Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Communication-Efficient Distributed Statistical Inference

Michael I. Jordan, Jason D. Lee & Yun Yang

To cite this article: Michael I. Jordan, Jason D. Lee & Yun Yang (2019) Communication-Efficient Distributed Statistical Inference, Journal of the American Statistical Association, 114:526, 668-681, DOI: 10.1080/01621459.2018.1429274

To link to this article: https://doi.org/10.1080/01621459.2018.1429274

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 114, NO. 526, 668–681: Theory and Methods
https://doi.org/10.1080/01621459.2018.1429274

Taylor & Francis
Taylor & Francis Group

Communication-Efficient Distributed Statistical Inference

Michael I. Jordan[a], Jason D. Lee[b], and Yun Yang[c]

[a]Department of Statistics, University of California Berkeley, Berkeley, CA; [b]Institute of Computational and Mathematical Engineering, Stanford University, Cupertino, CA; [c]Statistical Science, Duke University, Durham, NC

ABSTRACT

We present a *communication-efficient surrogate likelihood* (CSL) framework for solving distributed statistical inference problems. CSL provides a communication-efficient surrogate to the global likelihood that can be used for low-dimensional estimation, high-dimensional regularized estimation, and Bayesian inference. For low-dimensional estimation, CSL provably improves upon naive averaging schemes and facilitates the construction of confidence intervals. For high-dimensional regularized estimation, CSL leads to a minimax-optimal estimator with controlled communication cost. For Bayesian inference, CSL can be used to form a communication-efficient quasi-posterior distribution that converges to the true posterior. This quasi-posterior procedure significantly improves the computational efficiency of Markov chain Monte Carlo (MCMC) algorithms even in a nondistributed setting. We present both theoretical analysis and experiments to explore the properties of the CSL approximation. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received December 2016
Revised December 2017

KEYWORDS
Communication efficiency;
Distributed inference;
Likelihood approximation

1. Introduction

What is the relevance of the underlying computational architecture to statistical inference? Classically, the answer has been "not much"—the naive abstraction of a sequential program running on a single machine and providing instantaneous access to arbitrary data points has provided a standard conceptual starting point for statistical computing. In the modern era, however, it is commonplace for data analyses to run on hundreds or thousands of machines, with the data distributed across those machines and no longer available in a single central location. Moreover, the end of Moore's law has changed computer science—the focus is now on parallel, distributed architectures and, on the algorithmic front, on divide-and-conquer procedures. This has serious implications for statistical inference. Naively dividing datasets into subsets that are processed separately, with a naive merging of the results, can yield inference procedures that are highly biased and highly variable. Naive application of traditional statistical methodology can yield procedures that incur exorbitant communication costs.

Historically, the area in which statisticians have engaged most deeply with practical computing concerns has been in the numerical linear algebra needed to support regression and multivariate statistics, including both sparse and dense matrix algorithms. It is thus noteworthy that over the past decade there has been a revolution in numerical linear algebra in which new "communication-avoiding" algorithms have been developed to replace classical matrix routines (Demmel et al. 2012). The new algorithms can run significantly faster than classical algorithms on parallel, distributed architectures.

A statistical literature on parallel and distributed inference has begun to emerge, both in a frequentist setting (Duchi, Agarwal, and Wainwright 2012; Zhang, Duchi, and Wainwright 2013; Kannan, Vempala, and Woodruff 2014; Kleiner et al. 2014; Shamir, Srebro, and Zhang 2014; Mackey, Talwalkar, and Jordan 2015; Zhang and Lin 2015a; Lee et al. 2015a), and a Bayesian setting (Suchard et al. 2010; Cleveland and Hafen 2014; Maclaurin and Adams 2014; Wang and Dunson 2015; Neiswanger, Wang, and Xing 2015; Rabinovich, Angelino, and Jordan 2016; Scott et al. 2016; Terenin, Simpson, and Draper 2016). This literature has focused on data-parallel procedures in which the overall dataset is broken into subsets that are processed independently. Much of the current focus has been on "one-shot" or "embarrassingly parallel" approaches which only use one round of communication in which estimators or posterior samples are obtained in parallel on local machines, are communicated to a central node, and then combined to form a global estimator or approximation to the posterior distribution (Zhang, Duchi, and Wainwright 2013; Lee et al. 2015a; Wang and Dunson 2015; Neiswanger, Wang, and Xing 2015). In the frequentist setting, most one-shot approaches rely on averaging (Zhang, Duchi, and Wainwright 2013), where the global estimator is the average of the local estimators. Lee et al. (2015a) extended this idea to high-dimensional sparse linear regression by combining local debiased Lasso estimates (van de Geer et al. 2014). Recent work by Duchi et al. (2015) shows that under certain conditions, these averaging estimators can attain the information-theoretic complexity lower bound—for linear regression, at least $\mathcal{O}(dk)$ bits

CONTACT Yun Yang yy84@stat.duke.edu Statistical Science, Duke University, Box 90251, Durham, NC 27708.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

must be communicated to attain the minimax rate of parameter estimation, where d is the dimension of the parameter and k is the number of machines. This holds even in the sparse setting (Braverman et al. 2016).

These averaging-based, one-shot communication approaches suffer from several drawbacks. First, they have generally been limited to point estimation; it is not straightforward to create confidence intervals/regions and hypothesis tests based on the averaging estimator. Second, in order for the averaging estimator to achieve the minimax rate of convergence, each local machine must have access to at least $\Omega(\sqrt{N})$ samples, where N is the total sample size. In other words, the number of machines should be much smaller than $\sqrt{N}$; a highly restrictive assumption. Third, when the estimator is nonlinear, averaging can perform poorly, for example, our empirical study shows that even for small k, of order 10^1, the averaging estimator only exhibits a slight improvement over purely local estimators.

In the Bayesian setting, embarrassingly parallel approaches run Markov chain Monte Carlo (MCMC) algorithms in parallel across local machines and transmit the local posterior samples to a central node to produce an overall approximation to the global posterior distribution. Unfortunately, when the dimension d is high, the number of samples obtained locally must be large due to the curse of dimensionality, incurring significant communication costs. Also, when combining local posterior samples in the central node, existing approaches that approximate the global posterior distribution by a weighted empirical distribution of "averaging draws" (Wang and Dunson 2015; Neiswanger, Wang, and Xing 2015) tend to suffer from the weight-degeneracy issue (weights collapse to only a few samples) when k is large.

In this article, we formulate a unified framework for distributed statistical inference. We refer to our framework as the *communication-efficient surrogate likelihood* (CSL) framework. From the frequentist perspective, CSL provides a communication-efficient surrogate to the global likelihood function that can play the role of the global likelihood function in forming the maximum likelihood estimator (MLE) in regular parametric models or the penalized MLE in high-dimensional models. From a Bayesian perspective, CSL can be used to form a quasi-posterior distribution (Chernozhukov and Hong 2003) as a surrogate for the full posterior. The CSL approximation can be constructed efficiently by communicating $\mathcal{O}(dk)$ bits. After its construction, CSL can be efficiently evaluated by using the n samples in a single local machine. Even in a nondistributed Bayesian setting, CSL can be used as a computationally efficient surrogate to the likelihood function by predividing the dataset into k subsamples—the computational complexity of one iteration of MCMC is then reduced by a factor of k.

Our CSL-based distributed inference approach overcomes the aforementioned drawbacks associated with the one-shot and embarrassingly parallel approaches. In the frequentist framework, a CSL-based estimator can achieve the same rate of convergence as the global likelihood-based estimator while incurring a communication complexity of only $\mathcal{O}(dk)$. Moreover, the CSL framework can readily be applied iteratively, with the resulting multi-round algorithm achieving a geometric convergence rate with contraction factor $\mathcal{O}(n^{-1/2})$, where n is the number of samples in each local machine. This $\mathcal{O}(n^{-1/2})$ rate of convergence significantly improves on analyses based on condition-number contraction factors used to analyze methods that form the global gradient in each iteration by combining local gradients. As an implication, to achieve the same accuracy as the global likelihood-based estimator, we require $\mathcal{O}(\frac{\log N}{\log n})$ iterations, which is at most logarithmic in N and *constant* for $k \lesssim \text{poly}(n)$. In contrast, the averaging estimator requires $k \ll n$. Thus, due to the fast $\mathcal{O}(n^{-1/2})$ rate, usually two-three iterations suffice for our procedure to match the same accuracy of the global likelihood-based estimator even for relatively large k (see Section 4.1 for more details). Unlike bootstrap-based approaches (Zhang, Duchi, and Wainwright 2013) for boosting accuracy, the additional complexity of the iterative version of our approach grows only linearly in the number of iterations. Finally, our empirical study suggests that a CSL-based estimator may exhibit significant improvement over the averaging estimator for nonlinear distributed statistical inference problems.

For high-dimensional ℓ_1-regularized estimation, the CSL framework yields an algorithm that communicates $O(dk)$ bits, attaining the optimal communication/risk trade off (Garg, Ma, and Nguyen 2014). This improves over the averaging method of Lee et al. (2015a) because it requires d times less computation, and allows for iterative refinement to obtain arbitrarily low optimization error in a logarithmic number of rounds. (We note that during the preparation of this manuscript, we became aware of concurrent work of Wang et al. (2017) who also study a communication-efficient surrogate likelihood. Their focus is solely on the high-dimensional linear model setting, and includes the same results as Section 3.2, but not the results of Sections 3.1 and 3.3.) In the Bayesian framework, our method does not require transmitting local posterior samples and is thus free from the weight degeneracy issue. This makes the communication complexity of our approach considerably lower than competing embarrassingly parallel Bayesian computation approaches.

The remainder of this article is organized as follows. In Section 2, we informally present the motivation for CSL. Section 3 presents algorithms and theory for three different problems: parameter estimation in low-dimensional regular parametric models (Section 3.1), regularized parameter estimation in the high-dimensional problems (Section 3.2), and Bayesian inference in regular parametric models (Section 3.3). Section 4 presents experimental results in these three settings. Due to the space constraint, all proofs, some discussion about the literature, additional simulations, a real data application, and a discussion section are deferred to the Appendix.

2. Background and Problem Formulation

We begin by setting up our general framework for distributed statistical inference. We then turn to a description of the CSL methodology, demonstrating its application to both frequentist and Bayesian inference.

2.1. Statistical Models With Distributed Data

Let $Z_1^N := \{Z_{ij} : i = 1, \ldots, n,\ j = 1, \ldots, k\}$ denote $N = nk$ identically distributed observations with marginal distribution $\mathbb{P}_{\theta^*}$, where $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is a family of statistical models parameterized by $\theta \in \Theta \subset \mathbb{R}^d$, Θ is the parameter space and θ^* is the

true data-generating parameter. Suppose that the data points are stored in a distributed manner in which each machine stores a subsample of n observations. Let $Z_j := \{Z_{ij} : i = 1, \ldots, n\}$ denote the subsample that is stored in the jth machine $\mathcal{M}_j$, for $j = 1, \ldots, k$. Our goal is to conduct statistical inference on the parameter θ while taking into consideration the communication cost among the machines. For example, we may want to find a point estimator $\widehat{\theta}$ and an associated confidence interval (region).

Let $\mathcal{L} : \Theta \times \mathcal{Z} \mapsto \mathbb{R}$ be a twice-differentiable loss function such that the true parameter is a minimizer of the population risk $\mathcal{L}^*(\theta) := \mathbb{E}_{\theta^*}[\mathcal{L}(\theta;\ Z)]$, that is,

$$\theta^* \in \arg\min_{\theta \in \Theta} \mathbb{E}_{\theta^*}[\mathcal{L}(\theta;\ Z)]. \tag{1}$$

Define the local and global loss functions as

$$\mathcal{L}_j(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\theta; z_{ij}), \quad \text{for } j \in [k], \tag{2}$$

$$\mathcal{L}_N(\theta) = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{k} \mathcal{L}(\theta; z_{ij}) = \frac{1}{k} \sum_{j=1}^{k} \mathcal{L}_j(\theta). \tag{3}$$

Here, $\mathcal{L}_j(\theta)$ is the loss function evaluated at θ by using the local data stored in machine $\mathcal{M}_j$. The negative log-likelihood function is a standard example of the loss function $\mathcal{L}$.

