Private and Efficient Federated Statistical Learning

Jaemu Heo[1] **Xiwen Feng**[2] **Jeonghun Kang**[1] **Taehwan Kim**[1,*] **Changgee Chang**[2,*]

[1]Ulsan National Institute of Science Technology

[2]Indiana University School of Medicine

[*]Correspondence to: changgee@iu.edu taehwankim@unist.ac.kr

Abstract

Federated Learning (FL) enables collaborative model training across multiple data sources while preserving data privacy, and differential privacy (DP) provides a probabilistic framework to safeguard sensitive information when sharing output derived from data. While numerous DP-FL methods exist, achieving both DP and efficient utility in federated statistical learning remains a significant challenge. In this work, we propose a novel federated statistical learning framework that ensures efficient, robust, and privacy-preserving estimation. We introduce a new noising mechanism that encodes uncertainty along with the maximum likelihood estimate (MLE) by leveraging multiple noisy copies of the MLE. To calibrate noise effectively, we extend the smooth sensitivity to account for data-dependent correlations, ensuring strong DP guarantees while maintaining utility. Additionally, we develop INFEMBLER, an information-assembling algorithm that efficiently de-noises multiple noisy MLE copies using a hierarchical Bayesian model and via an expectation-maximization (EM) algorithm. INFEMBLER significantly enhances estimation efficiency over existing methods and is inherently robust, providing estimates at least as reliable as those derived from local data alone, thereby preserving the benefits of FL. We establish its asymptotic properties and validate its effectiveness through experiments on both simulated and real datasets, demonstrating its superior statistical efficiency and robustness.

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

Federated learning (FL) revolutionizes machine learning by enabling multiple parties to collaboratively build models while keeping their training data local and private. Unlike traditional centralized approaches, FL operates in a decentralized manner, with multiple sites communicating to exchange limited yet valuable information. This methodology enhances privacy and security while allowing learning from distributed data across diverse locations and devices, making it applicable to domains such as natural language processing, computer vision, and healthcare.

The federated learning literature primarily focuses on methods designed for large models, with an emphasis on improving prediction performance rather than refining the estimation of model parameters. In statistical learning, however, accurate model parameter estimation and proper inference are essential, particularly in scientific research. In biomedical studies, model parameters often represent effect sizes (e.g., the impact of age) or differences between groups (e.g., treatment vs. control), making their precise estimation essential for drawing meaningful scientific conclusions. Due to the limited number of patients in individual studies, combining and analyzing datasets from multiple cohorts or hospitals is necessary, making federated statistical learning that preserves privacy an indispensable tool. This work is dedicated to the private and efficient federated estimation of parameters in general statistical models.

Despite the inherent privacy protection offered by decentralization, ensuring data privacy in a formal manner remains a complex challenge. De-identification may mitigate risks to some extent, but it does not guarantee privacy protection. Differential privacy (DP) is a mathematical framework designed to ensure data privacy, while enabling the sharing of useful information. This is typically achieved by introducing random noise to the released output, making it difficult to infer the presence of specific individu-

als within the dataset (Dwork and Roth, 2014). The most widely used noising techniques include the Gaussian mechanism and the Laplace mechanism (Dwork et al., 2006). These methods are simple and easy to implement, as they involve adding noise proportional to the sensitivity of the output, making them applicable to statistical learning as well. However, when applied directly to data, these mechanisms distort the original data distribution, typically leading to biased estimates. To address this issue, Bi and Shen (2022) proposed a distribution-invariant privatization (DIP) mechanism, but it requires to accurately estimate multi-dimensional data distributions which itself can be quite challenging.

Instead of sharing the entire dataset, one approach is to release a noisy version of the local parameter estimates, such as the maximum likelihood estimate (MLE), with the average of these shared MLEs serving as the federated estimate. However, the averaging mechanism is not efficient at de-noising the DP noise added to the MLEs, and often the results from FL perform worse than those obtained using the locally available dataset only. Moreover, even in the absence of noise, averaging the MLEs is not efficient, in the sense that its asymptotic variance is higher than that of the global MLE, especially when there is heterogeneity across datasets.

Some private stochastic gradient descent (SGD) algorithms (Abadi et al., 2016; Agarwal et al., 2018) aim to obtain the lossless global optimizer by iteratively exchanging noisy gradients. While these methods have been successfully applied to large prediction models, they typically introduce additional complexities, such as allocating the privacy budget across multiple iterations and determining the optimal clipping threshold and learning rate—both of which are crucial for the convergence of the algorithm and the accuracy of the model estimate. In particular, the performance of DP-SGD is notoriously sensitive to the choice of the gradient clipping threshold. Properly tuning this hyperparameter is exceptionally difficult (Chen et al., 2020), and finding an optimal, adaptive clipping strategy without compromising the privacy budget or model utility remains a major practical hurdle (Andrew et al., 2021). As a result, federated learning based on these methods does not always lead to improved results.

