Communication-Efficient Distributed Statistical Inference

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

November 1, 2016

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 MCMC algorithms even in a non-distributed setting. We present both theoretical analysis and experiments to explore the properties of the CSL approximation.

KEYWORDS: Distributed inference, communication efficiency, 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 et al., 2012, Zhang et al., 2013, Kannan et al., 2014, Kleiner et al., 2014, Shamir et al., 2014, Mackey et al., 2015, Zhang and Lin, 2015, Lee et al., 2015], and Bayesian setting [Suchard et al., 2010, Cleveland and Hafen, 2014, Maclaurin and Adams, 2014, Wang and Dunson, 2015, Neiswanger et al., 2015, Rabinovich et al., 2016, Scott et al., 2016, Terenin et al., 2016]. This literature has focused on data-parallel procedures in which the overall dataset is broken into subsets that are processed independently. To the extent that communication-avoiding procedures have been discussed explicitly, the focus has been on "one-shot" or "embarrassingly parallel" approaches only use one round of communication in which estimators or posterior samples

are obtained in parallel on local machines, are communicated to a center node, and then combined to form a global estimator or approximation to the posterior distribution [Zhang et al., 2013, Lee et al., 2015, Wang and Dunson, 2015, Neiswanger et al., 2015]. In the frequentist setting, most one-shot approaches rely on averaging [Zhang et al., 2013], where the global estimator is the average of the local estimators. Lee et al. [2015] extends 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] show that under certain conditions, these averaging estimators can attain the information-theoretic complexity lower bound—for linear regression, at least $\mathcal{O}(dk)$ bits must be communicated in order 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., 2015].

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 statistical is nonlinear, averaging estimators 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 et al., 2015] tend to suffer from the weight-degeneracy issue (weights collapse to only a few samples) when k is large.

In this paper, 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 non-distributed Bayesian setting, CSL can be used as a computationally-efficient surrogate to the likelihood function by pre-dividing 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, in order to achieve the same accuracy as the global likelihood-based estimator, n can be independent of the total sample size N as long as $\mathcal{O}(\frac{\log N}{\log n})$ iterations are applied, which is *constant* for $n > k$. In contrast, the averaging estimator requires $n \gg \sqrt{N}$. 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 et al., 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 et al., 2014]. This improves over the averaging method of Lee et al. [2015] because it requires p times less computation, and allows for iterative refinement to obtain arbitrarily low optimization error in a logarithmic number of rounds[1].

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.

There is also relevant work in the distributed optimization literature, notably the distributed approximate Newton algorithm (DANE) proposed in Shamir et al. [2014]. Here the idea is to combine gradient descent with a local Newton method; as we will see, a similar algorithmic structure emerges from the CSL framework. DANE has been rigorously analyzed only for quadratic objectives, and indeed the analysis in Shamir et al. [2014] does not imply that DANE can improve over the gradient method for non-quadratic objectives. In contrast, our analysis demonstrates fast convergence rates for a broad class of regular parametric models and high-dimensional models and is not restricted to quadratic objectives. Another related line of work is the iterated Hessian sketch (IHS) algorithm Pilanci and Wainwright [2014] for constrained least-squares optimization. DANE applied to quadratic problems can be viewed as a special case of IHS by choosing the sketching to be a rescaled subsampling matrix; however, the analysis in Pilanci and Wainwright [2014] only applies to a class of low-incoherence sketching matrices that excludes subsampling.

The remainder of this paper 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 para-

[1]We note that during the preparation of this manuscript, we became aware of concurrent work of Wang et al. [2016] who also study a communication-efficient surrogate likelihood. Their focus is solely on the high-dimensional linear model setting.

metric models (Section 3.3). Section 4 presents experimental results in these three settings. All proofs are provided in 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 parametrized 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 is 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:

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

where the fact that $\left\| \nabla^2 \mathcal{L}_N(\overline{\theta}) - \nabla^2 \mathcal{L}_1(\overline{\theta}) \right\|_2 = O\big(n^{-1/2}\big)$ 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:

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

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 $\mathcal{M}_1$.

In anticipation of that theoretical result, we give a heuristic argument for why $\widetilde{\theta}$ is a good estimator. 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[2]

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

[2]There is no Taylor's theorem for vector-valued functions, but we formalize this heuristic in Section 3.1.

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 non-singular. Now let is 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,

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

which is of order $O_p(N^{-1/2})$ as long as $\|\theta^* - \overline{\theta}\|_2 = O_p\big(k^{-1/2}\big)$ 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 $n > k$.

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^*. Under such an assumption, regularized estimators are known to be effective for estimating θ. For concreteness, we focus on the sparsity assumption that most components of the d-dim vector θ^* is zero, and consider the ℓ_1-regularized estimator,

$$\widehat{\theta} := \arg\min_{\theta \in \Theta} \big\{ \mathcal{L}_N(\theta) + \lambda \|\theta\|_1 \big\}$$

, as the benchmark estimator that we want to approximate, where λ is the regularization parameter. In the distributed inference framework, we consider the following estimator obtained from the surrogate loss function $\widetilde{\mathcal{L}}$:

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

In Section 3.2, we show that $\widetilde{\theta}$ achieves the same rate of convergence as the benchmark estimator $\widehat{\theta}$ 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. For convenience, we also use the notation $\pi(\theta)$ to denote the pdf of the prior distribution at point θ. According to Bayes' theorem, the posterior distribution satisfies

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

The loss function $\mathcal{L}$ corresponds to the negative log-likelihood function for the statistical model $\{\mathbb{P}_\theta : \theta \in \Theta\}$ and $\mathcal{L}_N(\theta)$ is the global negative log-likelihood associated with the observations Z_1^N. The posterior distribution $\pi(\theta \,|\, Z_1^N)$ can be used to conduct statistical inference. For example, we may construct an estimator of θ as the posterior expectation and use the highest posterior region as a credible region for this estimator. Since the additive constant C in expression (6) can be absorbed into the normalizing constant, we may use the surrogate posterior distribution,

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

, to approximate the global posterior distribution $\pi(\theta \,|\, Z_1^N)$. In Section 3.3, we formalize this argument and show that this surrogate posterior gives a good approximation the global posterior.

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.

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 paper.

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 paper, 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, $\theta^* \in \mathrm{int}(\Theta)$ 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 $(\mu_-,\, \mu_+)$, 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 $U(\rho) = \{\theta \in \mathbb{R}^d \,|\, \|\theta - \theta^*\|_2 \leq \rho\} \subset \Theta$ 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

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

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 $\widetilde{\mathcal{L}}(\theta)$, i.e.,

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

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 3.1. *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*

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

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 3.1, 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

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

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 3.2. *Under the conditions of Theorem 3.1, we have*

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

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

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

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

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

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

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

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

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

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

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 $\|\overline{\theta} - \theta^*\|_2$ to be $\mathcal{O}(n^{-1/2})$.

Theorem 3.3. *Suppose that Assumptions PA-PD hold and the initial estimator $\overline{\theta}$ satisfies $\|\overline{\theta} - \theta^*\|_2 \leq \min\left\{\rho, (16M)^{-1}(1-\rho)\,\mu_-\right\}$. Then the local one-step estimator $\widetilde{\theta}^H$ satisfies*

$$\|\widetilde{\theta}^H - \widehat{\theta}\|_2 \leq C_2' \left(\|\overline{\theta} - \widehat{\theta}\|_2 + \|\widehat{\theta} - \theta^*\|_2 + \|\nabla^2 \mathcal{L}_1(\theta^*) - \nabla^2 \mathcal{L}_N(\theta^*)\|_2\right) \|\overline{\theta} - \widehat{\theta}\|_2,$$

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

The analogue of Corollary 3.2 can also be stated for $\widetilde{\theta}^H$.

Iterative local estimation algorithm: Theorem 3.1 (Theorem 3.3) 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 $\|\overline{\theta} - \widehat{\theta}\|_2 = O_p(n^{-1/2})$, or equivalently, $\|\overline{\theta} - \theta^*\|_2 = O_p(n^{-1/2})$. We refer to such an algorithm as an Iterative Local Estimation Algorithm (ILEA, see Algorithm 1). In each iteration of ILEA, we set $\overline{\theta}$ as 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 3.1 (or Theorem 3.3) guarantees, with high probability, the error bound

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

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, then 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.

Initialize $\theta^{(0)} = \overline{\theta}$;

for $t = 0, 1, \ldots, T-1$ **do**

  Transmit the current iterate $\theta^{(t)}$ to local machines $\{\mathcal{M}_j\}_{j=1}^k$;

  for $j = 1 : k$ **do**

    Compute the local gradient $\nabla \mathcal{L}_j(\theta^{(t)})$ at machine $\mathcal{M}_j$;

    Transmit the local gradient $\nabla \mathcal{L}_j(\theta^{(t)})$ to machine $\mathcal{M}_1$;

  end

  Calculate the global gradient $\nabla \mathcal{L}_N(\theta^{(t)}) = \frac{1}{k}\sum_{j=1}^k \nabla \mathcal{L}_j(\theta^{(t)})$ in Machine $\mathcal{M}_1$;

  Form the surrogate function $\widetilde{\mathcal{L}}^t(\theta) = \mathcal{L}_1(\theta) - \langle \theta, \nabla \mathcal{L}_1(\theta^{(t)}) - \nabla \mathcal{L}_N(\theta^{(t)}) \rangle$;

  Do one of the following in Machine $\mathcal{M}_1$:

  (1). Update $\theta^{(t+1)} \in \arg\min_{\theta \in \Theta} \widetilde{\mathcal{L}}^t(\theta)$; `// Exactly minimizing surrogate function` $\widetilde{\mathcal{L}}$

  (2). Update $\theta^{(t+1)} = \theta^{(t)} - \nabla^2 \mathcal{L}_1(\theta^{(t)})^{-1} \nabla \mathcal{L}_N(\theta^{(t)})$; `// One-step quadratic approximation`

end

return $\theta^{(T)}$

Algorithm 1: Iterative local estimation

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 3.4 in Section A.4)

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

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,

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

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^*.

The following 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 also have a consistent estimator $\widetilde{\Sigma}$ of Σ using only the local information in Machine M_1. Therefore, 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 3.4. *Under the same set of assumptions in Theorem 3.1, 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 $\|\overline{\theta} - \theta^*\|_2 = o_P\big(\sqrt{\frac{n}{N}}\big)$, then

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

Moreover, the following plug-in estimator:

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

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

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

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 3.3.

Corollary 3.4 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 $\widetilde{\mathcal{L}}(\theta)$ (cf. 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 $\widehat{\Sigma}$ and $\widehat{\Sigma}'$ 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^* is 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:

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

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

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

where δ is some positive constant independent of n.

As the name suggested, restricted strongly 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 Hessian at radius R: for all $\delta \in C(S) \cap B_R(\theta^*)$,

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

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

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

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 $\lambda \ge 2\|\nabla \mathcal{L}_N(\theta^*)\|_\infty$. Since the estimation error upper bound provided by Theorem 3.5 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

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

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 may also consider an iterative estimation procedure analogous to Algorithm 1 in order to provide higher-order estimation accuracy for the communication-efficient regularized estimator $\widetilde{\theta}$. The convergence rate can be analyzed by inducting on Theorem 3.5. We now apply Theorem 3.5 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

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

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

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

We consider a random design where x_{ij} is i.i.d. A-sub-Gaussian; that is, for all $\alpha \in \mathbb{R}^d$,

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

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 $n \ge C_0 s \log d$ for some constant $C_0 > 0$ depending on the minimal eigenvalue of Σ [Raskutti et al., 2010]. For linear models, the Lipschitz constant M in Assumption HB is zero and therefore HB is also satisfied.

Theorem 3.6. *If x_{ij} is A-sub-Gaussian, Σ is strictly positive definite and $n \ge C_0 s \log d$, then with probability at least $1 - c_1 \exp\{-c_2 n\}$, it holds that*

$$\|\widetilde{\theta} - \theta^*\|_2 \le C_1 A \sqrt{\frac{s \log d}{N}} + C_1 A \sqrt{\frac{s \log d}{n}}\, \|\overline{\theta} - \theta^*\|_1.$$

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

$$\|\widetilde{\theta} - \theta^*\|_2 \sim C_1 A \sqrt{\frac{s \log d}{N}} + C_3 \frac{s^{3/2} \log d}{n}$$

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 3.6 shows that our approximated estimator $\widetilde{\theta}$ is nearly minimax-optimal if $n \geq Cs\sqrt{N \log d}$ 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) to boost the estimation accuracy and obtain a minimax-optimal estimator as we remarked after Theorem 3.5.