2.2. Distributed Statistical Inference

In this subsection, we motivate the CSL methodology by constructing a surrogate loss $\widetilde{\mathcal{L}} : \Theta \times \mathcal{Z} \mapsto \mathbb{R}$ that approximates the global loss function $\mathcal{L}_N$ in a communication-efficient manner. We show that it can be constructed in any local machine $\mathcal{M}_j$ by communicating at most $(k-1)$ d-dim vectors. After the construction, $\widetilde{\mathcal{L}}$ can be used to replace the global loss function in various statistical inference procedures by only using the data in a local machine (see Sections 3.1–3.3). We aim to show that this distributed inference framework can simultaneously achieve high statistical accuracy and low communication cost. In this section, we motivate our construction using heuristic arguments; a rigorous analysis is provided in Section 3 to follow.

Our motivation starts from the Taylor series expansion of $\mathcal{L}_N$. Viewing $\mathcal{L}_N(\theta)$ as an analytic function, we expand it into an infinite series:

$$\mathcal{L}_N(\theta) = \mathcal{L}_N(\overline{\theta}) + \langle \nabla \mathcal{L}_N(\overline{\theta}), \theta - \overline{\theta} \rangle + \sum_{j=2}^{\infty} \frac{1}{j!} \nabla^j \mathcal{L}_N(\overline{\theta})\, (\theta - \overline{\theta})^{\otimes j}. \tag{4}$$

Here, $\overline{\theta}$ is any initial estimator of θ, for example, the local empirical loss minimizer $\arg\min_\theta \mathcal{L}_1(\theta)$ in the first machine $\mathcal{M}_1$. Because the data are split across machines, evaluating the derivatives $\nabla^j \mathcal{L}_N(\overline{\theta})(j \geq 1)$ requires one communication round. However, unlike the d-dim gradient vector $\nabla \mathcal{L}_N(\overline{\theta})$, the higher-order derivatives require communicating more than $O(d^2)$ bits from each machine. This reasoning motivates us to replace the global higher-order derivatives $\nabla^j \mathcal{L}_N(\overline{\theta})$ $(j \geq 2)$ with local derivatives, leading to the following approximation of $\mathcal{L}_N$:

$$\widetilde{\mathcal{L}}(\theta) = \mathcal{L}_N(\overline{\theta}) + \langle \nabla \mathcal{L}_N(\overline{\theta}), \theta - \overline{\theta} \rangle + \sum_{j=2}^{\infty} \frac{1}{j!} \nabla^j \mathcal{L}_1(\overline{\theta})\, (\theta - \overline{\theta})^{\otimes j}. \tag{5}$$

Comparing expressions (4) and (5), we see that the approximation error is

$$\begin{aligned}
\widetilde{\mathcal{L}}(\theta) - \mathcal{L}_N(\theta) &= \mathcal{L}_N(\overline{\theta}) + \langle \nabla \mathcal{L}_N(\overline{\theta}), \theta - \overline{\theta} \rangle + \sum_{j=2}^{\infty} \frac{1}{j!} \nabla^j \mathcal{L}_1(\overline{\theta})\, (\theta - \overline{\theta})^{\otimes j} \\
&\quad - \Big(\mathcal{L}_N(\overline{\theta}) + \langle \nabla \mathcal{L}_N(\overline{\theta}), \theta - \overline{\theta} \rangle + \sum_{j=2}^{\infty} \frac{1}{j!} \nabla^j \mathcal{L}_N(\overline{\theta})\, (\theta - \overline{\theta})^{\otimes j} \Big) \\
&= \frac{1}{2} \left\langle \theta - \overline{\theta},\ \left(\nabla^2 \mathcal{L}_1(\overline{\theta}) - \nabla^2 \mathcal{L}_N(\overline{\theta}) \right) (\theta - \overline{\theta}) \right\rangle + O\left(\|\theta - \overline{\theta}\|_2^3 \right) \\
&= O\left(\frac{1}{\sqrt{n}} \|\theta - \overline{\theta}\|_2^2 + \|\theta - \overline{\theta}\|_2^3 \right),
\end{aligned} \tag{6}$$

where the fact that $\left\| \nabla^2 \mathcal{L}_N(\overline{\theta}) - \nabla^2 \mathcal{L}_1(\overline{\theta}) \right\|_2 = O\left(n^{-1/2}\right)$ is a consequence of matrix concentration.

We now use a Taylor expansion of $\mathcal{L}_1(\theta)$ around $\overline{\theta}$ to replace the infinite sum of high-order derivatives in expression (5) with $\mathcal{L}_1(\theta) - \mathcal{L}_1(\overline{\theta}) - \langle \nabla \mathcal{L}_1(\overline{\theta}), \theta - \overline{\theta} \rangle$. This yields

$$\widetilde{\mathcal{L}}(\theta) = \mathcal{L}_N(\overline{\theta}) + \langle \nabla \mathcal{L}_N(\overline{\theta}), \theta - \overline{\theta} \rangle + \mathcal{L}_1(\theta) - \mathcal{L}_1(\overline{\theta}) - \langle \nabla \mathcal{L}_1(\overline{\theta}), \theta - \overline{\theta} \rangle. \tag{7}$$

Finally, we omit the additive constants in (7) and redefine $\widetilde{\mathcal{L}}(\theta)$ as follows:

$$\widetilde{\mathcal{L}}(\theta) := \mathcal{L}_1(\theta) - \left\langle \theta,\ \nabla \mathcal{L}_1(\overline{\theta}) - \nabla \mathcal{L}_N(\overline{\theta}) \right\rangle. \tag{8}$$

Henceforth, we refer to this expression for $\widetilde{\mathcal{L}}(\theta)$ as a *surrogate loss function*. In the remainder of the section, we present three examples of using this surrogate loss function for frequentist and Bayesian inference. A rigorous justification for $\widetilde{\mathcal{L}}(\theta)$, which effectively provides conditions under which the terms in (6) are small, follows in Section 3.

Example (M-estimator). In the low-dimensional regime where the dimensionality d of parameter space is $o(N)$, the global empirical loss minimizer,

$$\widehat{\theta} := \arg\min_{\theta \in \Theta} \mathcal{L}_N(\theta),$$

achieves a root-N rate of convergence under mild conditions. One may construct confidence regions associated with $\widehat{\theta}$ using the sandwiched covariance matrix (see, e.g., (11)). In our distributed inference framework, we aim to capture some of the desirable properties of $\widehat{\theta}$ by replacing the global loss function $\mathcal{L}_N(\theta)$ with the surrogate loss function $\widetilde{\mathcal{L}}$ and defining the following communication-efficient estimator:

$$\widetilde{\theta} = \arg\min_{\theta \in \Theta} \widetilde{\mathcal{L}}(\theta).$$

Indeed, in Section 3.1, we show that $\widetilde{\theta}$ is equivalent to $\widehat{\theta}$ up to higher-order terms, and we provide two ways to construct confidence regions for $\widetilde{\theta}$ using local observations stored in machine $\mathcal{M}_1$.

Example (High-dimensional regularized estimator). In the high-dimensional regime, where the dimensionality d can be much larger than the sample size N, we need to impose a low-dimensional structural assumption, such as sparsity, on the unknown true parameter θ^*. For concreteness, we focus on a sparsity assumption—that most components of the d-dim vector θ^* is zero—and consider the ℓ_1-regularized surrogate loss,

$$\widetilde{\theta} := \arg\min_{\theta\in\Theta} \left\{\widetilde{\mathcal{L}}_N(\theta) + \lambda\|\theta\|_1\right\}.$$

In Section 3.2, we show that $\widetilde{\theta}$ achieves the same rate of convergence as the benchmark lasso estimator $\widehat{\theta} = \arg\min_{\theta\in\Theta}\left\{\mathcal{L}_N(\theta) + \lambda\|\theta\|_1\right\}$ under a set of mild conditions. This idea of using the surrogate loss function to approximate the global regularized loss function is general and is applicable to other high-dimensional problems.

Example (Bayesian inference). In the Bayesian framework, viewing parameter θ as random, we place a prior distribution π over parameter space Θ and use the posterior distribution to conduct statistical inference. For convenience, we use the same notation $\pi(\theta)$ to denote the prior pdf at point θ. According to Bayes' theorem, the posterior pdf satisfies

$$\pi\left(\theta \mid Z_1^N\right) \propto \exp\{-N\mathcal{L}_N(\theta)\}\,\pi(\theta).$$

By viewing the loss function $\mathcal{L}$ as the negative log-likelihood function and approximate $\mathcal{L}_N$ by $\widetilde{\mathcal{L}}_N$, we obtain the following surrogate posterior distribution,

$$\widetilde{\pi}_N\left(\theta \mid Z_1^N\right) \propto \exp\{-N\widetilde{\mathcal{L}}(\theta)\}\,\pi(\theta),$$

for approximating the global posterior $\pi(\theta \mid Z_1^N)$. In Section 3.3, we formalize this argument and show that this surrogate posterior provides a good approximation to the global posterior.

We now give a heuristic argument for why $\widetilde{\theta}$ is a good estimator in the first example of M-estimator. A similar argument also applies to the other two examples. For convenience, we assume that the empirical risk function $\mathcal{L}_N(\theta)$ has a unique minimizer. First, consider the global empirical loss minimizer $\widehat{\theta}$. Under our assumption that the loss function is twice-differentiable, $\widehat{\theta}$ is the unique solution of equation (There is no Taylor's theorem for vector-valued functions, but we formalize this heuristic in Section 3.1.)

$$0 = \nabla\mathcal{L}_N(\widehat{\theta}) \approx \nabla\mathcal{L}_N(\theta^*) + \nabla^2\mathcal{L}_N(\theta^*)\,(\widehat{\theta} - \theta^*).$$

By solving this equation, we obtain $\|\widehat{\theta} - \theta^*\|_2 = O_p(\|\nabla\mathcal{L}_N(\theta^*)\|_2) = O_p(N^{-1/2})$, as long as the Hessian matrix $\nabla^2\mathcal{L}_N(\theta^*)$ is nonsingular. Now let us turn to the surrogate loss minimizer $\widetilde{\theta}$. An analogous argument leads to $\|\widetilde{\theta} - \theta^*\|_2 = O_p(\|\nabla\widetilde{\mathcal{L}}(\theta^*)\|_2)$ and we only need to show that $\|\nabla\widetilde{\mathcal{L}}(\theta^*)\|_2$ is of order $O_p(N^{-1/2})$. In fact, by our construction,

$$\begin{aligned}\nabla\widetilde{\mathcal{L}}(\theta^*) &= \left(\nabla\mathcal{L}_1(\theta^*) - \nabla\mathcal{L}_1(\overline{\theta})\right)\\ &\quad -\left(\nabla\mathcal{L}_N(\theta^*) - \nabla\mathcal{L}_N(\overline{\theta})\right) + \nabla\mathcal{L}_N(\theta^*)\\ &\approx \langle\nabla^2\mathcal{L}_1(\theta^*) - \nabla^2\mathcal{L}_N(\theta^*),\ \theta^* - \overline{\theta}\rangle + O_p(N^{-1/2})\\ &= O_p\left(n^{-1/2}\,\|\theta^* - \overline{\theta}\|_2\right) + O_p(N^{-1/2}),\end{aligned}$$

which is of order $O_p(N^{-1/2})$ as long as $\|\theta^* - \overline{\theta}\|_2 = O_p\left(k^{-1/2}\right)$ where $k = N/n$ is the number of machines. For example, this requirement on initial estimator $\overline{\theta}$ is satisfied by the minimizer $\widehat{\theta}_1$ of the subsample loss function $\mathcal{L}_1(\theta)$ when $n > k$.

From now on, we will refer the methodology of using the surrogate loss function $\widetilde{\mathcal{L}}(\cdot)$ to approximate the global loss function $\mathcal{L}(\cdot)$ for distributed statistical inference as a communication-efficient surrogate likelihood (CSL) method. Although our focus is on distributed inference, we also wish to note that the idea of computing the global likelihood function using subsamples may be useful not only in the distributed inference framework, but also in a single-machine setting in which the sample size is so large that the evaluation of the likelihood function or its gradient is unduly expensive. Using our surrogate loss function $\widetilde{\mathcal{L}}(\theta)$, we only need one pass over the entire dataset to construct $\widetilde{\mathcal{L}}(\theta)$. After its construction, $\widetilde{\mathcal{L}}(\theta)$ can be efficiently evaluated by using a small subset of the data.

