor. Undirected network predictors are often expressed in terms of symmetric adjacency matrices, with rows and columns of the matrix representing the nodes, and zero entries signifying no association between two corresponding nodes. Network predictor matrices are typically vectorized prior to any analysis, thus failing to account for the important structural information in the network. This results in poor inferential and predictive performance in presence of small sample sizes. Guha et al propose a novel class of network shrinkage priors for the coefficient corresponding to the undirected network predictor. The proposed framework is devised to detect both nodes and edges in the network predictive of the response. Their framework is implemented using an efficient Markov Chain Monte Carlo algorithm. Empirical results in simulation studies illustrate strikingly superior inferential and predictive gains of the proposed framework in comparison with the ordinary high dimensional Bayesian shrinkage priors and penalized optimization schemes. They apply their method to a brain connectome dataset that contains information on brain networks along with a measure of creativity for multiple individuals. Interest lies in building a regression model of the creativity measure on the network predictor to identify important regions and connections in the brain strongly associated with creativity. Their approach is the first principled Bayesian method that is able to detect scientifically interpretable regions and connections in the brain actively impacting the continuous

response (creativity) in the presence of a small sample size.

Luc Villandré gives an interesting and somewhat unique application talk with title "Challenges in the prediction of motor vehicle traffic collisions with GPS travel data." He mentions that in the field of road safety, crashes involving physical injuries typically occur on roadways, which constrain the events to lie along a linear network. Substantial research efforts have been devoted to the development of methods for point patterns on linear networks. In one such model, he and his coworkers assume that crash coordinates are produced by a Poisson point process whose domain corresponds to edges in the road network. His talk focuses on the analysis of geo-localised accident data in the context of a smart city initiative launched by the City of Quebec (Canada) aiming to identify crash hotspots on the road network based on covariates derived from GPS data. Data originate from three sources: i) a geo-localised traffic accident database whose entries are based on police reports, ii) GPS trajectories obtained from a study on 4,000 drivers involving 55,000 trips and iii) the structure of the road network obtained from the OpenStreetMap (OSM) database. He highlights challenges, both methodological and computational, with the use of those three data sources in producing sensible inference for the covariate effects.

Zhengwu Zhang gives a talk titled "Optimization Problems in Brain Connectome Analysis". He first gives an introduction to the different modalities of neuroimaging data and then presents some optimization problems arising from analysis of neuroimaging data. His talk is interesting in both scientific and computational points of view. He is able to connect his work with that of Joong-Ho Won and Sharmistha Guha. This workshop is beneficial as it fosters new collaboration between them who are across different fields.

All the above application talks are well presented and well accepted by the audience.

3.3 Optimization

Two presentations touch upon an important aspect of distributed inference – optimization and algorithms.

Joong-Ho Won's talk is titled "A Continuum of Optimal Primal-Dual Algorithms for Convex Composite Minimization Problems with Applications to Structured Sparsity." He states that many statistical learning problems can be posed as minimization of a sum of two convex functions, one typically a composition of non-smooth and linear functions. Examples include regression under structured sparsity assumptions. Popular algorithms for solving such problems, e.g., ADMM, often involve non-trivial optimization subproblems or smoothing approximation. He and his collaborators consider two classes of primal-dual algorithms that do not incur these difficulties, and unify them from a perspective of monotone operator theory. From this unification they propose a continuum of preconditioned forward-backward operator splitting algorithms amenable to parallel and distributed computing. For the entire region of convergence of the whole continuum of algorithms, they establish its rates of convergence. For some known instances of this continuum, their analysis closes the gap in theory. They further exploit the unification to propose a continuum of accelerated algorithms. They show that the whole continuum attains the theoretically optimal rate of convergence. The scalability of the proposed algorithms, as well as their convergence behavior, is demonstrated up to 1.2 million variables with a distributed implementation.

Josh Day, from Julia Computing, presents his work on online algorithms in the talk "Online Algorithms for Statistics". Traditional algorithms for calculating statistics and models are often infeasible when working with big data. A statistician will run into problems of not just scalability, but of handling data arriving in a continuous stream. Online algorithms, which update estimates one observation at a time, can naturally handle big and streaming data. Many traditional (offline) algorithms have online counterparts that produce exact estimates, but this is not always possible. There exists a variety of online (stochastic approximation, SA) algorithms for approximate solutions, but there is no universally "best" algorithm and convergence can be sensitive to the choice of learning rate, a decreasing sequence of step sizes. The current state-of-the-art SA algorithms are based entirely on first-order (gradient) information and therefore ignore potentially useful information in each update. Day and his collaborators derive two new algorithms for incorporating majorization-minimization (MM) concepts into SA that have strong stability properties. They analyze the new algorithms in a unified framework and offer stronger convergence results. The third algorithm (MSPI) incorporates the MM concept into Implicit Stochastic Gradient Descent (Toulis and Airoldi, 2015) and Stochastic Proximal Iteration (Ryu and Boyd, 2014). Compared to the algorithms on which it is based, MSPI can solve a wider class of problems and in many cases has a cheaper online update. For all three MM-based algorithms, it is typically straightforward to translate existing offline MM algorithms into an online counterpart,

particularly in the case of quadratic majorizations.

