Received: 16 April 2018 | Revised: 31 May 2018 | Accepted: 15 June 2018

DOI: 10.1002/wics.1451

WILEY WIREs COMPUTATIONAL STATISTICS

ADVANCED REVIEW

Aggregated inference

Xiaoming Huo | Shanshan Cao

School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia

Correspondence
School of Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Dr., Atlanta, GA 30332-0360.
Email: huo@gatech.edu

Funding information
National Science Foundation, Grant/Award Numbers: DMS-1613152, CCF-1740776; Transdisciplinary Research Institute for Advancing Data Science (TRIAD)

Aggregated inference on distributed data becomes more and more important due to the larger size of data collected in different industries. Modeling and inference are needed in the case where data cannot be obtained at a central location; aggregated statistical inference is a major tool to solve the aforementioned problems. In the literature, problems under the setting of regression model (more generally, M-estimator) are extensively studied. There are at least two popular techniques for distributed estimation: (a) averaging estimators from local locations and (b) the one-step approach, which combines the simple averaging estimator with a classical Newton's method (using the local Hessian matrices) to generate a "one-step" estimator. It is proved that under certain assumptions, the above constructed estimators enjoy the same asymptotic properties as the centralized estimator, which is obtained as if all data were available at a central location. We review the aforementioned two major estimations. It can be seen that, in Big-Data problems, dividing the data to multiple machines and then using the aggregation technique to solve the estimation problem in parallel can speed up the computation with little compromise of the quality of the estimators. We discuss potential extensions to other models, such as support vector machine, principle component analysis, and so on. Numerical examples are omitted due to the space limitation; they can be easily found in the literature.

This article is categorized under:
Statistical Learning and Exploratory Methods of the Data Sciences > Knowledge Discovery
Statistical Learning and Exploratory Methods of the Data Sciences > Modeling Methods
Statistical Models > Fitting Models
Statistical and Graphical Methods of Data Analysis > Modeling Methods and Algorithms

KEYWORDS
aggregated inference, averaging estimator, distributed statistical inference, M-estimation, one-step estimator

1 | INTRODUCTION

Aggregated inference is a relatively new area in statistical inference. It mainly deals with problems where the data is not available at a central location. It often happens due to different properties or background of the collected data. For example, a major web search engine has to save its data in a range of platforms, storage units, and even geographical locations (Corbett et al., 2013; Mitra et al., 2011); an international firm or organization might need to store data in different regions; a supply chain company, a government, hospitals, and the centers for disease control and prevention (CDC) all have tremendous amount of data, which are stored over different agencies or areas. In most of these cases, it is costly or prohibitive to transport all the data

to a central machine, even though a model fitting may need all the data at a central location. Furthermore, the speed of local processors can be thousands time faster than the transmission of the data to a central machine. A communication-efficient algorithm is advantageous to be developed. Additionally, one may hope that the produced aggregated solution is still consistent and has the same asymptotic efficiency as the centralized estimator that would be obtained as if all the data were available at a fusion center.

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.

In our terminology, *aggregated inference* indicates statistical inference and estimation on homogeneous and distributed data, while *distributed inference* is more inclusive, which indicates statistical inference and estimation on general distributed potentially heterogeneous data. One commonly method of deriving pooled results from heterogeneous datasets is the meta-analysis (Walker, Hernandez, & Kattan, 2008). In the meta-analysis, one believes that a common truth exists behind all conceptually similar scientific studies, but which has been measured with a certain error within individual studies. In the current work, we focus on a review of the aggregated inference.

Aggregated inference and communication-efficient algorithms are widely used in a variety of areas, such as regression, classification, network modeling, and so on. In this article, we focus on reviewing the main results under the M-estimator setting. Other areas will be briefly discussed in the end with references for further reading.

In the literature, there are at least two types of distributed data. One is that an observation X_i is observed completely at a particular local station. Different stations gather different subsets of the sample. One may even assume that the observations from different stations are independently and identically distributed (i.i.d.). For example, in customer segmentation, people would like to build a logistic regression model for segmentation. For each customer, it is assumed that they fall into one segmentation i.i.d. However, information from local stores (retailers) cannot be gathered at a central location. We mainly focus on problems with this type of data. The other type is that variables in each observation X_i are not observed at the same station. Different observations could also be available at different locations. This has been studied in papers such as Song and Liang (2015), El Gamal and Lai (2015), and so on.

This document is organized as follows. We first present a comprehensive review of the current literature on the recent progress of distributed optimization algorithms and aggregated statistical inference in Section 2, focusing on the high-level ideas of the existing works. We then describe the general formulation of the aggregated statistical inference problem under the setting of M-estimator (Section 3). Two representative the state-of-the-art estimators in the literature are thoroughly reviewed. Theoretical results regarding the consistency of these estimators are presented in Section 4. We discuss other types of distributed inference and some open problems related to distributed statistical inference in Section 5. This review is concluded in Section 6.

1.1 | Notations

In this subsection, we introduce some notations that are used. Let $\{m(x;\theta):\theta\in\Theta\subset\mathbb{R}^d\}$ denote a collection of criterion functions. The dataset consisting of $N=nk$ samples is denoted by $S=\{X_1,\cdots,X_N\}$, where the samples are drawn i.i.d. from the probability density function (PDF) $p(x)$. In the regression context, let $Y\in\mathbb{R}^N$ denote the response corresponding to the sample S. This dataset is divided evenly at random and stored in k machines. Let S_i denote the subset of data assigned to machine i, and $D_i\subset\{1,\cdots,N\}$ denote the index of data in S_i, $i=1,\ldots,k$. Let Y_{D_i} denote the subvector of responses assigned to machine i, $i=1,\ldots,k$. The generalized ℓ_q norm for $1\le q\le\infty$ is denoted by $\|\cdot\|_q$, where $\|a\|_q=\left(\sum_{j=1}^d a_j^q\right)^{\frac{1}{q}}$. Particularly, the Euclidean norm is denoted as $\|\cdot\|$. For $a\in\mathbb{R}^d$, $\|a\|=\left(\sum_{j=1}^d a_j^2\right)^{\frac{1}{2}}$. $\|a\|_0=|\mathrm{supp}(a)|$, where we have $\mathrm{supp}(a)=\{j|a_j\neq 0\}$, and $|\mathrm{supp}(a)|$ is the cardinality of the set $\mathrm{supp}(a)$. And we use $|||\cdot|||$ to denote a norm for matrix $A\in\mathbb{R}^{d\times d}$, which is defined as its maximal singular value, that is, we have

