Communication-efficient distributed statistical learning

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

May 26, 2016
Abstract

We present the Communication-efficient Surrogate Likelihood (CSL) framework for solving distributed statistical learning 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 the averaging schemes and facilitates the construction of confidence intervals. For high-dimensional regularized estimation, CSL leads to a minimax optimal estimator with minimal 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. The methods are illustrated through empirical studies.

KEYWORDS: Communication-efficient, Distributed Learning
1. INTRODUCTION

In modern statistical problems, people may not be able to store large datasets on a single machine, let alone conduct statistical inference via optimization or MCMC. In this big data paradigm, it is natural to consider methods that split the dataset across multiple machines and conduct statistical inference in a distributed manner, in either the frequentist framework of obtaining a point-estimator [Duchi et al. 2012, Zhang et al. 2013a, Zhang and Lin 2015, Kannan et al. 2014, Shamir et al. 2014, Lee et al. 2015], or in the Bayesian framework of sampling from the posterior distribution [Wang and Dunson 2015, Neiswanger et al. 2015]. Good distributed statistical learning procedures should give rise to statistically optimal estimators with low communication cost.

To achieve low communication cost, one-shot or embarrassingly parallel approaches [Zhang et al. 2013a, Lee et al. 2015, Wang and Dunson 2015, Neiswanger et al. 2015] only use one-round of communication by first sending estimators or posterior samples from local machines to a center node, and then combining them to form a global estimator or approximation to the posterior distribution.

In the frequentist framework, most one-shot approaches rely on averaging [Zhang et al. 2013a], where the global estimator is the average of the local estimators. Lee et al. 2015 extends this idea from low dimensional regular models [Zhang et al. 2013a] 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 $O(dk)$ bits must be communicated in order to attain the minimax rate of parameter estimation, in spite of the existence of low-dimensional structures such as sparsity [Braverman et al. 2015], where $d$ is the dimension of the parameter and $k$ is the number of machine. However, these averaging-based one-shot communication approaches tend to suffer from several drawbacks. First, it is not straightforward to conduct statistical inferences, such as creating confidence intervals/regions and doing hypothesis testing, 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 another word, the number of machines should be much smaller than $\sqrt{N}$. This could be restrictive when both $N$ and $k$ are large. Third, when the underlying
data generating model is non-linear, our empirical study shows that even for small $k$ of order $10^1$, the averaging estimator only exhibits a slight improvement over the local estimators and tends to have low estimation accuracy.

In the Bayesian framework, embarrassingly parallel approaches run a Markov chain Monte Carlo (MCMC) algorithm across local machines in parallel and transmit local posterior samples to a central node to produce an approximation to the global posterior distribution. Unfortunately, the number of local posterior samples must grow at least exponentially in the dimension $d$ to approximate a local $d$-dimensional posterior distributions due to the curse of dimensionality. As a consequence, the communication complexity can be prohibitive when $d$ is moderate to high.

When combining local posterior samples in the central node, existing approaches \cite{Wang2015, Neiswanger2015} that approximate the global posterior distribution by a weighted empirical distribution of “averaging draws” tend to suffer from the weight-degeneracy issue (weights collapse to only a few sample) when $k$ is large.

In this paper, we formulate a unified framework for frequentist and Bayesian distributed learning by introducing a Communication-efficient Surrogate likelihood (CSL) function. In the frequentist perspective, CSL can be viewed as a communication-efficient surrogate to the global likelihood function that uses all samples. For example, CSL can replace the role of the likelihood function in forming either the maximum likelihood estimator (MLE) in regular parametric models or the penalized MLE in high-dimensional models. In the Bayesian perspective, CSL can be used to form a quasi-posterior distribution \cite{Chernozhukov2003} as a surrogate to the full posterior. CSL can be constructed efficiently by communicating $O(dk)$ number of 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 computation environment, CSL can be used as 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 learning approach overcomes the aforementioned drawbacks suffered by the one-shot and embarrassingly parallel approaches. In the frequentist framework, CSL can be used in the same way as the global likelihood function to either form a point-estimator or conduct statistical inferences. Moreover, a CSL-based estimator can achieve the same rate of
convergence as the global likelihood-based estimator and has communication complexity $O(dk)$. On the computational side, we show that by applying our distributed estimation procedure in an iterative manner, the resulting multi-round algorithm has a geometric convergence rate with contraction factor $O(n^{-1/2})$, where $n$ is the number of samples in each local machine. This $O(n^{-1/2})$ rate of convergence significantly improves the condition number contraction factor attained by the naive distributed gradient method that forms the global gradient in each iteration by combining the 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 $O(\log N \log n)$ iterations are applied, which is constant for $n > k$. In contrast, the aforementioned averaging estimator requires $n \gg \sqrt{N}$. For instance, due to the fast $O(n^{-1/2})$ rate, usually 2 to 3 iterations suffices for our procedure to match the same accuracy of the global likelihood-based estimator even for reasonably large $k$ (See Section 4.1 for more details). Unlike the bootstrap-based approach Zhang et al. [2013a] for boosting the accuracy that has a high computational complexity, the additional complexity of the iterative version of our approach grows only linearly in the number of iterations. Last but not least, our empirical study suggests that a CSL-based estimator may exhibit significant improvement over the averaging estimator in non-linear distributed statistical learning problems, making CSL practically appealing for large distributed learning tasks.

For high-dimensional $\ell_1$-regularized estimator, we provide an algorithm that communicates $O(dk)$ bits that attains 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 logarithmic number of rounds. During the preparation of this manuscript, we became aware of the concurrent work of Wang et al. [2016] that also proposes the high-dimensional CSL-based estimator. The theoretical results for $\ell_1$-regularized least squares are equivalent. However, their focus is limited to the high-dimensional linear models, and do not consider parametric estimation in low-dimensional families, nor do they consider Bayesian posterior estimation.

In the Bayesian framework, our method does not require transmitting local posterior samples and is free from the weight degeneracy issue. This makes the communication complexity of our

\[\text{The contraction factor of gradient descent is } 1 - \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}}, \text{ where } \lambda \text{ are the eigenvalues of the Hessian at the optimum.}\]
approach considerably lower than those embarrassingly parallel Bayesian computation approaches.

In the optimization literature, the most related work to ours is the distributed approximate
Newton algorithm (DANE) proposed in Shamir et al. [2014]. Although they only rigorously anal-
ysed DANE for quadratic objectives, their idea of combining gradient descent with a local Newton
method also applies to non-quadratic objectives. Motivated by this optimization idea, we propose
the CSL framework that is specifically tailored for distributed statistical learning. CSL is a surro-
gate likelihood function that can be easily applied to M-estimation, high-dimensional regularized
estimation and Bayesian inference. On the theoretical side, 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 broader class of regular parametric
models and high-dimensional models and is not restricted to quadratic objectives. Another related
work to DANE and ours is the iterated Hessian sketch (IHS) algorithm Pilanci and Wainwright
[2014] that is specifically designed for constrained least-squares optimization. Although DANE ap-
piled to quadratic problems can be viewed as a special case of IHS by choosing the sketching matrix
as a rescaled sub-sampling matrix, the analysis in Pilanci and Wainwright [2014] only applies to a
class of low-incoherence sketching matrices that excludes sub-sampling.

Overall, we build a general framework for statistical inference in the distributed learning setting.
Our iterative distributed statistical learning procedure reveals a trade-off between communication
complexity and statistical accuracy. When applied to Bayesian inference, our approach has the same
flavour as Pereyra [2015], Amandine et al. [2015] that introduces modern optimization techniques
to improve the computational efficiency of sampling-based algorithms.

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 on three different problems: param-
eter 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-
metric models (Section 3.3). Section 4 shows our 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 learning: either in a frequentist or Bayesian framework. Hereafter, we describe the CSL methodology and explain the motivation.

2.1 Statistical models with distributed data

Let \( Z_N^1 := \{Z_{ij} : i = 1, \ldots, n, j = 1, \ldots, k\} \) denote \( N = nk \) identically distributed observations with marginal distribution \( P_{\theta^*} \), where \( \{P_\theta : \theta \in \Theta\} \) is a family of statistical models parametrized by \( \theta \in \Theta \subset \mathbb{R}^d \), \( \Theta \) is called the parameter space and \( \theta^* \) is the true data generating parameter.

Suppose that the data is stored in a distributed manner where each machine stores a sub-sample of \( n \) observations. Let \( Z_j := \{Z_{ij} : i = 1, \ldots, n\} \) denote the sub-sample that is stored in the \( j \)th machine \( M_j \) for \( j = 1, \ldots, k \). Our goal is to conduct statistical inference on the parameter \( \theta \) while taking into consideration the communication cost among the machines. For example, we may want to find a point estimator \( \hat{\theta} \) and an associated confidence interval (region).

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

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

Define the local and global loss functions as

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

\[
L_N(\theta) = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{k} L(\theta; z_{ij}) = \frac{1}{k} \sum_{j=1}^{k} L_j(\theta).
\] (3)

Here \( L_j(\theta) \) is the loss function evaluated at \( \theta \) by using the local data stored in machine \( M_j \).

Throughout the rest of the paper, the reader may consider the negative log-likelihood function as a typical candidate of loss function \( L \).

For large-scale problems, suppose that we estimate parameter \( \theta \) directly using the gradient method applied to the global loss function or Markov chain Monte Carlo (MCMC) algorithm for Bayesian inference. At every iteration, each machine communicates a vector of length \( d \), and these algorithms generally require hundreds to thousands of iterations. It is thus necessary to study
distributed statistical learning algorithms that has high statistical accuracy while requiring fairly
limited communication. In the next subsection, we motivate the CSL methodology by demonstrating
that our proposed surrogate loss function behaves like the global loss function $L_N$ in a suitable
sense that allows accurate statistical estimation.

2.2 Distributed statistical learning

In this subsection, we motivate the CSL methodology by constructing a surrogate loss $\tilde{L} : \Theta \times Z \mapsto \mathbb{R}$
that approximates the global loss function $L_N$ in a communication-efficient manner: it can be
constructed in any local machine $M_j$ by communicating at most $(k - 1)$ $d$-dim vectors. After
the construction, $\tilde{L}$ can be used to replace the global loss function in various statistical learning
approaches by only using the data in a local machine (see Sections 3.1-3.3). Thus, this distributed
learning approach can simultaneously achieve high statistical accuracy and low communication
cost. The mathematical development in this section is based on heuristic arguments for motivating
the methodology, and rigorous analysis is provided in Section 3 to follow.

Our motivation is from the Taylor series expansion of $L_N$. Viewing $L_N(\theta)$ as an analytic
function, we can expand it into the infinite series

$$L_N(\theta) = L_N(\bar{\theta}) + \langle \nabla L_{\bar{\theta}}(\theta - \bar{\theta}) \rangle + \sum_{j=2}^{\infty} \frac{1}{j!} \nabla^j L_N(\bar{\theta})(\theta - \bar{\theta})^j. \tag{4}$$

Here $\bar{\theta}$ is any initial estimator of $\theta$, for example, the local empirical loss minimizer $\arg \min_{\theta} L_1(\theta)$ in
the first machine $M_1$. Because the data is split across machines, evaluating the derivatives $\nabla^j L_N(\bar{\theta})$
($j \geq 1$) requires one communication round. However, unlike the $d$-dim gradient vector $\nabla L_N(\bar{\theta})$,
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 $\nabla^j L_N(\bar{\theta})$ ($j \geq 2$) with the
local ones, leading to the following modification of $L_N$,

$$L_N(\bar{\theta}) + \langle \nabla L_{\bar{\theta}}(\theta - \bar{\theta}) \rangle + \sum_{j=2}^{\infty} \frac{1}{j!} \nabla^j L_1(\bar{\theta})(\theta - \bar{\theta})^j = L_N(\bar{\theta}) + L_1(\theta) + \langle \nabla L_{\bar{\theta}} - \nabla L_1(\bar{\theta}), \theta - \bar{\theta} \rangle. \tag{5}$$

Therefore, we define the surrogate loss to be (ignoring some additive constants)

$$\tilde{L}(\theta) := L_1(\theta) - \langle \theta, \nabla L_1(\bar{\theta}) - \nabla L_N(\bar{\theta}) \rangle. \tag{6}$$
Comparing expressions (4) and (5), we see that the approximation error is (up to a constant)
\[
\tilde{L}(\theta) - L_N(\theta) = C + \frac{1}{2} \langle \theta - \overline{\theta}, (\nabla^2 L_1(\overline{\theta}) - \nabla^2 L_N(\overline{\theta})) (\theta - \overline{\theta}) \rangle + O(\|\theta - \overline{\theta}\|^2)
\]
\[
= C + O\left(\frac{1}{\sqrt{n}}\|\theta - \overline{\theta}\|^2 + \|\theta - \overline{\theta}\|^2\right),
\]
where we used the fact that \(\|\nabla^2 L_N(\theta) - \nabla^2 L_1(\overline{\theta})\|_2 = O(n^{-1/2})\) by using matrix concentration, and constant \(C = -L_N(\overline{\theta}) - \langle \overline{\theta}, \nabla L_1(\overline{\theta}) - \nabla L_N(\overline{\theta}) \rangle\). Let us consider three examples of using surrogate loss function \(\tilde{L}(\theta)\).