Other privacy-preserving frameworks specifically designed for statistical learning have been proposed in the literature (Chaudhuri et al., 2011; Wasserman and Zhou, 2010). However, none of the existing methods offer an efficient, robust, and tune-free federated estimator that achieves the optimal statistical efficiency obtainable from the original datasets. In this work, we introduce a novel framework that enables private, efficient, and robust federated statistical learning.

Our first contribution is the introduction of a novel noising mechanism that enables efficient federated learning for a wide range of statistical models. This mechanism is inspired by the inverse-variance weighted average estimator, which minimizes variance by assigning weights proportional to the inverse of variances when combining estimators with the same mean but different variances. This insight suggests that constructing an optimal estimator requires not only the location information of local MLEs but also their Fisher information, which quantifies uncertainty. To achieve this, we propose releasing multiple noisy copies of the MLE, where the added noise has a mean of zero and a variance proportional to the inverse of the Fisher information of the MLE. While this mechanism requires adding larger noise as the number of copies increases, the variance of their average remains nearly unchanged, allowing it to effectively encode Fisher information while preserving location information. Of course, there is an inherent trade-off, as the noise correlation now becomes data-dependent. Nevertheless, this approach offers greater efficiency and flexibility compared to directly sharing a noisy version of the Fisher information, especially when dealing with high-dimensional parameters.

We follow the smooth sensitivity approach, introduced by Nissim et al. (2007), to determine the noise level that ensures DP of the noisy MLE copies. A major challenge in applying DP to MLE is that the sensitivity of MLE tends to be large when the sample size is small, leading to large global sensitivity. Consequently, noise calibration based on global sensitivity can result in overly conservative noise levels and reduced utility. In contrast, local sensitivity, which measures sensitivity within a specific dataset's immediate neighborhood only, offers a more precise noise calibration. However, using local sensitivity directly compromises privacy, as it depends explicitly on the dataset. To address these limitations, smooth sensitivity refines local sensitivity by smoothing it across neighboring datasets, maintaining a similar level to local sensitivity while guaranteeing DP. Unfortunately, the original smooth sensitivity framework by Nissim et al. (2007) assumes independent noise or noise with data-independent correlation, making it inapplicable to our setting, where we introduce noise with data-dependent correlations. **Our second contribution** is an extension of smooth sensitivity that enables its application to our noising mechanism.

Our third contribution is the development of a novel algorithm, the Information Assembler (INFEM-

BLER), designed to effectively de-noise multiple noisy MLE copies and efficiently aggregate the information they carry. To achieve this, we construct a hierarchical Bayesian model, treating the MLEs and Fisher information matrices at all remote sites as latent variables while considering the noisy MLE copies as observed data. This approach leverages the large-sample theory on MLEs and the properties of our noising mechanism. Since the MLEs and Fisher information matrices at remote sites are unobserved, we develop a clever EM algorithm that efficiently computes the maximum a posteriori (MAP) estimator while marginalizing out the unobserved quantities. The resulting federated estimator, also called INFEMBLER, shows significantly improved accuracy over simple averaging and other approaches. What is striking about INFEMBLER is that it is tuning-free, yet robust—its results always perform as reliably as those obtained using the locally available dataset only. We examine the asymptotic properties of INFEMBLER, establishing its statistical efficiency, and validate its effectiveness through comprehensive experiments on both simulated and real datasets, showcasing its strong finite-sample performance, robustness, and broad applicability.

1.1 Related Works

Non-differentially private FL approaches which seek the loseless global optimizer via iterative communications have been proposed for linear regression Chen et al. (2006), logistic regression Wu et al. (2012), and survival data Lu et al. (2015). Communication efficient approaches, which entail minimal information loss, include methods like average mixture (AVGM, Zhang et al. (2013)), communication efficient surrogate loss (CSL, Jordan et al. (2019); Duan et al. (2022)), and CEDAR Chang et al. (2022). AVGM takes the average of the estimates from remote sites. CSL adjusts the central loss function using gradients of remote loss (log-likelihood) functions. CEDAR employs posterior samples to transfer uncertainty information but shares noise-free MLEs, thus lacking differential privacy. Additionally, CEDAR is limited to linear regression models.

For DP-based approaches, in addition to the previously mentioned mechanisms such as the Gaussian mechanism, the Laplace mechanism (Dwork et al., 2006), DP-SGD (Abadi et al., 2016), and the distribution-invariant privatization (DIP) mechanism (Bi and Shen, 2022), other methods, such as those proposed by Foulds et al. (2016); Heikkilä et al. (2017), involve sharing perturbed sufficient statistics instead of the entired perturbed dataset. However, perturbing sufficient statistics still introduces bias into estimators derived from them. Moreover, depending on the model, the sufficient statistic can encompass the entire dataset (e.g., order statistics), significantly limiting the applicability of these approaches.