Example (Generalized linear models): In this section, we apply Theorem 3.5 to generalized linear models with a ℓ_1-regularizer. We begin with some background on generalized linear models. Recall that the data is $z_{ij} = (x_{ij}, y_{ij})$, where y_{ij} is the response and x_{ij} is the d-dim covariate vector. A generalized linear model assumes the conditional distribution of y_{ij} given x_{ij} to be

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

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 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 3.7. *Assume that for some constants (A, B, m, L), x_{ij} is i.i.d. A-sub-Gaussian, $\|x_{ij}\|_\infty \leq B$, and $mI \preceq \Sigma = \mathbb{E}[x_{ij} x_{ij}^T] \preceq LI$. Then with probability at least $1 - c_1 \exp\{-c_2 n\}$, it holds that*

$$\|\widetilde{\theta} - \theta^*\|_2 \leq C_1 A \sqrt{\frac{s \log d}{N}} + C_1 A \sqrt{\frac{s \log d}{n}} \|\overline{\theta} - \theta^*\|_1 + C_1 A \sqrt{s} \|\overline{\theta} - \theta^*\|_2^2.$$

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

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

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

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

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 et al., 2013] 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 et al., 2015] in distributed Bayesian inference utilizes the decomposition

$$\pi(\theta \,|\, Z_1^N) = D \prod_{j=1}^{k} \exp\big\{ - n\mathcal{L}_j(\theta) \big\}, \tag{16}$$

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

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

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 $\pi_N(\theta) := \pi(\theta \,|\, Z_1^N)$ 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 $\overline{\theta}$, e.g. 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 $\widetilde{\pi}_N(\theta) \propto \exp\big(-N\widetilde{\mathcal{L}}(\theta)\big)\,\pi(\theta)$, 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 $\pi(\cdot \,|\, Z_1^N)$ as long as the initial estimator $\overline{\theta}$ is reasonably close to θ^*.

Theorem 3.8. *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*

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

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

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

This illustrates that we may choose $k = N/n$ up to $o\big(N^{1/2}(\log N)^{-1}\big)$ while still maintaining $\big\|\widetilde{\pi}_N - \pi_N\big\|_1 = o_p(1)$. 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

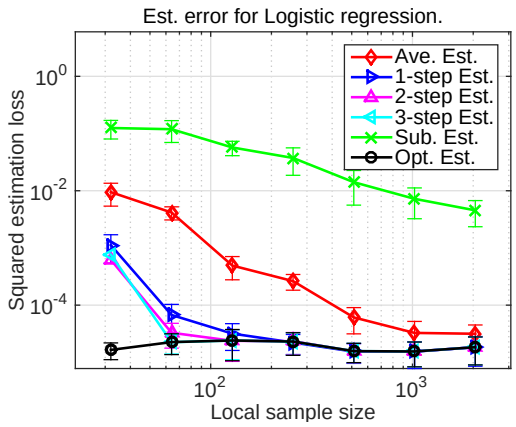
In logistic regression, i.i.d. observations $Z_1^N = \{Z_{ij} = (X_{ij}, Y_{ij}) : i = 1, \ldots, n. j = 1, \ldots, k\}$ are generated from the model

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

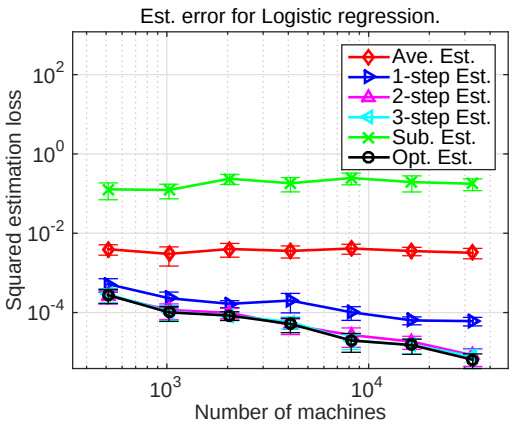
In our simulation, the true regression coefficient θ^* is a d-dim vector with $d \in \{2, 10, 50\}$ and the d-dim covariate vector X_{ij} is independently generated from $\mathcal{N}(0, I_d)$. For each replicate of the simulation, we uniformly sample the parameter θ^* from the d-dim unit cube $[0, 1]^d$.

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 $\overline{\theta}$. We also implement the iterative local estimation algorithm to produce 2-step and 3-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 ($d = 2$), 256 ($d = 10$) or 2048 ($d = 50$), and the number of machines k varies from 10^2 to 10^4.

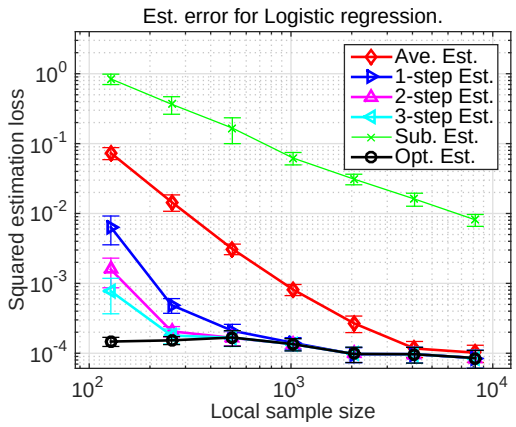
Figure 1 reports the results. In plots (a), (c) and (d), the total sample size N is fixed and therefore the estimation error associated with the global estimate θ^{global} remains approximately fixed as n varies. As expected, the remaining estimators exhibit a rapid decay in the estimation error as the local sample size n grows. Our communication-efficient estimators yield the best performance among the distributed estimators. When n is sufficiently large, the 1-step, 2-step and 3-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 3.3—the contraction coefficient $\|\theta^{(t+1)} - \theta^{global}\|_2 / \|\theta^{(t)} - \theta^{global}\|_2$ is dominated by the sum of two terms: the initial estimation error $\|\theta^{(t)} - \theta^{global}\|_2$ 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.



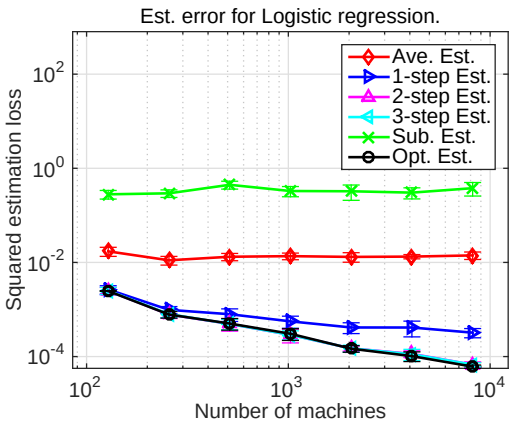
(a) $d = 2$ and $N = 524288$.



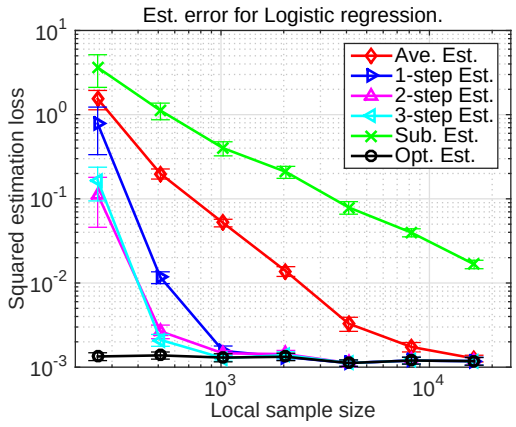
(b) $d = 2$ and $n = 64$.



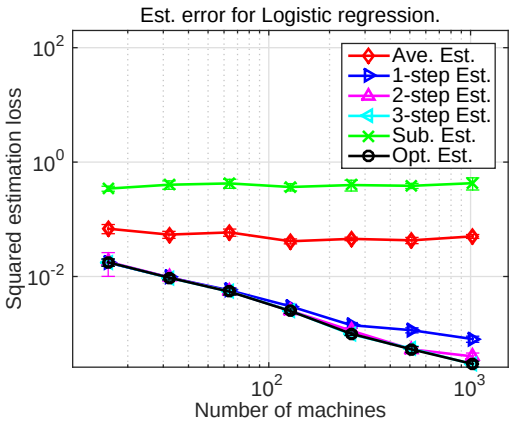
(c) $d = 10$ and $N = 524288$.



(d) $d = 10$ and $n = 256$.



(e) $d = 50$ and $N = 524288$.



(f) $d = 50$ 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), (c) and (e), we change the local sample size n while fixing the total sample size N (number of machines $k = N/n$) for dimension $d \in \{2, 10, 50\}$. In plots (b), (d) and (f), we change the number of machines k while fixing the local sample size n (total sample size $N = nk$) under dimension $d \in \{2, 10, 50\}$.

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 fails to predict, the 1-step, 2-step and 3-step estimators still have better performance than $\widehat{\theta}^A$ and θ^{sub}. In plots (b), (d) and (e), 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 is available. This is consistent with theoretical results in Zhang et al. [2013] that require $k \gg n$ 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 (d), even for k as large as 10^4 (much larger than the local sample size $n \sim 10^2$), the 2-step estimate $\theta^{(2)}$ already achieves the same level of estimation accuracy as the global estimator θ^{global}.

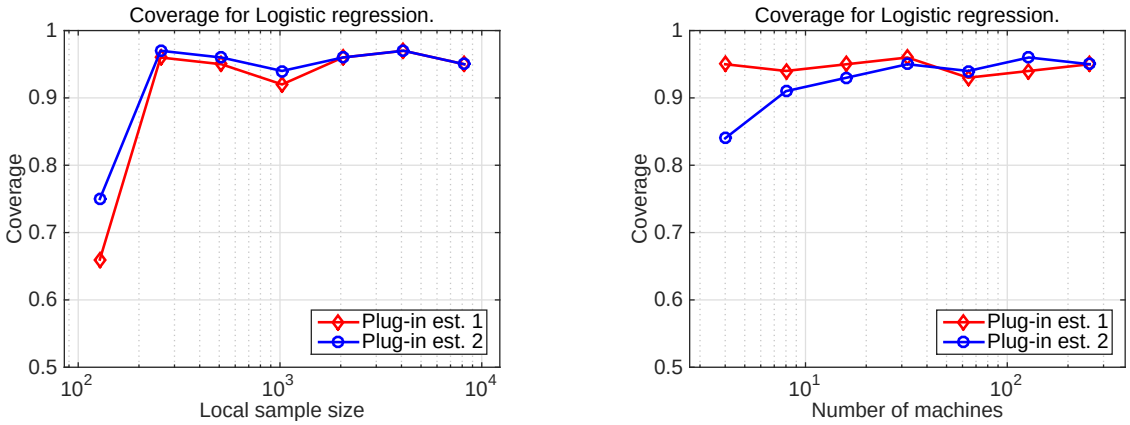
We now assess the performance of the inference procedures based on the plug-in estimators $\widetilde{\Sigma}$ and $\widetilde{\Sigma}'$ under the logistic model (17). We use $\widetilde{\Sigma}$ or $\widetilde{\Sigma}'$ and the 3-step estimator $\theta^{(3)}$ to construct a 95% confidence interval (CI)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

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

The coverage of the CI based on 100 trials is calculated. Figure 2 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 CI's 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 3.4.

4.2 Distributed sparse linear regression

We evaluate the CSL estimator on the sparse linear regression problem. The data is generated as $y_{ij} = X_{ij}^T\theta^* + \epsilon_{ij}$, where $i \in [n]$ and $j \in [k]$. The covariates X_{ij} are i.i.d. $\mathcal{N}(0, 1)$, the noise ϵ_{ij} is



(a) $d = 10$ and $N = 524288$.

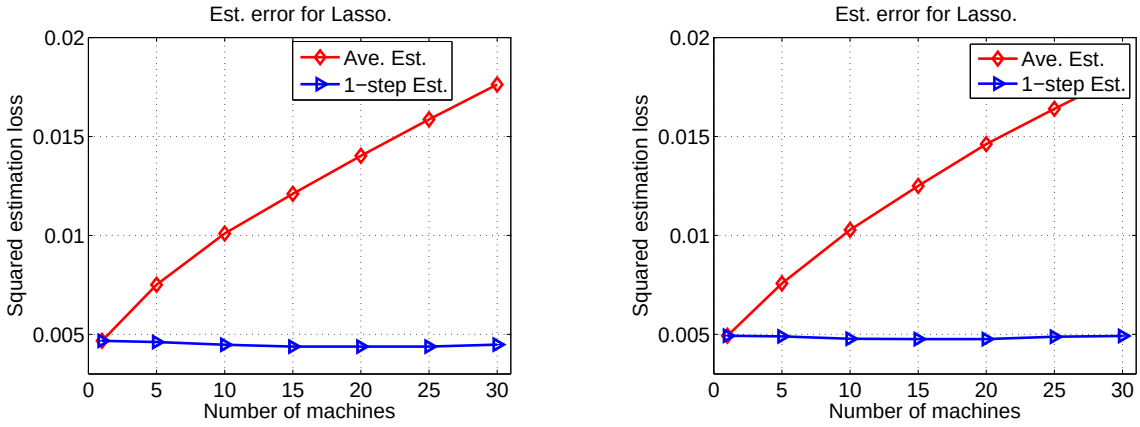
(b) $d = 10$ and $n = 256$.

Figure 2: 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 3-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 $N = nk$).

i.i.d. $\mathcal{N}(0,1)$, 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, 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 $\frac{s \log d}{nk}$.