**Example (M-estimator):** In the low-dimensional regime where the dimensionality \(d\) of parameter space is \(o(N)\), the global empirical loss minimizer \(\hat{\theta} = \arg\min_{\theta \in \Theta} L_N(\theta)\) is a good estimator of \(\theta\) that achieves the root-\(N\) rate of convergence under mild conditions. One may also construct confidence regions associated with \(\hat{\theta}\) using the sandwiched covariance matrix (for example, see formula (10)). In our distributed learning framework, we propose the following estimator by pretending the surrogate loss function \(\tilde{L}\) as the global loss function \(L_N(\theta)\),
\[
\tilde{\theta} = \arg\min_{\theta \in \Theta} \tilde{L}(\theta).
\]

In Section 3.1, we show that \(\tilde{\theta}\) is equivalent to \(\hat{\theta}\) up to higher-order terms, and provide two ways to construct confidence regions for \(\tilde{\theta}\) using local observations stored in machine \(M_1\).

Now we give a heuristic understanding of why \(\tilde{\theta}\) is a good estimator. For convenience, we may assume that the empirical risk function \(L_N(\theta)\) has a unique minimizer. First look at the global empirical loss minimizer \(\hat{\theta}\). Under our assumption that the loss function is twice-differentiable, \(\hat{\theta}\) is the unique solution of equation:

\[
0 = \nabla L_N(\hat{\theta}) \approx \nabla L_N(\theta^*) + \nabla^2 L_N(\theta^*) (\hat{\theta} - \theta^*)
\]

By solving this equation, we obtain \(\|\hat{\theta} - \theta^*\|_2 = O_p(\|\nabla L_N(\theta^*)\|_2) = O_p(N^{-1/2})\), as long as the Hessian matrix \(\nabla^2 L_N(\theta^*)\) is non-singular. Now let us turn to the surrogate loss minimizer \(\tilde{\theta}\).

---

2There is no Taylor’s theorem for vector-valued functions, but we formalize this heuristic in Section 3.1.
Similar argument leads to \(\|\tilde{\theta} - \theta^*\|_2 = O_p(\|\nabla L(\theta^*)\|_2)\) and we only need to show that \(\|\nabla \tilde{L}(\theta^*)\|_2\) is of order \(O_p(N^{-1/2})\). In fact, by our construction,

\[
\nabla \tilde{L}(\theta^*) = (\nabla L_1(\theta^*) - \nabla L_1(\overline{\theta})) - (\nabla L_N(\theta^*) - \nabla L_N(\overline{\theta})) + \nabla L_N(\theta^*) \\
\approx \langle \nabla^2 L_1(\theta^*) - \nabla^2 L_N(\theta^*), \theta^* - \overline{\theta} \rangle + O_p(N^{-1/2}) \\
= O_p(n^{-1/2} \|\theta^* - \overline{\theta}\|_2) + O_p(N^{-1/2}),
\]

which is of order \(O_p(N^{-1/2})\) as long as \(\|\theta^* - \overline{\theta}\|_2 = O_p(k^{-1/2})\) with \(k = N/n\) the number of machines. For example, this requirement on initial estimator \(\overline{\theta}\) is satisfied for the minimizer \(\hat{\theta}_1\) of the sub-sample loss function \(L_1(\theta)\) 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 some low-dimensional structural assumption, such as the sparsity assumption, on the unknown true parameter \(\theta^*\). Under such an assumption, regularized estimators are popular and widely used for estimating \(\theta\). For concreteness, we focus on the sparsity assumption that most components of the \(d\)-dim vector \(\theta^*\) is zero, and consider the \(\ell_1\)-regularized estimator

\[
\hat{\theta} := \arg \min_{\theta \in \Theta} \left\{ L_N(\theta) + \lambda \|\theta\|_1 \right\}
\]

as the benchmark estimator that we want to approximate. Here \(\lambda\) is the regularizing parameter. In the distributed learning framework, we consider the following estimator using surrogate loss function \(\tilde{L}\),

\[
\tilde{\theta} = \arg \min_{\theta \in \Theta} \left\{ \tilde{L}(\theta) + \lambda \|\theta\|_1 \right\}.
\]

In Section 3.2, we show that \(\tilde{\theta}\) achieves the same rate of convergence as the benchmark estimator \(\hat{\theta}\) under a set of mild conditions. This idea of using the surrogate loss function to approximate the global loss function is general and is applicable to other high-dimensional problems.

**Example (Bayesian inference):** In Bayesian framework, viewing parameter \(\theta\) as random, we place a prior distribution \(\pi\) over parameter space \(\Theta\). Without confusion, we also use the same
notation \( \pi(\theta) \) to denote the pdf of the prior distribution at point \( \theta \). According to Bayes’ rule, the posterior distribution satisfies

\[
\pi(\theta | Z_1^N) \propto \exp\{ -N L_N(\theta) \} \pi(\theta).
\]

The loss function \( L \) corresponds to the negative log-likelihood function of the statistical model \( \{P_\theta : \theta \in \Theta\} \) and \( L_N(\theta) \) is the global negative log-likelihood of 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 \( \theta \) as the posterior expectation and use the highest posterior region as a credible region for this estimator. Since the additive constant \( C \) in expression (7) can be absorbed into the normalizing constant, we may use the surrogate posterior distribution

\[
\tilde{\pi}_N(\theta | Z_1^N) \propto \exp\{ -N \tilde{L}(\theta) \} \pi(\theta)
\]

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 \( \tilde{L}(\cdot) \) to approximate the global loss function \( L(\cdot) \) for distributed statistical learning as a Communication-efficient Surrogate Likelihood (CSL) method. As a remark, the idea of computing the global likelihood-function using sub-samples is useful not only in the distributed learning framework, but also in a common learning framework if the sample size is so large that even a single evaluation of the likelihood function or its gradient can be expensive. Using our surrogate loss function \( \tilde{L}(\theta) \), we only need one pass over the entire dataset to construct \( \tilde{L}(\theta) \). After its construction, \( \tilde{L}(\theta) \) can be efficiently evaluated by using a small subset of the data.

### 3. MAIN RESULTS AND THEIR CONSEQUENCES

In this section, we elaborate on the three examples in Section 2.2 of applying the CSL method. For each of the examples, we provide explicit bound on either the estimation error \( \|\tilde{\theta} - \theta^*\|_2 \) of the resulting estimator \( \tilde{\theta} \) or the approximation error \( \|\tilde{\pi}_N - \pi_N\|_1 \) of the approximated posterior \( \tilde{\pi}_N(\cdot) \).
3.1 Communication-efficient \( M \)-estimators in low-dimensions

In this subsection, we consider low-dimensional parametric family \( \{ \mathbb{P}_\theta : \theta \in \Theta \} \), where the dimensionality \( d \) of \( \theta \) is much smaller than 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 \( \theta^* \) 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 of the observation under \( \mathbb{P}_\theta \).

Note that the developments in this subsection can also be extended to misspecified models 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 \( \theta^* \) 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 omitted in this paper.

For low-dimensional parametric models, we impose some regularity conditions on the parameter space, the loss function \( L \) and the associated population risk function \( L^* \). 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 \( \Theta \) and the true parameter \( \theta^* \).

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

The second assumption is a local identifiability condition, ensuring \( \theta^* \) to be a local minimum of \( L^* \).

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

Since the loss function is the negative log-likelihood function, the corresponding Hessian matrix is called the 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
\[
\liminf_{n \to \infty} P\left\{ \inf_{\|\theta - \theta^*\|_2 \geq \delta} (L_1(\theta) - L_1(\theta^*)) \geq \epsilon \right\} = 1.
\]

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

Assumption PD (Smoothness): There exist constants \((G, L)\) and a function \( M(z) \) such that
\[
E[\|\nabla L(\theta; Z)\|_2^{16}] \leq G^{16}, \quad E[\|\nabla^2 L(\theta; Z) - I(\theta)\|_2^{16}] \leq L^{16}, \quad \text{for all } \theta \in U,
\]
\[
\|\nabla^2 L(\theta; z) - \nabla^2 L(\theta'; z)\|_2 \leq M(z) \|\theta - \theta'\|_2, \quad \text{for all } \theta, \theta' \in U.
\]
Moreover, the function \( M(z) \) satisfies \( E[M(z)^{16}] \leq M^{16} \) for some constant \( M > 0 \).

Following the heuristic argument in Section 2.2, we use the surrogate function \( \tilde{L} \) defined in (6) 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 \( \bar{\theta} \), any minimizer \( \tilde{\theta} \) of \( \tilde{L}(\theta) \), i.e.
\[
\tilde{\theta} = \arg\min_{\theta \in \Theta} \tilde{L}(\theta),
\]
significantly boosts the accuracy in terms of the approximation error \( \|\tilde{\theta} - \hat{\theta}\|_2 \) to the global empirical risk minimizer \( \hat{\theta} = \arg\min_{\theta \in \Theta} \mathcal{L}_N(\theta) \).

**Theorem 3.1.** Suppose that Assumptions PA-PD hold and the initial estimator \( \bar{\theta} \) lies in the neighbour \( U(\rho) \) of \( \theta^* \). Then any minimizer \( \tilde{\theta} \) of surrogate loss function \( \tilde{L}(\theta) \) satisfies
\[
\|\tilde{\theta} - \hat{\theta}\|_2 \leq C_2 \|\tilde{\theta} - \bar{\theta}\|_2 + \|\tilde{\theta} - \theta^*\|_2 + \|\nabla^2 \mathcal{L}_1(\theta^*) - \nabla^2 \mathcal{L}_N(\theta^*)\|_2 \|\bar{\theta} - \tilde{\theta}\|_2
\]
with probability at least \( 1 - C_1 kn^{-8} \). Here constants \( (C_1, C_2) \) are independent of \((k, n, N)\).

Under the conditions of Theorem 3.1, it can be shown that \( \|\tilde{\theta} - \theta^*\|_2 = O_p(N^{-1/2}) \) 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
\[ \|\tilde{\theta} - \hat{\theta}\|_2 = (O_p(n^{-1/2}) + \|\bar{\theta} - \tilde{\theta}\|_2) \quad \|\hat{\theta} - \tilde{\theta}\|_2 = O_p(n^{-1/2}) \|\bar{\theta} - \hat{\theta}\|_2, \]
as long as \( \|\bar{\theta} - \hat{\theta}\|_2 = O_p(n^{-1/2}) \), which is true for \( \theta = \hat{\theta}_1 := \arg \min_{\theta} L_1(\theta) \), the empirical risk minimizer in local machine \( \mathcal{M}_1 \). To formalize this argument, we have the following corollary that provides an \( \ell_2 \) risk bound for \( \tilde{\theta} \).

**Corollary 3.2.** **Under conditions of Theorem 3.1, we have**
\[ \mathbb{E}[\|\hat{\theta} - \theta^*\|_2^2] \leq \frac{A}{N} + \frac{C}{N\sqrt{N}} + \frac{C}{\sqrt{NK}} \min \left\{ \frac{1}{\sqrt{n}}, \left( \mathbb{E}[\|\bar{\theta} - \hat{\theta}\|_2^4] \right)^{1/4} \right\} + \frac{C}{n^4} \sqrt{k}, \]
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 \( \hat{\theta}_N \) based on \( N \) samples,
\[ \lim_{c \to \infty} \lim_{N \to \infty} \sup_{\theta \in U(c/\sqrt{N})} N \mathbb{E}_\theta[\|\hat{\theta}_N - \theta\|_2^2] \geq A. \]
Therefore, the estimator \( \tilde{\theta} \) 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 \( \tilde{\mathcal{L}}(\theta) \) in (8) can be further reduced by using a local quadratic approximation to \( \mathcal{L} \). In fact, we have by Taylor’s theorem that
\[ \mathcal{L}_N(\theta) \approx \mathcal{L}_N(\bar{\theta}) + \langle \nabla \mathcal{L}_N(\bar{\theta}), \theta - \bar{\theta} \rangle + \frac{1}{2} \langle \theta - \bar{\theta}, \nabla^2 \mathcal{L}_N(\bar{\theta} - \theta) \rangle. \]
Similar to before, we replace the global gradient \( \nabla \mathcal{L}_N(\bar{\theta}) \) with the local one \( \nabla \mathcal{L}_1(\bar{\theta}) \), which leads to the quadratic surrogate loss,
\[ \tilde{\mathcal{L}}^H(\theta) := \langle \nabla \mathcal{L}_N(\bar{\theta}), \theta - \bar{\theta} \rangle + \frac{1}{2} \langle \theta - \bar{\theta}, \nabla^2 \mathcal{L}_1(\bar{\theta} - \theta) \rangle. \]
Because surrogate loss functions \( \tilde{\mathcal{L}}^H \) and \( \tilde{\mathcal{L}} \) agree up to the second-order Taylor expansion, they behave similarly when used as objective functions for constructing \( M \)-estimators. This motivates the estimator,
\[ \hat{\theta}^H := \arg \min_{\theta \in \Theta} \tilde{\mathcal{L}}^H(\theta) = \bar{\theta} - \nabla^2 \mathcal{L}_1(\bar{\theta})^{-1} \nabla \mathcal{L}_N(\bar{\theta}), \]
which can be computed in closed-form using matrix inversion. The next theorem shows that \(\tilde{\theta}^H\) satisfies the same estimation bound as \(\tilde{\theta}\). Unlike the classical one-step MLE that requires the initial estimator to be within an \(O(N^{-1/2})\) neighbourhood of the truth \(\theta^*\), we only require \(\|\tilde{\theta} - \theta^*\|_2\) to be \(O(n^{-1/2})\).