3.4 Students' presentations

The workshop become more energetic when three students gave their presentations on their works.

Bochao Jia talks about "Double-Parallel Monte Carlo for Bayesian Analysis of Big Data." He proposes a simple, practical and efficient MCMC algorithm for Bayesian analysis of big data. The proposed algorithm divides the big dataset into some smaller subsets and provides a simple method to aggregate the subset posteriors to approximate the full data posterior. To further speed up computation, the proposed algorithm employs the population stochastic approximation Monte Carlo (Pop-SAMC) algorithm, a parallel MCMC algorithm, to simulate from each subset posterior. Since this algorithm consists of two levels of parallel, data parallel and simulation parallel, it is coined as "Double Parallel Monte Carlo." The validity of the proposed algorithm is justified mathematically and numerically.

Emily Hector presents a talk titled "A Distributed and Integrated Method of Moments for High-Dimensional Correlated Data Analysis." She and her collaborators present a divide-and-conquer procedure implemented in a distributed and parallelized scheme for statistical estimation and inference of regression parameters with high-dimensional correlated responses with multi-level nested correlations. Despite significant efforts in the literature, the computational bottleneck associated with high-dimensional likelihoods prevents the scalability of existing methods. The proposed method addresses this challenge by dividing subjects into independent groups and responses into correlated subvectors to be analyzed separately and in parallel on a distributed platform. Theoretical challenges related to combining results from dependent data are overcome in a statistically efficient way using a meta-estimator derived from Hansens Generalized Method of Moments. They provide a rigorous theoretical framework for efficient estimation, inference, and goodness-of-fit tests. They develop an R package for ease of implementation. They illustrate their methods performance with simulations and the analysis of a complex neuroimaging motivating dataset from an association study of the effects of iron deficiency on auditory recognition memory.

Ms. Shanshan Cao gives a talk on nonconvex regularization in sparse regression. She shows how the nonconvex penalization can be unified under the framework of difference-of-convex (DC), which is a subfield of optimization. She then argue that many existing statistical procedures can be treated as special cases. Theory on both the numerical efficiency and the statistical optimality can be derived.

There were a lot of discussions during and after these students' presentations, which makes the alive.

3.5 Industrial presentations

Two presentations from industrial participants greatly enrich the content of this workshop.

Josh Day, from Julia Computing, gives an excellent tutorial on the new technical computing language Julia and how to use JuliaDB to analyze large, distributed data. He also presents OnlineStats, a package for computing statistics and models via online algorithms. It is designed for taking on big data and can naturally handle out-of-core processing, parallel/distributed computing, and streaming data. JuliaDB fully integrates OnlineStats for providing analytics on large persistent datasets. All these tools are open source and show impressive performance over current tools such as R and Python.

Sergio Adrin Lagunas Pinacho, a data scientist from Cognizant, presents an interesting talk "Good Practices to Write Good, Clean, and Collaborative Code". Being a competitive programmer himself, his advice on coding styles is tremendously helpful to distributed computing practitioners, including students, postdocs, and faculty in this workshop.

3.6 Local participants

Two local participants from Mexico add another layer of depth to this workshop.

Edgar Jimenez of CIMAT Unidad Monterrey gives a talk "Parallel Forecasts For Demand Planning of Perishable Processed Foods". Forecasting is a key activity in demand planning which is a pillar of supply chain management. In the present work they expose the results achieved by a system developed for a producer of processed foods located in Mexico. This system required a custom made architecture because of two key requirements: high speed of forecast generation and minimum possible error. The models were based on

weekly data for all clients and products. The forecast procedure was based on parallel processing, which in the latest iteration of the system also used hierarchical properties of the forecasts.

Ernesto Alverez Gonzalez, from Universidad Autnoma Benito Jurez de Oaxaca, gives probably the most abstract talk in this workshop, titled "Advances in computing parameters for JK69 triplets with fixed topology". Gonzalez demonstrates how the abstract algebraic geometry can guide the computation of parameters of some commonly used models in evolutionary biology.

4 Scientific Progress Made

Traditional statistical inference deals with data that are available at a central location. An objective function (which usually measures the goodness-of-fit of the model as a function of the unknown parameter) is minimized to solve the estimation problem. The objective function is usually the empirical loss function of the observed sample, which is the average of loss over the sample data. Due to the high volume of the observations, solving the optimization problem usually is time-consuming. Distributed formulation or parallel computation would become indispensable in these large-scale problems.