At the end of this section, we provide some remarks regarding the CSL method. First, the surrogate loss function $\widetilde{\mathcal{L}}(\theta)$ is constructed in a prespecified machine and is asymmetric in the use of local data across the machines. In principal, one could conduct the same procedure in each machine and average. According to our numerical experiments, this averaging procedure improves the statistical accuracy by a slight amount. However, this procedure increases the communication cost at least from $2kd$ (communicating the initial value $\overline{\theta}$ and k local d-dimensional gradient vectors across the k machines) to $4kd$ (by first communicating $\overline{\theta}$ and sending k local d-dimensional gradient vectors to one machine to compute the global gradient; then sending back the global gradient vector to local machines to find local estimators; and finally sending all local estimators to one machine to compute the average). With the same communication cost $4kd$, we may instead apply one more iteration in the iterative local estimation algorithm described in Section 3.1 that reduces the numerical error of approximating $\widehat{\theta}$ by a factor of $n^{-1/2}$. Second, by viewing the machine in which the surrogate loss function is constructed as the fusion center that is allowed to access to n data points, previous communication lower bounds obtained in Lee et al. (2017), Zhang and Lin (2015b), and Shamir, Srebro, and Zhang (2014) still apply, and imply the optimality of the CSL in terms of the trade-off between communication cost and statistical accuracy. Our last remark regards the homogeneity of the data over the machines. Due to the construction of $\widetilde{\mathcal{L}}$, we only require the data points in the central machine (the machine for constructing the surrogate loss function) to be uniformly drawn from the entire dataset. Data on all the other machines can be split in any arbitrary way.

3. Main Results and Their Consequences

In this section, we delve into the three examples in Section 2.2 of applying the CSL method. For each of the examples, we provide an explicit bound on either the estimation error $\|\widetilde{\theta} - \theta^*\|_2$ of the resulting estimator $\widetilde{\theta}$ or the approximation error $\|\widetilde{\pi}_N - \pi_N\|_1$ of the approximated posterior $\widetilde{\pi}_N(\cdot)$.

3.1. Communication-Efficient M-Estimators in Low Dimensions

In this subsection, we consider a low-dimensional parametric family $\{\mathbb{P}_\theta : \theta \in \Theta\}$, where the dimensionality d of θ is much

smaller than the sample size n. Under this setting, the minimizer of the population risk in optimization problem (1) is unique under the set of regularity conditions to follow and θ^* is identifiable. As a concrete example, we may consider the negative log-likelihood function $\ell(\theta; z) = -\log p(z; \theta)$ as the loss function, where $p(\cdot; \theta)$ is the pdf for $\mathbb{P}_\theta$. Note that the developments in this subsection can also be extended to misspecified families where the marginal distribution $\mathbb{P}$ of the observations is not contained in the model space $\{\mathbb{P}_\theta : \theta \in \Theta\}$. Under misspecification, we can view the parameter θ^* associated with the projection $\mathbb{P}_{\theta^*}$ of the true data-generating model $\mathbb{P}$ onto the misspecified model space, $\{\mathbb{P}_\theta : \theta \in \Theta\}$, as the "true parameter." The results under misspecification are similar to the well-specified case and are omitted in this article.

For low-dimensional parametric models, we impose some regularity conditions on the parameter space, the loss function $\mathcal{L}$, and the associated population risk function $\mathcal{L}^*$. These conditions are standard in classical statistical analysis of M-estimators. In the rest of the article, we call a parametric model that satisfies this set of regularity conditions a regular parametric model. Our first assumption describes the relationship of the parameter space Θ and the true parameter θ^*.

Assumption PA (Parameter space). The parameter space Θ is a compact and convex subset of $\mathbb{R}^d$. Moreover, $\theta^* \in \text{int}(\Theta)$ and $R := \sup_{\theta\in\Theta} \|\theta - \theta^*\|_2 > 0$.

The second assumption is a local identifiability condition, ensuring that θ^* is a local minimum of $\mathcal{L}^*$.

Assumption PB (Local convexity). The Hessian matrix $I(\theta) = \nabla^2\mathcal{L}^*(\theta)$ of the population risk function $\mathcal{L}^*(\theta)$ is invertible at θ^*: there exist two positive constants (μ_-, μ_+), such that $\mu_- I_d \preceq \nabla^2\mathcal{L}^*(\theta^*) \preceq \mu_+ I_d$.

When the loss function is the negative log-likelihood function, the corresponding Hessian matrix is an information matrix.

Our next assumption is a global identifiability condition, which is a standard condition for proving estimation consistency.

Assumption PC (Identifiability). For any $\delta > 0$, there exists $\epsilon > 0$, such that

$$\liminf_{n\to\infty} \mathbb{P}\Big\{ \inf_{\|\theta-\theta^*\|_2\geq\delta} (\mathcal{L}(\theta) - \mathcal{L}(\theta^*)) \geq \epsilon \Big\} = 1.$$

Our final assumption controls moments of higher-order derivatives of the loss function, and allows us to obtain high-probability bounds on the estimation error. Let $U(\rho) = \{\theta \in \mathbb{R}^d \mid \|\theta - \theta^*\|_2 \leq \rho\} \subset \Theta$ be a ball around the truth θ^* with radius $\rho > 0$.

Assumption PD (Smoothness). There exist constants (G, L) and a function $M(z)$ such that

$$\begin{aligned}\mathbb{E}\big[\|\nabla\mathcal{L}(\theta; Z)\|_2^{16}\big] \leq G^{16}, \mathbb{E}\big[\|\nabla^2\mathcal{L}(\theta; Z) - I(\theta)\|_2^{16}\big] \\ \leq L^{16}, \quad \text{for all } \theta \in U(\rho), \\ \times\|\nabla^2\mathcal{L}(\theta; z) - \nabla^2\mathcal{L}(\theta'; z)\|_2 \leq M(z) \\ \times\|\theta - \theta'\|_2, \quad \text{for all } \theta, \theta' \in U(\rho).\end{aligned}$$

Moreover, the function $M(z)$ satisfies $\mathbb{E}[M^{16}(Z)] \leq M^{16}$ for some constant $M > 0$.

Based on the heuristic argument in Section 2.2, we propose to use the surrogate function $\widetilde{\mathcal{L}}$ defined in (8) as the objective function for constructing an M-estimator in regular parametric models. Our first result shows that under Assumptions PA-PD, given any reasonably good initial estimator $\overline{\theta}$, any minimizer $\widetilde{\theta}$ of $\widetilde{\mathcal{L}}(\theta)$, that is,

$$\widetilde{\theta} \in \arg\min_{\theta\in\Theta} \widetilde{\mathcal{L}}(\theta), \tag{9}$$

significantly boosts the accuracy in terms of the approximation error $\|\widetilde{\theta} - \widehat{\theta}\|_2$ to the global empirical risk minimizer $\widehat{\theta} = \arg\min_{\theta\in\Theta} \mathcal{L}_N(\theta)$.

Theorem 1. Suppose that Assumptions PA-PD hold and the initial estimator $\overline{\theta}$ lies in the neighborhood $U(\rho)$ of θ^*. Then any minimizer $\widetilde{\theta}$ of the surrogate loss function $\widetilde{\mathcal{L}}(\theta)$ satisfies

$$\begin{aligned}\|\widetilde{\theta} - \widehat{\theta}\|_2 \leq C_2 \big(\|\overline{\theta} - \widehat{\theta}\|_2 + \|\widehat{\theta} - \theta^*\|_2 \\ + \|\nabla^2\mathcal{L}_1(\theta^*) - \nabla^2\mathcal{L}_N(\theta^*)\|_2\big) \|\overline{\theta} - \widehat{\theta}\|_2,\end{aligned} \tag{10}$$

with probability at least $1 - C_1 kn^{-8}$, where the constants C_1 and C_2 are independent of (k, n, N).

Under the conditions of Theorem 1, it can be shown that $\|\widehat{\theta} - \theta^*\|_2 = O_p(N^{-1/2})$ and $\|\nabla^2\mathcal{L}_1(\theta^*) - \nabla^2\mathcal{L}_N(\theta^*)\|_2 = O_p(n^{-1/2})$ (see Lemma B.1 and inequality (A.7) in Appendix B.1), and therefore

$$\begin{aligned}\|\widetilde{\theta} - \widehat{\theta}\|_2 &= \big(O_p(n^{-1/2}) + \|\overline{\theta} - \widehat{\theta}\|_2\big) \|\overline{\theta} - \widehat{\theta}\|_2 \\ &= O_p(n^{-1/2}) \|\overline{\theta} - \widehat{\theta}\|_2,\end{aligned}$$

as long as $\|\overline{\theta} - \widehat{\theta}\|_2 = O_p(n^{-1/2})$, which is true for $\overline{\theta} = \widehat{\theta}_1 := \arg\min_\theta \mathcal{L}_1(\theta)$, the empirical risk minimizer in local machine $\mathcal{M}_1$. To formalize this argument, we have the following corollary that provides an ℓ_2 risk bound for $\widetilde{\theta}$.

Corollary 2. Under the conditions of Theorem 1, we have

$$\begin{aligned}\mathbb{E}[\|\widetilde{\theta} - \theta^*\|_2^2] \leq \frac{A}{N} + \frac{C}{N\sqrt{N}} + \frac{C}{\sqrt{nN}} \\ \times \min\Big\{\frac{1}{\sqrt{n}}, \big(\mathbb{E}[\|\overline{\theta} - \widehat{\theta}\|_2^4]\big)^{1/4}\Big\} + \frac{C}{n^4}\sqrt{\frac{k}{N}},\end{aligned}$$

where $A = \mathbb{E}[\|I(\theta^*)^{-1}\nabla\mathcal{L}(\theta^*; Z)\|_2^2]$ and C is some constant independent of (n, k, N).

Note that the Hájek-Le Cam minimax theorem guarantees that for any estimator $\widehat{\theta}_N$ based on N samples, we have

$$\lim_{c\to\infty} \liminf_{N\to\infty} \sup_{\theta\in U(c/\sqrt{N})} N\,\mathbb{E}_\theta\big[\|\widehat{\theta}_N - \theta\|_2^2\big] \geq A.$$

Therefore, the estimator $\widetilde{\theta}$ is (first-order) minimax-optimal for $n = \Omega(\sqrt{n})$ and achieves the Cramér–Rao lower bound when the loss function $\mathcal{L}$ is the negative log-likelihood function.

The analysis of Theorem 1 focuses on the setting in which μ_- depends only on the data-generating distribution and is considered to be a fixed quantity. From the proof, we can also derive an explicit dependence on μ_- as $\|\widetilde{\theta} - \widehat{\theta}\|_2 \lesssim \mu_-^{-1}\big(\|\overline{\theta} - \widehat{\theta}\|_2 + (\sqrt{N}\mu_-)^{-1} + (\sqrt{n}\mu_-)^{-1}\big)\|\overline{\theta} - \widehat{\theta}\|_2 \lesssim (\sqrt{n}\mu_-^2)^{-1}\,\|\overline{\theta} - \widehat{\theta}\|_2$. Therefore, the contraction factor is of the order $(\sqrt{n}\mu_-^2)^{-1}$. However, this dependence is not sharp, at least in linear regression (i.e., for a quadratic objective). In fact, a more careful analysis (we omit this due to space constraints) leads to $\|\widetilde{\theta} - \widehat{\theta}\|_2 \leq C_{\text{abs}}\sqrt{d/n}\|\overline{\theta} - \widehat{\theta}\|_2$, which gives a contraction factor independent of μ_-, with C_{abs} depending only on the sub-Gaussian

parameter of the design matrix. We leave it as future work to generalize this analysis to the nonquadratic case to obtain sharper risk bounds in terms of μ_-.