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

\[
\|\tilde{\theta}^H - \hat{\theta}\|_2 \leq C_2' \left(\|\tilde{\theta} - \hat{\theta}\|_2 + \|\tilde{\theta} - \theta^*\|_2 + \|\nabla^2 L_1(\theta^*) - \nabla^2 L_N(\theta^*)\|_2\right) \|\tilde{\theta} - \hat{\theta}\|_2
\]

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

The analogue of Corollary 3.2 can also be stated for \(\tilde{\theta}^H\).

**Iterative local estimation algorithm:** It is worth mentioning that Theorem 3.1 (Theorem 3.3) suggests that an Iterative Local Estimation Algorithm (abbreviated as ILEA, see Algorithm 1) reduces the approximation error \(\|\tilde{\theta} - \hat{\theta}\|_2\) by a factor of \(n^{-1/2}\) in each iteration as long as the initial estimator satisfies \(\|\tilde{\theta} - \hat{\theta}\|_2 = O_p(n^{-1/2})\), or equivalently, \(\|\tilde{\theta} - \theta^*\|_2 = O_p(n^{-1/2})\). More precisely, in each iteration of ILEA, we set \(\tilde{\theta}\) as the current iterate \(\theta(t)\), construct the surrogate loss function \(\tilde{L}(t)\), and then solve the next iterate \(\theta(t+1)\) by either exactly minimizing the surrogate loss as

\[
\theta(t+1) \in \arg \min_{\theta \in \Theta} \tilde{L}(t)(\theta),
\]

or by the local one-step quadratic approximation as

\[
\theta(t+1) = \theta(t) - \nabla^2 L_1(\theta(t))^{-1} \nabla L_N(\theta(t)) = \arg \min_{\theta \in \Theta} \tilde{L}^{H}(t)(\theta).
\]

Theorem 3.1 (or Theorem 3.3) guarantees, with high probability, the error bound

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

where \(C_3\) is positive constant independent of \((n, k, N)\). If the desired accuracy is the statistical accuracy \(\|\tilde{\theta} - \theta^*\|_2\) of the MLE and our initial estimator is \(O(n^{-1/2})\) consistent, then we need proceed at most \(\left\lceil \frac{\log k}{\log n} \right\rceil\) 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 degenerates to a preconditioned gradient method. By appropriately choosing the sub-sample size \( n \), ILEA achieves a trade-off among storage, communication, and computational complexities, depending on specific constraints of computing resources.

\[
\begin{align*}
1 & \text{ Initialize } \theta^{(0)} = \bar{\theta}; \\
2 & \text{ for } t = 0, 1, \ldots, T - 1 \text{ do} \\
3 & \quad \text{ Transmit the current iterate } \theta^{(t)} \text{ to local machines } \{M_j\}_{j=1}^k; \\
4 & \quad \text{ for } j = 1 : k \text{ do} \\
5 & \quad \quad \text{ Compute the local gradient } \nabla L_j(\theta^{(t)}) \text{ at machine } M_j; \\
6 & \quad \quad \text{ Transmit the local gradient } \nabla L_j(\theta^{(t)}) \text{ to machine } M_1; \\
7 & \quad \text{ end} \\
8 & \quad \text{ Calculate the global gradient } \nabla L_N(\theta^{(t)}) = \frac{1}{k} \sum_{j=1}^k \nabla L_j(\theta^{(t)}) \text{ in Machine } M_1; \\
9 & \quad \text{ Form the surrogate function } \tilde{L}^t(\theta) = L_1(\theta) - \langle \theta, \nabla L_1(\theta^{(t)}) - \nabla L_N(\theta^{(t)}) \rangle; \\
10 & \quad \text{ Do one of the following in Machine } M_1: \\
11 & \quad \quad (1) \text{ Update } \theta^{(t+1)} \in \arg\min_{\theta \in \Theta} \tilde{L}^t(\theta); // \text{ Exactly minimizing surrogate function } \tilde{L} \\
12 & \quad \quad (2) \text{ Update } \theta^{(t+1)} = \theta^{(t)} - \nabla^2 L_1(\theta^{(t)})^{-1} \nabla L_N(\theta^{(t)}); // \text{ One-step quadratic approximation} \\
13 & \text{ end} \\
14 & \text{ return } \theta^{(T)}
\end{align*}
\]

**Algorithm 1:** Iterative local estimation

**Confidence region construction:** Next, we consider a natural class of local statistical inference procedures based on the surrogate function \( \tilde{L}(\theta) \) that only uses the sub-sample \( \{z_i\}_{i=1}^n \) in Machine \( M_1 \). It is a classical result that under Assumptions PA-PD, the global empirical risk minimizer \( \hat{\theta} \) satisfies (see the proof of Corollary 3.4 in Section A.4)

\[
\begin{align*}
\hat{\theta} - \theta^* = -I(\theta^*)^{-1} \nabla L_N(\theta^*) + O_p(N^{-1}), \quad & \text{ and} \\
\sqrt{N}(\hat{\theta} - \theta^*) \to \mathcal{N}(0, \Sigma) \quad & \text{ in distribution as } N \to \infty,
\end{align*}
\]

16
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

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

is a consistent estimator of the asymptotic covariance matrix $\Sigma$, that is, $\hat{\Sigma} \to \Sigma$ in probability as $N \to \infty$. Based on the limiting distribution of $\sqrt{N} (\hat{\theta} - \theta^*)$ and the plug-in estimator $\hat{\Sigma}$, we can conduct statistical inference, for example, constructing confidence intervals for $\theta$.

The following corollary shows that for any reasonably good initial estimator $\bar{\theta}$, the asymptotic distribution of either the minimizer $\tilde{\theta}$ of the surrogate function $\tilde{\mathcal{L}}(\theta)$ or the local one-step quadratic approximated estimator $\tilde{\theta}$ matches that of the global empirical risk minimizer $\hat{\theta}$. Moreover, we also have a consistent estimator $\tilde{\Sigma}$ of $\Sigma$ only using the local information in Machine $M_1$. Therefore, we can conduct statistical inference locally without access to the entire data while achieving the same power as global statistical inference procedures.

**Corollary 3.4.** Under the same set of assumptions in Theorem 3.1, if the initial estimator $\bar{\theta}$ satisfies $\|\bar{\theta} - \theta^*\|_2 = O_p(n^{-1/2})$, then the surrogate minimizer $\tilde{\theta}$ satisfies

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

and if $\|\bar{\theta} - \theta^*\|_2 = o_P(\sqrt{N})$, then

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

Moreover, the following plug-in estimator

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

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

$$
\hat{\Sigma}' := \nabla^2 \tilde{\mathcal{L}}(\hat{\theta})^{-1} \left( \frac{1}{k} \sum_{j=1}^{k} \nabla \mathcal{L}_j(\hat{\theta}) \nabla \mathcal{L}_j(\hat{\theta})^T \right) \nabla^2 \tilde{\mathcal{L}}(\hat{\theta})^{-1}
$$

is also a consistent estimator for $\Sigma$ as $(n, k) \to \infty$. Similar results hold for the local one-step quadratic approximated estimator $\tilde{\theta}^H$ under the assumptions of Theorem 3.3.
Corollary 3.4 illustrates that we may pretend \( \tilde{L}(\theta) \) as the global loss function and use it for statistical inference—\( \tilde{\Sigma} \) is precisely the plug-in estimator of the sandwiched covariance matrix using surrogate loss function \( \tilde{L}(\theta) \) (cf. equation (11)). In the special case when \( L(\theta) \) is the negative log-likelihood function, we may instead use \( \nabla^2 \tilde{L}(\tilde{\theta})^{-1} \) as our plug-in estimator for \( \Sigma = I(\theta^*)^{-1} = E[\nabla^2 L(\theta^*)^{-1}] \). \( \tilde{\Sigma}' \) tends to be a better estimator than \( \tilde{\Sigma} \) when \( k \gg n \), since variance \( \mathcal{O}(k^{-1}) \) of the middle term in equation (13) is much smaller than variance \( \mathcal{O}(n^{-1}) \) of the middle term in equation (12). See Section 4.1 for an empirical comparison of using \( \hat{\Sigma} \) and \( \hat{\Sigma}' \) for constructing confidence intervals.

3.2 Communication-efficient regularized estimators with \( \ell_1 \)-regularizer

In this subsection, we consider high-dimensional estimation problems where the dimensionality \( d \) of parameter \( \theta \) can be much larger than the sample size \( n \). Although the development here can apply to a broader class of problems, we focus on \( \ell_1 \)-regularized procedures. \( \ell_1 \)-regularized estimators work well under the sparsity assumption that most components of the true parameter \( \theta^* \) is zero. Let \( S = \text{supp}(\theta^*) \) be a subset of \( \{1, \ldots, d\} \) that encodes the sparsity pattern of \( \theta^* \) and \( s = |S| = \sum_{j=1}^{d} \mathbb{I}(\theta^*_j \neq 0) \). Applying the same idea of using the surrogate loss function \( \tilde{L}(\theta) \) as a proxy to the global likelihood function in \( \ell_1 \)-regularized estimation procedures, we propose the following communication-efficient regularized estimator

\[
\tilde{\theta} \in \arg \min_{\theta \in \Theta} \{ \tilde{L}(\theta) + \lambda \|\theta\|_1 \}.
\]

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

We first present a theorem on the statistical error bound \( \|\tilde{\theta} - \theta^*\|_2 \) of estimator \( \tilde{\theta} \) for general loss function \( L \). Then we will verify the conditions in the theorem for high-dimensional linear models and generalized linear models. To begin with, we state our assumptions.

**Assumption HA (Restricted strongly convexity):** Local loss function \( L_1(\theta) \) at machine \( L_1 \) is restricted strongly convex over \( S \): for all \( \delta \in C(S) := \{ v : \|v_S\|_1 \leq 3 \|v_{\bar{S}}\|_1 \} \),

\[
L_1(\theta^* + \delta) - L_1(\theta^*) - \nabla L_1(\theta^*)^T \delta \geq \mu \|\delta\|_2^2,
\]

where \( \delta \) is some positive constant independent of \( n \).
As the name suggested, restricted strongly convexity requires the global loss function $L_n(\theta)$ to be a strongly convex function when restricted to the cone $C(S)$.

Assumption HB (Restricted Lipschitz Hessian): Both local and global loss function $L_1(\theta)$ and $L_N(\theta)$ have restricted Lipschitz Hessian at radius $R$: for all $\delta \in C(S) \cap B_R(\theta^*)$,

$$
\| (\nabla^2 L_1(\theta^* + \delta) - \nabla^2 L_1(\theta^*)) \|_\infty \leq M \| \delta \|_2^2, \quad \text{and}
$$

$$
\| (\nabla^2 L_N(\theta^* + \delta) - \nabla^2 L_N(\theta^*)) \|_\infty \leq M \| \delta \|_2^2,
$$

where $M$ is some positive constant independent of $N$.

The restricted Lipschitz Hessian condition is always satisfied for linear models where the Hessian $\nabla^2 L_N(\theta)$ is a constant function of $\theta$.