$$|||A|||=\sup_{u:u\in R^d,\,\|u\|\le 1}\|Au\|.$$

2 | OVERVIEW

In the literature, the distributed problem has been studied for a long time. Plenty of work has been done in the area of distributed optimization for large-scale problems. Distributed and aggregated statistical inference has also been studied in many existing works. Due to the high volume of the data, statistical estimation has to be obtained in a distributed manner through

communication-efficient algorithms. On the other hand, consistency and asymptotic efficiency (in relative to the estimator that would be obtained via the centralized data) are desired properties in such distributed algorithms. In this section, we review the literature on distributed optimization algorithms and the aggregated statistical inference in the Section 2.1 and 2.2, respectively.

2.1 | Distributed optimization

In this subsection, we review the literature on existing distributed optimization algorithms. Alternating direction method of multipliers (ADMM) is advocated by Boyd et al. (2011) to solve distributed optimization problems in statistics and machine learning. The idea of ADMM was first introduced by Gabay and Mercier (1976). Then it was well studied during 1980s and 1990s for theoretical properties. Boyd et al. (2011) redevelop the ADMM algorithm for the distributed version of the least absolute shrinkage and selection operator (LASSO) estimators (Tibshirani, 1996), the distributed logistic regression estimators, covariance selection, support vector machines (SVMs), and many more. ADMM is an iterative algorithm that is feasible for a wide range of distributed problems. However, the communication between the central machine and the local machines in each iteration can be inefficient. In another line of the literature on aggregated statistical inference, the multiplicity in communication can be avoided by communicating with the local processors only once or twice, on information like local estimators, local gradients, and local Hessian matrices.

Zinkevich, Weimer, Li, and Smola (2010) propose to use the parallelized stochastic gradient descent method to minimize the empirical risk. Convergence results are presented. Shamir, Srebro, and Zhang (2014) develop the distributed approximation Newton-type method (DANE) for distributed statistical problems. They propose an iterative algorithm where the average of the local gradient is computed, followed by averaging the local estimators at each iteration. They prove the linear convergence rate in quadratic problems. The algorithm still requires communication at each iteration by transmitting the local gradient and the local estimators.

Jaggi et al. (2014) develop a communication-efficient algorithm for distributed optimization in machine learning. Local computation is combined with the randomized dual coordinate descent in the primal-dual setting. Their algorithm is proven to have the geometric convergence rate.

All the above are methods to solve for the numerical solution for an optimization problem in statistics and machine learning, where statistical inference such as the convergence in probability, asymptotic normality, Fisher information bound, is untouched. Furthermore, most of the above algorithms require communication between local and central machines at each iteration. Below, we review the literature on aggregated statistical inference, which focuses on the inference of the estimators and the reduction in the communication requirements between the local and center machines.

2.2 | Aggregated statistical inference

In this subsection, we briefly review the aggregated statistical inference at the high level in the setting of the M-estimators, more specifically, the penalized M-estimator setting. Aggregated inference is used to analyze the estimator drawn from different distributed algorithms. We focus on the theoretical properties, such as consistency, asymptotic normality, and so on, of the estimator. The study on aggregated inference provides theoretical guarantees on estimators obtained from the communication-efficient distributed algorithms. Below, we review the main ideas of aggregated inference in the literature. The problem formulation and consistency analysis are discussed.

Y. Zhang, Wainwright, and Duchi (2012) present an intuitive approach to solve large-scale statistical optimization problem. They use the average of the local empirical risk minimizers as their estimator, where they split the N sample data evenly into k subsets. Under some regular assumptions, for example, the convexity of the parameter space and the local strong convexity of the convex loss functions, and so on, they show that the averaging estimator achieves mean squared errors (MSE) that decays as $O(N^{-1} + (N/k)^{-2})$. They also show that the MSE could be even reduced to $O(N^{-1} + (N/k)^{-3})$ with one additional bootstrapping subsampling step.

Liu and Ihler (2014) propose an inspiring two-step approach: the first step is to find the local maximum-likelihood estimators $\hat{\theta}_i$; the second step is to combine them by minimizing the total Kullback–Leibler divergence (KL-divergence) between the local distributions and the desired distribution. They prove the exactness of their estimator as the global maximal-likelihood estimator in the exponential family. They also estimate the MSE of the proposed estimator for a curved exponential family. Due to the adoption of the KL-divergence, the effectiveness of this approach heavily depends on the parametric form of the underlying model.

Chen and Xie (2014) propose to use a split-and-conquer approach for a penalized regression problem (in the generalized linear model [GLM] setting) in the case of extraordinarily large dataset. They apply majority voting to decide the support for the estimated parameter, followed by estimating the parameter using a weighted average of the local estimators. They show

that the resulting estimator enjoys the same oracle property as the method that uses the entire dataset in a single machine. They require the number of local machines to satisfy $k \leq O\left(N^{\frac{1}{3}}\right)$.

Rosenblatt and Nadler (2016) analyze the error of the averaging estimator in distributed statistical learning under two settings. In the first setting, the number of machines is fixed and in the second one, the number of machines grows in the same order with the number of samples per machine. The asymptotically exact expression for estimation errors in both scenarios are provided, where the error grows linearly with the number of machines in the latter case.

Battey, Fan, Liu, Lu, and Zhu (2015) study the distributed parameter estimation method for penalized regression and establish the oracle asymptotic property of the averaging debiased estimator. Precise upper bounds on the errors of their proposed estimator have been derived. Hypotheses' testing is also considered in their paper.