One-step approximation. The computational complexity of exactly minimizing the surrogate loss $\widetilde{\mathcal{L}}(\theta)$ in (9) can be further reduced by using a local quadratic approximation to $\mathcal{L}$. In fact, we have by Taylor's theorem that

$$\mathcal{L}_N(\theta) \approx \mathcal{L}_N(\bar{\theta}) + \langle \nabla \mathcal{L}_N(\bar{\theta}), \theta - \bar{\theta} \rangle + \frac{1}{2} \langle \theta - \bar{\theta}, \nabla^2 \mathcal{L}_N(\bar{\theta} - \theta) \rangle.$$

As before, we replace the global Hessian $\nabla^2 \mathcal{L}_N(\bar{\theta})$ with the local Hessian $\nabla^2 \mathcal{L}_1(\bar{\theta})$, which leads to the following quadratic surrogate loss:

$$\widetilde{\mathcal{L}}^H(\theta) := \langle \nabla \mathcal{L}_N(\bar{\theta}), \theta - \bar{\theta} \rangle + \frac{1}{2} \langle \theta - \bar{\theta}, \nabla^2 \mathcal{L}_1(\bar{\theta} - \theta) \rangle.$$

Because the surrogate loss functions $\widetilde{\mathcal{L}}^H$ and $\widetilde{\mathcal{L}}$ agree up to the second-order Taylor expansion, they behave similarly when used as objective functions for constructing M-estimators. This motivates the closed-form estimator

$$\widetilde{\theta}^H := \arg\min_{\theta \in \Theta} \widetilde{\mathcal{L}}^H(\theta) = \bar{\theta} - \nabla^2 \mathcal{L}_1(\bar{\theta})^{-1} \nabla \mathcal{L}_N(\bar{\theta}).$$

The next theorem shows that $\widetilde{\theta}^H$ satisfies the same estimation bound as $\widetilde{\theta}$. Unlike the classical one-step MLE that requires the initial estimator to be within an $\mathcal{O}(N^{-1/2})$ neighborhood of the truth θ^*, we only require $\|\bar{\theta} - \theta^*\|_2$ to be $\mathcal{O}(n^{-1/2})$.

Theorem 3. Suppose that Assumptions PA-PD hold and the initial estimator $\bar{\theta}$ satisfies $\|\bar{\theta} - \theta^*\|_2 \le \min\{\rho, (16M)^{-1}(1-\rho)\mu_-\}$. Then the local one-step estimator $\widetilde{\theta}^H$ satisfies

$$\|\widetilde{\theta}^H - \widehat{\theta}\|_2 \le C_2' \big(\|\bar{\theta} - \widehat{\theta}\|_2 + \|\widehat{\theta} - \theta^*\|_2 + \|\nabla^2 \mathcal{L}_1(\theta^*) - \nabla^2 \mathcal{L}_N(\theta^*)\|_2 \big) \|\bar{\theta} - \widehat{\theta}\|_2,$$

with probability at least $1 - C_1' k n^{-8}$, where C_1' and C_2' are independent of (k, n, N).

The analog of Corollary 2 can also be stated for $\widetilde{\theta}^H$.

Iterative local estimation algorithm. Theorem 1 (Theorem 3) suggests that an iterative algorithm may reduce the approximation error $\|\widetilde{\theta} - \widehat{\theta}\|_2$ by a factor of $n^{-1/2}$ in each iteration as long as the initial estimator satisfies $\|\bar{\theta} - \widehat{\theta}\|_2 = O_p(n^{-1/2})$, or equivalently, $\|\bar{\theta} - \theta^*\|_2 = O_p(n^{-1/2})$. We refer to such an algorithm as an iterative local estimation algorithm (ILEA, see Algorithm 1). In each iteration of ILEA, we set $\bar{\theta}$ to be the current iterate $\theta^{(t)}$, construct the surrogate loss function $\widetilde{\mathcal{L}}^{(t)}(\theta)$, and then solve for the next iterate $\theta^{(t+1)}$ by either exactly minimizing the surrogate loss:

$$\theta^{(t+1)} \in \arg\min_{\theta \in \Theta} \widetilde{\mathcal{L}}^{(t)}(\theta),$$

or by forming a local one-step quadratic approximation:

$$\theta^{(t+1)} = \theta^{(t)} - \nabla^2 \mathcal{L}_1(\theta^{(t)})^{-1} \nabla \mathcal{L}_N(\theta^{(t)}) = \arg\min_{\theta \in \Theta} \widetilde{\mathcal{L}}^{H,(t)}(\theta).$$

Theorem 1 (or Theorem 3) guarantees, with high probability, the error bound

$$\|\theta^{(t+1)} - \widehat{\theta}\|_2 \le \frac{C_3}{\sqrt{n}} \|\theta^{(t)} - \widehat{\theta}\|_2, \quad \text{for each } t \ge 0,$$

where C_3 is positive constant independent of (n, k, N). If the desired accuracy is the statistical accuracy $\|\widehat{\theta} - \theta^*\|_2$ of the MLE and our initial estimator is $n^{-1/2}$-consistent, then we need to conduct at most $\lceil \frac{\log k}{\log n} \rceil$ iterations. ILEA interpolates between the gradient method and Newton's algorithm. When n is large relative to k, ILEA behaves like Newton's algorithm, and we achieve the optimal statistical accuracy in one iteration. If n is a fixed constant size, then ILEA reduces to a preconditioned gradient method. By appropriately choosing the subsample size n, ILEA achieves a trade-off among storage, communication, and computational complexities, depending on specific constraints of computing resources.

Algorithm 1: Iterative local estimation.

```
Initialize θ^(0) = θ̄;
for t = 0, 1, ..., T − 1 do
    Transmit the current iterate θ^(t) to local machines {M_j}_{j=1}^k;
    for j = 1 : k do
        Compute the local gradient ∇L_j(θ^(t)) at machine M_j;
        Transmit the local gradient ∇L_j(θ^(t)) to machine M_1;
    end
    Calculate the global gradient ∇L_N(θ^(t)) = (1/k) Σ_{j=1}^k ∇L_j(θ^(t)) in machine M_1;
    Form the surrogate function L̃^t(θ) = L_1(θ) − ⟨θ, ∇L_1(θ^(t)) − ∇L_N(θ^(t))⟩;
    Do one of the following in machine M_1:
    (1). Update θ^(t+1) ∈ arg min_{θ∈Θ} L̃^t(θ) ; // Exactly minimizing surrogate function L̃
    (2). Update θ^(t+1) = θ^(t) − ∇²L_1(θ^(t))^{-1} ∇L_N(θ^(t)) ; // One-step quadratic approximation
end
return θ^(T)
```

Confidence region construction. We now consider a natural class of local statistical inference procedures based on the surrogate function $\widetilde{\mathcal{L}}(\theta)$ that only uses the subsample $\{z_{i1}\}_{i=1}^n$ in machine $\mathcal{M}_1$. It is a classical result that under Assumptions PA-PD, the global empirical risk minimizer $\widehat{\theta}$ satisfies (see the proof of Corollary 4 in Section A.4)

$$\widehat{\theta} - \theta^* = -I(\theta^*)^{-1} \nabla \mathcal{L}_N(\theta^*) + O_p(N^{-1}), \quad \text{and} \quad \sqrt{N}(\widehat{\theta} - \theta^*) \to \mathcal{N}(0, \Sigma) \text{ in distribution as } N \to \infty, \tag{11}$$

where $\Sigma := I(\theta^*)^{-1} \mathbb{E}[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^T] I(\theta^*)^{-1}$ is the so-called sandwich covariance matrix. For example, when $\mathcal{L}$ corresponds to the negative log-likelihood function, $\Sigma = I(\theta^*)^{-1}$ will be the inverse of the information matrix. It is easy to see that the plug-in estimator,

$$\widehat{\Sigma} := \nabla^2 \mathcal{L}_N(\widehat{\theta})^{-1} \left(\frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{k} \nabla \mathcal{L}(\widehat{\theta}; z_{ij}) \nabla \mathcal{L}(\widehat{\theta}; z_{ij})^T \right) \times \nabla^2 \mathcal{L}_N(\widehat{\theta})^{-1}, \tag{12}$$

is a consistent estimator of the asymptotic covariance matrix Σ, that is, $\widehat{\Sigma} \to \Sigma$ in probability as $N \to \infty$. Based on the limiting distribution of $\sqrt{N}(\widehat{\theta} - \theta^*)$ and the plug-in estimator $\widehat{\Sigma}$, we can conduct statistical inference, for example, constructing confidence intervals for θ^*. However, this plug-in estimator is not convenient in our distributed learning setting since it needs access to all data. In the following corollary, we construct two communication-efficient plug-in estimators for Σ, one which only uses local data in one machine, and the other which uses all data but only requires other machines to send their local gradient evaluations at the estimate of θ.

The next corollary shows that for any reasonably good initial estimator $\overline{\theta}$, the asymptotic distribution of either the minimizer $\widetilde{\theta}$ of the surrogate function $\widetilde{\mathcal{L}}(\theta)$ or the local one-step quadratic approximated estimator $\widetilde{\theta}$ matches that of the global empirical risk minimizer $\widehat{\theta}$. Moreover, we construct two consistent estimators $\widetilde{\Sigma}$ and $\widetilde{\Sigma}'$ of Σ. In particular, by using $\widetilde{\Sigma}$, we can conduct statistical inference locally without access to the entire data while achieving the same asymptotic inferential accuracy as global statistical inference procedures.

Corollary 4. Under the same set of assumptions in Theorem 1, if the initial estimator $\overline{\theta}$ satisfies $\|\overline{\theta} - \theta^*\|_2 = O_p(n^{-1/2})$, then the surrogate minimizer $\widetilde{\theta}$ satisfies

$$\widetilde{\theta} - \theta^* = -I(\theta^*)^{-1} \nabla \mathcal{L}_N(\theta^*) + O_p(N^{-1} + n^{-1/2} \|\overline{\theta} - \theta^*\|_2),$$

and if $\|\overline{\theta} - \theta^*\|_2 = o_P\left(\sqrt{\frac{n}{N}}\right)$, then

$$\sqrt{N}(\widetilde{\theta} - \theta^*) \to \mathcal{N}(0,\ \Sigma) \quad \text{in distribution as } N \to \infty.$$

Moreover, the following local plug-in estimator,

$$\widetilde{\Sigma} := \nabla^2 \widetilde{\mathcal{L}}(\widetilde{\theta})^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \nabla \widetilde{\mathcal{L}}(\widetilde{\theta};\ z_{i1}) \nabla \widetilde{\mathcal{L}}(\widetilde{\theta};\ z_{i1})^T \right) \nabla^2 \widetilde{\mathcal{L}}(\widetilde{\theta})^{-1}, \tag{13}$$

is a consistent estimator for Σ as $n \to \infty$. If we also have $k \to \infty$, then the following global plug-in estimator,

$$\widetilde{\Sigma}' := \nabla^2 \widetilde{\mathcal{L}}(\widetilde{\theta})^{-1} \left(\frac{n}{k} \sum_{j=1}^{k} \nabla \mathcal{L}_j(\widetilde{\theta}) \nabla \mathcal{L}_j(\widetilde{\theta})^T \right) \nabla^2 \widetilde{\mathcal{L}}(\widetilde{\theta})^{-1}, \tag{14}$$

is also a consistent estimator for Σ as $(n, k) \to \infty$. Similar results hold for the local one-step quadratic approximated estimator $\widetilde{\theta}^H$ under the assumptions of Theorem 3.

Corollary 4 illustrates that we may substitute $\widetilde{\mathcal{L}}(\theta)$ as the global loss function and use it for statistical inference—$\widetilde{\Sigma}$ is precisely the plug-in estimator of the sandwiched covariance matrix using the surrogate loss function $\widetilde{\mathcal{L}}(\theta)$ (see Equation (12)). In the special case when $\mathcal{L}(\theta)$ is the negative log-likelihood function, we may instead use $\nabla^2 \widetilde{\mathcal{L}}(\widetilde{\theta})^{-1}$ as our plug-in estimator for $\Sigma = I(\theta^*)^{-1} = \mathbb{E}[\nabla^2 \mathcal{L}(\theta^*)^{-1}]$. $\widetilde{\Sigma}'$ tends to be a better estimator than $\widetilde{\Sigma}$ when $k \gg n$, since the variance $\mathcal{O}(k^{-1})$ of the middle term in Equation (14) is much smaller than variance $\mathcal{O}(n^{-1})$ of the middle term in Equation (13). See Section 4.1 for an empirical comparison of using $\widetilde{\Sigma}$ and $\widetilde{\Sigma}'$ for constructing confidence intervals.