Theorem 3.5. Suppose that Assumption HA and Assumption HB at radius $R > \| \bar{\theta} - \theta^* \|_2$ are true. If regularization parameter $\lambda$ satisfies $\lambda \geq 2 \| \nabla L_N(\theta^*) \|_\infty + 2 \| \nabla^2 L_N(\theta^*) - \nabla^2 L_1(\theta^*) \|_\infty \| \bar{\theta} - \theta^* \|_1 + 4M \| \bar{\theta} - \theta^* \|_2^2$, then

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

The lower bound condition on the regularization parameter $\lambda$ for $\tilde{\theta}$ is slightly stronger than that for the estimator $\hat{\theta}$ based on the global loss function, which is $\lambda \geq 2 \| \nabla L_N(\theta^*) \|_\infty$. Since the estimation error upper bound provided by Theorem 3.5 is proportional to the regularization parameter, it is reasonable to expect that $\tilde{\theta}$ will have a slight larger error than $\hat{\theta}$, depending on how good the initial estimator $\bar{\theta}$ is. For example, in generalized linear models, if smallest regularization parameters $\lambda$ are chosen for $\tilde{\theta}$ and $\hat{\theta}$, then the estimation error of $\tilde{\theta}$ will be greater than that of $\hat{\theta}$ by an amount of

$$
\frac{6s \sqrt{s}}{\sqrt{\mu}} \left( \| \nabla^2 L_N(\theta^*) - \nabla^2 L_1(\theta^*) \|_\infty \| \bar{\theta} - \theta^* \|_1 + 2M \| \bar{\theta} - \theta^* \|_2^2 \right)
$$

$$
\sim \sqrt{s \log d \over n} \| \bar{\theta} - \theta^* \|_1 + M \sqrt{s} \| \bar{\theta} - \theta^* \|_2^2.
$$

19
As long as $\|\theta - \theta^*\|_1$ and $\|\theta - \theta^*\|_2$ are sufficiently small, this difference will be negligible with respect to the estimation error bound of $\bar{\theta}$, which is $\sqrt{s \log d / N}$. For example, we may choose $\bar{\theta}$ to be the local $\ell_1$ regularized estimator $\hat{\theta}_1 := \arg \min \theta \{ L_1(\theta) + \lambda_1 \|\theta\| \}$ with estimation error $\sqrt{s \log d / n}$, so that

$$\|\hat{\theta}_1 - \theta^*\|_1 \leq Cs \sqrt{\log d / n} \quad \text{and} \quad \|\hat{\theta}_1 - \theta^*\|_2 \leq C \sqrt{s \log d / n}.$$ 

We may also consider an iterative estimation procedure analogous to Algorithm 1 for improving the higher-order estimation accuracy of the communication-efficient regularized estimator $\tilde{\theta}$. The convergence rate can be analyzed by inducting on Theorem 3.5. Now we apply Theorem 3.5 to two examples.

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

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

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 $s = \sum_{j=1}^d I(\theta_j^* \neq 0) = o(n)$. For linear regression, the global loss function takes the form as

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

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

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

Let $\Sigma = \mathbb{E}[x_{ij} x_{ij}^T]$ be the covariance matrix of the design. For this class of design, it is known that Assumption HA is satisfied with high probability as long as $\Sigma$ is strictly positive definite and $n \geq C_0 s \log d$ for some constant $C_0 > 0$ depending on the minimal eigenvalue of $\Sigma$ Raskutti et al. [2010]. For linear model, 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, $\Sigma$ is strictly positive definite and $n \geq C_0 s \log d$, then with probability at least $1 - c_1 \exp\{-c_2 n\}$, it holds that

$$\|\bar{\theta} - \theta^*\|_2 \leq C_1 A \sqrt{s \log d / N} + C_1 A \sqrt{s \log d / n} \|\theta - \theta^*\|_1.$$ 

20
If the initial estimator satisfies \( \| \tilde{\theta} - \theta^* \|_1 \leq C_2 s \sqrt{\frac{\log d}{n}} \), then with the same probability, it holds that
\[
\| \tilde{\theta} - \theta^* \|_2 \sim C_1 A \sqrt{\frac{s \log d}{N}} + C_3 \frac{s^{3/2} \log d}{n}.
\]
The constants \((c_1, c_2, C_0, C_1, C_2, C_3)\) are independent of \((n, k, d, s)\).

For sparse linear regression under the sparsity condition, the minimax rate of estimating \( \theta \) is \( \sqrt{\frac{s \log d}{N}} \). Therefore, Theorem 3.6 shows that our approximated estimator \( \tilde{\theta} \) 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 \( \ell_1 \)-regularizer. We begin with some background on generalized linear models. Recall that the data \( z_{ij} = (x_{ij}, y_{ij}) \), 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
\[
\mathbb{P}(y_{ij} | x_{ij}, \theta, \sigma) \propto \exp \left\{ \frac{y_{ij} x_{ij}^T \theta - \phi(x_{ij}^T \theta)}{\sigma} \right\},
\]
where \( \sigma \) is a scalar parameter, \( \theta \) is the unknown \( d \)-dim parameter to be estimated and \( \phi \) 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 is given by
\[
\mathcal{L}_N(\theta) = \frac{1}{N} \sum_{j=1}^k \sum_{i=1}^n -y_{ij} x_{ij}^T \theta + \phi(x_{ij}^T \theta), \quad \text{and}
\]
\[
\nabla \mathcal{L}_N(\theta) = \frac{1}{N} \sum_{j=1}^k \sum_{i=1}^n (\phi'(x_{ij}^T \theta) - y_{ij}) x_{ij}.
\]

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
\[
\| \tilde{\theta} - \theta^* \|_2 \leq C_1 A \sqrt{\frac{s \log d}{N}} + C_1 A \sqrt{\frac{s \log d}{n}} \| \tilde{\theta} - \theta^* \|_1 + C_1 A \sqrt{s} \| \tilde{\theta} - \theta^* \|_2^2.
\]
If $||\theta - \theta^*||_1 \leq C_2 s \sqrt{\frac{\log d}{n}}$ and $||\theta - \theta^*||_2 \leq C_2 \sqrt{\frac{s \log d}{n}}$, then with the same probability, we have

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

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 a Bayesian framework of distributed statistical learning for regular parametric models. In a standard Bayesian framework, one specifies a prior distribution $\pi$ over the parameter space $\Theta$ and then conducts statistical inference by sampling from the global posterior distribution

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

where $L : \Theta \times Z \rightarrow \mathbb{R}$ is the negative log-likelihood function and $D$ is the normalizing constant. In the rest of this subsection, we tacitly assume that the loss function $L$ introduced in Section 2.1 is the negative log-likelihood function. Extensions to the Gibbs posterior [Bissiri Pier and Stephen 2013] where $L$ is replaced with a generic loss function $\mathcal{L}$ in posterior (14) is also straightforward. Most existing literature [Zhang et al. 2013a, Lee et al. 2015, Wang and Dunson 2015, Neiswanger et al. 2015] in distributed Bayesian learning framework utilizes the decomposition

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

so that the global posterior $\pi(\theta \mid Z_1^N)$ can be written as the product of sub-sample posteriors

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

where the prior is raised to power $k^{-1}$ so that it is appropriately weighted in product (15) and $D_j$ is the normalizing constant. In a typical MapReduce framework, separate Markov chains are run in machines $\{M_j\}_{j=1}^{k}$ based on the local data in an embarrassingly parallel manner so that the draws can be collected from each sub-sample posterior. After running individual Markov chains, all local posterior draws are transmitted to a central node. In the central node, an approximation $\tilde{\pi}_N(\theta)$ to the global posterior $\pi_N(\theta) := \pi(\theta \mid Z_1^N)$ is formed. A main drawback of these approaches is that
the communication cost can be extremely high—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.

Following the heuristic in Section 2.2, we propose a different sampling scheme for distributed Bayesian inference using the surrogate loss function $\tilde{L}(\theta)$. Our sampling scheme is communication efficient and requires 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 $\hat{\theta}^H$ in Section 3.1.
2. For $j = 1, \ldots, k$, compute the local gradient $\nabla L_j(\overline{\theta})$ in machine $M_j$.
3. Transmit all local gradients to Machine $M_1$ and form the global gradient $\nabla L_N(\overline{\theta}) = \frac{1}{k} \sum_{j=1}^{k} \nabla L_j(\overline{\theta})$.
4. Machine $M_1$ constructs the surrogate function $\tilde{L}(\theta)$ as (6).
5. Machine $M_1$ runs a Markov chain to sample from the surrogate posterior $\tilde{\pi}_N(\theta) \propto \exp(\sum_{j=1}^{k} \nabla L_j(\overline{\theta})) \pi(\theta)$, and uses the draws to conduct statistical inference.

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

**Theorem 3.8.** If Assumption PA-PD hold and $\| \overline{\theta} - \hat{\theta} \|_2 = O_p(N^{-1/2})$, then the approximate posterior $\tilde{\pi}_N(\theta)$ satisfies

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

where $\| P - Q \|_1 = \int |P(d\theta) - Q(d\theta)|$ is the $\ell_1$ distance between two distributions $P$ and $Q$.

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

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

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

4. SIMULATIONS

In this section, we conduct simulations for the three examples of using CSL methodology 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

$$Y_{ij} \sim \text{Ber}(P_{ij}), \quad \text{with} \quad \log \frac{P_{ij}}{1-P_{ij}} = \langle X_{ij}, \theta^* \rangle.$$ (16)

In our simulation, the true regression coefficient $\theta^*$ 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 $\theta^*$ from the $d$-dim unit cube $[0, 1]^d$.

We implement the one-step Newton-Raphson estimator $\theta^{(1)}$ with the averaging estimator $\hat{\theta}^A$ 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 the our communication estimators with the (optimal) global $M$-estimator $\theta^{\text{global}}$ and the sub-sample estimator $\theta^{\text{sub}}$ 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 $\theta^{\text{global}}$ approximately remains fixed as $n$ varies. As expected, the rest estimators exhibit a rapid decay in the estimation error as the local sample size $n$ grows. Except for the global estimate $\theta^{\text{global}}$, our communication-efficient estimators
Figure 1: Squared estimation error $\|\hat{\theta} - \theta^*\|^2_2$ 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 $k \cdot 25 N/n$) under 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 $d \in \{2, 10, 50\}$. 
have the best performance. When $n$ is sufficiently large, the 1-step, 2-step and 3-step estimators have almost the same performance as $\theta^{\text{global}}$. 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 $\|\theta^{(t+1)} - \theta^{\text{global}}\|_2 / \|\theta^{(t)} - \theta^{\text{global}}\|_2$ is dominated by the sum of two terms: the initial estimation error $\|\theta^{(t)} - \theta^{\text{global}}\|_2$ and the local Hessian approximation error $| | |\nabla L_1(\theta^*) - \nabla L_N(\theta^*)| | |_2$. 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 applying the iterative procedure. We remark that the condition that the local size $n$ should exceed a $d$-dependent threshold is a mild requirement in practice. Indeed, the local machine storage limit in reality is often large enough to ensure $n \gg d$. Even under the scenario (small $n$) where our theory fails to predict, the 1-step, 2-step and 3-step estimators still have better performance than $\hat{\theta}^A$ and $\theta^{\text{sub}}$. 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 $\hat{\theta}^A$ does not improve as more data is available. This is consistent with the theoretical results in [Zhang et al. 2013b] that requires $k \gg n$ for $\hat{\theta}^A$ to have comparable performance as $\theta^{\text{global}}$. By using our approach, even a single step application of Algorithm 1 significantly improves the accuracy of $\hat{\theta}^A$. Moreover, $\theta^{(2)}$ and $\theta^{(3)}$ achieve almost the same accuracy as $\theta^{\text{global}}$. Consistent with our theory, for a fixed number of steps $t$, the $t$-step estimate $\theta^{(t)}$ tends to have larger estimation error than $\theta^{\text{global}}$ 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 $\theta^{\text{global}}$. These simulation results illustrate the potential of our communication-efficient estimation procedures for large-scale distributed statistical learning.

Now we assess the performance of the inference procedures based on the plug-in estimators $\tilde{\Sigma}$ and $\tilde{\Sigma}'$ under the logistic model (16). We use $\tilde{\Sigma}$ or $\tilde{\Sigma}'$ and the 3-step estimator $\theta^{(3)}$ to construct a 95% confidence interval (CI) for the first component $\theta_1$ of $\theta$ as

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

The coverage of the CI based on 100 trials is calculated. Figure 2 shows the result. In plot (a), coverage based on both plug-in estimators are low at $n = 2^7$ because the sample size is so small.
Figure 2: Coverage of the confidence interval for the first component of $\beta$ 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, the plug-in est. 1 (2) corresponds to the confidence interval constructed based on the plug-in estimator $\tilde{\Sigma} (\tilde{\Sigma}')$ and the 3-step estimator $\theta^{(3)}$. 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 $N = nk$).
that the center $\theta^{(3)}$ of the CI has a large bias (see Figure 1(c)). In plot (b), the CI based on $\Sigma'$ 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 have coverage that is close to the nominal level 95%. Moreover, the CI based on $\Sigma'$ is slightly better than the one based on $\Sigma$ for large $k$, which empirically justifies our intuition in the discussion right 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 $\epsilon_{ij}$ is i.i.d. $\mathcal{N}(0, 1)$, and $\theta^*$ is $s$-sparse with signal-to-noise ratio $|\theta^*|_2 = 5$.