Lee, Sun, Liu, and Taylor (2015) use a one-shot approach, where the averaging "debiased" LASSO estimators is developed, to distributed sparse regression in the high-dimensional setting. It is shown that their approach enjoys the same convergence rate as the LASSO estimator when the dataset is not split across too many machines. Some intermediate quantities in the centralized fashion need to be distributed to local machines.

Another work that applies the idea of one additional updating in the aggregated inference literature is Huang and Huo (2015). They propose a one-step estimator, which utilizes the Hessian matrix to update the simple-averaging estimator for one extra step. They show that the proposed one-step estimator enjoys the same asymptotic properties (including convergence and asymptotic normality) as the centralized estimator that would utilize the entire data. The upper bound for the MSE of the proposed one-step estimator can be slightly better than those in the existing literature.

Most recently, Jordan, Lee, and Yang (2018) propose a unified framework for distributed statistical inference called communication-efficient surrogate-likelihood (CSL) framework. They first find a surrogate-likelihood function using the gradient function only or with the Hessian matrix, the latter would require more transmission of information. Then they illustrate the distributed statistical inference problem using the surrogate-likelihood function in the low-dimensional estimation, high-dimensional regularized estimation, and the Bayesian inference. They provide some theoretical results on the error bounds in the ℓ_2 norm of the global estimator in two scenarios: (a) all the data are available at a central machine and (b) the data are distributed and the estimator is based on the surrogate-likelihood function approach. They argue that iteratively updating the surrogate function can produce an optimal estimator.

In summary, there are at least two lines in the distributed statistical inference and estimation problems: the simple averaging estimators and the one-step estimator, which is based on the former. Iterative updating algorithms, which require more 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 al. (2018) study distributed statistical inference comprehensively in all the aforementioned settings. Chen and Xie (2014), Battey et al. (2015), and Lee et al. (2015) consider a high-dimensional however sparse parameter vector estimation problem, where they adopt the penalized M-estimator setting.

The one-step method, which could make a consistent estimator as efficient as the maximum-likelihood estimation (MLE) or M-estimators, with a single Newton–Raphson iteration, has been in the literature for a long time, where the efficiency means that the relative efficiency converging to 1. More details on the general one-step approach can be found in Van der Vaart (2000). Examples where the one-step method is used can be found in Bickel (1975), Fan and Chen (1999), and Zou and Li (2008).

In the following review, we adopt the formulation in Battey et al. (2015) and Huang and Huo (2015), to demonstrate the distributed M-estimator, with and without a penalty, respectively.

3 | GENERAL FRAMEWORK

In this section, we describe the general formulation of the distributed estimation and inference in the setting of the M-estimator, which is a generalization of the 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 1.

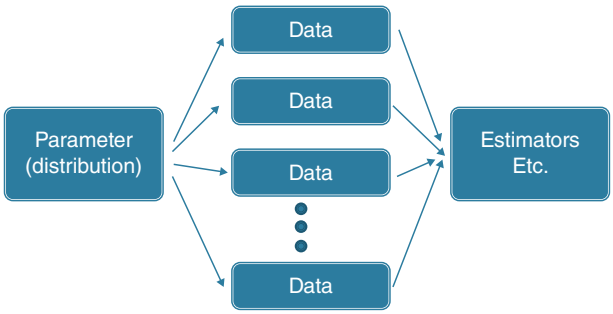

FIGURE 1 Diagram of aggregated statistical estimation: The first step is obtaining the local parameter estimation using the local data and the second step is computing the centralized estimation using the local estimations

The remaining of this section is organized as follows. We first review the M-estimator in Section 3.1. The aggregated estimators that are studied in Huang and Huo (2015) and Battey et al. (2015) are reviewed in details in Section 3.2 and Section 3.3, respectively.

3.1 | M-estimator

Let $\hat{\theta}$ denote the M-estimators, which is a generalization of the MLE and can be obtained by maximizing the empirical criterion (the negative loss) function. We have

$$\hat{\theta} = \arg\max_{\theta\in\Theta} M(\theta) = \arg\max_{\theta\in\Theta} \frac{1}{|S|}\sum_{x\in S} m(x;\theta).$$

The MLE is obtained when the criterion function is the log-likelihood function, that is, $m(x;\theta) = \log p(x;\theta)$, where $p(x;\theta)$ is the probability density function/probability mass function (PMF) of the observations $X_i \in S$. Let $M_0(\theta) = \int_{\mathcal{X}} m(x;\theta)p(x)dx$ denote the population criterion function and $\theta_0 = \arg\max_{\theta\in\Theta} M_0(\theta)$ denote the maximizer of the population criterion function. It is well known that under certain conditions, the aforementioned estimator, $\hat{\theta}$, is a consistent estimator for θ_0, that is, $\hat{\theta} - \theta_0 \to^P 0$, seeing Chapter 5 of Van der Vaart (2000).

3.2 | Distributed one-step M-estimator

In this subsection, we mainly review the construction of the one-step distributed M-estimator that is proposed in Huang and Huo (2015). We first formulate the problem mathematically. Then we derive the estimator. The theoretical properties are reviewed in Section 4.

For each $i \in \{1, \ldots, k\}$, the local empirical criterion function that is based on the local dataset S_i on machine i and the corresponding maximizer are denoted by

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

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 consistent with the notations in Section 3.1, we have

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

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

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

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

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

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

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

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 $\theta^{(0)}$, that is, we have

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

where $M(\theta)=\frac{1}{k}\sum_{i=1}^{k}M_i(\theta)$ is the global empirical criterion function, $M(\theta)$ 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 dataset,

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

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,

$$M\left(\theta^{(0)}\right)=\frac{1}{k}\sum_{i=1}^{k}M_i\left(\theta^{(0)}\right),\ \ddot{M}\left(\theta^{(0)}\right)=\frac{1}{k}\sum_{i=1}^{k}\ddot{M}_i\left(\theta^{(0)}\right).$$
$$\theta^{(1)}=\theta^{(0)}-\left[\ddot{M}\left(\theta^{(0)}\right)\right]^{-1}\left[M\left(\theta^{(0)}\right)\right].$$

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

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

where the covariance matrix, Σ, will be specified later.