3.2. Communication-Efficient Regularized Estimators With ℓ_1-Regularizer

In this subsection, we consider high-dimensional estimation problems where the dimensionality d of parameter θ can be much larger than the sample size n. Although the development here applies to a broader class of problems, we focus on ℓ_1-regularized procedures. ℓ_1-regularized estimators work well under the sparsity assumption that most components of the true parameter θ^* are zero. Let $S = \text{supp}(\theta^*)$ be a subset of $\{1, \ldots, d\}$ that encodes the sparsity pattern of θ^* and let $s = |S| = \sum_{j=1}^{d} \mathbb{I}(\theta_j^* \neq 0)$. Using the surrogate loss function $\widetilde{\mathcal{L}}(\theta)$ as a proxy to the global likelihood function in ℓ_1-regularized estimation procedures, we obtain the following communication-efficient regularized estimator:

$$\widetilde{\theta} \in \arg\min_{\theta \in \Theta} \left\{ \widetilde{\mathcal{L}}(\theta) + \lambda \|\theta\|_1 \right\}.$$

We study the statistical precision of this estimator in the high-dimensional regime.

We first present a theorem on the statistical error bound $\|\widetilde{\theta} - \theta^*\|_2$ of the estimator $\widetilde{\theta}$ for general loss function $\mathcal{L}$. We then illustrate the use of the theorem in the settings of high-dimensional linear models and generalized linear models. We begin by stating our assumptions.

Assumption HA (Restricted strong convexity). The local loss function $\mathcal{L}_1(\theta)$ at machine $\mathcal{L}_1$ is restricted strongly convex over S: for all $\delta \in C(S) := \{v : \|v_S\|_1 \leq 3\|v_{S^c}\|_1\}$,

$$\mathcal{L}_1(\theta^* + \delta) - \mathcal{L}_1(\theta^*) - \nabla \mathcal{L}_1(\theta^*)^T \delta \geq \mu \|\delta\|_2^2, \tag{4}$$

where μ is some positive constant independent of n.

As the name suggests, restricted strong convexity requires the global loss function $\mathcal{L}_n(\theta)$ to be a strongly convex function when restricted to the cone $C(S)$.

Assumption HB (Restricted Lipschitz Hessian). Both the local and global loss function $\mathcal{L}_1(\theta)$ and $\mathcal{L}_N(\theta)$ have restricted Lipschitz Hessians at radius R: for all $\delta \in C(S) \cap B_R(\theta^*)$,

$$\left\| (\nabla^2 \mathcal{L}_1(\theta^* + \delta) - \nabla^2 \mathcal{L}_1(\theta^*))\, \delta \right\|_\infty \leq M \|\delta\|_2^2, \quad \text{and}$$
$$\left\| (\nabla^2 \mathcal{L}_N(\theta^* + \delta) - \nabla^2 \mathcal{L}_N(\theta^*))\, \delta \right\|_\infty \leq M \|\delta\|_2^2,$$

where M is some positive constant independent of N.

The restricted Lipschitz Hessian condition is always satisfied for linear models where the Hessian $\nabla^2 \mathcal{L}_N(\theta)$ is a constant function of θ.

Theorem 5. Suppose that Assumption HA and Assumption HB at radius $R > \left\|\overline{\theta} - \theta^*\right\|_2$ are true. If the regularization parameter λ satisfies $\lambda \geq 2\|\nabla \mathcal{L}_N(\theta^*)\|_\infty + 2\left\|\nabla^2 \mathcal{L}_N(\theta^*) - \nabla^2 \mathcal{L}_1(\theta^*)\right\|_\infty \|\overline{\theta} - \theta^*\|_1 + 4M\|\overline{\theta} - \theta^*\|_2^2$, then

$$\|\widetilde{\theta} - \theta^*\|_2 \leq \frac{3\sqrt{s}\lambda}{\sqrt{\mu}},$$

where μ is the restricted strong convexity parameter of Assumption 3.2.

The lower bound condition on the regularization parameter λ for $\widetilde{\theta}$ is slightly stronger than that for the estimator $\widehat{\theta}$ based on the global loss function, which is $\lambda \geq 2\|\nabla \mathcal{L}_N(\theta^*)\|_\infty$. Since the estimation error upper bound provided by Theorem 5 is proportional to the regularization parameter, it is reasonable to

expect that $\widetilde{\theta}$ will yield a slightly larger error than $\widehat{\theta}$, depending on how good the initial estimator $\overline{\theta}$ is. For example, in generalized linear models, if small values of the regularization parameters are chosen for $\widetilde{\theta}$ and $\widehat{\theta}$, then the estimation error of $\widetilde{\theta}$ will be greater than that of $\widehat{\theta}$ by an amount

$$\frac{6\sqrt{s}}{\sqrt{\mu}}\left(\left\|\nabla^2\mathcal{L}_N(\theta^*)-\nabla^2\mathcal{L}_1(\theta^*)\right\|_\infty\|\overline{\theta}-\theta^*\|_1+2M\|\overline{\theta}-\theta^*\|_2^2\right)$$
$$\sim\sqrt{\frac{s\log d}{n}}\,\|\overline{\theta}-\theta^*\|_1+M\sqrt{s}\,\|\overline{\theta}-\theta^*\|_2^2.$$

As long as $\|\overline{\theta}-\theta^*\|_1$ and $\|\overline{\theta}-\theta^*\|_2$ are sufficiently small, this difference will be negligible with respect to the estimation error bound of $\widehat{\theta}$, which is $\sqrt{\frac{s\log d}{N}}$. For example, we may choose $\overline{\theta}$ to be the local ℓ_1-regularized estimator $\widehat{\theta}_1:=\arg\min_\theta\left\{\mathcal{L}_1(\theta)+\lambda_1\|\theta\|\right\}$, with estimation error $\sqrt{\frac{s\log d}{n}}$, so that

$$\|\widehat{\theta}_1-\theta^*\|_1\le Cs\sqrt{\frac{\log d}{n}}\quad\text{and}\quad\|\widehat{\theta}_1-\theta^*\|_2\le C\sqrt{\frac{s\log d}{n}}.$$

We now apply Theorem 5 to two examples.

Example (Sparse linear regression). In sparse linear regression, observations $\{z_{ij}=(x_{ij},y_{ij}):1\le i\le n,\ 1\le j\le k\}$ satisfy

$$y_{ij}=x_{ij}^T\beta+\epsilon_{ij},\qquad\epsilon\sim\mathcal{N}(0,\sigma^2),$$

where x_{ij} is a d-dimensional covariate vector, y_{ij} is the response, and $\beta\in\mathbb{R}^d$ is the unknown regression coefficient to be estimated. Recall the sparsity assumption that $s=\sum_{j=1}^d\mathbb{I}(\theta_j^*\neq 0)=o(n)$. For linear regression, the global loss function takes the form

$$\mathcal{L}_N(\theta)=\frac{1}{N}\sum_{i=1}^n\sum_{j=1}^k\left(y_{ij}-x_{ij}^T\beta\right)^2.$$

We consider a random design where the x_{ij} are iid and A-sub-Gaussian, that is, for all $\alpha\in\mathbb{R}^d$,

$$\mathbb{E}[\exp(\alpha^Tx_{ij})]\le\exp\left(A^2\|\alpha\|_2^2/2\right).$$

Let $\Sigma=\mathbb{E}[x_{ij}x_{ij}^T]$ be the covariance matrix of the design. For this class of design, it is known that Assumption HA is satisfied with high probability as long as Σ is strictly positive definite and $n\ge C_0s\log d$ for some constant $C_0>0$ depending on the minimal eigenvalue of Σ (Raskutti, Wainwright, and Yu 2010). For linear models, the Lipschitz constant M in Assumption HB is zero and therefore HB is also satisfied.

Theorem 6. If x_{ij} is A-sub-Gaussian, Σ is strictly positive definite and $n\ge C_0s\log d$, then with probability at least $1-c_1\exp\{-c_2n\}$, it holds that

$$\|\widetilde{\theta}-\theta^*\|_2\le C_1A\sqrt{\frac{s\log d}{N}}+C_1A\sqrt{\frac{s\log d}{n}}\,\|\overline{\theta}-\theta^*\|_1.$$

If the initial estimator satisfies $\|\overline{\theta}-\theta^*\|_1\le C_2\,s\sqrt{\frac{\log d}{n}}$, then with the same probability, it holds that

$$\|\widetilde{\theta}-\theta^*\|_2\sim C_1A\sqrt{\frac{s\log d}{N}}+C_3\frac{s^{3/2}\log d}{n}.$$

The constants $(c_1,c_2,C_0,C_1,C_2,C_3)$ are independent of (n,k,d,s).

For linear regression under the sparsity condition, the minimax rate of estimating θ is $\sqrt{\frac{s\log\frac{d}{s}}{N}}$. Therefore, Theorem 6 shows that our approximated estimator $\widetilde{\theta}$ is nearly minimax-optimal and communication-optimal (Braverman et al. 2016) if $n\ge Cs\sqrt{N\log d}$ for some constant $C>0$. When this lower bound on the local sample size n fails, we may still apply the iterative estimation procedure (Algorithm 1) for a logarithmic number of rounds to obtain a nearly minimax-optimal estimator. The convergence rate can be analyzed by inducting on Theorem 5.

Example (Generalized linear models). In this section, we apply Theorem 5 to generalized linear models with an ℓ_1-regularizer. We begin with some background on generalized linear models. Recall that the data points are of the form $z_{ij}=(x_{ij},y_{ij})$, where y_{ij} is the response and x_{ij} is the d-dimensional covariate vector. A generalized linear model assumes the conditional distribution of y_{ij} given x_{ij} to be

$$\mathbb{P}(y_{ij}\,|\,x_{ij},\ \theta,\ \sigma)\propto\exp\left\{\frac{y_{ij}x_{ij}^T\theta-\phi\left(x_{ij}^T\theta\right)}{c(\sigma)}\right\},$$

where σ is a scalar parameter, θ is the unknown d-dim parameter to be estimated, and ϕ is a link function. For example, $\phi(x)=\log(1+e^x)$ in logistic regression, and $\phi(x)=e^x$ in Poisson regression. We still assume sparsity, such that $s=\sum_{j=1}^d\mathbb{I}(\theta_j^*\neq 0)=o(n)$. Now the global loss function and its gradient are given by

$$\mathcal{L}_N(\theta)=\frac{1}{N}\sum_{j=1}^k\sum_{i=1}^n-y_{ij}x_{ij}^T\theta+\phi(x_{ij}^T\theta),\qquad\text{and}$$
$$\nabla\mathcal{L}_N(\theta)=\frac{1}{N}\sum_{j=1}^k\sum_{i=1}^n(\phi'(x_{ij}^T\theta)-y_{ij})\,x_{ij}.$$

Under a random design assumption, we verify Assumptions HA and HB, and obtain the following result.

Theorem 7. Assume that for some constants (A,B,m,L), x_{ij} is iid. A-sub-Gaussian, $\left\|x_{ij}\right\|_\infty\le B$, and $mI\preceq\Sigma=\mathbb{E}[x_{ij}x_{ij}^T]\preceq LI$. Then with probability at least $1-c_1\exp\{-c_2n\}$, it holds that

$$\|\widetilde{\theta}-\theta^*\|_2\le C_1A\sqrt{\frac{s\log d}{N}}+C_1A\sqrt{\frac{s\log d}{n}}\|\overline{\theta}-\theta^*\|_1$$
$$+C_1A\sqrt{s}\|\overline{\theta}-\theta^*\|_2^2.$$

If $\|\overline{\theta}-\theta^*\|_1\le C_2\,s\sqrt{\frac{\log d}{n}}$ and $\|\overline{\theta}-\theta^*\|_2\le C_2\sqrt{\frac{s\log d}{n}}$, then with the same probability, we have

$$\|\widetilde{\theta}-\theta^*\|_2\le C_3\sqrt{\frac{s\log d}{N}}+C_3\frac{s^{3/2}\log d}{n}.$$

The constants $(c_1,c_2,C_0,C_1,C_2,C_3)$ are independent of (n,k,d,s).