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 the distributed 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 only 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 $\frac{s \log d}{nk}$.

4.3 Distributed Bayesian inference

Our synthetic dataset is generated from the logistic model (16) under dimension $d \in \{2, 10, 50\}$. We use the 3-step estimator $\theta^{(3)}$ in Section 3.1 as the initial estimator $\theta$ and implement the Bayesian procedures based on the (approximated) posterior distribution $\pi_n(\theta)$ and $\tilde{\pi}_N(\theta)$ by sampling using Markov Chain Monte Carlo algorithm. We use the Metropolis algorithm, where at each iteration the proposal distribution for $\theta$ 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 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.

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 \( k \).
Figure 5: Marginal posterior distribution of the first component $\theta_1$ of $\theta$ for logistic regression under 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 randomly splits the data into $k$ subsamples. The vertical dotted line indicates the location of the truth $\theta^*_1$. 

(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 plots the (approximated) marginal posterior distributions of the first component $\theta_1$ of $\theta$ under different $(d,n,k)$ combinations ($n$ is chosen so that $\theta^{(3)}$ is a good approximation to the global estimator $\hat{\theta}$, see Figure 1). Consistent with our theoretical prediction, $\tilde{\pi}_N(\theta)$ provides a good approximation to $\pi_N(\theta)$ as long as the initial estimator $\overline{\theta}$ is sufficiently close to $\hat{\theta}$, even when $k$ is much larger than $n$ (see plot (b)). Since the computation of the approximate posterior distribution $\tilde{\pi}_N(\theta)$ only uses the local data in Machine $\mathcal{M}_1$, the computation of the acceptance ratio using $\tilde{\pi}_N(\theta)$ 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 the skin dataset [Rajen and Abhinav, 2012]. The goal is to predict whether a given color sample described by its B, G, R values (each ranges from 0 – 255) corresponds to a skin sample or non-skin sample. The dataset is generated using skin textures from face images of diversity of age, gender, and race people. The total sample size is 245,057, out of which 50,859 is the skin samples and 194,198 is non-skin samples. The skin 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 non-linear dependence between the response and features. Therefore, the dimension of the covariate $X$ is $d = 45$.

We randomly split the entire 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. Now, we exactly minimize the surrogate function to form the 1-step estimator $\theta^{(1)}$ with the averaging estimator $\hat{\theta}^A$ [Zhang et al., 2013b] 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 gains significant improvement on the 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 already has reasonably good performance.

![Graph showing test error for Logistic regression.](image)

Figure 6: Distributed logistic regression for the skin dataset.

6. CONCLUDING REMARKS

In this paper, we presented the Communication-efficient Surrogate Likelihood (CSL) framework for solving distributed statistical learning problems. We applied this methodology to three examples: low-dimensional $M$-estimation, high-dimensional regularized estimation and low-dimensional Bayesian inference. We believe that the general idea of constructing a surrogate function to the negative log-likelihood function (or general loss function) that can be efficient evaluated using only a sub-sample of the observations is also useful in big-data problems where the sample size is so large that even the calculation of the likelihood function is expensive. One future direction is to investigate how this framework can be applied to high-dimensional Bayesian inference with strong theoretical supports. Another direction is to find a sharp theoretical lower bound on the local sample size $n$ for the final estimator to remain optimal, for example, in the minimax sense.
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”:

\[
E_0 := \left\{ \|\hat{\theta} - \theta^*\|_2 \leq \min \left\{ \frac{\rho \mu_-}{8M}, \frac{(1-\rho)\mu_- \delta_\rho}{8\mu_+}, \sqrt{\frac{(1-\rho)\mu_- \delta_\rho}{16M}} \right\} \right\}, \quad \text{and} \quad \mathcal{E}_j := \left\{ M_j \leq 2M, \|\nabla^2 \mathcal{L}_j(\theta^*) - I(\theta^*)\|_2 \leq \frac{\rho \mu_-}{4}, \|\nabla \mathcal{L}_j(\theta^*)\|_2 \leq \frac{(1-\rho)\mu_- \delta_\rho}{4} \right\}.
\]

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

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

\[
P\left( \bigcup_{j=0}^{k} \mathcal{E}_j \right) \leq \left( c_1 + c_2 (\log 2d)^{16} L^{16} + c_3 G_{16} \right) \frac{k}{n^8}.
\]

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

The second result characterizes the error bound \( \|\tilde{\theta} - \hat{\theta}\|_2 \) in terms of the gradient norm \( \|\nabla \tilde{\mathcal{L}}(\hat{\theta})\|_2 \) at \( \hat{\theta} \), 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

\[
\|\tilde{\theta} - \hat{\theta}\|_2 \leq \frac{2 \|\nabla \tilde{\mathcal{L}}(\hat{\theta})\|_2}{(1-\rho)\mu_-}.
\]

Therefore, it remains to prove a high probability upper bound of the gradient norm \( \|\nabla \tilde{\mathcal{L}}(\hat{\theta})\|_2 \).

Simple calculation yields

\[
\nabla \tilde{\mathcal{L}}(\hat{\theta}) = \nabla \mathcal{L}_1(\hat{\theta}) - \nabla \mathcal{L}_1(\hat{\theta}) + \nabla \mathcal{L}_N(\hat{\theta}). \tag{A.1}
\]

By the optimality of the global empirical risk minimizer \( \hat{\theta} \), we have

\[
\nabla \mathcal{L}_N(\hat{\theta}) = 0.
\]

By adding and subtracting \( \nabla N(\hat{\theta}) \) in equation (A.1), we obtain

\[
\nabla \tilde{\mathcal{L}}(\hat{\theta}) = (\nabla \mathcal{L}_1(\hat{\theta}) - \nabla \mathcal{L}_1(\hat{\theta})) - (\nabla \mathcal{L}_N(\hat{\theta}) - \nabla \mathcal{L}_N(\hat{\theta})). \tag{A.2}
\]
By the integral form of Taylor’s expansion, we have that for any $j \in \{1, \ldots, k\}$,

$$
\nabla L_j(\hat{\theta}) - \nabla L_j(\overline{\theta}) = H_j(\hat{\theta} - \overline{\theta}),
$$

where $H_j = \int_0^1 \nabla^2 L_j(\overline{\theta} + t(\hat{\theta} - \overline{\theta})) \, dt$ satisfies

$$
\|H_j - \nabla^2 L_j(\theta^*)\|_2 \leq 2M (\|\overline{\theta}\|_2 + \|\hat{\theta} - \theta^*\|_2)
$$

under event $E_j$. Combining the three preceding displays, we obtain that under event $\bigcap_{j=0}^k E_k$,

$$
\|\nabla \tilde{L}(\hat{\theta})\|_2 \leq \|H_1 - \nabla^2 L_1(\theta^*)\|_2 \|\hat{\theta} - \overline{\theta}\|_2 + \frac{1}{k} \sum_{j=1}^k \|H_j - \nabla^2 L_j(\theta^*)\|_2 \|\hat{\theta} - \overline{\theta}\|_2
$$

$$
+ \|\nabla^2 L_1(\theta^*) - \nabla^2 L_N(\theta^*)\|_2 \|\hat{\theta} - \overline{\theta}\|_2
$$

$$
\leq (2M \|\hat{\theta} - \overline{\theta}\|_2 + 2M \|\hat{\theta} - \theta^*\|_2 + \|\nabla^2 L_1(\theta^*) - \nabla^2 L_N(\theta^*)\|_2) \|\hat{\theta} - \overline{\theta}\|_2.
$$

Combining Lemma A.2 and the above display yields the claimed error bound on $\|\hat{\theta} - \overline{\theta}\|_2$.

### A.2 Proof of Corollary 3.2

Recall definitions of events $\{E_j\}_{j=0}^k$ 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 $\bigcap_{j=0}^k E_j$, we have

$$
\|\hat{\theta} - \theta^* - I(\theta^*)^{-1} \nabla L_N(\theta^*)\|_2
$$

$$
\leq \frac{2}{(1 - \rho)\mu_-} \|\nabla^2 L_N(\theta^*) - I(\theta^*)\|_2 \|\nabla L_N(\theta^*)\|_2 + \frac{8M}{(1 - \rho)^2\mu_-^2} \|\nabla L_N(\theta^*)\|_2^2.
$$

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

$$
\|\hat{\theta} - \theta^* - I(\theta^*)^{-1} \nabla L_N(\theta^*)\|_2 \leq \|\hat{\theta} - \theta^* - I(\theta^*)^{-1} \nabla L_N(\theta^*)\|_2 + \|\overline{\theta} - \hat{\theta}\|_2
$$

$$
\leq C \|\nabla^2 L_N(\theta^*) - I(\theta^*)\|_2 \|\nabla L_N(\theta^*)\|_2 + C \|\nabla L_N(\theta^*)\|_2^2
$$

$$
+ C (\|\overline{\theta} - \hat{\theta}\|_2 + \|\nabla L_N(\theta^*)\|_2 + \|\nabla^2 L_1(\theta^*) - \nabla^2 L_N(\theta^*)\|_2) \|\overline{\theta} - \hat{\theta}\|_2.
$$

34
Now by applying Hölder’s inequality and Lemma B.1 in Appendix B.1 we obtain
\[
\mathbb{E} \left[ \left\| \left( \hat{\theta} - \theta^* - I(\theta^*)^{-1} \nabla L_N(\theta^*) \right) \mathcal{I} \left( \bigcap_{j=0}^{k} \mathcal{E}_j \right) \right\|^2 \right] \\
\leq C \sqrt{\mathbb{E}[\|\nabla^2 L_N(\theta^*) - I(\theta^*)\|_2^2]} \mathbb{E}[\|\nabla L_N(\theta^*)\|_2^2] + C \mathbb{E}[\|\nabla L_N(\theta^*)\|_2^2] \\
+ C \mathbb{E}[\|\bar{\theta} - \hat{\theta}\|_2^2] + C \sqrt{\mathbb{E}[\|\nabla^2 L_1(\theta^*) - \nabla^2 L_N(\theta^*)\|_2^2]} \mathbb{E}[\|\nabla L_N(\theta^*)\|_2^2] \sqrt{\mathbb{E}[\|\bar{\theta} - \hat{\theta}\|_2^2]} \\
\leq \frac{C}{N^2} + \frac{C}{n} \min \left\{ \frac{1}{n}, \sqrt{\mathbb{E}[\|\bar{\theta} - \hat{\theta}\|_2^2]} \right\}.
\]
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,
\[
\mathbb{E} \left[ \left\| \tilde{\theta} - \theta^* - I(\theta^*)^{-1} \nabla L_N(\theta^*) \right\|^2 \right] \leq \frac{C}{N^2} + \frac{C}{n} \min \left\{ \frac{1}{n}, \sqrt{\mathbb{E}[\|\bar{\theta} - \hat{\theta}\|_2^2]} \right\} + \frac{Ck}{n^8},
\]
which implies the claimed bound on \(\mathbb{E}[\|\bar{\theta} - \theta^*\|_2^2]\).

A.3 Proof of Theorem 3.3

Before analysing the one-step Newton-Raphson estimator \(\theta^H\), we prove some auxiliary results.