The above procedure is only applicable in cases where the dimension of the observations is moderately large. When the dimension is high, the transmission of the Hessian matrices can be costly.

3.3 | Distributed high-dimensional sparse estimator

The content in this section follows the work of Battey et al. (2015), which deals with the distributed inference on high-dimensional sparse estimation in general likelihood-based framework. This is a generalization of the M-estimator with a

penalty to promote the sparsity of the parameter in both the linear regression and the generalized linear regression (GLM) settings. Given the observations S and Y, where the response Y_i is distributed as F_{β^*}, for $i = 1, \cdots, N$, which is the cumulative distribution function conditioned on the explanatory variables X_i. Let f_{β^*} be the corresponding probability density or mass function (PDF or PMF) of the distribution. Note that in order to differentiate with the previous setting, we denote the parameter of interest with β in this subsection. The negative log-likelihood function of the data, $\ell_N(\beta)$, is defined as

$$\ell(\beta) = -\frac{1}{N}\sum_{i=1}^{N} \log f_\beta(Y_i|X_i).$$

And the penalized M-estimator is defined as

$$\hat{\beta} = \arg\min\ \ell(\beta) + P_\lambda(\beta),$$

where the function $P_\lambda(\beta)$ is a sparsity-inducing penalty function. Examples for the penalty function $P_\lambda(\beta)$ include the well-known convex ℓ_1 penalty, $P_\lambda(\beta) = \lambda\|\beta\|_1$ (Tibshirani, 1996), and some folded concave penalties such as the smoothly clipped absolute deviation (SCAD; Fan & Li, 2001) and the minimax concave penalty (MCP) (C.-H. Zhang, 2010), and many more. Similarly, the negative log-likelihood function $\ell_j(\beta)$, $j = 1, \cdots, k$, of the data on each machine can be defined as follows

$$\ell_j(\beta) = -\frac{1}{n}\sum_{i\in D_j} \log f_\beta(Y_i|X_i). \tag{10}$$

The corresponding estimation problem at each local machine is

$$\hat{\beta}_j = \arg\min \ell_j(\beta) + P_\lambda(\beta). \tag{11}$$

For linear model

$$Y_i = X_i^T\beta^* + \epsilon_i,$$

where $\{\epsilon_i\}_{i=1}^N$ are i.i.d with mean 0 and variance σ^2, the procedure to estimate the parameter β from the distributed data is as follows.

1. Compute the debiased estimator $\hat{\beta}_j^d$ for LASSO estimator at each local machine j, for $j = 1, \cdots, k$.
2. The final estimator is the average of the local debiased estimators: $\overline{\beta}^d = \frac{1}{k}\sum_{j=1}^k \hat{\beta}_j^d$.

In the first step, in order to obtain the debiased estimator, $\hat{\beta}_j^d$, for the LASSO estimator at each local machine j, $j = 1, \cdots, k$, one can follow the proposed method in Javanmard and Montanari (2014). Let $\hat{\beta}_j^l$ denote the local LASSO estimator from machine j, $j = 1, \cdots, k$. The debiased estimator is

$$\hat{\beta}_j^d = \hat{\beta}_j^l + \frac{1}{n}M_j\left(X_{D_j}\right)^T\left(Y_{D_j} - X_{D_j}\hat{\beta}_j^l\right),$$

where $M_j = \left[m_1^{(j)}, \cdots, m_d^{(j)}\right]^T$ and $m_v^{(j)}$, $v = 1, \cdots, d$, is obtained by solving the following

$$\begin{aligned} m_v^{(j)} &= \arg\min_{b\in\mathcal{R}^d} m^T\hat{\Sigma}^{(j)}m \\ \text{subject to } &\left\|\hat{\Sigma}^{(j)}m - e_v\right\|_\infty \le \nu_1, \\ &\left\|X_{D_j}m\right\|_\infty \le \nu_2, \end{aligned} \tag{12}$$

where $\hat{\Sigma}^{(j)} = \frac{1}{n}\left(X_{D_j}\right)^T X_{D_j}$ is the sample covariance matrix at the jth machine, and $\{e_1, \cdots, e_d\}$ denotes the canonical basis for $\mathcal{R}^d$. The choice of ν_1 and ν_2 is discussed in Javanmard and Montanari (2014). From the constraints in problem (12), M_j can be interpreted as the regularized inverse of the population covariance matrix $\Sigma = \mathbb{E}\left[X_1X_1^T\right]$. Battey et al. (2015) propose the following as a substitute of the above procedure

$$\hat{\beta}_j^d = \hat{\beta}_j^l + \frac{1}{n}B_j^T\left(Y_{D_j} - X_{D_j}\hat{\beta}_j^l\right),$$

where the matrix $B_j \in \mathscr{R}^{n \times d}$ is defined as $B_j = \left(b_1^j, \cdots, b_d^j\right)$, with

$$\begin{aligned} b_v^j = & \arg\min_{b \in \mathscr{R}^n} \frac{b^T b}{n}, \\ \text{subject to } & \left\| \frac{X_{D_j}^T b}{n} - e_v \right\|_\infty \le \nu_1, \\ & \|b\|_\infty \le \nu_2, \end{aligned} \tag{13}$$

for $v = 1, \cdots, d$.