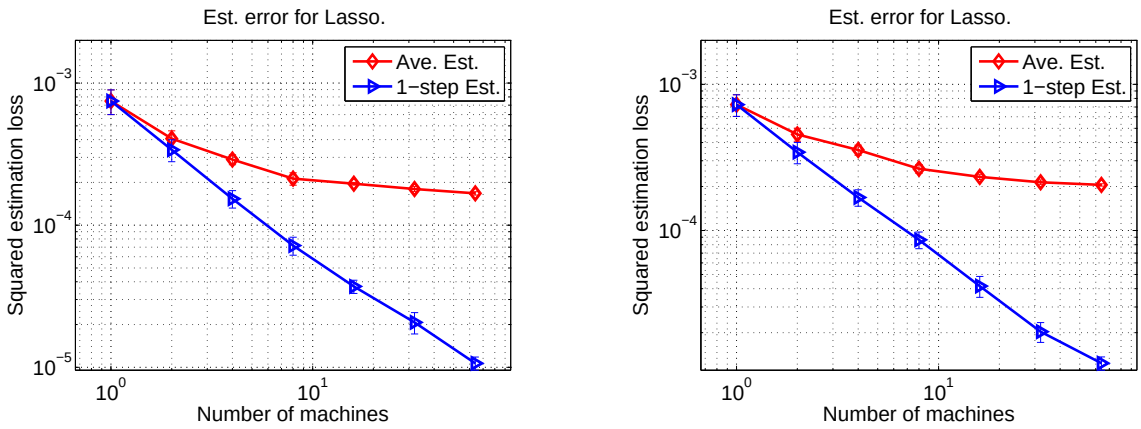


$(N, d, s) = (21600, 5000, 20)$ $(N, d, s) = (21600, 10000, 20)$

Figure 3: As $k \in \{1, 5, 10, 15, 20, 25, 30\}$ increases, the local data size n decreases, but the one-step CSL estimator has constant error. The averaging estimator error increases, since k decreases.

4.3 Distributed Bayesian inference

Our synthetic dataset is generated from the logistic model (17) for dimension $d \in \{2, 10, 50\}$. We use the 3-step estimator $\theta^{(3)}$ in Section 3.1 as the initial estimator $\overline{\theta}$ and implement the Bayesian



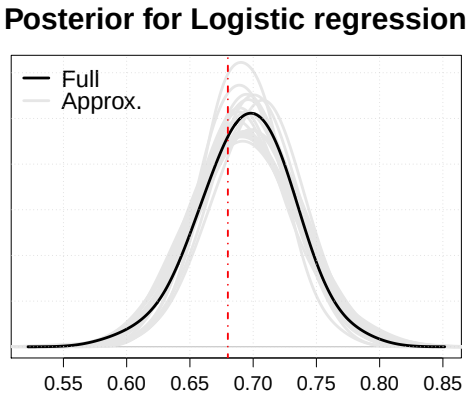
$(n, d, s) = (800, 1000, 20)$ $(n, d, s) = (800, 2000, 20)$

Figure 4: (a) As $k \in \{1, 2, 4, 8, 16, 32, 64\}$ increases, the mean-squared error of the one-step CSL estimator decreases. For the averaging estimator, the mean-squared error does not decrease significantly for large values of k.

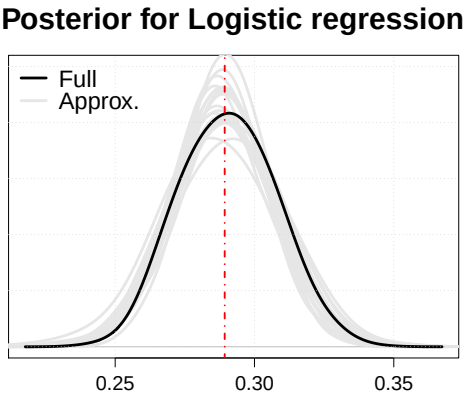
procedures based on the (approximated) posterior distribution $\pi_n(\theta)$ and $\widetilde{\pi}_N(\theta)$ by sampling a Markov Chain Monte Carlo algorithm. We use the Metropolis algorithm, where at each iteration the proposal distribution for θ is a d-dim Gaussian distribution centered at the current iterate $\theta^{(t)}$. In each case, we run the Markov chain for 20000 iterations and treat the first half as burn-in. Figure 5 plots the (approximated) marginal posterior distributions 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). 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.

5. REAL DATA APPLICATION

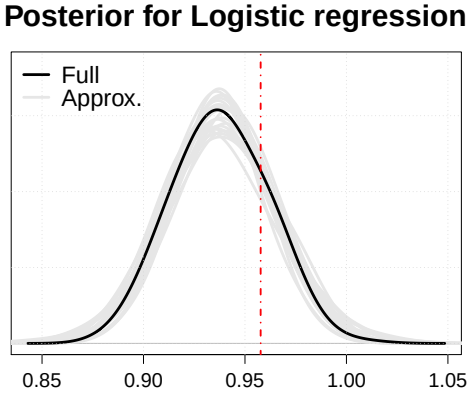
We apply distributed logistic regression to a computer-vision dataset [Rajen and Abhinav, 2012]. The goal is to predict whether a given color sample described by its B, G, R values (each ranges



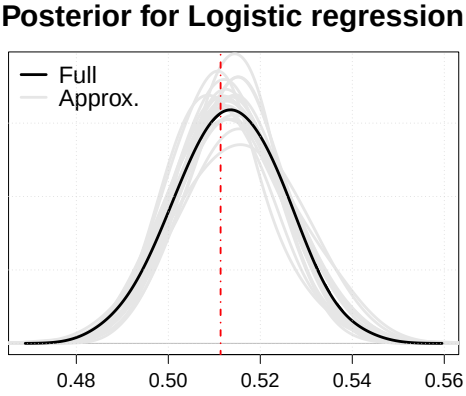
(a) $(d,\, n,\, k) = (2,\, 64,\, 64)$.



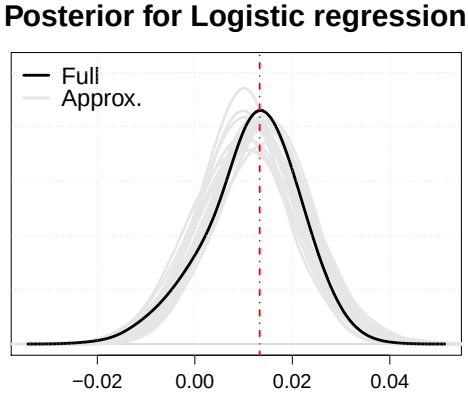
(b) $(d,\, n,\, k) = (2,\, 64,\, 256)$.



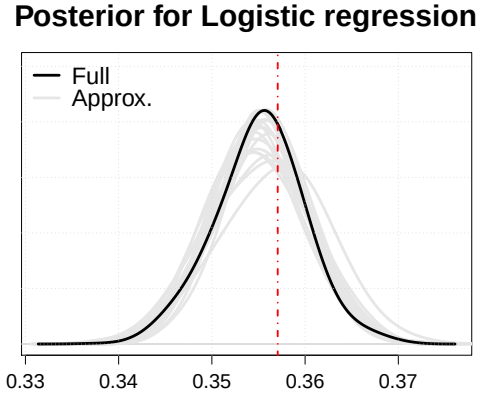
(c) $(d,\, n,\, k) = (10,\, 256,\, 64)$.



(d) $(d,\, n,\, k) = (10,\, 256,\, 256)$.



(e) $(d,\, n,\, k) = (50,\, 2048,\, 64)$.



(f) $(d,\, n,\, k) = (50,\, 2048,\, 256)$.

Figure 5: Marginal posterior distribution of the first component θ_1 of θ for logistic regression for dimension $d \in \{2, 10, 50\}$ are shown. In each plot, 20 approximations (grey 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^*.

from $0-255$) corresponds to a skin sample or non-skin sample. The dataset is generated using skin textures from face images of diverse of age, gender, and race. The total sample size is $245,057$, out of which $50,859$ images are skin samples and $194,198$ are non-skin samples. The dataset contains three features—B, G, R values of the color, and a 0-1 response variable indicates whether the sample is non-skin (0) or skin (1). We randomly split the dataset into a training set of size $N = 200,000$ and a testing set $N_0 = 45057$, and use B-spline transforms (df= 15) for each feature as predictors to allow a nonlinear dependence between the response and features. Therefore, the dimension of the covariate X is $d = 45$.

We randomly split the training set into $k_0 = 100$ subsets, each of size $n = 2000$. We apply our distributed M-estimation method for logistic regression to a training set with $k \in \{20, 40, 60, 80, 100\}$ subsets, and test the fitted model to the testing set. We then exactly minimize the surrogate function to form the 1-step estimator $\theta^{(1)}$, using the averaging estimator $\widehat{\theta}^A$ [Zhang et al., 2013] as our initial estimator. We also implement the iterative local estimation algorithm to produce 2-step and 3-step estimators $\theta^{(2)}$ and $\theta^{(3)}$ by iteratively applying the one-step estimation procedure. Figure 6 plots the misclassification rate versus the number of subsets used. As we can see, the 1-step estimator yields significant gains in prediction performance over the initial averaging estimator, and both the 2-step and 3-step estimators have similar prediction performance as the 1-step estimator. This suggests that for the skin dataset and our split setting, the one-step approximation of the likelihood function is already adequate.

6. DISCUSSION

We have presented a Communication-efficient Surrogate Likelihood (CSL) framework for solving distributed statistical inference problems. We applied this methodology to three problem domains: low-dimensional M-estimation, high-dimensional regularized estimation and low-dimensional Bayesian inference. Our results demonstrate that the general idea of constructing a surrogate to the negative log-likelihood function (or general loss function) is viable for communication-limited statistical inference. We also believe that the approach can prove useful for "big-data" problems on a single machine, when the sample size is large and the calculation of the likelihood function is expensive.

There are several directions for future research in this area. We would like to find methods

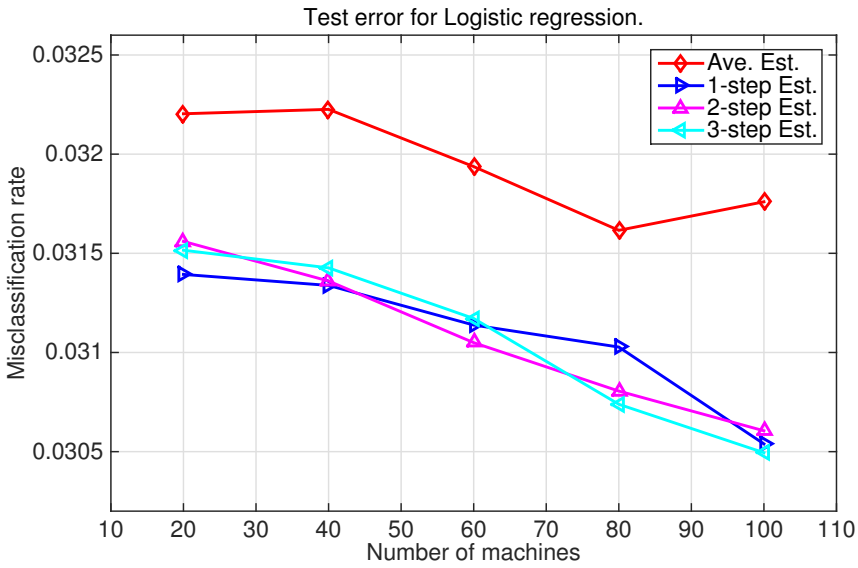


Figure 6: Distributed logistic regression for the skin dataset.

that permit high-dimensional distributed Bayesian inference. We would also like to find a sharp theoretical lower bound on the local sample size n needed for the final estimator to remain optimal (for example, in the minimax sense). It is also worthwhile to consider ensemble and hierarchical versions of the CSL method, in which multiple machines aggregate local results.

REFERENCES

P. Bissiri, C. Holmes, and S. Walker. A general framework for updating belief distributions. *arXiv:1306.6430*, 2013.

Mark Braverman, Ankit Garg, Tengyu Ma, Huy Nguyen, and David Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. *arXiv:1506.07216*, 2015.

S. Bubeck. Theory of convex optimization for machine learning. *Foundations and Trends in Machine Learning*, 8, 2015.

V. Chernozhukov and H. Hong. An MCMC approach to classical estimation. *Journal of Econometrics*, 115(2):293 – 346, 2003.

W. Cleveland and R. Hafen. Divide and recombine (D&R): Data science for large complex data. *Statistical Analysis and Data Mining*, 7:425–433, 2014.

J. Demmel, L. Grigori, M. Hoemmen, and J. Langou. Communication-optimal parallel and sequential QR and LU factorizations. *SIAM Journal on Scientific Computing*, 34:206–239, 2012.

J. Duchi, M. Jordan, M. Wainwright, and Y. Zhang. Optimality guarantees for distributed statistical estimation. *arXiv:1405.0782*, 2015.

John C. Duchi, Alekh Agarwal, and Martin J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57:592–606, 2012.

A. Garg, T. Ma, and H. Nguyen. On communication cost of distributed statistical estimation and dimensionality. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2014.

R. Kannan, S. Vempala, and D. Woodruff. Principal component analysis and higher correlations for distributed data. In *the 27th Conference on Learning Theory*, pages 1040–1057, 2014.