Recall that for \(j = 1, \ldots, k\), let \(M_j = \frac{1}{n} \sum_{i=1}^{n} M(z_{ij})\) and \(\delta_\rho = \min\{\rho, \rho \mu_- / 4M\}\). Similar to events \(\mathcal{E}_j\ (j = 0, 1, \ldots, k)\) in Section A.1, we define the following “good events”:

\[
\mathcal{E}_0^c : = \left\{ \|\tilde{\theta} - \theta^*\|_2 \leq \frac{\mu_-}{4M} \right\}, \quad \text{and}
\mathcal{E}_j^c : = \left\{ M_j \leq 2M, \|\nabla^2 L_j(\theta^*) - I(\theta^*)\|_2 \leq \frac{\rho \mu_-}{4}, \|\nabla L_j(\theta^*)\|_2 \leq \frac{(1 - \rho) \mu_- \delta_\rho}{4} \right\}.
\]

Then similar to Lemma A.1 we have that under Assumptions PA-PD,
\[
\mathbb{P} \left( \bigcup_{j=0}^{k} \mathcal{E}_j^c \right) \leq \left( \epsilon'_1 + \epsilon'_2 (\log 2d) \right) 16 + \epsilon'_3 G 16 \frac{k}{n^8},
\]
where \(\epsilon'_j\ (j = 1, 2, 3)\) 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 \(\bigcap_{j=0}^{k} \mathcal{E}_j^c\) we have
\[
\lambda_{\min}[\nabla^2 L_N(\tilde{\theta})] \geq \frac{1}{2} (1 - \rho) \mu_-, \quad \|\tilde{\theta} - \hat{\theta}\|_2 \leq \Delta := \frac{(1 - \rho) \mu}{8M},
\]
\[
U_N : = \max_{\theta \in (\tilde{\theta} - \Delta, \tilde{\theta} + \Delta)} \|\nabla^2 L_N(\theta)\|_2 \leq U : = 2M \Delta + \frac{\rho \mu}{4} + \mu_+, \quad \text{and}
\|\nabla^2 L_N(\tilde{\theta})^{-1} - \nabla^2 L_1(\tilde{\theta})^{-1}\|_2 \leq \left( \frac{2M \rho}{\mu^2_-} + \frac{\rho + 4}{4 \mu_-} \right) \left( \|\nabla^2 L_N(\theta^*) - \nabla^2 L_1(\theta^*)\|_2 + 4M \|\tilde{\theta} - \theta^*\|_2 \right).
\]

35
The proof of this lemma is provided in Appendix A.4.

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

The error can be decomposed as

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

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

$$\theta^H - \theta^N = (\bar{\theta} - \nabla^2 L_1(\bar{\theta})^{-1}\nabla L_N(\bar{\theta})) - (\bar{\theta} - \nabla^2 L_N(\bar{\theta})^{-1}\nabla L_N(\bar{\theta}))$$

$$= (\nabla^2 L_N(\bar{\theta})^{-1} - \nabla^2 L_1(\bar{\theta})^{-1}) \nabla L_N(\bar{\theta})$$

$$= (\nabla^2 L_N(\bar{\theta})^{-1} - \nabla^2 L_1(\bar{\theta})^{-1}) (\nabla L_N(\bar{\theta}) - \nabla L_N(\hat{\theta})),$$

which yields the bound

$$\|\theta^H - \theta^N\|_2 \leq U_N \|\nabla^2 L_N(\bar{\theta})^{-1} - \nabla^2 L_1(\bar{\theta})^{-1}\|_2 \|\bar{\theta} - \hat{\theta}\|_2.$$

The second term can be analysed by using Theorem 5.3 in Bubeck [2014], which guarantees that under the assumption $\|\bar{\theta} - \hat{\theta}\|_2 \leq \frac{\mu_N}{2M_N}$, it holds that

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

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

$$\|\theta^H - \hat{\theta}\|_2 \leq U_N \|\nabla^2 L_N(\bar{\theta})^{-1} - \nabla^2 L_1(\bar{\theta})^{-1}\|_2 \|\bar{\theta} - \hat{\theta}\|_2 + \frac{M_N}{\mu_N} \|\bar{\theta} - \hat{\theta}\|_2^2.$$

Now the claimed bounds on $\|\theta^H - \hat{\theta}\|_2$ 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 $\Sigma$ is standard by using the consistency of $\bar{\theta}$ 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 $\bar{\theta}$. Based on Theorem 3.1...
we only need to establish the asymptotic expansion (10) of the global empirical risk minimizer $\hat{\theta}$.

By the integral form of Taylor’s expansion, we have

$$0 = \nabla L_N(\hat{\theta}) = \nabla L_N(\theta^*) + H_N(\hat{\theta} - \theta^*),$$

where $H_N = \int_0^1 \nabla^2 L_N(\theta^* + t(\hat{\theta} - \theta^*)) \, dt$. Then simple linear algebra yields

$$\hat{\theta} - \theta^* = -I(\theta^*)^{-1} \nabla L_N(\theta^*) - U_N(\hat{\theta} - \theta^*) - V_N(\hat{\theta} - \theta^*), \tag{A.3}$$

where $U_N = H_N - \nabla^2 L_N(\theta^*)$ and $V_N = \nabla^2 L_N(\theta^*) - I(\theta^*)$. 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 $\tilde{L}(\theta)$ satisfies the restricted strong convexity condition.

The restricted strong convexity of $\tilde{L}$ is implied by the same property of $L_1$, since

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

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

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

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

$$\nabla \tilde{L}(\theta^*) = \nabla L_1(\theta^*) - \nabla L_1(\overline{\theta}) + L_N(\overline{\theta})$$

$$= (\nabla L_N(\overline{\theta}) - \nabla L_N(\theta^*)) - (\nabla L_1(\overline{\theta}) - \nabla L_1(\theta^*)) + \nabla L_N(\theta^*)$$

$$= \nabla^2 L_N(\theta^*)(\overline{\theta} - \theta^*) - \nabla^2 L_1(\theta^*)(\overline{\theta} - \theta^*)$$

$$+ \int_{s=0}^{s=1} ds(\nabla^2 L_N(\theta^* + s(\overline{\theta} - \theta^*)) - \nabla^2 L_N(\theta^*)(\overline{\theta} - \theta^*))$$

$$- \int_{s=0}^{s=1} ds(\nabla^2 L_1(\theta^* + s(\overline{\theta} - \theta^*)) - \nabla^2 L_1(\theta^*)(\overline{\theta} - \theta^*) + \nabla L_N(\theta^*))$$

Using Assumption HB,

$$\|\nabla \tilde{L}(\theta^*)\|_\infty \leq \|\nabla^2 L_N(\theta^*) - \nabla^2 L_1(\theta^*)\|_\infty \|\overline{\theta} - \theta^*\|_1 + \|\nabla L_N(\theta^*)\|_\infty$$

$$+ 2M \|\overline{\theta} - \theta\|_2$$
Proof of Theorem 3.6. To apply Theorem 3.5, we have to compute \( \| \nabla^2 L_N(\theta^*) - \nabla^2 L_1(\theta^*) \|_\infty \). Let \( \Sigma = E[xx^T] \).

\[
\| \nabla^2 L_N(\theta^*) - \nabla^2 L_1(\theta^*) \|_\infty = \left\| (\Sigma - \frac{1}{N}X^T X) + \frac{1}{n}X_1^T X_1 - \Sigma \right\|_\infty + \sigma \sqrt{\frac{2 \log d}{N}}
\]

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

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

Thus let \( t = C \sqrt{\frac{\log d}{N}} \), then \( \left\| \frac{1}{N}X^T X - \Sigma \right\|_{\max} < C \sqrt{\frac{\log d}{N}} \) with probability greater than \( 1 - 1/p^C \).

By a similar argument, \( \left\| \frac{1}{n}X_1^T X_1 - \Sigma \right\|_{\max} < C \sqrt{\frac{\log d}{n}} \).

Since \( \nabla^2 L \) is a constant in linear regression, \( M = 0 \). Thus

\[
\left\| \tilde{\theta} - \theta^* \right\|_2 \leq \sqrt{\frac{s \log d}{n}} \left\| \bar{\theta} - \theta^* \right\|_1 + \sqrt{\frac{s \log d}{N}}.
\]

Proof of Theorem 3.7. To apply Theorem 3.5, we need to compute verify Assumptions HA and HB. The restricted strong convexity of \( L \) is verified in Proposition 1 of Negahban et al. [2012]. Next we verify Assumption HB.

\[
\nabla^2 L_1(\theta^* + \delta) - \nabla^2 L_1(\theta^*) = \frac{1}{n} \sum_{i=1}^{n} (\phi''(x_{ij}^T \theta^* + x_{ij}^T \delta) - \phi''(x_{ij}^T \theta^*))x_{ij}x_{ij}^T
\]

\[
\quad = \frac{1}{n} \sum_{i=1}^{n} \phi''(x_{ij}^T \theta^* + s_{ij}x_{ij}^T \delta)x_{ij}(x_{ij}^T \delta)^2.
\]

Thus,

\[
\left\| \nabla^2 L_1(\theta^* + s\delta) - \nabla^2 L_1(\theta^*) \right\|_{\infty} \leq \left\| \frac{1}{n} \sum_{i=1}^{n} \phi''(x_{ij}^T \theta^* + s_{ij}x_{ij}^T \delta)x_{ij}(x_{ij}^T \delta)^2 \right\|_1
\]

\[
\quad \leq L_\phi B \left\| \frac{1}{n} \sum_{i=1}^{n} (x_{ij}^T \delta)^2 \right\|
\]

\[
\quad \leq L_\phi BL \| \delta \|_2^2,
\]

38
Thus $M = L_\phi BL$, where $L_\phi$ is a local upper bound on $\phi'''$, $L$ is the upper restricted eigenvalue of $X$, and $B = \max \|x\|_\infty$.

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

$$
\|\nabla^2 \mathcal{L}_N(\theta^*) - \nabla^2 \mathcal{L}_1(\theta^*)\|_\infty = \left( \frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{n} \phi''(x_{ij}^T \theta^*) x_{ij} x_{ij}^T - A \right) + \left( A - \frac{1}{n} \sum_{i=1}^{n} \phi''(x_{i1}^T \theta^*) x_{i1} x_{i1}^T \right)
$$

$$
\leq C \sqrt{\frac{\log d}{N}} + C \sqrt{\frac{\log d}{n}},
$$

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

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

Thus by Theorem 3.5, we have shown

$$
\|\tilde{\theta} - \theta^*\|_2 \leq C \sqrt{\frac{s \log d}{n}} \|\theta - \theta^*\|_1 + \sqrt{\frac{s \log d}{N} + C \|\theta - \theta^*\|_2^2}.
$$

\[ \Box \]

### A.6 Proof of Theorem 3.8

Recall the definition of the “good events” $E_j$ for $j = 1, \ldots, k$ in Section A.1 as

$$
E_j := \left\{ M_j \leq 2M, \|\nabla^2 \mathcal{L}_j(\theta^*) - I(\theta^*)\|_2 \leq \frac{\rho \mu_-}{4}, \|\nabla \mathcal{L}_j(\theta^*)\|_2 \leq \frac{(1 - \rho) \mu_\delta}{4} \right\}.
$$

Moreover, we define events

\begin{align*}
A_n &= \left\{ \inf_{\|\theta - \theta^*\|_2 \geq \delta} \frac{1}{n} (\mathcal{L}_1(\theta) - \mathcal{L}_1(\theta^*)) \geq 3\epsilon \right\}, \\
B_1 &= \left\{ \|\tilde{\theta} - \theta^*\|_2 \leq \frac{\epsilon}{4R} \min \left\{ \frac{1}{2\sqrt{M}}, \frac{1}{\rho \mu_- + 2\mu_+} \right\} \right\}, \quad \text{and} \\
B_2 &= \left\{ \sqrt{N} \mu_+ \|\tilde{\theta} - \theta\|_2 + 2M \sqrt{N} \|\tilde{\theta} - \hat{\theta}\|_2 + 2M \sqrt{N} \|\tilde{\theta} - \hat{\theta}\|_2 + M \|\hat{\theta} - \theta^*\|_2 + M \|\tilde{\theta} - \theta^*\|_2 \leq \mu_- / 16 \right\},
\end{align*}

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

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

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

$$
P \left( A_n^c \cup B_1^c \cup B_2^c \cup \bigcup_{j=1}^{k} E_j^c \right) \to 1, \quad \text{as } n \to \infty.
$$

(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 $\tilde{\mathcal{L}}(\theta)$ around the global empirical loss minimizer $\hat{\theta}$. 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 $\theta \in U(\rho)$,

$$\left| \mathcal{L}_N(\theta) - \mathcal{L}_N(\hat{\theta}) - \frac{1}{2} \langle \theta - \hat{\theta}, I(\theta^*) (\theta - \hat{\theta}) \rangle \right| \leq \left( M \| \theta - \theta^* \|_2 + \frac{1}{2k} \sum_{j=1}^k \| \nabla^2 \mathcal{L}_j(\theta^*) - I(\theta^*) \|_2 \right) \| \theta - \hat{\theta} \|_2^2 + M \| \theta - \hat{\theta} \|_2^3,$$

and

$$\left| \tilde{\mathcal{L}}(\theta) - \tilde{\mathcal{L}}(\hat{\theta}) - \frac{1}{2} \langle \theta - \hat{\theta}, I(\theta^*) (\theta - \hat{\theta}) \rangle \right| \leq A_n \| \theta - \hat{\theta} \|_2 + B_n \| \theta - \hat{\theta} \|_2^2 + M \| \theta - \hat{\theta} \|_2^3,$$

where $A_n := \mu_+ \| \theta - \theta^* \|_2 + 2M \| \theta - \theta^* \|_2^3 + 2M \| \theta - \theta^* \|_2 \| \theta - \theta^* \|_2$ and $B_n := M \| \theta - \theta^* \|_2 + \frac{1}{2k} \sum_{j=1}^k \| \nabla \mathcal{L}_j(\theta^*) - I(\theta^*) \|_2$.