Comparable estimation and theoretical results under the GLM have been developed in Battey et al. (2015). In the GLM setting, the PDF is of the form

$$f\left(Y_i; X_i, \beta^*\right) = f\left(Y_i; \eta_i^*\right) = c(Y_i) \exp\left\{ \frac{Y_i \eta_i^* - b\left(\eta_i^*\right)}{\phi} \right\},$$

where $\eta_i^* = X_i^T \beta^*$. Thus, the negative log-likelihood function can be written as

$$\ell(\beta) = \frac{1}{N} \sum_{i=1}^N -Y_i X_i^T \beta + b\left(X_i^T \beta\right) = \frac{1}{N} \sum_{i=1}^N -Y_i \eta_i + b(\eta_i).$$

The gradient and Hessian of $\ell(\beta)$ can be expressed as follows

$$\nabla \ell(\beta) = -\frac{1}{N} X^T (Y - \mu(X\beta)), \quad \text{and} \quad \nabla^2 \ell(\beta) = \frac{1}{N} X^T D(X\beta) X,$$

where we have $\mu(X\beta) = (b'(\eta_1), \cdots, b'(\eta_N))^T$ and $D(X\beta) = \text{diag}\{b''(\eta_1), \cdots, b''(\eta_N)\}$. Thus, the information matrix at the ground truth is $J^* = \mathbb{E}\left[b''\left(X_1^T \beta^*\right) X_1 X_1^T\right]$. The debiased estimator in the GLM is borrowed from Van de Geer et al. (2014). We have

$$\hat{\beta}_j^d = \hat{\beta}_j - \hat{\Theta}_j \nabla \ell_j\left(\hat{\beta}_j\right),$$

where $\hat{\beta}_j$ is the optimal solution to the local problem (11) at the jth machine, $\nabla \ell_j\left(\hat{\beta}_j\right)$ is the gradient of $\ell_j(\beta)$ at $\hat{\beta}_j$, $\hat{\Theta}_j$ is a regularized inverse of the Hessian matrix (the second-order derivatives of $\ell_j(\beta)$) at $\hat{\beta}_j$, for $j = 1, \cdots, k$. Similarly, the debiased estimator $\hat{\beta}^d$ with all data can also be derived from the estimator $\hat{\beta}$.

Hypothesis testing and accuracy of the distributed estimation are studied in Battey et al. (2015). Specifically, the authors discuss the distributed Score test and the distributed Wald test in both the linear regression case and the GLM case. They provide some bias bounds in the ℓ_∞ norm and ℓ_2 norm.

Next, we review the consistency results in Section 4.

4 | CONSISTENCY THEORIES

In this section, we discuss the theoretical properties of the estimators constructed in Section 3. The proofs are omitted due to the space limitation and can be found in Huang and Huo (2015).

4.1 | Theoretical results on the distributed one-step M-estimator

In this section, we review the theoretical results on the one-step estimator discussed in Section 3.2. It is easily seen that the one-step estimator $\theta^{(1)}$ defined in Equation (9) is not necessarily the maximizer of the empirical criterion function $M(\theta)$, but it shares the same asymptotic properties with the corresponding global maximizer (M-estimator) under some mild conditions. Below, we first provide the assumptions for the proof of the theorems. Theoretical results are formally stated after the assumptions. Succinct discussion is also provided thereafter.

Assumption 4.1: ***(parameter space)*** *The parameter space $\Theta \in \mathbb{R}^d$ is a compact convex set.*

Assumption 4.2: ***(concave criterion function)*** *$m(x;\theta)$ is concave with respect to θ.*

Assumption 4.3: ***(invertibility)*** *The Hessian of the population criterion function $M_0(\theta)$ at θ_0 is a nonsingular matrix, which means $\ddot{M}(\theta_0)$ is negative definite and there exists some $\lambda > 0$ such that* $\sup_{u\in\mathbb{R}^d:\|u\|<1} u^t\ddot{M}(\theta_0)u \le -\lambda$.

Let $B_\delta = \{\theta \in \Theta: \|\theta - \theta_0\| \le \delta\}$.

Assumption 4.4: ***(smoothness)*** *There exist some constants G and H such that*

$$\mathbb{E}\left[\|m(X;\theta)\|^8\right] \le G^8 \text{ and } \mathbb{E}\left[\left|\left|\left|\ddot{m}(X;\theta)-\ddot{M}_0(\theta)\right|\right|\right|^8\right] \le H^8, \forall\theta\in B_\delta.$$

For any $x \in \mathcal{X}$, the Hessian matrix $\ddot{m}(x;\theta)$ is L(x)-Lipschitz continuous,

$$|||\ddot{m}(x;\theta)-\ddot{m}(x;\theta')||| \le L(x)\|\theta-\theta'\|, \forall\theta,\theta'\in B_\delta,$$

where L(x) satisfies

$$\mathbb{E}\left[L(X)^8\right] \le L^8 \text{ and } \mathbb{E}\left[(L(X)-\mathbb{E}[L(X)])^8\right] \le L^8,$$

for some finite constant $L > 0$.

The main result is that the one-step estimator enjoys the oracle asymptotic normality properties and has the MSE of $O(N^{-1})$ under some mild conditions.

Theorem 4.1: *Let* $\Sigma = \ddot{M}_0(\theta_0)^{-1}\mathbb{E}\left[m(x;\theta_0)m(x;\theta_0)^t\right]\ddot{M}_0(\theta_0)^{-1}$, *where the expectation is taken with respect to p(x). Under Assumptions 4.1, 4.2, 4.3, and 4.4, when the number of machines k satisfies* $k = O(\sqrt{N})$, $\theta^{(1)}$ *is consistent and asymptotically normal, that is, we have*

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

Theorem 4.2: *Under Assumptions 4.1, 4.3, and 4.4, the MSE of the one-step estimator* $\theta^{(1)}$ *is bounded by*

$$\mathbb{E}\left[\left\|\theta^{(1)}-\theta_0\right\|^2\right] \le \frac{2\mathrm{Tr}[\Sigma]}{N} + O\left(N^{-2}\right) + O\left(k^4N^{-4}\right).$$

When the number of machines k satisfies $k = O(\sqrt{N})$, we have

$$\mathbb{E}\left[\left\|\theta^{(1)}-\theta_0\right\|^2\right] \le \frac{2\mathrm{Tr}[\Sigma]}{N} + O\left(N^{-2}\right).$$

Particularly, the one-step estimator has the same asymptotic properties as the MLE when the criterion function is exactly the log likelihood function, that is, $m(x; \theta) = \log f(x; \theta)$.