3.3. Communication-Efficient Bayesian Inference

In this subsection, we consider distributed Bayesian in the setting of regular parametric models. We place a prior distribution π on the parameter space Θ and form the global posterior distribution

$$\pi\left(\theta \mid Z_1^N\right) = D \exp\left\{-\sum_{i=1}^{n}\sum_{j=1}^{k}\mathcal{L}(\theta;\, z_{ij})\right\}\pi(\theta), \qquad (15)$$

where D is the normalizing constant. In the rest of this subsection, we tacitly assume that the loss function $\mathcal{L}$ is the negative log-likelihood function. Extensions to the Gibbs posterior (Bissiri, Holmes, and Walker 2016) where $\mathcal{L}$ is replaced with a generic loss function $\mathcal{L}$ in posterior (15) is straightforward.

Most existing literature (Wang and Dunson 2015; Neiswanger, Wang, and Xing 2015) in distributed Bayesian inference uses the decomposition

$$\pi\left(\theta \mid Z_1^N\right) = D\prod_{j=1}^{k}\exp\left\{-n\mathcal{L}_j(\theta)\right\}, \qquad (16)$$

such that the global posterior $\pi(\theta \mid Z_1^N)$ can be written as the product of subsample posteriors

$$\pi(\theta \mid Z_j) = D_j \exp\left\{-n\mathcal{L}_j(\theta)\right\}\pi^{1/k}(\theta), \quad j = 1, \ldots, k,$$

where the prior is raised to power k^{-1} so that it is appropriately weighted in product (16) and D_j is the normalizing constant. This decomposition motivates a MapReduce computational framework in which separate Markov chains are run in machines $\{\mathcal{M}_j\}_{j=1}^k$ based on the local data on that machine. After running these Markov chains in parallel, all local posterior draws are transmitted to a central node, where an approximation $\widetilde{\pi}_N(\theta)$ to the global posterior $\pi_N(\theta) := \pi(\theta \mid Z_1^N)$ is formed. A main drawback of these approaches is that the communication cost can be exorbitant—for example, exponentially large in the dimension d—since the number of draws from each local posterior must be large enough to be representative of the local posterior distribution.

Our approach to distributed Bayesian inference is based on using the surrogate function $\widetilde{\mathcal{L}}(\theta)$. Our sampling scheme is communication efficient, requiring running one single Markov chain in a local machine. Here is an outline of the algorithm:

1. Compute a good initial estimate $\overline{\theta}$, for example, the one-step estimate $\widetilde{\theta}^H$ in Section 3.1.
2. For $j = 1, \ldots, k$, compute the local gradient $\nabla\mathcal{L}_j(\overline{\theta})$ in machine $\mathcal{M}_j$.
3. Transmit all local gradients to machine $\mathcal{M}_1$ and form the global gradient $\nabla\mathcal{L}_N(\overline{\theta}) = \frac{1}{k}\sum_{j=1}^{k}\nabla\mathcal{L}_j(\overline{\theta})$.
4. Machine $\mathcal{M}_1$ constructs the surrogate function $\widetilde{\mathcal{L}}(\theta)$ as (8).
5. Machine $\mathcal{M}_1$ runs a Markov chain to sample from the surrogate posterior $\widetilde{\pi}_N(\theta) \propto \exp\left(-N\widetilde{\mathcal{L}}(\theta)\right)\pi(\theta)$, and uses the draws to conduct statistical inference.

The following result shows that the surrogate posterior $\widetilde{\pi}_N(\cdot)$ is close to the global posterior $\pi(\cdot \mid Z_1^N)$ as long as the initial estimator $\overline{\theta}$ is reasonably close to θ^*.

Theorem 8. If Assumption PA-PD hold and $\|\overline{\theta} - \widehat{\theta}\|_2 = o_p(N^{-1/2})$, then the approximate posterior $\widetilde{\pi}_N(\theta)$ satisfies

$$\left\|\widetilde{\pi}_N - \pi_N\right\|_1 = O_p\left(\sqrt{N}\log N\,\|\overline{\theta} - \widehat{\theta}\|_2 + \frac{(\log N)^2}{\sqrt{n}}\right),$$

where $\|P - Q\|_1 = \int |P(d\theta) - Q(d\theta)|$ is the variation distance between the distributions P and Q.

If we use the local one-step estimator $\widetilde{\theta}^H$ as the initial estimator $\overline{\theta}$, then the approximation error becomes

$$\left\|\widetilde{\pi}_N - \pi_N\right\|_1 = O_p\left(\frac{\sqrt{N}\log N}{n}\right) + \left(\frac{(\log N)^2}{\sqrt{n}}\right).$$

This illustrates that we may choose $k = N/n$ up to $o\big(N^{1/2}(\log N)^{-1}\big)$ while still maintaining $\left\|\widetilde{\pi}_N - \pi_N\right\|_1 = o_p(1)$. The overall communication requirements of this procedure are two passes over the entire dataset (one for computing $\widetilde{\theta}^H$ and one for constructing $\widetilde{\mathcal{L}}(\theta)$). To allow larger k, we may apply the iterative algorithm in Section 3.1 to improve the accuracy of the initial estimator $\overline{\theta}$. Note that our theory only covers low-dimensional regular parameter models; it is still an open problem to design theoretically sound communication-efficient Bayesian procedures for high-dimensional problems.

4. Simulations

In this section, we present examples of simulation experiments using the CSL methodology developed in Section 2.2.

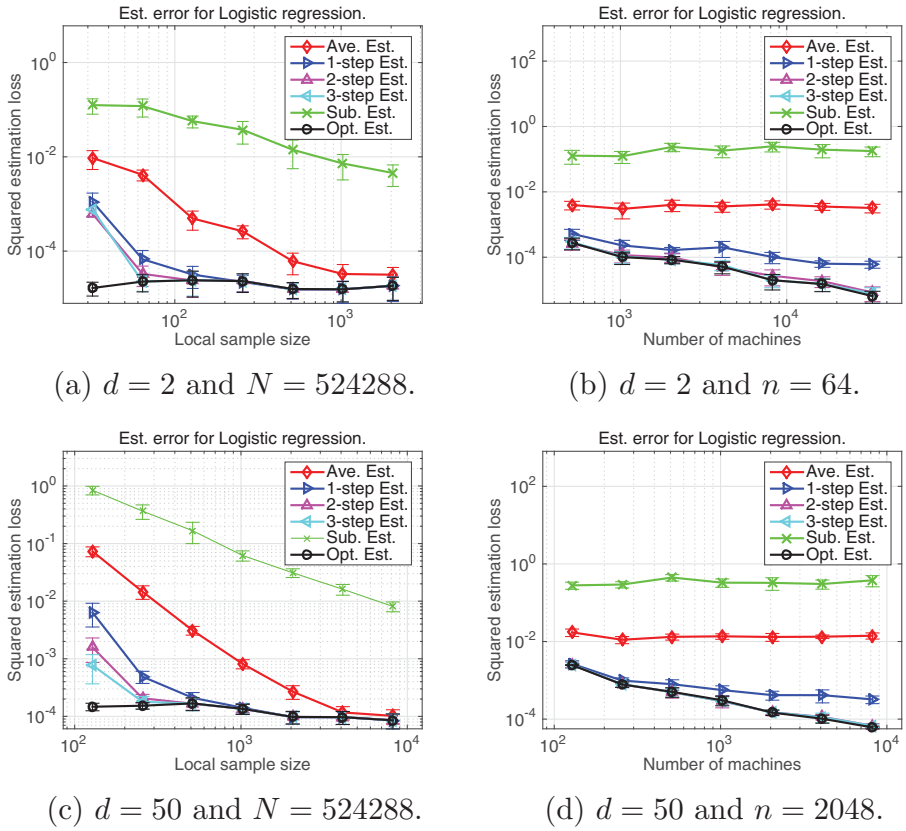
4.1. Distributed M-Estimation in Logistic Regression

In logistic regression, iid observations $Z_1^N = \{Z_{ij} = (X_{ij}, Y_{ij}) : i = 1, \ldots, n.\ j = 1, \ldots, k\}$ are generated from the model

$$Y_{ij} \sim \mathrm{Ber}(P_{ij}), \quad \text{with} \quad \log\frac{P_{ij}}{1 - P_{ij}} = \langle X_{ij}, \theta^*\rangle. \qquad (17)$$

In our simulation, the true regression coefficient θ^* is a d-dim vector with $d \in \{2, 50\}$ and the d-dim covariate vector X_{ij} is independently generated from $\mathcal{N}(0, I_d)$ (due to space constraints, plots under $d = 10$ are provided in the Appendix). For each replicate of the simulation, we uniformly sample the parameter θ^* from the d-dim unit cube $[0, 1]^d$.

We implement the one-step CSL estimator $\theta^{(1)}$ with the averaging estimator $\widehat{\theta}^A$ (based on simply averaging the local estimators) as our initial estimator $\overline{\theta}$ for easier illustration and comparison, since it is asymptotically equivalent to the one-step CSL estimator with a local estimator as the initialization. We also implement the iterative local estimation algorithm to produce two-step and three-step estimators $\theta^{(2)}$ and $\theta^{(3)}$ by iteratively applying the one-step estimation procedure. We compare our communication-efficient estimators with the (optimal) global M-estimator θ^{global} and the subsample estimator θ^{sub} that only uses the local data in machine $\mathcal{M}_1$. Two different regimes are considered: (1) the total sample size N is fixed at $N = 2^{19} \approx 10^6$, and the local sample size n varies from 10^2 to 10^4; (2) the local sample size n is fixed at 64 ($d = 2$) or 2048 ($d = 50$), and the number of machines k varies from 10^2 to 10^4.



(a) $d = 2$ and $N = 524288$. (b) $d = 2$ and $n = 64$.

(c) $d = 50$ and $N = 524288$. (d) $d = 50$ and $n = 2048$.

Figure 1. Squared estimation error $\|\widehat{\theta} - \theta^*\|_2^2$ versus local sample size n and number of machines k for logistic regression. In all cases, each point corresponds to the average of 100 trials, with standard errors also shown. In plots (a) and (c), we change the local sample size n while fixing the total sample size N (number of machines $k = N/n$) for dimension $d \in \{2, 50\}$. In plots (b) and (d), we change the number of machines k while fixing the local sample size n (total sample size $N = nk$) under dimension $d \in \{2, 50\}$.

Figure 1 reports the results. In plots (a) and (c), the total sample size N is fixed and therefore the estimation error associated with the global estimate θ^{global} remains approximately fixed as n varies. As expected, the remaining estimators exhibit a rapid decay in the estimation error as the local sample size n grows. Our communication-efficient estimators yield the best performance among the distributed estimators. When n is sufficiently large, the one-step, two-step, and three-step estimators have almost the same performance as θ^{global}. However, as n becomes small, further application of the iterative local estimation procedure in Algorithm 1 does not improve the statistical accuracy. This is in fact consistent with Theorem 3—the contraction coefficient $\|\theta^{(t+1)} - \theta^{\text{global}}\|_2 / \|\theta^{(t)} - \theta^{\text{global}}\|_2$ is dominated by the sum of two terms: the initial estimation error $\|\theta^{(t)} - \theta^{\text{global}}\|_2$ and the local Hessian approximation error $\|\nabla \mathcal{L}_1(\theta^*) - \nabla \mathcal{L}_N(\theta^*)\|_2$. Even though the initial estimation error can be reduced to a small level, the local Hessian approximation error still persists for small n and prevents further improvement from application of the iterative procedure. We remark that the condition that the local size n should exceed a d-dependent threshold is a mild requirement in practice. Indeed, the local machine storage limit in reality is often large enough to ensure $n \gg d$. Even under the scenario (small n) where our theory does not apply, the one-step, two-step, and three-step estimators still have better performance than $\widehat{\theta}^A$ and θ^{sub}. In plots (b) and (d), we fix the local sample size n under different d such that n exceeds the d-dependent threshold, and gradually increase the number of machines k. In our regime, k is comparable or even much larger than n, and therefore the averaging estimator $\widehat{\theta}^A$ does not improve as more data are available. This is consistent with theoretical results in Zhang, Duchi, and Wainwright (2013) that require $k \ll n$ for $\widehat{\theta}^A$ to have comparable performance as θ^{global}. By using our approach, even a single step of Algorithm 1 significantly improves the accuracy of $\widehat{\theta}^A$. Moreover, $\theta^{(2)}$ and $\theta^{(3)}$ achieve almost the same accuracy as θ^{global}. Consistent with our theory, for a fixed number of steps t, the t-step estimate $\theta^{(t)}$ tends to have larger estimation error than θ^{global} as k grows. In plot (b), even for k as large as 10^5 (much larger than the local sample size $n \sim 10^1$), the two-step estimate $\theta^{(2)}$ already achieves the same level of estimation accuracy as the global estimator θ^{global}.