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

J. Lee, Y. Sun, Q. Liu, and J. Taylor. Communication-efficient sparse regression: a one-shot approach. *arXiv:1503.04337*, 2015.

L. Mackey, A. Talwalkar, and M. I. Jordan. Distributed matrix completion and robust factorization. *Journal of Machine Learning Research*, 16:913–960, 2015.

D. Maclaurin and R. Adams. Firefly Monte Carlo: Exact MCMC with subsets of data. *arXiv:1403.5693*, 2014.

S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

W. Neiswanger, C. Wang, and E. Xing. Asymptotically exact, embarrassingly parallel MCMC. *arXiv:1311.4780*, 2015.

M. Pilanci and M. Wainwright. Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares. *arXiv:1411.0347*, 2014.

M. Rabinovich, E. Angelino, and M. Jordan. Variational consensus Monte Carlo. In *Advances in Neural Information Processing Systems*, Red Hook, NY, 2016. Curran Associates.

G. Rajen and D. Abhinav. Skin segmentation dataset. *UCI Machine Learning Repository*, 2012.

G. Raskutti, M. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

S. Scott, A. Blocker, F. Bonassi, H. Chipman, E. George, and R. McCulloch. Bayes and big data: the consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11:78–88, 2016.

O. Shamir, N. Srebro, and T. Zhang. Communication-efficient distributed optimization using an approximate Newton-type method. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1000–1008, 2014.

M. Suchard, Q. Wang, C. Chan, J. Frelinger, M. Cron, and M. West. Understanding gpu programming for statistical computation: Studies in massively parallel massive mixtures. *Journal of Computational and Graphical Statistics*, 19:419–438, 2010.

A. Terenin, D. Simpson, and D. Draper. Asynchronous Gibbs sampling. *arXiv:1509.08999*, 2016.

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

J. Wang, M. Kolar, N. Srebro, and T. Zhang. Efficient distributed learning with sparsity. *arXiv:1605.07991*, 2016.

X. Wang and D. Dunson. Parallelizing MCMC via Weierstrass sampler. *arXiv:1312.4605*, 2015.

Y. Zhang and X. Lin. Communication-efficient distributed optimization of self-concordant empirical loss. *arXiv:1501.00263*, 2015.

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

APPENDIX A. PROOFS OF MAIN RESULTS

A.1 Proof of Theorem 3.1

For $j = 1, \ldots, k$, let $M_j = \frac{1}{n}\sum_{i=1}^{n} M(z_{ij})$ and $\delta_\rho = \min\{\rho, \rho\mu_-/4M\}$. Consider the following "good events":

$$\mathcal{E}_0 := \left\{ \|\widehat{\theta} - \theta^*\|_2 \le \min\left\{ \frac{\rho\mu_-}{8M}, \frac{(1-\rho)\mu_-\delta_\rho}{8\mu_+}, \sqrt{\frac{(1-\rho)\mu_-\delta_\rho}{16M}} \right\} \right\}, \qquad \text{and}$$
$$\mathcal{E}_j := \left\{ M_j \le 2M, \, \|\nabla^2 \mathcal{L}_j(\theta^*) - I(\theta^*)\|_2 \le \frac{\rho\mu_-}{4}, \, \|\nabla \mathcal{L}_j(\theta^*)\|_2 \le \frac{(1-\rho)\mu_-\delta_\rho}{4} \right\}.$$

Before proving the claimed error bound for $\widetilde{\theta}$, we state two auxiliary results that are used in the proof. The first result provides control on the probability of a bad event $\bigcup_{j=0}^{k} \mathcal{E}_j^c$, which is proved in Appendix B.1.

Lemma A.1. *Under Assumptions PA-PD, we have*

$$\mathbb{P}\Big(\bigcup_{j=0}^{k} \mathcal{E}_j^c \Big) \le \left(c_1 + c_2 \left(\log 2d\right)^{16} L^{16} + c_3 \, G^{16}\right) \frac{k}{n^8}.$$

Here c_j $(j = 1, 2, 3)$ are constants independent of (n, k, N, d, G, L).

The second result characterizes the error bound $\|\widetilde{\theta} - \widehat{\theta}\|_2$ in terms of the gradient norm $\|\nabla\widetilde{\mathcal{L}}(\widehat{\theta})\|_2$ at $\widehat{\theta}$, which formalizes the heuristic argument in Section 2.2. Its proof is provided in Appendix B.2.

Lemma A.2. *Suppose that Assumptions PA-PD hold. Then under event $\mathcal{E}_0 \cap \mathcal{E}_1$ we have*

$$\|\widetilde{\theta} - \widehat{\theta}\|_2 \le \frac{2\,\|\nabla\widetilde{\mathcal{L}}(\widehat{\theta})\|_2}{(1-\rho)\mu_-}.$$

Therefore, it remains to prove a high-probability upper bound for the gradient norm $\|\nabla\widetilde{\mathcal{L}}(\widehat{\theta})\|_2$. A simple calculation yields

$$\nabla\widetilde{\mathcal{L}}(\widehat{\theta}) = \nabla\mathcal{L}_1(\widehat{\theta}) - \nabla\mathcal{L}_1(\overline{\theta}) + \nabla\mathcal{L}_N(\overline{\theta}). \tag{A.1}$$

By the optimality of the global empirical risk minimizer $\widehat{\theta}$, we have

$$\nabla\mathcal{L}_N(\widehat{\theta}) = 0.$$

By adding and subtracting $\nabla_N(\widehat{\theta})$ in equation (A.1), we obtain

$$\nabla\widetilde{\mathcal{L}}(\widehat{\theta}) = \big(\nabla\mathcal{L}_1(\widehat{\theta}) - \nabla\mathcal{L}_1(\overline{\theta})\big) - \big(\nabla\mathcal{L}_N(\widehat{\theta}) - \nabla\mathcal{L}_N(\overline{\theta})\big). \tag{A.2}$$

By the integral form of Taylor's expansion, we have that for any $j \in \{1, \ldots, k\}$,

$$\nabla \mathcal{L}_j(\widehat{\theta}) - \nabla \mathcal{L}_j(\overline{\theta}) = H_j \, (\widehat{\theta} - \overline{\theta}),$$

where $H_j = \int_0^1 \nabla^2 \mathcal{L}_j\big(\overline{\theta} + t(\widehat{\theta} - \overline{\theta})\big) \, dt$ satisfies

$$\|\!| H_j - \nabla^2 \mathcal{L}_j(\theta^*) |\!\|_2 \leq 2M \, (\|\overline{\theta} - \widehat{\theta}\|_2 + \|\widehat{\theta} - \theta^*\|_2)$$

under event $\mathcal{E}_j$. Combining the three preceding displays, we obtain that under event $\bigcap_{j=0}^k \mathcal{E}_k$,

$$\begin{aligned}
\|\nabla \widetilde{\mathcal{L}}(\widehat{\theta})\|_2 &\leq \|\!| H_1 - \nabla^2 \mathcal{L}_1(\theta^*) |\!\|_2 \, \|\widehat{\theta} - \overline{\theta}\|_2 + \frac{1}{k} \sum_{j=1}^k \|\!| H_j - \nabla^2 \mathcal{L}_j(\theta^*) |\!\|_2 \, \|\widehat{\theta} - \overline{\theta}\|_2 \\
&\quad + \|\!| \nabla^2 \mathcal{L}_1(\theta^*) - \nabla^2 \mathcal{L}_N(\theta^*) |\!\|_2 \, \|\widehat{\theta} - \overline{\theta}\|_2 \\
&\leq \big(2M \, \|\widehat{\theta} - \overline{\theta}\|_2 + 2M \, \|\widehat{\theta} - \theta^*\|_2 + \|\!| \nabla^2 \mathcal{L}_1(\theta^*) - \nabla^2 \mathcal{L}_N(\theta^*) |\!\|_2\big) \, \|\widehat{\theta} - \overline{\theta}\|_2.
\end{aligned}$$

Combining Lemma A.1 and the above display yields the claimed error bound on $\|\widetilde{\theta} - \widehat{\theta}\|_2$.

A.2 Proof of Corollary 3.2

Recall the definitions of the events $\{\mathcal{E}_j\}_{j=0}^k$ in Section A.1. In the remaining of this proof, we use C to denote some constant independent of (n, k, N), whose magnitude may change from line to line. We need the following auxiliary result, whose proof is provided in Appendix B.3.

Lemma A.3. *Under event $\bigcap_{j=0}^k \mathcal{E}_j$, we have*

$$\begin{aligned}
&\Big\| \widehat{\theta} - \theta^* - I(\theta^*)^{-1} \nabla \mathcal{L}_N(\theta^*) \Big\|_2 \\
&\quad \leq \frac{2}{(1-\rho)\mu_-} \|\!| \nabla^2 \mathcal{L}_N(\theta^*) - I(\theta^*) |\!\|_2 \, \|\nabla \mathcal{L}_N(\theta^*)\|_2 + \frac{8M}{(1-\rho)^2 \mu_-^2} \, \|\nabla \mathcal{L}_N(\theta^*)\|_2^2.
\end{aligned}$$

Combining this Lemma, inequality (A.7) in Appendix B.1 and Theorem 3.1, we obtain that under event $\bigcap_{j=0}^k \mathcal{E}_j$,

$$\begin{aligned}
\Big\| \widetilde{\theta} - \theta^* - I(\theta^*)^{-1} \nabla \mathcal{L}_N(\theta^*) \Big\|_2 &\leq \Big\| \widehat{\theta} - \theta^* - I(\theta^*)^{-1} \nabla \mathcal{L}_N(\theta^*) \Big\|_2 + \|\widetilde{\theta} - \widehat{\theta}\|_2 \\
&\leq C \, \|\!| \nabla^2 \mathcal{L}_N(\theta^*) - I(\theta^*) |\!\|_2 \, \|\nabla \mathcal{L}_N(\theta^*)\|_2 + C \, \|\nabla \mathcal{L}_N(\theta^*)\|_2^2 \\
&\quad + C \, \big(\|\overline{\theta} - \widehat{\theta}\|_2 + \|\nabla \mathcal{L}_N(\theta^*)\|_2 + \|\!| \nabla^2 \mathcal{L}_1(\theta^*) - \nabla^2 \mathcal{L}_N(\theta^*) |\!\|_2\big) \, \|\overline{\theta} - \widehat{\theta}\|_2.
\end{aligned}$$

Now by applying Hölder's inequality and Lemma B.1 in Appendix B.1, we obtain

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\Big(\widetilde{\theta}-\theta^*-I(\theta^*)^{-1}\,\nabla\mathcal{L}_N(\theta^*)\Big)\,I\Big(\bigcap_{j=0}^{k}\mathcal{E}_j\Big)\Big\|_2^2\Big]\\
\le\,& C\,\sqrt{\mathbb{E}[\|\nabla^2\mathcal{L}_N(\theta^*)-I(\theta^*)\|_2^4]\;\mathbb{E}[\|\nabla\mathcal{L}_N(\theta^*)\|_2^4]}+C\,\mathbb{E}[\|\nabla\mathcal{L}_N(\theta^*)\|_2^4]\\
&+C\,\mathbb{E}[\|\overline{\theta}-\widehat{\theta}\|_2^4]+C\,\sqrt{\mathbb{E}[\|\nabla^2\mathcal{L}_1(\theta^*)-\nabla^2\mathcal{L}_N(\theta^*)\|_2^4]+\mathbb{E}[\|\nabla\mathcal{L}_N(\theta^*)\|_2^4]}\,\sqrt{\mathbb{E}[\|\overline{\theta}-\widehat{\theta}\|_2^4]}\\
\le\,&\frac{C}{N^2}+\frac{C}{n}\,\min\Big\{\frac{1}{n},\,\sqrt{\mathbb{E}[\|\overline{\theta}-\widehat{\theta}\|_2^4]}\Big\}.
\end{aligned}
$$

Combining this with bound (A.8) in Appendix B.1 on $\mathbb{P}(\bigcup_{j=0}^{k}\mathcal{E}_j^c)$, we obtain that under Assumption PA,

$$
\mathbb{E}\Big[\Big\|\widetilde{\theta}-\theta^*-I(\theta^*)^{-1}\,\nabla\mathcal{L}_N(\theta^*)\Big\|_2^2\Big]\le\frac{C}{N^2}+\frac{C}{n}\,\min\Big\{\frac{1}{n},\,\sqrt{\mathbb{E}[\|\overline{\theta}-\widehat{\theta}\|_2^4]}\Big\}+\frac{C\,k}{n^8},
$$

which implies the claimed bound on $\mathbb{E}[\|\widetilde{\theta}-\theta^*\|_2^2]$.