Corollary 4.1: *If* $m(x; \theta) = \log f(x; \theta)$ *and* $k = O(\sqrt{N})$, *the one-step estimator* $\theta^{(1)}$ *is a consistent and asymptotic efficient estimator of* θ_0,

$$\theta^{(1)}-\theta_0 \xrightarrow{P} 0 \text{ and } \sqrt{N}\left(\theta^{(1)}-\theta_0\right) \xrightarrow{d} \mathbf{N}\left(0, I(\theta_0)^{-1}\right), as\, N \to \infty,$$

where $I(\theta_0)$ is the Fisher's information at $\theta = \theta_0$. And the MSE of $\theta^{(1)}$ is bounded as follows:

$$\mathbb{E}\left[\left\|\theta^{(1)}-\theta_0\right\|^2\right] \le \frac{2Tr[I^{-1}(\theta_0)]}{N} + O\left(N^{-2}\right) + O\left(k^4N^{-4}\right).$$

In comparison to Y. Zhang et al. (2012), where it is shown that there exists some constant C_1, $C_2 > 0$ such that

$$\mathbb{E}\left[\left\|\theta^{(0)}-\theta_0\right\|^2\right] \le \frac{C_1}{N} + \frac{C_2k^2}{N^2} + O\left(kN^{-2}\right) + O\left(k^3N^{-3}\right),$$

the one-step estimator $\theta^{(1)}$ achieves a lower upper bound of MSE with only one additional step.

Below, we briefly review the solution under the scenario when there are communication failures. Communication failure happens in reality, where the information (local estimator, local gradient and local Hessian) from a local machine cannot be received by the central machine. We discuss the distributed estimation and inference under the condition that the communication failure happens to each machine independently. Let $a_i \in \{0, 1\}$, i = 1, …, k, denote the status of local machines: $a_i = 1$ when machine i successfully sends all its local information to the central machine; otherwise we have $a_i = 0$. The corresponding simple-averaging estimator is thus computed as

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

The corresponding one-step estimator is as follows

$$\theta^{(1)} = \theta^{(0)} - \left[\sum_{i=1}^{k} a_i \ddot{M}_i\left(\theta^{(0)}\right)\right]^{-1}\left[\sum_{i=1}^{k} a_i M_i\left(\theta^{(0)}\right)\right],$$

which is the one-step distributed estimator using only the datasets from the working machines.

Corollary 4.2: *Suppose r is the probability (or rate) that a local machine fails to send its information to the central machine. When $n = N/k \to \infty$, $k \to \infty$ and $k = O\left(\sqrt{N}\right)$, the one-step estimator is asymptotically normal:*

$$\sqrt{(1-r)N}\left(\theta^{(1)} - \theta_0\right) \xrightarrow{d} \mathbf{N}(0, \Sigma).$$

And more precisely, unless all machines fail, we have

$$\mathbb{E}\left[\left\|\theta^{(1)} - \theta_0\right\|^2\right] \le \frac{2\mathrm{Tr}[\Sigma]}{N(1-r)} + \frac{6\mathrm{Tr}[\Sigma]}{Nk(1-r)^2} + O\left(N^{-2}(1-r)^{-2}\right) + O\left(k^2 N^{-2}\right).$$

4.2 | Theoretical results on the distributed sparse estimator

In this section, we review the theoretical results on the sparse distributed estimator described in Section 3.3. The estimation accuracy results are discussed in both the linear model and the GLM settings. We focus on the linear case in the following. Similar asymptotic normality results, and accuracy error bound are also available under the GLM setting. We first list the necessary assumptions. The theorems regarding the statistical properties are then presented without proof.

Assumption 4.5: *For any $\delta \in (0, 1)$, if $\lambda = O\left(\sqrt{\log(d/\delta)/N}\right)$,*

$$P\left(\left\|\hat{\beta} - \beta^*\right\|_1 > Cs\sqrt{\log(d/\delta)/N}\right) \le \delta,$$

where s is the sparsity of the ground truth β^*, that is, $s = \|\beta^*\|_0$.

Let Σ denote the population covariance matrix of the samples X_i, that is, $\Sigma = \mathbb{E}\left[X_1 X_1^T\right]$.

Assumption 4.6: *$\{Y_i, X_i\}_{i=1}^N$ are i.i.d. and the population covariance matrix Σ satisfies $0 < C_{min}$ $\lambda_{min}(\Sigma)$ $\lambda_{max}(\Sigma)$ C_{max}.*

Let $\|\cdot\|_{\psi_2}$ denote the sub-Gaussian norm, that is, for a random variable X, $\|X\|_{\psi_2} = \sup_{q \ge 1} q^{-\frac{1}{2}}(\mathbb{E}|X|^q)^{1/q}$, and for a random vector $X \in \mathscr{R}^d$, $\|X\|_{\psi_2} = \sup_{x \in S^{d-1}} \|\langle X, x\rangle\|_{\psi_2}$, where S^{d-1} denotes the unit sphere in $\mathscr{R}^d$.

Assumption 4.7: The rows of X are sub-Gaussian with $\|X_i\|_{\psi_2} \le \kappa$, for $i = 1, \cdots, N$.

The following theorems can be established.

Theorem 4.3: *Under Assumptions 4.5, 4.6, and 4.7, if* $\mathbb{E}[\epsilon^4] < \infty$, *and* ν_1, ν_2 *and* k *are chosen as* $\nu_1 = O\left(\sqrt{k \log d/N}\right)$, $\nu_2 n^{-1/2} = o(1)$ *and* $k = o\left((s \log d)^{-1}\sqrt{N}\right)$, *for any* $v = 1, \cdots, d$, *we have*

$$\sqrt{N}\frac{1}{k}\sum_{j=1}^{k}\frac{\left(\hat{\beta}_j^d\right)_v - \beta_v^*}{\hat{Q}_v^{(j)}} \to N(0, \sigma^2),$$

where $\hat{Q}_v^{(j)} = \left(m_v^{(j)T}\hat{\Sigma}^{(j)}m_v^{(j)}\right)^{1/2}$.