We also compare our CSL method with some state-of-the-art distributed learning methods: the full gradient method (FGD, Nesterov 2013), the elastic averaging stochastic gradient method (EASGD, Zhang, Choromanska, and LeCun 2015), and the stochastic variance-reduced gradient method (SVRG, Johnson and Zhang 2013) (the latter two methods treat data in local machines as mini-batches, and update the parameter in a cyclic manner). We compare these methods by reporting the squared estimation loss versus the number of sweeps over the entire dataset as shown in Figure 2, under $(d, n, k) = (10, 256, 512)$ and $N = 524, 288$. In particular, the i-step CSL estimator needs i sweeps over the dataset for $i = 1, \ldots, 9$. As we expected, these competitors are based on first-order information (gradients)

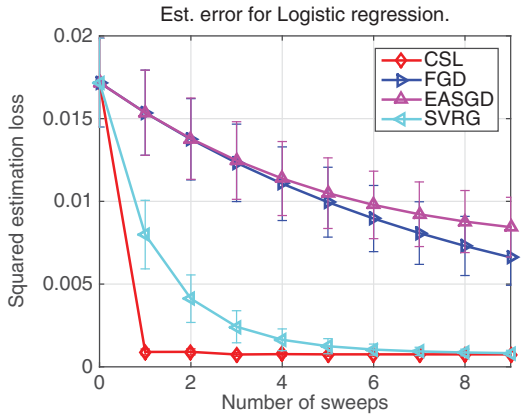


Figure 2. Squared estimation error $\|\widehat{\theta} - \theta^*\|_2^2$ versus number of sweeps over the entire dataset in logistic regression with $(d, n, k) = (10, 256, 512)$. CSL: the proposed method; FGD: full gradient descent method; EASGD: elastic averaging stochastic gradient method; SVRG: stochastic variance reduced gradient method. Each point corresponds to the average of 100 trials, with standard errors also shown.

and exhibit relative slow convergence. In comparison, the CSL uses some noisy second-order information (local Hessian), and tends to converge within the first iteration. Therefore, under a given communication constraint (in each sweep, CSL, FGD, and SVRG communicate dk many numbers, and EASGD $2dk$ numbers), CSL tends to achieve the best estimation accuracy among these competitors.

We now assess the performance of the inference procedures based on the plug-in estimators $\widetilde{\Sigma}$ and $\widetilde{\Sigma}'$ under the logistic model (17). We use $\widetilde{\Sigma}$ or $\widetilde{\Sigma}'$ and the three-step estimator $\theta^{(3)}$ to construct a 95% confidence interval (CI) for the first component θ_1 of θ as

$$\left[\theta_1^{(3)} - 1.96\,\widetilde{\Sigma}_{11}/\sqrt{N},\ \theta_1^{(3)} + 1.96\,\widetilde{\Sigma}_{11}/\sqrt{N}\right] \quad \text{or}$$
$$\left[\theta_1^{(3)} - 1.96\,\widetilde{\Sigma}'_{11}/\sqrt{N},\ \theta^{(3)} + 1.96\,\widetilde{\Sigma}'_{11}/\sqrt{N}\right].$$

The coverage of the CI based on 100 trials is calculated. Figure 3 shows the results. In plot (a), coverage based on both plug-in estimators is low at $n = 2^7$ because the sample size is so small that the center $\theta^{(3)}$ of the CI has a large bias (see Figure 1 (c)). In plot (b), the CI based on $\widetilde{\Sigma}'$ has low coverage when the number k of machines is small, which is consistent with our theory. In all other regimes of (n, k), both CIs have coverage that is close to the nominal level 95%. Moreover, the CI based on $\widetilde{\Sigma}'$ is slightly better than the one based on $\widetilde{\Sigma}$ for large k, which empirically supports our intuition in the discussion after Corollary 4.

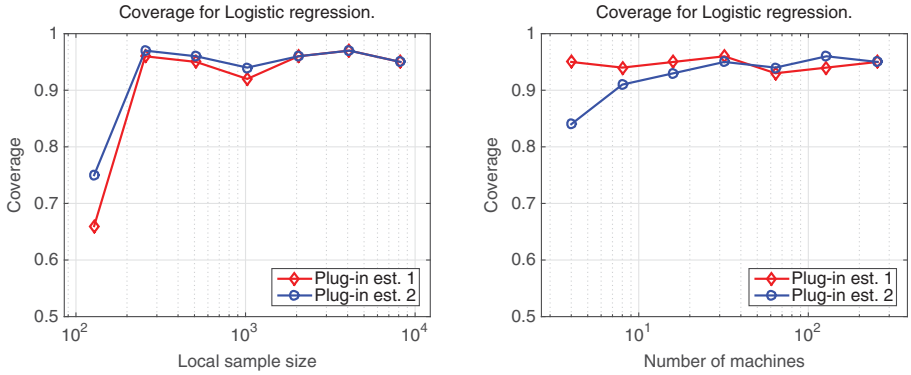
4.2. Distributed Sparse Linear Regression

We evaluate the CSL estimator on the sparse linear regression problem. The data are generated as $y_{ij} = X_{ij}^T\theta^* + \epsilon_{ij}$, where $i \in [n]$ and $j \in [k]$ with n ranging from 10^2 to 10^4 and k from 1 to 32. The covariate vector X_{ij} is iid $\mathcal{N}(0, I_d)$ with dimension $d \in \{5000, 10{,}000\}$, the noise ϵ_{ij} is iid $\mathcal{N}(0, \sigma^2)$ with $\sigma = 1$, and θ^* is s-sparse with signal-to-noise ratio $\frac{|\theta_i|}{\sigma} = 5$.

In the first experiment, we keep the total data size N fixed, and increase the number of machines k. This corresponds to each machine having a smaller local sample size n as k increases. We observe that the one-step CSL estimator has nearly constant error, even though each machine has less local data. In fact at $k = 30$, the local data size is $n = 720$, which is much smaller than $d = 5000$, yet the CSL estimator achieves the same mean-square error as lasso on all N points. The error of the averaging estimator increases dramatically as n decreases, since the mean-squared error is $\frac{s\log d}{n}$, showing that the averaging algorithm is not suitable in this setting.

In the second experiment, we keep n fixed and increase k and N. As predicted by our theory, the one-step CSL estimator has error that is linear on the log-log scale because the mean-squared error scales as $\frac{s\log d}{nk}$. The averaging estimator has error that slowly decreases with the increased sample size, due to the bias induced by regularization. The averaging estimator does not attain mean-square error of $\frac{s\log d}{nk}$.

Next, we study the variable selection performance by comparing the number of true discoveries and false discoveries. The experimental setup is with $(n, s, d, k) = (800, 20, 2000, 4)$ and the data are generated from the same linear model as above. We compare three algorithms across a range of signal-to-noise ratios: (1) local lasso on the center machine (local), (2) one-step CSL estimator (one-step CSL), and (3) global lasso on the entire dataset (global). From Table 1, we see that the global lasso performs the best, closely followed by one-step CSL, and both



(a) $d = 10$ and $N = 524288$. (b) $d = 10$ and $n = 256$.

Figure 3. Coverage of the confidence interval for the first component of β versus local sample size n and number of machines k for logistic regression under $d = 10$. In all cases, the coverage probability is computed based on 100 trials. Here, "plug-in est. 1" corresponds to the confidence interval constructed based on the plug-in estimator $\widetilde{\Sigma}$ and the three-step estimator $\theta^{(3)}$, whereas "plug-in est. 2" is based on $\widetilde{\Sigma}'$. In plots (a), we change the local sample size n while fixing the total sample size N (number of machines $k = N/n$). In plots (b), we change the number of machines k while fixing the local sample size n (total sample size $N = nk$).

Table 1. Variable selection performance in the high-dimensional sparse linear regression. The signal-to-noise ratio SNR takes values in {0.1, . . . , 0.6}.

SNR	Local	One-step CSL	Global	SNR	Local	One-step CSL	Global
0.1	5.5	10.3	10.9	0.1	30.9	14.8	10.1
0.2	11.1	17.4	18.6	0.2	11.2	1.7	0.2
0.3	15.6	19.9	20.0	0.3	3.2	0.2	0.0
0.4	17.1	20.0	20.0	0.4	1.1	0.0	0.0
0.5	19.4	20.0	20.0	0.5	0.1	0.0	0.0
0.6	20.0	20.0	20.0	0.6	0.0	0.0	0.0

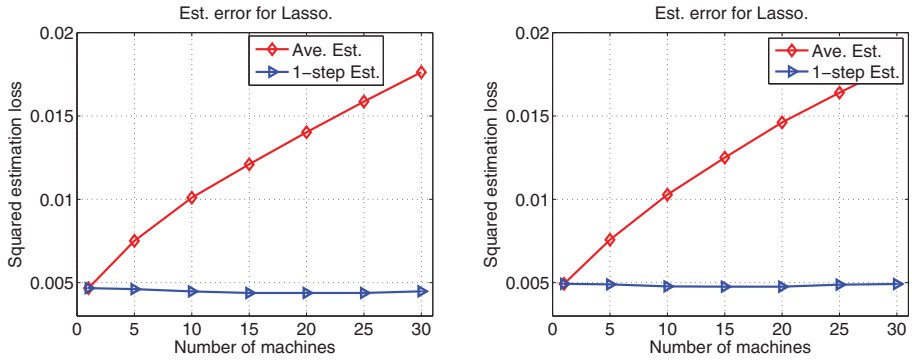
Table 2. Posterior quantile difference between the approximated posterior distribution and the exact posterior distribution for θ_1 at level $\alpha \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$. All numbers are based on averaging over 20 random splits of the data.

	$(d, n) = (2, 64)$		$(d, n) = (10, 256)$		$(d, n) = (50, 2048)$	
α	$k = 64$	$k = 256$	$k = 64$	$k = 256$	$k = 64$	$k = 256$
0.1	2.2×10^{-2}	1.1×10^{-2}	10.3×10^{-3}	5.2×10^{-3}	5.6×10^{-3}	2.7×10^{-3}
0.25	1.6×10^{-2}	0.9×10^{-2}	7.4×10^{-3}	3.9×10^{-3}	3.8×10^{-3}	1.7×10^{-3}
0.5	1.1×10^{-2}	0.6×10^{-2}	5.5×10^{-3}	3.1×10^{-3}	3.3×10^{-3}	1.4×10^{-3}
0.75	1.6×10^{-2}	0.8×10^{-2}	7.2×10^{-3}	4.2×10^{-3}	3.8×10^{-3}	1.8×10^{-3}
0.9	2.3×10^{-2}	1.1×10^{-2}	9.9×10^{-3}	5.4×10^{-3}	5.5×10^{-3}	2.8×10^{-3}

greatly outperform local lasso. The one-step CSL and global lasso make approximately the same number of true discoveries, and the local lasso makes much fewer true discoveries especially for low signal-to-noise ratio. The global lasso makes the fewest number of false discoveries, which is closely followed by one-step CSL, and the local lasso makes an order of magnitude more false discoveries.