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

**Lemma A.6.** Under the joint event $A_n \cap B_1 \cap \bigcap_{j=1}^k \mathcal{E}_j$, we have

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

Our last lemma shows that if the results in the previous two lemma holds, then we a Bernstein-von mises result for the approximated posterior $\bar{\pi}_N$. 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 $B_2$, we have

$$\left\| \bar{\pi}_N(\theta) - \mathcal{N}_d(\hat{\theta}, I(\theta^*)^{-1} (\theta)) \right\|_1 \leq CR,$$

(A.5)

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

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

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 $\tilde{\pi}_N(\theta)$ with $n = N$ and $k = 1$, we obtain that

$$\left\| \pi_N(\theta) - N_d(\tilde{\theta}, I(\theta^*)^{-1}) (\theta) \right\|_1 \leq C R. \quad (A.6)$$

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. [2013b], we obtain that under the event $\cap_{j=1}^k E_j$,

$$\|\tilde{\theta} - \theta^*\|_2 \leq \frac{2\|\nabla L_N(\theta^*)\|_2}{(1 - \rho)\mu_-}, \quad (A.7)$$

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

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

$$E[\|\nabla L_j(\theta^*)\|_2^{2\nu}] \leq \frac{c G^{2\nu}}{n^{\nu}}$$

and

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

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

$$\mathbb{P}\left(\bigcup_{j=0}^k E_j \right) \leq (c_1 + c_2 (\log 2d)^1 L^{16} + c_3 G^{16}) \frac{k}{n^8}. \quad (A.8)$$

B.2 Proof of Lemma A.2

We will apply Lemma 6 in Zhang et al. [2013b] with $\theta^* = \tilde{\theta}$ and $F_1 = \tilde{L}$ in the notation therein. Since the Hessian of $\tilde{L}$ is the same as that of $L_1$, in order to apply their result, we only need to verify that under event $E_0 \cap E_1$, it holds that

$$\|\nabla^2 L_1(\tilde{\theta}) - I(\theta^*)\|_2 \leq \frac{\rho\mu}{2}, \quad \text{and} \quad \|\nabla L_1(\tilde{\theta})\|_2 \leq \frac{(1 - \rho)\mu - \delta_\rho}{2}.$$
The first inequality is true since under event $\mathcal{E}_0 \cap \mathcal{E}_1$ and Assumption D, we have
\[
\|\nabla^2 L_1(\hat{\theta}) - I(\theta^*)\|_2 \leq 2M \|\hat{\theta} - \theta^*\|_2 + \|\nabla^2 L_1(\theta^*) - I(\theta^*)\|_2 \leq \frac{\rho \mu - }{4} + \frac{\rho \mu - }{4} = \frac{\rho \mu - }{2}.
\]
To prove the second inequality, we apply the integral form of Taylor’s expansion to obtain that
\[
\nabla L_1(\hat{\theta}) - \nabla L_1(\theta^*) = H_1(\hat{\theta} - \theta^*),
\]
where matrix $H_1 = \int_0^1 \nabla^2 L_1(\theta^* + t(\hat{\theta} - \theta^*)) dt$ satisfies
\[
\|H_1 - I(\theta^*)\|_2 \leq 2M \|\hat{\theta} - \theta^*\|_2
\]
under event $\mathcal{E}_1$. Therefore, triangle’s inequality yields that under event $\mathcal{E}_0 \cap \mathcal{E}_1$,
\[
\|\nabla L_1(\hat{\theta})\|_2 \leq \|\nabla L_1(\theta^*)\|_2 + H_1 - I(\theta^*)\|_2 \|\hat{\theta} - \theta^*\|_2 + \|I(\theta^*)\|_2 \|\hat{\theta} - \theta^*\|_2
\]
\[
\leq \frac{(1 - \rho)\mu - \delta_\rho}{4} + 2M \|\hat{\theta} - \theta^*\|_2 + \mu_+ \|\hat{\theta} - \theta^*\|_2
\]
\[
\leq \frac{(1 - \rho)\mu - \delta_\rho}{2}.
\]
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 $\nabla^2 L_N(\hat{\theta})$ as
\[
\lambda_{\min}[\nabla^2 L_N(\hat{\theta})] \geq \lambda_{\min}[I(\theta^*)] - \|\nabla^2 L_N(\theta^*) - I(\theta^*)\|_2 - \|\nabla^2 L_N(\theta^*) - \nabla^2 L_N(\theta^*)\|_2
\]
\[
\geq \mu_- - \frac{\rho \mu - }{2} - 2M \|\hat{\theta} - \theta^*\|_2 \geq \frac{1}{2} (1 - \rho)\mu_-.
\]
This proves the first claimed inequality.

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

Under event $\bigcap_{j=0}^k \mathcal{E}_j'$, the third inequality can be proved as
\[
U_N \leq \max_{\theta \in (-\Delta, \hat{\theta} + \Delta)} \|\nabla^2 L_N(\theta) - \nabla^2 L_N(\theta^*)\|_2 + \|\nabla^2 L_N(\theta^*) - I(\theta^*)\|_2 + \|I(\theta^*)\|_2
\]
\[
\leq 2M \Delta + \frac{\rho \mu - }{4} + \mu_+.
\]
To bound the term \( \| \nabla^2 L_N(\theta)^{-1} - \nabla^2 L_1(\theta)^{-1} \|_2 \), we make use of the following inequality: for any matrix \( A \in \mathbb{R}^{d \times d} 
abla \)

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

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

\[
\|\nabla^2 L_N(\theta) - I(\theta^*)\|_2 \leq \|\nabla^2 L_N(\theta) - \nabla^2 L_N(\theta^*)\|_2 + \|\nabla^2 L_N(\theta^*) - I(\theta^*)\|_2
\]

\[
\leq 2M \|\theta - \theta^*\|_2 + \|\nabla^2 L_N(\theta^*) - I(\theta^*)\|_2.
\]

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

\[
\|\nabla^2 L_N(\theta)^{-1}\|_2 \leq \|\nabla^2 L_N(\theta)^{-1} - I(\theta^*)^{-1}\|_2 + \mu_\perp^{-1}
\]

\[
\leq 2M \mu_\perp^{-1} \|\theta - \theta^*\|_2 + \mu_\perp^{-2} \|\nabla^2 L_N(\theta^*) - I(\theta^*)\|_2 + \mu_\perp^{-1}
\]

\[
\leq 2M \mu_\perp^{-2} \rho + \mu_\perp^{-1} + \mu_\perp^{-1} \rho/4,
\]

where in the last step we used the assumption that \( \|\theta - \theta^*\|_2 \leq \rho \) and the definition of events \( \mathcal{E}'_j \)’s.

Now choosing \( A = \nabla^2 L_N(\theta) \) and \( \Delta A = \nabla^2 L_1(\theta) - \nabla^2 L_N(\theta) \) in inequality (A.9), we obtain

\[
\|\nabla^2 L_N(\theta)^{-1} - \nabla^2 L_1(\theta)^{-1}\|_2
\]

\[
\leq \|\nabla^2 L_N(\theta)^{-1}\|_2 \|\nabla^2 L_N(\theta) - \nabla^2 L_1(\theta)\|_2
\]

\[
\leq \|\nabla^2 L_N(\theta)^{-1}\|_2 \left( \|\nabla^2 L_N(\theta)^{-1}\|_2 + \|\nabla^2 L_N(\theta) - \nabla^2 L_1(\theta^*)\|_2 + \|\nabla^2 L_1(\theta) - \nabla^2 L_N(\theta^*)\|_2 \right)
\]

\[
\leq \|\nabla^2 L_N(\theta^*) - \nabla^2 L_1(\theta^*)\|_2 + 4M \|\theta - \theta^*\|_2.
\]

Putting pieces together, we can prove the last claimed inequality.

B.5 Proof of Lemma A.5

To prove the first expansion for \( L_N(\theta) \), it suffices to prove the following inequality by using the fact that \( \nabla L_N(\theta) = 0 \):

\[
\begin{align*}
\left| L_j(\theta) - L_j(\hat{\theta}) - \langle \nabla L_j(\hat{\theta}), \theta - \hat{\theta} \rangle - \frac{1}{2} \langle \theta - \hat{\theta}, I(\theta^*) (\theta - \hat{\theta}) \rangle \right| \\
\leq \left( M \|\theta - \theta^*\|_2 + \frac{1}{2} \|\nabla^2 L_j(\theta^*) - I(\theta^*)\|_2 \right) \|\theta - \hat{\theta}\|_2 + M \|\theta - \hat{\theta}\|^2
\end{align*}
\]

\[
\text{ (A.10)}
\]
for $j = 1, \ldots, k$. In fact, by Taylor’s theorem, we have
\[
\mathcal{L}_j(\theta) - \mathcal{L}_j(\hat{\theta}) = \langle \nabla \mathcal{L}_j(\hat{\theta}), \theta - \hat{\theta} \rangle + \frac{1}{2} \langle \theta - \hat{\theta}, I(\theta^*) (\theta - \hat{\theta}) \rangle + \frac{1}{2} \langle \theta - \hat{\theta}, (\widetilde{H}_j - I(\theta^*)) (\theta - \hat{\theta}) \rangle,
\]
where $\widetilde{H}_j = \nabla^2 \mathcal{L}_j(\hat{\theta} + t_j(\theta - \hat{\theta}))$ for some $t_j \in [0,1]$. Under event $E_j$, we can bound the last remainder term for $\theta \in U(\rho)$ by
\[
\frac{1}{2} \| \theta - \hat{\theta} \|^2 \left( \| \widetilde{H}_j - \nabla^2 \mathcal{L}_j(\theta^*) \|_2 + \| \nabla^2 \mathcal{L}_j(\theta^*) - I(\theta^*) \|_2 \right)
\leq M \| \theta - \hat{\theta} \|^2 + \left( M \| \theta - \theta^* \|_2 + \frac{1}{2} \| \nabla^2 \mathcal{L}_j(\theta^*) - I(\theta^*) \|_2 \right) \| \theta - \hat{\theta} \|^2,
\]
which yields the expansion (A.10).

To prove the expansion for $\tilde{\mathcal{L}}(\theta)$, we note that simple calculation yields
\[
\tilde{\mathcal{L}}(\theta) - \tilde{\mathcal{L}}(\hat{\theta}) = \mathcal{L}_1(\theta) - \mathcal{L}_1(\hat{\theta}) + \langle \nabla \mathcal{L}_N(\hat{\theta}), \theta - \hat{\theta} \rangle.
\]
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 $\bigcap_{j=1}^k E_j$,
\[
\left| \langle \nabla \mathcal{L}_N(\hat{\theta}), \theta - \hat{\theta} \rangle \right| + \left| \langle \nabla \mathcal{L}_1(\hat{\theta}), \theta - \hat{\theta} \rangle \right| \leq A_n \| \theta - \hat{\theta} \|_2,
\]
because $\nabla \mathcal{L}_N(\hat{\theta}) = 0$. 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
\[
\| \tilde{H}_N \|_2 \| \theta - \hat{\theta} \|_2 + \| \tilde{H}_j \|_2 \| \theta - \hat{\theta} \|_2 \| \theta - \hat{\theta} \|_2
\leq 2(\mu + 2M \| \theta - \hat{\theta} \|_2 + 2M \| \theta - \theta^* \|_2) \| \theta - \hat{\theta} \|_2 \| \theta - \hat{\theta} \|_2 = A_n \| \theta - \hat{\theta} \|_2,
\]
where $\tilde{H}_N = \int_0^1 \nabla^2 \mathcal{L}_N(\hat{\theta} + t(\theta - \hat{\theta})) dt$, $\tilde{H}_j = \int_0^1 \nabla^2 \mathcal{L}_j(\hat{\theta} + t(\theta - \hat{\theta})) dt$ and in the last step we used the fact that under the joint event $\bigcap_{j=1}^k E_j$, $\| \tilde{H}_j - I(\theta^*) \|_2 \leq 2M (\| \theta - \hat{\theta} \|_2 + \| \theta - \theta^* \|_2)$ for each $j$ and $\| \tilde{H}_N - I(\theta^*) \|_2 \leq k^{-1} \sum_{j=1}^k \| \tilde{H}_j - I(\theta^*) \|_2$. 