Theorem 4.4: *Under Assumptions 4.6, 4.7, if* λ, ν_1, *and* ν_2 *are chosen as* $\lambda = O\left(\sqrt{k \log d/N}\right)$, $\nu_1 = O\left(\sqrt{k \log d/N}\right)$, *and* $\nu_2 n^{-1/2} = o(1)$, *we have with probability* $1 - c/d$,

$$\left\|\overline{\beta}^d - \hat{\beta}^d\right\|_\infty \le C\frac{sk \log d}{N} \text{ and } \left\|\hat{\beta}^d - \beta^*\right\|_\infty \le C\left(\sqrt{\frac{\log d}{N}} + \frac{sk \log d}{N}\right).$$

From the above results, $\overline{\beta}^d$ achieves the same rate as the LASSO estimator under the infinity norm, while this results does not hold for the ℓ_2 norm since the debiased estimator is no longer sparse. For any $\beta \in \mathcal{R}^d$, let the hard threshold operator $\mathcal{T}_v$ be such that the jth entry of $\mathcal{T}_v(\beta)$ is

$$[\mathcal{T}_v(\beta)]_j = \beta_j 1\{|\beta_j| \ge v\}, \text{ for } j = 1, \cdots, d.$$

Theorem 4.5: *Under Assumptions 4.6, 4.7, if* λ, ν_1, *and* ν_2 *are chosen as* $\lambda = O\left(\sqrt{k \log d/N}\right)$, $\nu_1 = O\left(\sqrt{k \log d/N}\right)$, *and* $\nu_2 n^{-1/2} = o(1)$. *Take the parameter in the hard threshold operator* $\mathcal{T}_v$ *as* $v = C_0\sqrt{\log d/N}$ *for some large enough constant* C_0. *If the number of local machine satisfies* $k = O\left(\sqrt{N/(s^2 \log d)}\right)$, *then we have with probability* $1 - c/d$,

$$\left\|\mathcal{T}_v\left(\overline{\beta}^d\right) - \mathcal{T}_v\left(\hat{\beta}^d\right)\right\|_2 \le C\frac{s^{3/2}k \log d}{N}, \left\|\mathcal{T}_v\left(\hat{\beta}^d\right) - \beta^*\right\|_\infty \le C\sqrt{\frac{\log d}{N}} \text{ and } \left\|\mathcal{T}_v\left(\hat{\beta}^d\right) - \beta^*\right\|_2 \le C\sqrt{\frac{s \log d}{N}}.$$

5 | RELATED WORKS AND OPEN QUESTIONS

There is a rich literature on distributed statistical inference in both the regression setting and the M-estimation setting. Major related works are reviewed in Section 2. In this section, we briefly mention other works and some open questions in the field of aggregated and distributed inference.

Nowak (2003) proposes to use a distributed expectation–maximization (EM) algorithm for density estimation and clustering in sensor networks. Although the studied problem is different from aggregated and distributed statistical inference, it provides an inspiring historic perspective: distributed inference has been studied more than 15 years ago.

Y. Zhang, Duchi, Jordan, and Wainwright (2013) study the combination of the distributed statistical inference and information theory in communication, where the results rely on either the special uniform location family or the Gaussian location families. It will be interesting to see whether or not more general results are feasible.

Wang, Peng, and Dunson (2014) develop a distributed variable selection algorithm, where a variable is accepted if more than half of the machines choose that variable. They derive the upper bounds for the success probability and the MSE of the estimator.

Song and Liang (2015) propose the split-and-merge Bayesian approach for variable selection for linear models, where they split the ultrahigh dimensional dataset into a number of lower dimensional subsets. Variable selection is performed in each of the subsets, which are then aggregated as a set of possible variables. The relevant variables are finally selected from the aggregated dataset. The consistency result is proved under mild conditions.

Arjevani and Shamir (2015) study the fundamental limits to communication-efficient distributed methods for convex learning and optimization, under different assumptions on the information available to individual machines, and the types of functions considered. The current problem formulation is more for numerical algorithms than for statistical properties analysis. Their idea may lead to interesting counterparts in statistical inference.

Zhao, Cheng, and Liu (2016) consider a partially linear framework for massive heterogeneous data and propose an aggregation type estimator for the commonality parameter that possesses the minimax optimal bound and asymptotic distribution when the number of subpopulations does not grow too fast.

More details on communication-efficient algorithms for statistical estimation and statistical inference, which have been well studied or discussed, can also be found in the references of this review. However, aggregated and distributed inference is underdeveloped in other learning problems, beyond regression. We review a few possible directions of research along this line, and describe some potential future works.

Considering the well developed body of theory for bounding and/or computing the minimax risk for various statistical estimation problems, for example, see Yang and Barron (1999) and references therein, deriving the optimal minimax rate for estimators under the distributed inference setting will be an interesting future research direction.

Besides estimation, other distributed statistical technique may be of interests, such as the distributed principal component analysis (Balcan, Kanchanapally, Liang, & Woodruff, 2014), consensus-based distributed SVMs (Forero, Cano, & Giannakis, 2010), which utilizes ADMM (Boyd et al., 2011), and so on. Distributed version of topics like nonnegative matrix factorization, as a data analysis technique, high-dimensional structured nonparametric model, which is the sparse additive model (Fan, Feng, & Song, 2011; Ravikumar, Lafferty, Liu, & Wasserman, 2009), are also of interest. Distributed inference on mixture model for Big Data is also an important topic. A stylized feature of Big Data is that they are often comprised of many heterogeneous subgroups (Fan, Han, & Liu, 2014) and are often modeled by a mixture model (Fan et al., 2014; Städler, Bühlmann, & Van De Geer, 2010). This involves optimizing a nonconvex objective function and requires intensive computation, which makes it very compelling for the development of distributed inference. Although various distributed algorithms have been introduced in most of the aforementioned scenarios, there is a lack of work focus on theoretical guarantees of the above topics. Extending the theoretical analysis from regression with low-dimensional (i.e., sparse) underlying structure to the above cases is a compelling research topic. Note that in the central estimation setting, most of the above approaches have corresponding supporting statistical theory. For example, it is known that the centralized SVMs achieve statistical consistency (in the sense of approaching to the Bayes classifier) at an optimal rate, if the underlying distribution satisfies certain properties (T. Zhang, 2004). These potential research topics demonstrate the broadness of aggregated and distributed inference problems.