A.3 Proof of Theorem 3.3

Before analyzing the one-step Newton-Raphson estimator θ^H, we establish some auxiliary results. Recall that for $j=1,\ldots,k$, let $M_j=\frac{1}{n}\sum_{i=1}^{n}M(z_{ij})$ and $\delta_\rho=\min\{\rho,\,\rho\mu_-/4M\}$. Analogously to the definition of the events $\mathcal{E}_j$ $(j=0,1\ldots,k)$ in Section A.1, we define the following "good events":

$$
\begin{aligned}
\mathcal{E}_0'&:=\Big\{\|\widehat{\theta}-\theta^*\|_2\le\frac{\mu_-}{4M}\Big\},\qquad\text{and}\\
\mathcal{E}_j'&:=\Big\{M_j\le 2M,\,\|\nabla^2\mathcal{L}_j(\theta^*)-I(\theta^*)\|_2\le\frac{\rho\mu_-}{4},\,\|\nabla\mathcal{L}_j(\theta^*)\|_2\le\frac{(1-\rho)\mu_-\delta_\rho}{4}\Big\}.
\end{aligned}
$$

We then have that under Assumptions PA-PD,

$$
\mathbb{P}\Big(\bigcup_{j=0}^{k}\mathcal{E}_j'^{c}\Big)\le\big(c_1'+c_2'\,(\log 2d)^{16}\,L^{16}+c_3'\,G^{16}\big)\,\frac{k}{n^8},
$$

where c_1',c_2' and $c_3'j$ are constants independent of (n,k,N,d,G,L).

Use $\lambda_{\min}(A)$ to denote the minimal eigenvalue of a symmetric matrix A.

Lemma A.4. *Assume that the conditions in Theorem 3.3 are true. Then under event $\bigcap_{j=0}^{k}\mathcal{E}_j'$ we have*

$$
\begin{aligned}
&\lambda_{\min}[\nabla^2\mathcal{L}_N(\widehat{\theta})]\ge\frac{1}{2}\,(1-\rho)\mu_-,\quad\|\overline{\theta}-\widehat{\theta}\|_2\le\Delta:=\frac{(1-\rho)\mu}{8M},\\
&U_N:=\max_{\theta\in(\widehat{\theta}-\Delta,\,\widehat{\theta}+\Delta)}\|\nabla^2\mathcal{L}_N(\theta)\|_2\le U:=2M\Delta+\frac{\rho\mu}{4}+\mu_+,\quad\textit{and}\\
&\|\nabla^2\mathcal{L}_N(\overline{\theta})^{-1}-\nabla^2\mathcal{L}_1(\overline{\theta})^{-1}\|_2\le\Big(\frac{2M\rho}{\mu_-^2}+\frac{\rho+4}{4\mu_-}\Big)\Big(\|\nabla^2\mathcal{L}_N(\theta^*)-\nabla^2\mathcal{L}_1(\theta^*)\|_2+4M\,\|\overline{\theta}-\theta^*\|_2\Big).
\end{aligned}
$$

The proof of this lemma is provided in Appendix B.4.

Now we proceed to prove Theorem 3.3. For the purpose of analysis, we define the global one-step Newton-Raphson estimator $\theta^N := \overline{\theta} - \nabla^2 \mathcal{L}_N(\overline{\theta})$.

The error can be decomposed as

$$\theta^H - \widehat{\theta} = (\theta^H - \theta^N) + (\theta^N - \widehat{\theta}).$$

We analyze the two terms respectively. The first term can be expressed as

$$\begin{aligned}
\theta^H - \theta^N &= \big(\overline{\theta} - \nabla^2 \mathcal{L}_1(\overline{\theta})^{-1} \nabla \mathcal{L}_N(\overline{\theta})\big) - \big(\overline{\theta} - \nabla^2 \mathcal{L}_N(\overline{\theta})^{-1} \nabla \mathcal{L}_N(\overline{\theta})\big) \\
&= \big(\nabla^2 \mathcal{L}_N(\overline{\theta})^{-1} - \nabla^2 \mathcal{L}_1(\overline{\theta})^{-1}\big) \nabla \mathcal{L}_N(\overline{\theta}) \\
&= \big(\nabla^2 \mathcal{L}_N(\overline{\theta})^{-1} - \nabla^2 \mathcal{L}_1(\overline{\theta})^{-1}\big) \big(\nabla \mathcal{L}_N(\overline{\theta}) - \nabla \mathcal{L}_N(\widehat{\theta})\big),
\end{aligned}$$

which yields the bound

$$\|\theta^H - \theta^N\|_2 \leq U_N \, |\!|\!| \nabla^2 \mathcal{L}_N(\overline{\theta})^{-1} - \nabla^2 \mathcal{L}_1(\overline{\theta})^{-1} |\!|\!|_2 \, \|\overline{\theta} - \widehat{\theta}\|_2.$$

The second term can be analyzed by using Theorem 5.3 in Bubeck [2015], which guarantees that under the assumption $\|\overline{\theta} - \widehat{\theta}\|_2 \leq \frac{\mu_N}{2M_N}$, it holds that

$$\|\theta^N - \widehat{\theta}\|_2 \leq \frac{M_N}{\mu_N} \|\overline{\theta} - \widehat{\theta}\|_2^2,$$

where $\mu_N := \lambda_{\min}[\nabla^2 \mathcal{L}_N(\widehat{\theta})]$ and M_N is the Lipschitz constant of the Hessian $\nabla^2 \mathcal{L}_N(\theta)$, that is $\big\|\nabla^2 \mathcal{L}_N(\theta_1) - \nabla^2 \mathcal{L}_N(\theta_2)\big\|_2 \leq M_N \, \|\theta_1 - \theta_2\|_2$ for all $\theta_1, \theta_2 \in U(\rho)$. Putting the pieces together, we obtain

$$\|\theta^H - \widehat{\theta}\|_2 \leq U_N \, |\!|\!| \nabla^2 \mathcal{L}_N(\overline{\theta})^{-1} - \nabla^2 \mathcal{L}_1(\overline{\theta})^{-1} |\!|\!|_2 \, \|\overline{\theta} - \widehat{\theta}\|_2 + \frac{M_N}{\mu_N} \, \|\overline{\theta} - \widehat{\theta}\|_2^2.$$

Now the claimed bound on $\|\theta^H - \widehat{\theta}\|_2$ is a direct consequence of the preceding display and Lemma A.4.

A.4 Proof of Corollary 3.4

The proof of the second part on the consistency of plug-in estimators for Σ is standard by using the consistency of $\widetilde{\theta}$ implied by the first part, the central limit theorem and Slutsky's theorem. Therefore we only prove the first part on the asymptotic expansion of $\widetilde{\theta}$. Based on Theorem 3.1,

we only need to establish the asymptotic expansion (11) of the global empirical risk minimizer $\widehat{\theta}$. By the integral form of Taylor's expansion, we have

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

where $H_N = \int_0^1 \nabla^2 \mathcal{L}_N\big(\theta^* + t(\widehat{\theta} - \theta^*)\big)\,dt$. Then simple linear algebra yields

$$\widehat{\theta} - \theta^* = -I(\theta^*)^{-1}\,\nabla \mathcal{L}_N(\theta^*) - U_N\,(\widehat{\theta} - \theta^*) - V_N\,(\widehat{\theta} - \theta^*), \tag{A.3}$$

where $U_N = H_N - \nabla^2 \mathcal{L}_N(\theta^*)$ and $V_N = \nabla^2 \mathcal{L}_N(\theta^*) - I(\theta^*)$. Then, the claimed expansion is an easy consequence of inequality (A.7) and Assumption D.

A.5 Proofs for regularized M-estimators

Proof of Theorem 3.5. This theorem follows from applying Corollary 1 of Negahban et al. [2012] to the objective $F(\theta)$. We check that $\widetilde{\mathcal{L}}(\theta)$ satisfies the restricted strong convexity condition.

The restricted strong convexity of $\widetilde{\mathcal{L}}$ is implied by the same property of $\mathcal{L}_1$, since

$$\widetilde{\mathcal{L}}(\theta^* + \delta) - \widetilde{\mathcal{L}}(\theta^*) - \nabla \widetilde{\mathcal{L}}(\theta^*)^T \delta = \mathcal{L}_1(\theta^* + \delta) - \mathcal{L}_1(\theta^*) - \nabla \mathcal{L}_1(\theta^*)^T \delta.$$

Thus by Corollary 1 of Negahban et al. [2012], we have established

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

for $\lambda > 2\left\|\nabla \widetilde{\mathcal{L}}(\theta^*)\right\|_\infty$. We can upper bound $\left\|\nabla \widetilde{\mathcal{L}}(\theta^*)\right\|_\infty$ as follows:

$$\begin{aligned}
\nabla \overline{\mathcal{L}}(\theta^*) &= \nabla \mathcal{L}_1(\theta^*) - \nabla \mathcal{L}_1(\overline{\theta}) + \mathcal{L}_N(\overline{\theta}) \\
&= (\nabla \mathcal{L}_N(\overline{\theta}) - \nabla \mathcal{L}_N(\theta^*)) - (\nabla \mathcal{L}_1(\overline{\theta}) - \nabla \mathcal{L}_1(\theta^*)) + \nabla \mathcal{L}_N(\theta^*) \\
&= \nabla^2 \mathcal{L}_N(\theta^*)(\overline{\theta} - \theta^*) - \nabla^2 \mathcal{L}_1(\theta^*)(\overline{\theta} - \theta^*) \\
&+ \int_{s=0}^{s=1} ds(\nabla^2 \mathcal{L}_N(\theta^* + s(\overline{\theta} - \theta^*)) - \nabla^2 \mathcal{L}_N(\theta^*))(\overline{\theta} - \theta^*) \\
&- \int_{s=0}^{s=1} ds(\nabla^2 \mathcal{L}_1(\theta^* + s(\overline{\theta} - \theta^*)) - \nabla^2 \mathcal{L}_1(\theta^*))(\overline{\theta} - \theta^*) + \nabla \mathcal{L}_N(\theta^*)
\end{aligned}$$

Using Assumption HB,

$$\begin{aligned}
\left\|\nabla \widetilde{\mathcal{L}}(\theta^*)\right\|_\infty &\leq \left\|\nabla^2 \mathcal{L}_N(\theta^*) - \nabla^2 \mathcal{L}_1(\theta^*)\right\|_\infty \left\|\overline{\theta} - \theta^*\right\|_1 + \left\|\nabla \mathcal{L}_N(\theta^*)\right\|_\infty \\
&\quad + 2M \left\|\overline{\theta} - \theta\right\|_2^2
\end{aligned}$$

□

Proof of Theorem 3.6. To apply Theorem 3.5, we have to compute $\left\|\nabla^2\mathcal{L}_N(\theta^*) - \nabla^2\mathcal{L}_1(\theta^*)\right\|_\infty$. Let $\Sigma = E[xx^T]$.

$$\left\|\nabla^2\mathcal{L}_N(\theta^*) - \nabla^2\mathcal{L}_1(\theta^*)\right\|_\infty = \left\|(\Sigma - \frac{1}{N}X^TX\right\|_\infty + \left\|\frac{1}{n}X_1^TX_1 - \Sigma\right\|_\infty + \sigma\sqrt{\frac{2\log d}{N}}.$$

By applying the sub-exponential concentration inequality, we have $\Pr(\frac{1}{N}\sum_{i=1}^N(|x_{ij}x_{ik} - \Sigma_{jk}| > t) \leq \exp(-c_\Sigma \min(t^2, t)N)$, where c_Σ is a constant that depends on Σ. By a union bound over all (j,k) pairs,

$$\Pr(\left\|\frac{1}{N}X^TX - \Sigma\right\|_{\max} > t) \leq \exp(2\log d - c_\Sigma \min(t^2, t)N).$$

Thus, letting $t = C\sqrt{\frac{\log d}{N}}$, we have $\left\|\frac{1}{N}X^TX - \Sigma\right\|_{\max} < C\sqrt{\frac{\log d}{N}}$ with probability greater than $1 - 1/p^{C'}$. By a similar argument, $\left\|\frac{1}{n}X_1^TX_1 - \Sigma\right\|_{\max} < C\sqrt{\frac{\log d}{n}}$.

Since $\nabla^2\mathcal{L}$ is a constant in linear regression, $M = 0$. Thus

$$\left\|\widetilde{\theta} - \theta^*\right\|_2 \leq \sqrt{\frac{s\log d}{n}}\left\|\overline{\theta} - \theta^*\right\|_1 + \sqrt{\frac{s\log d}{N}}.$$

□

Proof of Theorem 3.7. To apply Theorem 3.5, we need to verify Assumptions HA and HB. The restricted strong convexity of $\mathcal{L}$ is established in Proposition 1 of Negahban et al. [2012]. Next we verify Assumption HB:

$$\begin{aligned}\nabla^2\mathcal{L}_1(\theta^* + \delta) - \nabla^2\mathcal{L}_1(\theta^*) &= \frac{1}{n}\sum_{i=1}^n(\phi''(x_{ij}^T\theta^* + x_{ij}^T\delta) - \phi''(x_{ij}^T\theta^*))x_{ij}x_{ij}^T \\ &= \frac{1}{n}\sum_{i=1}^n \phi'''(x_{ij}^T\theta^* + s_{ij}x_{ij}^T\delta)x_{ij}(x_{ij}^T\delta)^2.\end{aligned}$$

Thus,

$$\begin{aligned}\left\|(\nabla^2\mathcal{L}_1(\theta^* + s\delta) - \nabla^2\mathcal{L}_1(\theta^*))(\delta)\right\|_\infty &\leq \left\|\frac{1}{n}\sum_{i=1}^n \phi'''(x_{ij}^T\theta^* + s_{ij}x_{ij}^T\delta)x_{ij}(x_{ij}^T\delta)^2\right\| \\ &\leq L_\phi B|\frac{1}{n}\sum_{i=1}^n(x_{ij}^T\delta)^2| \\ &\leq L_\phi BL\left\|\delta\right\|_2^2,\end{aligned}$$

Thus $M = L_\phi B L$, where L_ϕ is a local upper bound on ϕ''', L is the upper restricted eigenvalue of X, and $B = \max \|x\|_\infty$.