Iterative updating algorithms, which require inter-nodes communication, are popular in the distributed optimization problems. Different settings are studied, such as the sparse high dimensional estimation in the linear regression setting, the M-estimator setting, the Bayesian framework, the scenario where the number of machines increases as the sample size increases, and many more. Jordan et, at. [**?**] study distributed statistical inference comprehensively in all the aforementioned settings. Chen and Xie [**?**], Battey et al. [**?**], and Lee et al. [**?**] consider a high-dimensional however sparse parameter vector estimation problem, where they adopt the penalized M-estimator setting.

We describe the general formulation of the distributed estimation and inference in the setting of the M-estimator, which is a generalization of the Maximum Likelihood Estimation (MLE). Estimators in statistical inference are to infer some unknown parameters. It is a mapping from the sample space (observations) to the parameter space Θ. For an M-estimator, the objective is to maximize the average sample criterion function, which is equivalent to minimizing the sum of loss. In the distributed data case, where data are distributed at different stations, one of the popular method to obtain an estimation is to first compute the estimator at each local station $\hat{\theta}_i$; then transfer to the central station the local optimal estimators; the final estimator is the "average" of these local estimators. A diagram associated with such a procedure can be seen in Figure **??**.

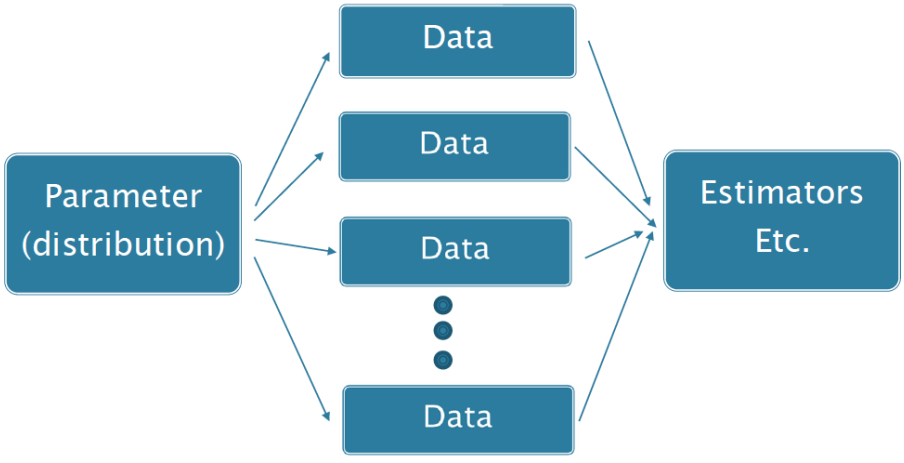

Figure 1: Diagram of aggregated statistical estimation: the first step is obtaining the local parameter estimation using the local data; the second step is computing the centralized estimation using the local estimations.

4.1 Distributed one-step M-estimator

For illustration purpose, we review the construction of the one-step distributed M-estimator that is proposed in Huang and Huo (2015) [**?**]. For each $i \in \{1, \ldots, k\}$, the local empirical criterion function that is based on the local data set S_i on machine i and the corresponding maximizer are denoted by

$$M_i(\theta) = \frac{1}{|S_i|} \sum_{x \in S_i} m(x; \theta) \quad \text{and} \quad \theta_i = \arg\max_{\theta \in \Theta} M_i(\theta). \tag{1}$$

Thus the global empirical criterion function can be denoted by

$$M(\theta) = \frac{1}{k} \sum_{i=1}^{k} M_i(\theta). \tag{2}$$

Let the population criterion function and its maximizer be xxx, we have

$$M_0(\theta) = \int_{\mathcal{X}} m(x; \theta) p(x) dx \quad \text{and} \quad \theta_0 = \arg\max_{\theta \in \Theta} M_0(\theta), \tag{3}$$

where $\mathcal{X}$ is the sample space, θ_0 is the parameter of interest. We further denote the gradient and the Hessian of $m(x; \theta)$ with respect to θ by

$$\dot{m}(x; \theta) = \frac{\partial m(x; \theta)}{\partial \theta}, \text{ and } \ddot{m}(x; \theta) = \frac{\partial^2 m(x; \theta)}{\partial \theta \, \partial \theta^T}. \tag{4}$$