6 | CONCLUSION

Aggregated inference (or distributed estimation) deals with a class of problems where the data are not available at a central location. This article reviews the major results in both the classical M-estimator setting and the high-dimensional sparse estimation setting. Many existing work focuses on the averaging estimator, for example, Y. Zhang et al. (2012) together with many others, in the aforementioned settings. The averaging debiased estimator studied in Battey et al. (2015) is particularly useful as a remedy of the bias introduced through solving the regularized loss functions, in both the estimation problem and the hypothesis testing problem. The one-step estimation in Huang and Huo (2015) suggests another approach to enhance a simple-averaging-based distributed estimator. References are listed for further reading. Prospective topics, such as the inference problem in distributed classification, are raised for future research.

ACKNOWLEDGMENTS

This project is partially supported by the Transdisciplinary Research Institute for Advancing Data Science (TRIAD), http://triad.gatech.edu, which is a part of the TRIPODS program at NSF and locates at Georgia Tech, enabled by the NSF grant CCF-1740776. Both authors are also partially supported by the NSF grant DMS-1613152.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

RELATED WIREs ARTICLES

Generalized linear models
Multivariate methods

REFERENCES

Arjevani, Y., & Shamir, O. (2015). Communication complexity of distributed convex learning and optimization. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28* (pp. 1756–1764). Red Hook, NY: Curran Associates, Inc.

Balcan, M.-F., Kanchanapally, V., Liang, Y., & Woodruff, D. (2014, August). *Improved distributed principal component analysis.* Technical Report. ArXiv. Retrieved from http://arxiv.org/abs/1408.5823

Battey, H., Fan, J., Liu, H., Lu, J., & Zhu, Z. (2015). Distributed estimation and inference with statistical guarantees. *arXiv preprint arXiv:1509.05457*.

Bickel, P. J. (1975). One-step huber estimates in the linear model. *Journal of the American Statistical Association*, *70*(350), 428–434.

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, *3*(1), 1–122.

Chen, X., & Xie, M.-g. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, *24*(4), 1655–1684.

Corbett, J. C., Dean, J., Epstein, M., Fikes, A., Frost, C., Furman, J. J., … others. (2013). Spanner: Google's globally distributed database. *ACM Transactions on Computer Systems (TOCS)*, *31*(3), 8.

El Gamal, M., & Lai, L. (2015). Are slepian-wolf rates necessary for distributed parameter estimation? In *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference* (pp. 1249–1255).

Fan, J., & Chen, J. (1999). One-step local quasi-likelihood estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(4), 927–943.

Fan, J., Feng, Y., & Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, *106*(494), 544–557.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review*, *1*, 293–314.

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

Forero, P. A., Cano, A., & Giannakis, G. B. (2010). Consensus-based distributed support vector machines. *The Journal of Machine Learning Research*, *11*, 1663–1707.

Gabay, D., & Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, *2*(1), 17–40.

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

Jaggi, M., Smith, V., Takác, M., Terhorst, J., Krishnan, S., Hofmann, T., & Jordan, M. I. (2014). Communication-efficient distributed dual coordinate ascent. In *Advances in Neural Information Processing Systems* (pp. 3068–3076).

Javanmard, A., & 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., & Yang, Y. (2018). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2018.1429274

Lee, J. D., Sun, Y., Liu, Q., & Taylor, J. E. (2015). Communication-efficient sparse regression: a one-shot approach. *arXiv preprint arXiv:1503.04337*.

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

Mitra, S., Agrawal, M., Yadav, A., Carlsson, N., Eager, D., & Mahanti, A. (2011). Characterizing web-based video sharing workloads. *ACM Transactions on the Web (TWEB)*, *5*(2), 8.

Nowak, R. D. (2003). Distributed EM algorithms for density estimation and clustering in sensor networks. *IEEE Transactions on Signal Processing*, *51*(8), 2245–2253.

Ravikumar, P., Lafferty, J., Liu, H., & Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(5), 1009–1030.

Rosenblatt, J. D., & 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., & Zhang, T. (2014). Communication-efficient distributed optimization using an approximate Newton-type method. In *International Conference on Machine Learning* (pp. 1000–1008).

Song, Q., & Liang, F. (2015). A split-and-merge Bayesian variable selection approach for ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *77*(5), 947–972.

Städler, N., Bühlmann, P., & Van De Geer, S. (2010). ℓ1-penalization for mixture regression models. *TEST*, *19*(2), 209–256.

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

Van de Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, *42*(3), 1166–1202.

Van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge, UK: Cambridge University Press.

Walker, E., Hernandez, A. V., & Kattan, M. W. (2008). Meta-analysis: Its strengths and limitations. *Cleveland Clinic Journal of Medicine*, *75*(6), 431–439.

Wang, X., Peng, P., & Dunson, D. B. (2014). Median selection subset aggregation for parallel inference. In *Advances in Neural Information Processing Systems* (pp. 2195–2203).

Yang, Y., & Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, *27*(5), 1564–1599.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, *32*(1), 56–85.

Zhang, Y., Duchi, J., Jordan, M. I., & Wainwright, M. J. (2013). Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems* (pp. 2328–2336).

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

Zhao, T., Cheng, G., & Liu, H. (2016). A partially linear framework for massive heterogeneous data. *Annals of Statistics*, *44*(4), 1400–1437.

Zinkevich, M., Weimer, M., Li, L., & Smola, A. J. (2010). Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems* (pp. 2595–2603).

Zou, H., & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, *36*(4), 1509–1533.

How to cite this article: Huo X, Cao S. Aggregated inference. *WIREs Comput Stat*. 2018;e1451. https://doi.org/10.1002/wics.1451