We also need to compute an upper bound on $\left\|\nabla^2 \mathcal{L}_N(\theta^*) - \nabla^2 \mathcal{L}_1(\theta^*)\right\|_\infty$. Define $A = E[\phi''(x^T\theta^*)xx^T]$.

$$\begin{aligned}\left\|\nabla^2 \mathcal{L}_N(\theta^*) - \nabla^2 \mathcal{L}_1(\theta^*)\right\|_\infty &= \left(\frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{n}\phi''(x_{ij}^T\theta^*)x_{ij}x_{ij}^T - A\right) + \left(A - \frac{1}{n}\sum_{i=1}^{n}\phi''(x_{i1}^T\theta^*)x_{i1}x_{i1}^T\right) \\ &\leq C\sqrt{\frac{\log d}{N}} + C\sqrt{\frac{\log d}{n}},\end{aligned}$$

where we used the same argument as in the proof of Theorem 3.6.

By Lemma 6 of Negahban et al. [2012], we know $\|\nabla \mathcal{L}_N(\theta^*)\|_\infty \leq C\sqrt{\frac{\log d}{N}}$.

Thus by Theorem 3.5, we have shown

$$\left\|\widetilde{\theta} - \theta^*\right\|_2 \leq C\sqrt{\frac{s\log d}{n}}\left\|\overline{\theta} - \theta^*\right\|_1 + \sqrt{\frac{s\log d}{N}} + C\left\|\overline{\theta} - \theta\right\|_2^2).$$

□

A.6 Proof of Theorem 3.8

Recall the definition of the "good events" $\mathcal{E}_j$ for $j = 1, \ldots, k$ in Section A.1 as

$$\mathcal{E}_j := \Big\{M_j \leq 2M,\, \|\!|\nabla^2 \mathcal{L}_j(\theta^*) - I(\theta^*)|\!\|_2 \leq \frac{\rho\mu_-}{4},\, \|\nabla \mathcal{L}_j(\theta^*)\|_2 \leq \frac{(1-\rho)\mu_-\delta_\rho}{4}\Big\}.$$

Moreover, we define events

$$\begin{aligned}\mathcal{A}_n &= \big\{\inf_{\|\theta-\theta^*\|_2 \geq \delta}\frac{1}{n}\left(\mathcal{L}_1(\theta) - \mathcal{L}_1(\theta^*)\right) \geq 3\epsilon\big\}, \\ \mathcal{B}_1 &= \Big\{\|\overline{\theta} - \theta^*\|_2 \leq \frac{\epsilon}{4R}\min\big\{\frac{1}{2\sqrt{M}}, \frac{1}{\rho\mu_- + 2\mu_+}\big\}\Big\}, \quad \text{and} \\ \mathcal{B}_2 &= \big\{\sqrt{N}\,\mu_+\|\overline{\theta} - \widehat{\theta}\|_2 + 2M\sqrt{N}\,\|\overline{\theta} - \widehat{\theta}\|_2^2 + 2M\sqrt{N}\,\|\overline{\theta} - \widehat{\theta}\|_2\|\widehat{\theta} - \theta^*\|_2 + M\|\widehat{\theta} - \theta^*\|_2 \leq \mu_-/16\big\},\end{aligned}$$

where $\delta = \min\{\rho/2, (4M)^{-1}\mu_-\}$ and

$$\epsilon = 4R\min\Big\{\frac{(1-\rho)\mu_-\delta_\rho}{2}, \frac{(1-\rho)^3\mu_-^2}{8M}, \frac{(1-\rho)^{3/2}\mu_-^{3/2}}{8M}\Big\}.$$

Then under the assumptions of the theorem and our previous developments in Section A.1, we have

$$\mathbb{P}\Big(\mathcal{A}_n^c \cup \mathcal{B}_1^c \cup \mathcal{B}_2^c \cup \bigcup_{j=1}^{k}\mathcal{E}_j^c\Big) \to 1, \quad \text{as } n \to \infty. \tag{A.4}$$

To prove the claimed result, we need three auxiliary lemmas. The first lemma provides the local expansions of global loss function $\mathcal{L}_N(\theta)$ and surrogate function $\widetilde{\mathcal{L}}(\theta)$ around the global empirical loss minimizer $\widehat{\theta}$. The proof is provided in Appendix B.5.

Lemma A.5. *Under event $\bigcap_{j=1}^k \mathcal{E}_j$, we have that for all $\theta \in U(\rho)$,*

$$\begin{aligned}
&\left|\mathcal{L}_N(\theta) - \mathcal{L}_N(\widehat{\theta}) - \frac{1}{2}\left\langle \theta - \widehat{\theta}, I(\theta^*)\left(\theta - \widehat{\theta}\right)\right\rangle\right| \\
&\qquad \leq \left(M\left\|\widehat{\theta} - \theta^*\right\|_2 + \frac{1}{2k}\sum_{j=1}^k \left\|\nabla^2 \mathcal{L}_j\left(\theta^*\right) - I(\theta^*)\right\|_2\right)\left\|\theta - \widehat{\theta}\right\|_2^2 + M\left\|\theta - \widehat{\theta}\right\|_2^3, \quad \textit{and} \\
&\left|\widetilde{\mathcal{L}}\left(\theta\right) - \widetilde{\mathcal{L}}\left(\widehat{\theta}\right) - \frac{1}{2}\left\langle \theta - \widehat{\theta}, I(\theta^*)\left(\theta - \widehat{\theta}\right)\right\rangle\right| \\
&\qquad \leq A_n\left\|\theta - \widehat{\theta}\right\|_2 + B_n\left\|\theta - \widehat{\theta}\right\|_2^2 + M\left\|\theta - \widehat{\theta}\right\|_2^3,
\end{aligned}$$

where $A_n := \mu_+\left\|\overline{\theta} - \widehat{\theta}\right\|_2 + 2M\left\|\overline{\theta} - \widehat{\theta}\right\|_2^2 + 2M\left\|\overline{\theta} - \widehat{\theta}\right\|_2\left\|\widehat{\theta} - \theta^*\right\|_2$ and $B_n := M\left\|\widehat{\theta} - \theta^*\right\|_2 + \frac{1}{2k}\sum_{j=1}^k \left\|\mathcal{L}_j\left(\theta^*\right) - I(\theta^*)\right\|_2$.

Our second lemma shows that the global identifiability assumption PC for $\mathcal{L}_1(\theta)$ implies the identifiability for the surrogate loss $\widetilde{\mathcal{L}}(\theta)$. The proof is provided in Appendix B.6.

Lemma A.6. *Under the joint event $\mathcal{A}_n \cap \mathcal{B}_1 \cap \bigcap_{j=1}^k \mathcal{E}_j$, we have*

$$\inf_{\|\theta - \theta^*\|_2 \geq \delta} (\widetilde{\mathcal{L}}(\theta) - \widetilde{\mathcal{L}}(\theta^*)) \geq 2\epsilon.$$

Our final lemma shows that if the results in the previous two lemma hold, then we obtain a Bernstein-von Mises result for the approximated posterior $\widetilde{\pi}_N$. The proof is provided in Appendix B.7.

Lemma A.7. *Suppose that the conclusions of Lemma A.5 and Lemma A.6 are true. Then under the event $\mathcal{B}_2$, we have*

$$\left\|\widetilde{\pi}_N(\theta) - \mathcal{N}_d\left(\widehat{\theta},\, I(\theta^*)^{-1}\right)(\theta)\right\|_1 \leq C\,R, \tag{A.5}$$

where $\mathcal{N}_d(\mu, \Sigma)(\cdot)$ is the pdf of a d-dim Gaussian distribution with mean vector μ and covariance matrix Σ, and the remainder term

$$R := A_n\sqrt{N}\,\log N + B_n\,(\log N)^2 + M\,N^{-1/2}\,(\log N)^3.$$

Here C is a constant independent of (n, k, N).

Combining the three lemmas and the high-probability bound (A.4), we obtain that with probability tending to one, bound (A.5) holds. Similarly, by considering the global posterior $\pi_N(\theta)$ as the approximated posterior $\widetilde{\pi}_N(\theta)$ with $n = N$ and $k = 1$, we obtain that

$$\left\| \pi_N(\theta) - \mathcal{N}_d\big(\widehat{\theta},\, I(\theta^*)^{-1}\big)\,(\theta) \right\|_1 \leq C\,R. \tag{A.6}$$

Combining (A.5) and (A.6) yields a proof of the claimed result.

APPENDIX B. PROOF OF THE AUXILIARY RESULTS IN THE PROOFS OF THEOREM ?? AND THEOREM ??

B.1 Proof of Lemma A.1

Apply Lemma 6 in Zhang et al. [2013], we obtain that under the event $\bigcap_{j=1}^{k} \mathcal{E}_j$,

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{2\|\nabla \mathcal{L}_N(\theta^*)\|_2}{(1-\rho)\mu_-}, \tag{A.7}$$

where $\nabla \mathcal{L}_N(\theta^*) = \frac{1}{k}\sum_{j=1}^{k} \nabla \mathcal{L}_j(\theta^*)$. In order to obtain high-probability bounds for $\nabla \mathcal{L}_j(\theta^*)$ and $|||\nabla^2 \mathcal{L}_j(\theta^*) - I(\theta^*)|||_2$ for $j = 1, \ldots, k$, we apply the following result.

Lemma B.1 (Zhang et al. [2013], Lemma 7)**.** *Under Assumption PB and PD, there exist universal constants c, c' such that for $\nu \in \{1, \ldots, 8\}$,*

$$\mathbb{E}[\|\nabla \mathcal{L}_j(\theta^*)\|_2^{2\nu}] \leq \frac{c\,G^{2\nu}}{n^{\nu}},$$
$$\mathbb{E}[|||\nabla^2 \mathcal{L}_j(\theta^*) - I(\theta^*)|||_2^{2\nu}] \leq \frac{c'\,(\log 2d)^{\nu}\,L^{2\nu}}{n^{\nu}}.$$

Now we apply Markov's inequality, Jensen's inequality and the union bound to obtain that there exist constants c_1, c_2, c_3 independent of (n, k, N, d, G, L) such that

$$\mathbb{P}\Big(\bigcup_{j=0}^{k} \mathcal{E}_j^c\Big) \leq \big(c_1 + c_2\,(\log 2d)^{16}\,L^{16} + c_3\,G^{16}\big)\,\frac{k}{n^8}. \tag{A.8}$$

B.2 Proof of Lemma A.2

We will apply Lemma 6 in Zhang et al. [2013] with $\theta^* = \widehat{\theta}$ and $F_1 = \widetilde{\mathcal{L}}$ in the notation therein. Since the Hessian of $\widetilde{\mathcal{L}}$ is the same as that of $\mathcal{L}_1$, in order to apply their result, we only need to verify that under event $\mathcal{E}_0 \cap \mathcal{E}_1$, it holds that

$$|||\nabla^2 \mathcal{L}_1(\widehat{\theta}) - I(\theta^*)|||_2 \leq \frac{\rho\mu}{2}, \quad \text{and} \quad \|\nabla \mathcal{L}_1(\widehat{\theta})\|_2 \leq \frac{(1-\rho)\mu_-\delta_\rho}{2}.$$

The first inequality is true since under event $\mathcal{E}_0 \cap \mathcal{E}_1$ and Assumption D, we have

$$\|\nabla^2 \mathcal{L}_1(\widehat{\theta}) - I(\theta^*)\|_2 \leq 2M \|\widehat{\theta} - \theta^*\|_2 + \|\nabla^2 \mathcal{L}_1(\theta^*) - I(\theta^*)\|_2 \leq \frac{\rho\mu_-}{4} + \frac{\rho\mu_-}{4} = \frac{\rho\mu_-}{2}.$$

To prove the second inequality, we apply the integral form of Taylor's expansion to obtain that

$$\nabla \mathcal{L}_1(\widehat{\theta}) - \nabla \mathcal{L}_1(\theta^*) = H_1 (\widehat{\theta} - \theta^*),$$

where matrix $H_1 = \int_0^1 \nabla^2 \mathcal{L}_1\big(\theta^* + t(\widehat{\theta} - \theta^*)\big)\, dt$ satisfies

$$\|H_1 - I(\theta^*)\|_2 \leq 2M \|\widehat{\theta} - \theta^*\|_2$$

under event $\mathcal{E}_1$. Therefore, the triangle inequality yields that under event $\mathcal{E}_0 \cap \mathcal{E}_1$,

$$\begin{aligned}
\|\nabla \mathcal{L}_1(\widehat{\theta})\|_2 &\leq \|\nabla \mathcal{L}_1(\theta^*)\|_2 + \|H_1 - I(\theta^*)\|_2 \|\widehat{\theta} - \theta^*\|_2 + \|I(\theta^*)\|_2 \|\widehat{\theta} - \theta^*\|_2 \\
&\leq \frac{(1-\rho)\mu_- \delta_\rho}{4} + 2M \|\widehat{\theta} - \theta^*\|_2^2 + \mu_+ \|\widehat{\theta} - \theta^*\|_2 \\
&\leq \frac{(1-\rho)\mu_- \delta_\rho}{2}.
\end{aligned}$$

This proves the second inequality and therefore the claimed result.