44
B.6 Proof of Lemma A.6

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

\[
\inf_{\|\theta - \theta^*\|_2 \geq \delta} \left( \hat{L}(\theta) - \hat{L}(\theta^*) \right) \\
\geq \inf_{\|\theta - \theta^*\|_2 \geq \delta} \left( L_1(\theta) - L_1(\theta^*) \right) - \sup_{\theta \in \Theta} \langle \theta - \theta^*, \nabla L_1(\theta) - \nabla L_1(\theta^*) \rangle \\
\geq 3\epsilon - R \left( \|\nabla L_1(\theta) - \nabla L_1(\theta^*)\|_2 + \|\nabla L_1(\theta^*)\|_2 + \|\nabla L_1(\theta)\|_2 + \|\nabla L_1(\theta) - \nabla L_1(\theta^*)\|_2 \right).
\]

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

\[
\|\nabla L_1(\theta) - \nabla L_1(\theta^*)\|_2 \leq \max_{\theta \in U(\rho)} \|\nabla^2 L_1(\theta)\|_2 \|\theta - \theta^*\|_2 \\
\leq (M \|\theta - \theta^*\|_2 + \frac{\rho \mu_+}{2} + \mu_+) \|\theta - \theta^*\|_2 \leq \frac{\epsilon}{4R},
\]

\[
\|\nabla L_1(\theta^*)\|_2 \leq \min \left\{ \frac{(1 - \rho) \mu_- \delta^\rho}{2}, \frac{(1 - \rho) \mu^2 \delta^\rho}{8M}, \frac{(1 - \rho)^{3/2} \mu^3 \delta^\rho}{8M} \right\} \leq \frac{\epsilon}{4R},
\]

\[
\|\nabla L_1(\theta^*)\|_2 \leq \frac{1}{k} \sum_{j=1}^k \|\nabla L_j(\theta^*)\|_2 \leq \frac{\epsilon}{4R}, \quad \text{and}
\]

\[
\|\nabla L_1(\theta) - \nabla L_1(\theta^*)\|_2 \leq \frac{1}{k} \sum_{j=1}^k \|\nabla L_j(\theta) - \nabla L_j(\theta^*)\|_2 \leq \frac{\epsilon}{4R}.
\]

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

\[
\inf_{\|\theta - \theta^*\|_2 \geq \delta} \frac{1}{n} \left( \hat{L}(\theta) - \hat{L}(\theta^*) \right) \geq 3\epsilon - R \cdot 4 \cdot \frac{\epsilon}{4R} = 2\epsilon,
\]

which completes the proof.

B.7 Proof of Lemma A.1

The approximated posterior can be expressed by

\[
\tilde{\pi}_N(\theta) = \frac{\pi(\theta) e^{-N(\hat{L}(\theta) - \hat{L}(\theta))}}{\int_{\Theta} \pi(\theta) e^{-N(\hat{L}(\theta) - \hat{L}(\theta))} d\theta}.
\]

We claim that it suffices to prove

\[
\left| \pi(\theta) e^{-N(\hat{L}(\theta) - \hat{L}(\theta))} - \pi(\hat{\theta}) e^{-N(\theta - \hat{\theta})} \right| \\
\leq CR \pi(\hat{\theta}) e^{-N(\theta - \hat{\theta})} + \pi(\theta) e^{-N\epsilon}.
\]  

(A.11)
In fact, if (A.11) holds, then by integrating $\theta$ over $\mathbb{R}^d$, we obtain
\[
\left| N^{d/2} \int_{\mathbb{R}^d} \pi(\theta) e^{-N(\tilde{\mathcal{L}}(\theta) - \tilde{\mathcal{L}}(\tilde{\theta}))} \, d\theta - \pi(\tilde{\theta}) \frac{(2\pi)^{d/2}}{\sqrt{\det I(\tilde{\theta})}} \right| \leq C R + C N^{(d-1)/2} e^{-N\epsilon} \leq C' R.
\]
Then, by combining all three preceding displays, we obtain
\[
\int_{\mathbb{R}^d} \left| \tilde{\pi}_N(\theta) - \frac{N^{d/2}}{(2\pi)^{d/2}} \sqrt{\det I(\tilde{\theta})} e^{-\frac{N}{2} (\theta - \tilde{\theta}, I(\tilde{\theta})^{-1}(\theta - \tilde{\theta}))} \right| \, d\theta \leq C'' R,
\]
which is the claimed Bernstein-von Mises result for $\tilde{\pi}_N$. The remainder of the proof focuses on proving (A.11).

Let $s = \sqrt{N}(\theta - \tilde{\theta})$ be the localized parameter. Then (A.11) is equivalent to
\[
\left| \pi(\tilde{\theta} + s/\sqrt{N}) e^{-N(\tilde{\mathcal{L}}(\tilde{\theta} + s/\sqrt{N}) - \tilde{\mathcal{L}}(\tilde{\theta}))} - \pi(\tilde{\theta}) e^{-\frac{1}{2} \langle s, I(\tilde{\theta})^{-1} s \rangle} \right| \leq C R \pi(\tilde{\theta}) e^{-\frac{1}{2} \langle s, I(\tilde{\theta})^{-1} s \rangle} + \pi(\tilde{\theta} + s/\sqrt{N}) e^{-N\epsilon}.
\]

Corollary A.5 guarantees that for all $\|s\| \leq \delta \sqrt{N}$,
\[
\left| N(\tilde{\mathcal{L}}(\tilde{\theta} + s/\sqrt{N}) - \tilde{\mathcal{L}}(\tilde{\theta})) - \frac{1}{2} \langle s, I(\tilde{\theta}) s \rangle \right| 
\leq A_n \sqrt{N} \|s\|_2 + B_n \|s\|_2^2 + \frac{M}{\sqrt{N}} \|s\|_2^3.
\]

We prove (A.12) by considering $s$ in the following three subsets separately:
\[
S_1 : = \{ s : \|s\|_2 \leq c \log N \}
\]
\[
S_2 : = \{ s : c \log N \leq \|s\|_2 \leq \delta \sqrt{N} \}
\]
\[
S_3 : = \{ s : \|s\|_2 > \delta \sqrt{N} \}.
\]

We begin with $s \in S_1$. Using (A.13), we obtain that
\[
\left| \pi(\tilde{\theta} + s/\sqrt{N}) e^{-N(\tilde{\mathcal{L}}(\tilde{\theta} + s/\sqrt{N}) - \tilde{\mathcal{L}}(\tilde{\theta}))} - \pi(\tilde{\theta}) e^{-\frac{1}{2} \langle s, I(\tilde{\theta})^{-1} s \rangle} \right| 
\leq \left| \pi(\tilde{\theta} + s/\sqrt{N}) e^{-N(\tilde{\mathcal{L}}(\tilde{\theta} + s/\sqrt{N}) - \tilde{\mathcal{L}}(\tilde{\theta}))} - \pi(\tilde{\theta} + s/\sqrt{N}) e^{-\frac{1}{2} \langle s, I(\tilde{\theta})^{-1} s \rangle} \right| 
\leq C \pi(\tilde{\theta}) e^{-\frac{1}{2} \langle s, I(\tilde{\theta})^{-1} s \rangle} \left( A_n \sqrt{N} \log N + B_n (\log N)^2 + M N^{-1/2} (\log N)^3 \right)
\leq C R \pi(\tilde{\theta}) e^{-\frac{1}{2} \langle s, I(\tilde{\theta})^{-1} s \rangle}.
\]
Next consider $s \in S_2$. Then $\|s\|_2 \leq \|s\|_2^2$ for sufficiently small constant $c$. Under the event $B_2$, we have $A_n \sqrt{N} \|s\|_2^2 + B_n \|s\|_2^2 + M N^{-1/2} \|s\|_2 \leq \langle s, I(\theta^*) s \rangle / 4$. Then using (A.13), we obtain
\[
N(\hat{\mathcal{L}}(\hat{s} + \sqrt{N}) - \mathcal{L}(\theta^*)) - \frac{1}{2} \langle s, I(\theta^*) s \rangle \leq \frac{1}{4} \langle s, I(\theta^*) s \rangle.
\]
Therefore, we have
\[
\left| \pi(\hat{s} + s/\sqrt{N}) e^{-N(\hat{\mathcal{L}}(\hat{s} + \sqrt{N}) - \mathcal{L}(\theta^*))} - \pi(\hat{\theta}) e^{-\frac{1}{2} \langle s, I(\theta^*) s \rangle} \right| \leq \pi(\hat{\theta} + s/\sqrt{N}) e^{-N\epsilon} e^{-N(\hat{\mathcal{L}}(\hat{s} + \sqrt{N}) - \mathcal{L}(\theta^*))}
\]
\[
\leq C (R_1 + 1) \pi(\hat{\theta}) e^{-\frac{1}{4} \langle s, I(\theta^*) s \rangle} \leq C R \pi(\hat{\theta}) e^{-\frac{1}{4} \langle s, I(\theta^*) s \rangle},
\]
where in step (i) we used Lemma A.6 and the optimality of $\hat{\theta}$ that implies $\mathcal{L}(\theta^*) - \mathcal{L}(\hat{\theta}) \geq 0$.

Therefore, we have
\[
\left| \pi(\hat{s} + s/\sqrt{N}) e^{-N(\hat{\mathcal{L}}(\hat{s} + \sqrt{N}) - \mathcal{L}(\theta^*))} - \pi(\hat{\theta}) e^{-\frac{1}{2} \langle s, I(\theta^*) s \rangle} \right| \leq \pi(\hat{\theta} + s/\sqrt{N}) e^{-N\epsilon} e^{-N(\hat{\mathcal{L}}(\hat{s} + \sqrt{N}) - \mathcal{L}(\theta^*))} \leq e^{-N\epsilon},
\]
where in step (i) we used Lemma A.6 and the optimality of $\hat{\theta}$ that implies $\mathcal{L}(\theta^*) - \mathcal{L}(\hat{\theta}) \geq 0$.

Then we obtain
\[
e^{-N(\hat{\mathcal{L}}(\hat{s} + \sqrt{N}) - \mathcal{L}(\hat{\theta}))} \leq e^{-N(\hat{\mathcal{L}}(\hat{s} + \sqrt{N}) - \mathcal{L}(\hat{\theta}))} \leq e^{-N(\hat{\mathcal{L}}(\hat{s} + \sqrt{N}) - \mathcal{L}(\theta^*))} e^{-N(\mathcal{L}(\theta^*) - \mathcal{L}(\hat{\theta}))} \leq C R \pi(\hat{\theta}) e^{-\frac{1}{4} \langle s, I(\theta^*) s \rangle} + \pi(\hat{\theta} + s/\sqrt{N}) e^{-N\epsilon}.
\]

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

47
REFERENCES

Schreck Amandine, Fort Gersende, Le Corff Sylvain, and Moulines Eric. A shrinkage-thresholding metropolis adjusted langevin algorithm for bayesian variable selection. *arXiv:1312.5658*, 2015.

Holmes Chris Bissiri Pier and Walker Stephen. A general framework for updating belief distributions. *arXiv preprint 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.

Sébastien Bubeck. Theory of convex optimization for machine learning. *arXiv preprint arXiv:1405.4980*, 2014.

Victor Chernozhukov and Han Hong. An \{MCMC\} approach to classical estimation. *Journal of Econometrics*, 115(2):293 – 346, 2003.

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

John C. Duchi, Michael Jordan, Martin Wainwright, and Yuchen Zhang. Optimality guarantees for distributed statistical estimation. *arXiv:1405.0782*, 2015.

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

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

Jason D Lee, Yuekai Sun, Qiang Liu, and Jonathan E. Taylor. Communication- efficient sparse regression: a one-shot approach. *arXiv preprint arXiv:1503.04337*, 2015.
Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Willie Neiswanger, Chong Wang, and Eric Xing. Asymptotically exact, embarrassingly parallel mcmc. *arXiv:1311.4780*, 2015.

Marcelo Pereyra. Proximal markov chain monte carlo algorithms. *Statistics and Computing*, pages 1–16, 2015.

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

Bhatt Rajen and Dhall Abhinav. Skin segmentation dataset. *UCI Machine Learning Repository*, 2012.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *J. Mach. Learn. Res.*, 11:2241–2259, 2010.

Ohad Shamir, Nati Srebro, and Tong 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.

Sara van de Geer, Peter Bühlmann, Ya’acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42:1166–1202, 2014.

Jialei Wang, Mladen Kolar, Nati Srebro, and Tong Zhang. Efficient distributed learning with sparsity. *Personal Communication*, 2016.

Xiangyu Wang and David B. Dunson. Parallelizing mcmc via weierstrass sampler. *arXiv:1312.4605*, 2015.

Yuchen Zhang and Xiao Lin. Communication-efficient distributed optimization of self-concordant empirical loss. *arXiv:1501.00263*, 2015.
Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363, 2013a.

Yuchen Zhang, John C Duchi, and Martin J Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363, 2013b.