The gradient and Hessian of the local empirical criterion function thus can be denoted by

$$\dot{M}_i(\theta) = \frac{\partial M_i(\theta)}{\partial \theta} = \frac{1}{|S_i|} \sum_{x \in S_i} \frac{\partial m(x; \theta)}{\partial \theta}, \text{ and } \ddot{M}_i(\theta) = \frac{\partial^2 M_i(x; \theta)}{\partial \theta \, \partial \theta^T} = \frac{1}{|S_i|} \sum_{x \in S_i} \frac{\partial^2 m(x; \theta)}{\partial \theta \, \partial \theta^T}, \tag{5}$$

where $i \in \{1, 2, \ldots, k\}$. The gradient and Hessian of the global empirical criterion function can be denoted by

$$\dot{M}(\theta) = \frac{\partial M(\theta)}{\partial \theta}, \text{ and } \ddot{M}(\theta) = \frac{\partial^2 M(\theta)}{\partial \theta \, \partial \theta^T}. \tag{6}$$

Similarly, the gradient and Hessian of the population criterion function are denoted by

$$\dot{M}_0(\theta) = \frac{\partial M_0(\theta)}{\partial \theta}, \text{ and } \ddot{M}_0(\theta) = \frac{\partial^2 M_0(\theta)}{\partial \theta \, \partial \theta^T}. \tag{7}$$

In order to derive the one-step estimator, let $\theta^{(0)}$ denote the average of these local M-estimators, we have

$$\theta^{(0)} = \frac{1}{k} \sum_{i=1}^{k} \theta_i. \tag{8}$$

The one-step estimator $\theta^{(1)}$ is obtained by performing a single Newton-Raphson update based on the simple averaging estimator $\theta^{(0)}$, i.e., we have

$$\theta^{(1)} = \theta^{(0)} - [\ddot{M}(\theta^{(0)})]^{-1} [\dot{M}(\theta^{(0)})] \tag{9}$$

where $M(\theta) = \frac{1}{k} \sum_{i=1}^{k} M_i(\theta)$ is the global empirical criterion function, $\dot{M}(\theta)$ and $\ddot{M}(\theta)$ are the gradient and Hessian of $M(\theta)$, respectively. In Huang and Huo (2015) [**?**], the dimension d of the parameter space, Θ, is assumed to be at most moderate. Consequently, the Hessian matrix $\ddot{M}(\theta)$, which should be $d \times d$, is not considered to be large. The process of computing the one-step estimator can be summarized as follows.

1. For each $i \in \{1, 2, \ldots, k\}$, machine i computes the local M-estimator with its local data set,

$$\theta_i = \arg\max_{\theta \in \Theta} M_i(\theta) = \arg\max_{\theta \in \Theta} \frac{1}{|S_i|} \sum_{x \in S_i} m(x; \theta).$$

2. The simple averaging estimator is obtained as follows,

$$\theta^{(0)} = \frac{1}{k}\sum_{i=1}^{k}\theta_i\,.$$

   Then $\theta^{(0)}$ is sent back to each local machine.

3. For each $i \in \{1, 2, \ldots, k\}$, the gradient and the Hessian matrix of its local empirical criterion function $M_i(\theta)$ at $\theta = \theta^{(0)}$ are first computed by machine i and then sent back to the central machine.

4. At the central machine, the one-step estimator is then computed as follows,

$$\dot{M}(\theta^{(0)}) = \frac{1}{k}\sum_{i=1}^{k}\dot{M}_i(\theta^{(0)}),\ \ddot{M}(\theta^{(0)}) = \frac{1}{k}\sum_{i=1}^{k}\ddot{M}_i(\theta^{(0)}).$$

$$\theta^{(1)} = \theta^{(0)} - [\ddot{M}(\theta^{(0)})]^{-1}[\dot{M}(\theta^{(0)})].$$

4.2 Theoretical Guarantees

Under some regularity conditions, the consistency results as below are proved in Huang and Huo (2015) [**?**]:

$$\theta^{(1)} \xrightarrow{P} \theta_0, \quad \sqrt{N}(\theta^{(1)} - \theta_0) \xrightarrow{d} \mathbf{N}(0, \Sigma), \quad \text{as } N \to \infty,$$

where Σ is the covariance matrix.

5 Outcome of the Meeting

The previous section gives one example of the distributed inference algorithm. An interesting aspect is that one can derive adequate theoretical guarantees. In particular, for the algorithm that is described in Section **??**, one actually shows that the derived algorithm can be as good as the oracle algorithm (that is the best possible method when the data were not distributed in different locations).

References

[1] Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed estimation and inference with statistical guarantees. *arXiv preprint arXiv:1509.05457*, 2015.

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

[3] Cheng Huang and Xiaoming Huo. A distributed one-step estimator. *arXiv preprint arXiv:1511.01443*, 2015.

[4] Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, (just-accepted), 2018.

[5] Jason D Lee, Yuekai Sun, Qiang Liu, and Jonathan E Taylor. Communication-efficient sparse regression: a one-shot approach. *arXiv preprint arXiv:1503.04337*, 2015.