B.3 Proof of Lemma B.3

The claimed inequality is a immediate consequence of equation (A.3) in Section A.4 and inequality (A.7) in Appendix B.1.

B.4 Proof of Lemma A.4

Under Assumption D and event $\bigcap_{j=0}^k \mathcal{E}_j'$, we can bound $\nabla^2 \mathcal{L}_N(\widehat{\theta})$ as

$$\begin{aligned}
\lambda_{\min}[\nabla^2 \mathcal{L}_N(\widehat{\theta})] &\geq \lambda_{\min}[I(\theta^*)] - \|\nabla^2 \mathcal{L}_N(\theta^*) - I(\theta^*)\|_2 - \|\nabla^2 \mathcal{L}_N(\widehat{\theta}) - \nabla^2 \mathcal{L}_N(\theta^*)\|_2 \\
&\geq \mu_- - \frac{\rho\mu_-}{2} - 2M \|\widehat{\theta} - \theta^*\|_2 \geq \frac{1}{2}(1-\rho)\mu_-.
\end{aligned}$$

This proves the first claimed inequality.

The claimed inequality $\|\overline{\theta} - \widehat{\theta}\|_2 \leq \frac{(1-\rho)\mu}{8M}$ is immediate under the definition of $\mathcal{E}_0$ and the condition $\|\widehat{\theta} - \theta^*\|_2 \leq \frac{(1-\rho)\mu}{16M}$.

Under event $\bigcap_{j=0}^k \mathcal{E}_j'$, the third inequality can be proved as

$$\begin{aligned}
U_N &\leq \max_{\theta \in (\widehat{\theta}-\Delta, \widehat{\theta}+\Delta)} \|\nabla^2 \mathcal{L}_N(\theta) - \nabla^2 \mathcal{L}_N(\theta^*)\|_2 + \|\nabla^2 \mathcal{L}_N(\theta^*) - I(\theta^*)\|_2 + \|I(\theta^*)\|_2 \\
&\leq 2M\Delta + \frac{\rho\mu}{4} + \mu_+.
\end{aligned}$$

To bound the term $\|\nabla^2\mathcal{L}_N(\overline{\theta})^{-1} - \nabla^2\mathcal{L}_1(\overline{\theta})^{-1}\|_2$, we make use of the following inequality: for any matrix $A \in \mathbb{R}^{d\times d}$,

$$\|(A+\Delta A)^{-1} - A^{-1}\|_2 \leq \|A^{-1}\|_2^2 \|\Delta A\|_2. \tag{A.9}$$

First, choose $A = I(\theta^*)$ and $\Delta A = \nabla^2\mathcal{L}_N(\overline{\theta}) - I(\theta^*)$ in (A.9). Note that under the event $\bigcap_{j=0}^k \mathcal{E}_j'$, we have

$$\begin{aligned}\|\nabla^2\mathcal{L}_N(\overline{\theta}) - I(\theta^*)\|_2 &\leq \|\nabla^2\mathcal{L}_N(\overline{\theta}) - \nabla^2\mathcal{L}_N(\theta^*)\|_2 + \|\nabla^2\mathcal{L}_N(\theta^*) - I(\theta^*)\|_2 \\ &\leq 2M\,\|\overline{\theta} - \theta^*\|_2 + \|\nabla^2\mathcal{L}_N(\theta^*) - I(\theta^*)\|_2.\end{aligned}$$

Therefore, we have that under the event $\bigcap_{j=0}^k \mathcal{E}_j'$

$$\begin{aligned}\|\nabla^2\mathcal{L}_N(\overline{\theta})^{-1}\|_2 &\leq \|\nabla^2\mathcal{L}_N(\overline{\theta})^{-1} - I(\theta^*)^{-1}\|_2 + \mu_-^{-1} \\ &\leq 2M\,\mu_-^{-2}\,\|\overline{\theta} - \theta^*\|_2 + \mu_-^{-2}\,\|\nabla^2\mathcal{L}_N(\theta^*) - I(\theta^*)\|_2 + \mu_-^{-1} \\ &\leq 2M\,\mu_-^{-2}\,\rho + \mu_-^{-1} + \mu_-^{-1}\,\rho/4,\end{aligned}$$

where in the last step we used the assumption that $\|\overline{\theta} - \theta^*\|_2 \leq \rho$ and the definition of events $\mathcal{E}_j'$'s. Now choosing $A = \nabla^2\mathcal{L}_N(\overline{\theta})$ and $\Delta A = \nabla^2\mathcal{L}_1(\overline{\theta}) - \nabla^2\mathcal{L}_N(\overline{\theta})$ in inequality (A.9), we obtain

$$\begin{aligned}&\|\nabla^2\mathcal{L}_N(\overline{\theta})^{-1} - \nabla^2\mathcal{L}_1(\overline{\theta})^{-1}\|_2 \\ &\leq \|\nabla^2\mathcal{L}_N(\overline{\theta})^{-1}\|_2^2\,\|\nabla^2\mathcal{L}_N(\overline{\theta}) - \nabla^2\mathcal{L}_1(\overline{\theta})\|_2 \\ &\leq \|\nabla^2\mathcal{L}_N(\overline{\theta})^{-1}\|_2^2 \Big(\|\nabla^2\mathcal{L}_N(\theta^*) - \nabla^2\mathcal{L}_1(\theta^*)\|_2 + \|\nabla^2\mathcal{L}_N(\overline{\theta}) - \nabla^2\mathcal{L}_N(\theta^*)\|_2 + \|\nabla^2\mathcal{L}_1(\overline{\theta}) - \nabla^2\mathcal{L}_1(\theta^*)\|_2\Big) \\ &\leq \|\nabla^2\mathcal{L}_N(\overline{\theta})^{-1}\|_2^2 \Big(\|\nabla^2\mathcal{L}_N(\theta^*) - \nabla^2\mathcal{L}_1(\theta^*)\|_2 + 4M\,\|\overline{\theta} - \theta^*\|_2\Big).\end{aligned}$$

Putting the pieces together, we have proved the final claimed inequality.

B.5 Proof of Lemma A.5

To prove the first expansion for $\mathcal{L}_N(\theta)$, it suffices to prove the following inequality by using the fact that $\nabla\mathcal{L}_N(\widehat{\theta}) = 0$:

$$\begin{aligned}&\Big|\mathcal{L}_j(\theta) - \mathcal{L}_j(\widehat{\theta}) - \langle\nabla\mathcal{L}_j(\widehat{\theta}), \theta - \widehat{\theta}\rangle - \frac{1}{2}\,\langle\theta - \widehat{\theta}, I(\theta^*)\,(\theta - \widehat{\theta})\rangle\Big| \\ &\qquad \leq \Big(M\,\|\widehat{\theta} - \theta^*\|_2 + \frac{1}{2}\,\|\nabla^2\mathcal{L}_j(\theta^*) - I(\theta^*)\|_2\Big)\,\|\theta - \widehat{\theta}\|_2^2 + M\,\|\theta - \widehat{\theta}\|_2^3\end{aligned} \tag{A.10}$$

for $j = 1, \ldots, k$. In fact, by Taylor's theorem, we have

$$\mathcal{L}_j(\theta) - \mathcal{L}_j(\widehat{\theta}) = \langle \nabla \mathcal{L}_j(\widehat{\theta}), \theta - \widehat{\theta} \rangle + \frac{1}{2} \langle \theta - \widehat{\theta}, I(\theta^*)\,(\theta - \widehat{\theta}) \rangle + \frac{1}{2} \langle \theta - \widehat{\theta}, \big(\widetilde{H}_j - I(\theta^*)\big)\,(\theta - \widehat{\theta}) \rangle,$$

where $\widetilde{H}_j = \nabla^2 \mathcal{L}_j\big(\widehat{\theta} + t_j(\theta - \widehat{\theta})\big)$ for some $t_j \in [0,1]$. Under event $\mathcal{E}_j$, we can bound the last remainder term for $\theta \in U(\rho)$ by

$$\begin{aligned} &\frac{1}{2} \|\theta - \widehat{\theta}\|_2^2 \big(\|\widetilde{H}_j - \nabla^2 \mathcal{L}_j(\theta^*)\|_2 + \|\nabla^2 \mathcal{L}_j(\theta^*) - I(\theta^*)\|_2\big) \\ &\quad \leq M \|\theta - \widehat{\theta}\|_2^3 + \Big(M \|\widehat{\theta} - \theta^*\|_2 + \frac{1}{2} \|\nabla^2 \mathcal{L}_j(\theta^*) - I(\theta^*)\|_2\Big) \|\theta - \widehat{\theta}\|_2^2, \end{aligned}$$

which yields the expansion (A.10).

To prove the expansion for $\widetilde{\mathcal{L}}(\theta)$, we note that simple calculation yields

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

Given the first expansion for $\mathcal{L}_N(\theta)$ and the expansion (A.10) for $\mathcal{L}_1(\theta)$, we only need to show that under the joint event $\bigcap_{j=1}^k \mathcal{E}_j$,

$$\Big|\langle \nabla \mathcal{L}_N(\overline{\theta}) - \nabla \mathcal{L}_N(\widehat{\theta}), \theta - \widehat{\theta} \rangle\Big| + \Big|\langle \nabla \mathcal{L}_1(\overline{\theta}) - \nabla \mathcal{L}_1(\widehat{\theta}), \theta - \widehat{\theta} \rangle\Big| \leq A_n \|\theta - \widehat{\theta}\|_2,$$

because $\nabla \mathcal{L}_N(\widehat{\theta}) = 0$. This is true since by using the integral form of Taylor's expansion and Cauchy-Schwarz inequality, the left-hand side in the preceding display can be bounded by

$$\begin{aligned} &\|\widehat{H}_N\|_2 \|\overline{\theta} - \widehat{\theta}\|_2 \|\theta - \widehat{\theta}\|_2 + \|\widehat{H}_j\|_2 \|\overline{\theta} - \widehat{\theta}\|_2 \|\theta - \widehat{\theta}\|_2 \\ &\quad \leq 2\big(\mu_+ + 2M \|\overline{\theta} - \widehat{\theta}\|_2 + 2M \|\widehat{\theta} - \theta^*\|_2\big) \|\overline{\theta} - \widehat{\theta}\|_2 \|\theta - \widehat{\theta}\|_2 = A_n \|\theta - \widehat{\theta}\|_2, \end{aligned}$$

where $\widehat{H}_N = \int_0^1 \nabla^2 \mathcal{L}_N\big(\widehat{\theta} + t(\overline{\theta} - \widehat{\theta})\big)\, dt$, $\widehat{H}_j = \int_0^1 \nabla^2 \mathcal{L}_j\big(\widehat{\theta} + t(\overline{\theta} - \widehat{\theta})\big)\, dt$ and in the last step we used the fact that under the joint event $\bigcap_{j=1}^k \mathcal{E}_j$, $\|\widehat{H}_j - I(\theta^*)\|_2 \leq 2M\,(\|\overline{\theta} - \widehat{\theta}\|_2 + \|\widehat{\theta} - \theta^*\|_2)$ for each j and $\|\widehat{H}_N - I(\theta^*)\|_2 \leq k^{-1} \sum_{j=1}^k \|\widehat{H}_j - I(\theta^*)\|_2$.