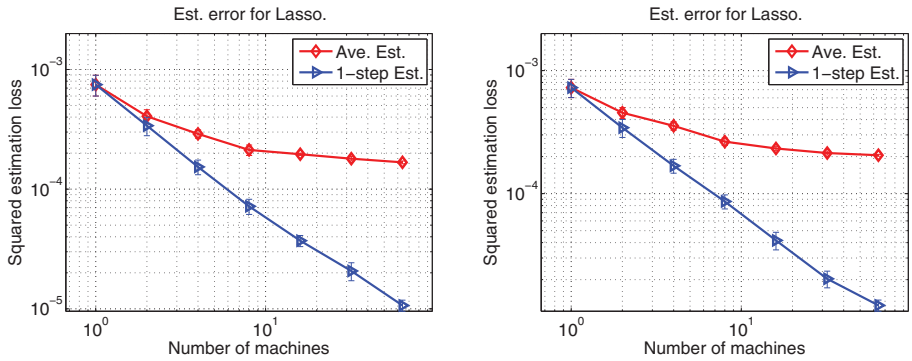
4.3. Distributed Bayesian Inference

Our synthetic dataset is generated from the logistic model (17) for dimension $d \in \{2, 50\}$ (due to space constraints, plots under $d = 10$ are provided in the Appendix). We use the three-step estimator $\theta^{(3)}$ in Section 3.1 as the initial estimator $\overline{\theta}$ and implement the Bayesian procedures based on the (approximated) posterior distribution $\pi_n(\theta)$ and $\widetilde{\pi}_N(\theta)$ by sampling a Markov chain Monte Carlo algorithm. We use the Metropolis algorithm, where at each iteration the proposal distribution for θ is a d-dim Gaussian distribution centered at the current iterate $\theta^{(t)}$. In each case, we run the Markov chain for 20,000 iterations and treat the first half as burn-in. Figure 6 plots the (approximated) marginal posterior density of the first component θ_1 of θ under different (d, n, k) combinations (n is chosen so that $\theta^{(3)}$ is a good approximation to the global estimator $\widehat{\theta}$, see Figure 1). Table 2 reports the root-mean-squared distance between quantiles of the approximated marginal posterior and the exact marginal posterior for the first component θ_1. As we can see, all numbers in this table are significantly smaller than the lengths of the span of corresponding the posterior distributions in Figure 6,



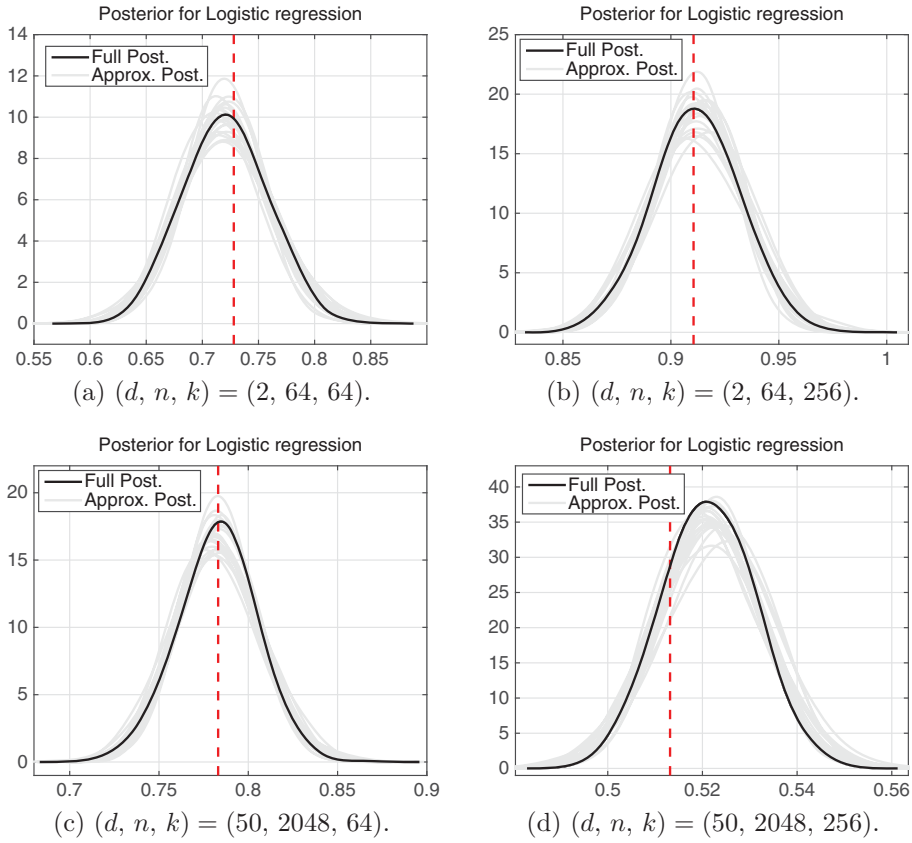
(a) $(N, d, s) = (21600, 5000, 20)$ (b) $(N, d, s) = (21600, 10000, 20)$

Figure 4. Squared estimation error $\|\widehat{\theta} - \theta^*\|_2^2$ versus the number of machines k for high-dimensional sparse linear regression with entire sample size being fixed at $N = 21{,}600$. The results show that as $k \in \{1, 5, 10, 15, 20, 25, 30\}$ increases, the local data size $n = 21{,}600/k$ decreases, and as expected, the one-step CSL estimator has constant error. In contrast, the error of the averaging estimator increases as the local sample size decreases.



(a) $(n, d, s) = (800, 1000, 20)$ (b) $(n, d, s) = (800, 2000, 20)$

Figure 5. Squared estimation error $\|\widehat{\theta} - \theta^*\|_2^2$ versus the number of machines k for high-dimensional sparse linear regression with local sample size being fixed at $N = 800$. Similar to the results in Figure 4, as $k \in \{1, 2, 4, 8, 16, 32, 64\}$ increases, the mean-squared error of the one-step CSL estimator decreases linearly in the log-log scale. However, the mean-squared error of the averaging estimator exhibits a sub-linear decrease for large values of k.



(a) $(d, n, k) = (2, 64, 64)$. (b) $(d, n, k) = (2, 64, 256)$.

(c) $(d, n, k) = (50, 2048, 64)$. (d) $(d, n, k) = (50, 2048, 256)$.

Figure 6. Marginal posterior distribution of the first component θ_1 of θ for logistic regression for dimension $d \in \{2, 50\}$ is shown. In each plot, 20 approximations (gray curves) to the full posterior (black curve) are shown based on random splits of the data into k subsamples. The vertical dotted line indicates the location of the truth θ_1^*.

indicating closeness between the approximated posterior distribution and the exact posterior distribution. Consistent with our theoretical prediction, $\widetilde{\pi}_N(\theta)$ provides a good approximation to $\pi_N(\theta)$ as long as the initial estimator $\overline{\theta}$ is sufficiently close to $\widehat{\theta}$, even when k is much larger than n (see plot (b)). Since the computation of the approximate posterior distribution $\widetilde{\pi}_N(\theta)$ only uses the local data in machine $\mathcal{M}_1$, the computation of the acceptance ratio using $\widetilde{\pi}_N(\theta)$ is k times as fast as that using the full data posterior $\pi_N(\theta)$ in each iteration of the Metropolis algorithm.

Supplementary Materials

The supplementary materials contain the appendices for the article.

References

Bissiri, P., Holmes, C., and Walker, S. (2016), "A General Framework for Updating Belief Distributions," *Journal of the Royal Statistical Society*, Series B, 78, 1103–1130. [676]

Braverman, M., Garg, A., Ma, T., Nguyen, H., and Woodruff, D. (2016), "Communication Lower Bounds for Statistical Estimation Problems via a Distributed Data Processing Inequality," in *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, New York: ACM, pp. 1011–1020. [669,675]

Chernozhukov, V., and Hong, H. (2003), "An MCMC Approach to Classical Estimation," *Journal of Econometrics*, 115, 293–346. [669]

Cleveland, W., and Hafen, R. (2014), "Divide and Recombine (D&R): Data Science for Large Complex Data," *Statistical Analysis and Data Mining*, 7, 425–433. [668]

Demmel, J., Grigori, L., Hoemmen, M., and Langou, J. (2012), "Communication-Optimal Parallel and Sequential QR and LU Factorizations," *SIAM Journal on Scientific Computing*, 34, 206–239. [668]

Duchi, J., Agarwal, A., and Wainwright, M. (2012), "Dual Averaging for Distributed Optimization: Convergence Analysis and Network Scaling," *IEEE Transactions on Automatic Control*, 57, 592–606. [668]

Duchi, J., Jordan, M., Wainwright, M., and Zhang, Y. (2015), "Optimality Guarantees for Distributed Statistical Estimation," arXiv:1405.0782. [668]

Garg, A., Ma, T., and Nguyen, H. (2014), "On Communication Cost of Distributed Statistical Estimation and Dimensionality," in *Advances in Neural Information Processing Systems*, pp. 2726–2734. [669]

Johnson, R., and Zhang, T. (2013), "Accelerating Stochastic Gradient Descent Using Predictive Variance Reduction," in *Advances in Neural Information Processing Systems*, pp. 315–323. [677]

Kannan, R., Vempala, S., and Woodruff, D. (2014), "Principal Component Analysis and Higher Correlations for Distributed Data," in *The 27th Conference on Learning Theory*, pp. 1040–1057. [668]

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, 76, 795–816. [668]

Lee, J., Sun, Y., Liu, Q., and Taylor, J. (2015a), "Communication-Efficient Sparse Regression: A One-Shot Approach," arXiv:1503.04337. [668,669]

Lee, J. D., Lin, Q., Ma, T., and Yang, T. (2017), "Distributed Stochastic Variance Reduced Gradient Methods and a Lower Bound for Communication Complexity," *Journal of Machine Learning Research*, 18, 1–43. [671]

Mackey, L., Talwalkar, A., and Jordan, M. I. (2015), "Distributed Matrix Completion and Robust Factorization," *Journal of Machine Learning Research*, 16, 913–960. [668]

Maclaurin, D., and Adams, R. (2014), "Firefly Monte Carlo: Exact MCMC with Subsets of Data," arXiv:1403.5693. [668]

Neiswanger, W., Wang, C., and Xing, E. (2015), "Asymptotically Exact, Embarrassingly Parallel MCMC," in *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, Arlington, VA: AUAI Press, pp. 623–632. [668,669,676]

Nesterov, Y. (2013), *Introductory Lectures on Convex Optimization: A Basic Course* (Vol. 87), New York: Springer Science & Business Media. [677]

Rabinovich, M., Angelino, E., and Jordan, M. (2016), "Variational Consensus Monte Carlo," in *Advances in Neural Information Processing Systems*, Red Hook, NY: Curran Associates. [668]

Raskutti, G., Wainwright, M., and Yu, B. (2010), "Restricted Eigenvalue Properties for Correlated Gaussian Designs," *Journal of Machine Learning Research*, 11, 2241–2259. [675]

Scott, S., Blocker, A., Bonassi, F., Chipman, H., George, E., and McCulloch, R. (2016), "Bayes and Big Data: The Consensus Monte Carlo Algorithm," *International Journal of Management Science and Engineering Management*, 11, 78–88. [668]

Shamir, O., Srebro, N., and Zhang, T. (2014), "Communication-efficient Distributed Optimization using an Approximate Newton-type Method," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1000–1008. [668,671]

Suchard, M., Wang, Q., Chan, C., Frelinger, J., Cron, M., and West, M. (2010), "Understanding GPU Programming for Statistical Computation: Studies in Massively Parallel Massive Mixtures," *Journal of Computational and Graphical Statistics*, 19, 419–438. [668]

Terenin, A., Simpson, D., and Draper, D. (2016), "Asynchronous Gibbs Sampling," arXiv:1509.08999. [668]

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-dimensional Models," *Annals of Statistics*, 42, 1166–1202. [668]

Wang, J., Kolar, M., Srebro, N., and Zhang, T. (2017), "Efficient Distributed Learning with Sparsity," *Proceedings of the 34th International Conference on Machine Learning*, 70, 3636–3645. [669]

Wang, X., and Dunson, D. (2015), "Parallelizing MCMC via Weierstrass Sampler," arXiv:1312.4605. [668,669,676]

Zhang, S., Choromanska, A., and LeCun, Y. (2015), "Deep Learning with Elastic Averaging Sgd," in *Advances in Neural Information Processing Systems*, pp. 685–693. [677]

Zhang, Y., Duchi, J., and Wainwright, M. (2013), "Communication-efficient Algorithms for Statistical Optimization," *Journal of Machine Learning Research*, 14, 3321–3363. [668,669,677]

Zhang, Y., and Lin, X. (2015a), "Communication-efficient Distributed Optimization of Self-concordant Empirical Loss," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, pp. 362–370. [668]

Zhang, Y., and Lin, X. (2015b), "Disco: Distributed Optimization for Self-concordant Empirical Loss," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 362–370. [671]