B.6 Proof of Lemma A.6

By Assumption PA, $\|\theta - \theta^*\|_2 \le R$ for all $\theta \in \Theta$. Therefore, by the definition of events $\mathcal{E}_j$ and $\mathcal{A}_n$, we obtain

$$\begin{aligned}
&\inf_{\|\theta-\theta^*\|_2\ge\delta}(\tilde{\mathcal{L}}(\theta) - \tilde{\mathcal{L}}(\theta^*)) \\
&\ge \inf_{\|\theta-\theta^*\|_2\ge\delta}(\mathcal{L}_1(\theta) - \mathcal{L}_1(\theta^*)) - \sup_{\theta\in\Theta}\langle\theta-\theta^*, \nabla\mathcal{L}_1(\overline{\theta}) - \nabla\mathcal{L}_N(\overline{\theta})\rangle \\
&\ge 3\epsilon - R\left(\|\nabla\mathcal{L}_1(\overline{\theta}) - \nabla\mathcal{L}_1(\theta^*)\|_2 + \|\nabla\mathcal{L}_1(\theta^*)\|_2 + \|\nabla\mathcal{L}_N(\theta^*)\|_2 + \|\nabla\mathcal{L}_N(\overline{\theta}) - \nabla\mathcal{L}_N(\theta^*)\|_2\right).
\end{aligned}$$

Now we bound the four terms inside the brackets under the event $\mathcal{B}_n$, respectively, as

$$\begin{aligned}
\|\nabla\mathcal{L}_1(\overline{\theta}) - \nabla\mathcal{L}_1(\theta^*)\|_2 &\le \max_{\theta\in U(\rho)} |\!|\!|\nabla^2\mathcal{L}_1(\theta)|\!|\!|_2 \, \|\overline{\theta} - \theta^*\|_2 \\
&\le (M\,\|\overline{\theta} - \theta^*\|_2 + \frac{\rho\mu_-}{2} + \mu_+)\,\|\overline{\theta} - \theta^*\|_2 \le \frac{\epsilon}{4R}, \\
\|\nabla\mathcal{L}_1(\theta^*)\|_2 &\le \min\Big\{\frac{(1-\rho)\mu_-\delta_\rho}{2}, \frac{(1-\rho)^3\mu_-^2}{8M}, \frac{(1-\rho)^{3/2}\mu_-^{3/2}}{8M}\Big\} \le \frac{\epsilon}{4R}, \\
\|\nabla\mathcal{L}_N(\theta^*)\|_2 &\le \frac{1}{k}\sum_{j=1}^{k}\|\nabla\mathcal{L}_j(\theta^*)\|_2 \le \frac{\epsilon}{4R}, \quad \text{and} \\
\|\nabla\mathcal{L}_N(\overline{\theta}) - \nabla\mathcal{L}_N(\theta^*)\|_2 &\le \frac{1}{k}\sum_{j=1}^{k}\|\nabla\mathcal{L}_j(\overline{\theta}) - \nabla\mathcal{L}_j(\theta^*)\|_2 \le \frac{\epsilon}{4R}.
\end{aligned}$$

Putting the pieces together, we obtain that under the joint event $\mathcal{A}_n \cap \mathcal{B}_1 \cap \bigcap_{j=1}^k \mathcal{E}_j$,

$$\inf_{\|\theta-\theta^*\|_2\ge\delta}\frac{1}{n}\left(\tilde{\mathcal{L}}(\theta) - \tilde{\mathcal{L}}(\theta^*)\right) \ge 3\epsilon - R\cdot 4\cdot\frac{\epsilon}{4R} = 2\epsilon,$$

which completes the proof.

B.7 Proof of Lemma A.7

The approximated posterior can be expressed by

$$\widetilde{\pi}_N(\theta) = \frac{\pi(\theta)\, e^{-N(\tilde{\mathcal{L}}(\theta)-\tilde{\mathcal{L}}(\widehat{\theta}))}}{\int_\Theta \pi(\theta)\, e^{-N(\tilde{\mathcal{L}}(\theta)-\tilde{\mathcal{L}}(\widehat{\theta}))}\, d\theta}.$$

We claim that it suffices to prove

$$\begin{aligned}
&\Big|\pi(\theta)\, e^{-N(\tilde{\mathcal{L}}(\theta)-\tilde{\mathcal{L}}(\widehat{\theta}))} - \pi(\widehat{\theta})\, e^{-\frac{N}{2}\langle\theta-\widehat{\theta}, I(\theta^*)(\theta-\widehat{\theta})\rangle}\Big| \\
&\qquad\le C\,R\,\pi(\widetilde{\theta})\, e^{-\frac{N}{4}\langle\theta-\widetilde{\theta}, I(\theta^*)(\theta-\widetilde{\theta})\rangle} + \pi(\theta)\, e^{-N\epsilon}.
\end{aligned} \tag{A.11}$$

In fact, if (A.11) holds, then by integrating θ over $\mathbb{R}^d$, we obtain

$$\Big| N^{d/2} \int_\Theta \pi(\theta)\, e^{-N(\tilde{\mathcal{L}}(\theta)-\tilde{\mathcal{L}}(\widehat{\theta}))}\, d\theta - \pi(\widehat{\theta}\,) \frac{(2\pi)^{d/2}}{\sqrt{\det I(\theta^*)}} \Big| \leq C\, R + C\, N^{(d-1)/2} e^{-N\epsilon} \leq C' R.$$

Then, by combining all three preceding displays, we obtain

$$\int_{\mathbb{R}^d} \Big| \widetilde{\pi}_N(\theta) - \frac{N^{d/2}\sqrt{\det I(\theta^*)}}{(2\pi)^{d/2}}\, e^{-\frac{N}{2}\,\langle \theta-\widehat{\theta}, I(\theta^*)\,(\theta-\widehat{\theta})\rangle} \Big|\, d\theta \leq C''\, R,$$

which is the claimed Bernstein-von Mises result for $\widetilde{\pi}_N$. The remainder of the proof focuses on proving (A.11).

Let $s = \sqrt{N}\,(\theta - \widehat{\theta})$ be the localized parameter. Then (A.11) is equivalent to

$$\begin{aligned} &\Big| \pi(\widehat{\theta} + s/\sqrt{N})\, e^{-N(\tilde{\mathcal{L}}(\widehat{\theta}+s/\sqrt{N})-\tilde{\mathcal{L}}(\widehat{\theta}))} - \pi(\widehat{\theta}\,)\, e^{-\frac{1}{2}\,\langle s, I(\theta^*)\, s\rangle} \Big| \\ &\qquad \leq C\, R\, \pi(\widehat{\theta}\,)\, e^{-\frac{1}{4}\,\langle s, I(\theta^*)\, s\rangle} + \pi(\widehat{\theta} + s/\sqrt{N})\, e^{-N\epsilon}. \end{aligned} \tag{A.12}$$

Corollary A.5 guarantees that for all $\|s\| \leq \delta\sqrt{N}$,

$$\begin{aligned} &\Big| N(\widetilde{\mathcal{L}}(\widehat{\theta} + s/\sqrt{N}) - \widetilde{\mathcal{L}}(\theta^*)) - \frac{1}{2}\,\langle s, I(\theta^*)\, s\rangle \Big| \\ &\qquad \leq A_n \sqrt{N}\, \|s\|_2 + B_n\, \|s\|_2^2 + \frac{M}{\sqrt{N}}\, \|s\|_2^3. \end{aligned} \tag{A.13}$$

We prove (A.12) by considering s in the following three subsets separately:

$$\begin{aligned} S_1 &:= \{s :\ \|s\|_2 \leq c\, \log N\} \\ S_2 &:= \{s :\ c\, \log N \leq \|s\|_2 \leq \delta\sqrt{N}\} \\ S_3 &:= \{s :\ \|s\|_2 > \delta\sqrt{N}\}. \end{aligned}$$

We begin with $s \in S_1$. Using (A.13), we obtain that

$$\begin{aligned} &\Big| \pi(\widehat{\theta} + s/\sqrt{N})\, e^{-N(\tilde{\mathcal{L}}(\widehat{\theta}+s/\sqrt{N})-\tilde{\mathcal{L}}(\theta^*))} - \pi(\widehat{\theta}\,)\, e^{-\frac{1}{2}\,\langle s, I(\theta^*)\, s\rangle} \Big| \\ &\leq \Big| \pi(\widehat{\theta} + s/\sqrt{N})\, e^{-N(\tilde{\mathcal{L}}(\widehat{\theta}+s/\sqrt{N})-\tilde{\mathcal{L}}(\theta^*))} - \pi(\widehat{\theta} + s/\sqrt{N})\, e^{-\frac{1}{2}\,\langle s, I(\theta^*)\, s\rangle} \Big| \\ &\qquad + \big| \pi(\widehat{\theta} + s/\sqrt{N}) - \pi(\widehat{\theta}) \big|\, e^{-\frac{1}{2}\,\langle s, I(\theta^*)\, s\rangle} \\ &\leq C\, \pi(\widehat{\theta})\, e^{-\frac{1}{2}\,\langle s, I(\theta^*)\, s\rangle} \left(A_n \sqrt{N}\, \log N + B_n (\log N)^2 + M\, N^{-1/2}\, (\log N)^3 \right) \\ &\qquad + C\, \frac{\log N}{\sqrt{N}}\, e^{-\frac{1}{2}\,\langle s, I(\theta^*)\, s\rangle} \\ &\leq C\, R\, \pi(\widetilde{\theta}) e^{-\frac{1}{4}\,\langle s, I(\theta^*)\, s\rangle}. \end{aligned}$$

Next consider $s \in S_2$. Then $\|s\|_2 \le \|s\|_2^2$ for sufficiently small constant c. Under the event $\mathcal{B}_2$, we have $A_n \sqrt{N}\, \|s\|_2^2 + B_n \, \|s\|_2^2 + M \, N^{-1/2} \, \|s\|_2^3 \le \langle s, I(\theta^*)\, s\rangle/4$. Then using (A.13), we obtain

$$\Big| N(\widetilde{\mathcal{L}}(\widehat{\theta} + s/\sqrt{N}) - \widetilde{\mathcal{L}}(\theta^*)) - \frac{1}{2}\, \langle s, I(\theta^*)\, s\rangle \Big| \le \frac{1}{4}\, \langle s, I(\theta^*)\, s\rangle.$$

Therefore, we have

$$\begin{aligned}
&\Big| \pi(\widehat{\theta} + s/\sqrt{N})\, e^{-N(\tilde{\mathcal{L}}(\widehat{\theta}+s/\sqrt{N}) - \tilde{\mathcal{L}}(\theta^*))} - \pi(\widehat{\theta}\,)\, e^{-\frac{1}{2}\,\langle s, I(\theta^*)\, s\rangle} \Big| \\
&\le \pi(\widehat{\theta} + s/\sqrt{N})\, e^{-N(\tilde{\mathcal{L}}(\widetilde{\theta}+s/\sqrt{N}) - \tilde{\mathcal{L}}(\theta^*))} + \pi(\widehat{\theta}\,)\, e^{-\frac{1}{2}\,\langle s, I(\theta^*)\, s\rangle} \\
&\le C\, (R_1 + 1)\, \pi(\widehat{\theta}) e^{-\frac{1}{4}\,\langle s, I(\theta^*)\, s\rangle} \\
&\le C\, R\, \pi(\widehat{\theta}\,)\, e^{-\frac{1}{4}\,\langle s, I(\theta^*)\, s\rangle}.
\end{aligned}$$

For $s \in S_3$, we have $\|\widehat{\theta} + s/\sqrt{N} - \theta^*\|_2 = \|\theta - \theta^*\|_2 \ge \|\theta - \widehat{\theta}\|_2 - \|\widehat{\theta} - \theta^*\|_2 \ge \delta$. The proof of Lemma A.6 shows that under the joint event $\mathcal{A}_n \cap \mathcal{B}_n \cap \bigcap_{j=1}^k \mathcal{E}_j$, we have

$$\sup_{\theta, \theta' \in \Theta} \Big| \tilde{\mathcal{L}}(\theta) - \tilde{\mathcal{L}}(\theta') - \mathcal{L}_1(\theta) - \mathcal{L}_1(\theta') \Big| \le \epsilon.$$

Then we obtain

$$\begin{aligned}
e^{-N(\tilde{\mathcal{L}}(\widehat{\theta}+s/\sqrt{N}) - \tilde{\mathcal{L}}(\widehat{\theta}))} &\le e^{N\epsilon}\, e^{-N(\mathcal{L}(\widehat{\theta}+s/\sqrt{N}) - \mathcal{L}(\widehat{\theta}))} \\
&= e^{N\epsilon}\, e^{-N(\mathcal{L}(\widehat{\theta}+s/\sqrt{N}) - \mathcal{L}(\theta^*)}\, e^{-N(\mathcal{L}(\theta^*) - \mathcal{L}(\widehat{\theta}))} \overset{(i)}{\le} e^{-N\epsilon},
\end{aligned}$$

where in step (i) we used Lemma A.6 and the optimality of $\widehat{\theta}$ that implies $\mathcal{L}(\theta^*) - \mathcal{L}(\widehat{\theta}\,) \ge 0$. Therefore, we have

$$\begin{aligned}
&\Big| \pi(\widehat{\theta} + s/\sqrt{N})\, e^{-N(\tilde{\mathcal{L}}(\widehat{\theta}+s/\sqrt{N}) - \tilde{\mathcal{L}}(\theta^*))} - \pi(\widehat{\theta}\,)\, e^{-\frac{1}{2}\,\langle s, I(\theta^*)\, s\rangle} \Big| \\
&\le \pi(\widehat{\theta} + s/\sqrt{N})\, e^{-N\epsilon} + C\, e^{-\mu\, \delta^2\, N/2}\, \pi(\widehat{\theta})\, e^{-\frac{1}{4}\,\langle s, I(\theta^*)\, s\rangle} \\
&\le C\, R\, \pi(\widehat{\theta})\, e^{-\frac{1}{4}\,\langle s, I(\theta^*)\, s\rangle} + \pi(\widehat{\theta} + s/\sqrt{N})\, e^{-N\epsilon}.
\end{aligned}$$

Putting the pieces together, we can prove (A.13) and therefore the claimed Bernstein-von Mises result for $\widetilde{\pi}_N$.