Inspired by works such as Avella-Medina (2021); Cai et al. (2021), we consider the AVG approach, which ensures differential privacy while maintaining efficiency. To illustrate, suppose there are M data sites, with site 1 designated as the central site responsible for conducting the analysis, while the remaining sites $(2, \ldots, M)$ act as remote sites that share information with the central site. Let n_m and $\boldsymbol{\theta}_m$ represent the sample size and MLE obtained from the dataset at site m, respectively. The federated average estimator is defined as:

$$\widehat{\boldsymbol{\theta}}_{AVG} = \frac{1}{N} \sum_{m=1}^{M} n_m \widetilde{\boldsymbol{\theta}}_m, \tag{1}$$

where the noisy MLEs are given by $\widetilde{\boldsymbol{\theta}}_m = \boldsymbol{\theta}_m + \mathbf{e}_m$ with $\mathbf{e}_1 = \mathbf{0}$ and $\mathbf{e}_m \sim \mathcal{N}(\mathbf{0}, \tau_m^2 I)$ for $m \geq 2$. Here, τ_m^2 is determined by the desired privacy protection level.

2 METHOD

2.1 Multiple Noisy MLEs with Dependent Error

Consider a scenario with M sites, where the m-th site holds the dataset D_m. We designate site 1 as the central site responsible for analysis, while the remaining sites $2, \ldots, M$ serve as remote sites, transmitting information to the central site. We examine a model parameterized by $\boldsymbol{\theta} \in \Theta$, with the full likelihood given by:

$$L(\boldsymbol{\theta}|D_{1:M}) = \prod_{m=1}^{M} \pi(D_m|\boldsymbol{\theta}),$$

where $\pi(D|\boldsymbol{\theta})$ denotes the density function for data D given parameter $\boldsymbol{\theta}$. This global likelihood function is not directly useful, as the central site does not have access to all datasets. Alternatively, each remote site generates multiple noisy copies of the MLE as follows to share with the central site. Let $\ell_m(\boldsymbol{\theta}) = \log \pi(D_m|\boldsymbol{\theta})$ denote the log-likelihood function at site m, with $\boldsymbol{\theta}_m$ and F_m representing the MLE and the associated empirical Fisher information, respectively, defined as:

$$\boldsymbol{\theta}_m = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \ell_m(\boldsymbol{\theta}), \quad F_m = -\ddot{\ell}_m(\boldsymbol{\theta}_m),$$

where $\ddot{\ell}(\cdot)$ denotes the second derivative of $\ell(\cdot)$. Each remote site m then generates K_m independent noisy copies of the MLE from a Normal distribution with mean $\boldsymbol{\theta}_m$ and covariance $\psi_m F_m^{-1}$, given by:

$$\widetilde{\boldsymbol{\theta}}_{mk}|\boldsymbol{\theta}_m, F_m \sim \mathcal{N}(\boldsymbol{\theta}_m, \psi_m F_m^{-1}),\ k = 1, \ldots, K_m, \tag{2}$$

and transmits the matrix of noisy copies $\widetilde{\Theta}_m = \widetilde{\boldsymbol{\theta}}_{m,1:K_m}$ to the central site. The dispersion parameter ψ_m controls the noise variance, thereby regulating the level of privacy protection. To determine ψ_m, we employ the modified smooth sensitivity approach, as introduced in Section 3.1.

2.2 Information Assembler

The central site receives the noisy MLE copies from the remote sites and integrates them with the MLE and Fisher information from the central site using the following Bayesian hierarchical model. While equation (2) provides the likelihood for the noisy copies, the quantities $\boldsymbol{\theta}_m$ and F_m remain unobserved. Therefore, we construct a hierarchical prior for $\boldsymbol{\theta}_m$ and F_m. Drawing upon standard large sample theory for the MLE, we assign the following priors:

$$\begin{aligned} \boldsymbol{\theta}_m|\boldsymbol{\theta}, F_m &\sim \mathcal{N}(\boldsymbol{\theta}, F_m^{-1}), \\ F_m|F &\sim \mathcal{W}(F, n_m), \end{aligned} \qquad m = 1, \ldots, M,$$

where $\mathcal{W}(F, n)$ denotes the Wishart distribution with scale matrix F and degrees of freedom n. Here, $\boldsymbol{\theta}$ and F are the model parameters. By specifying the MLEs and empirical Fisher information $(\boldsymbol{\theta}_m, F_m)$ as realizations of the Normal-Wishart distribution parameterized by $(\boldsymbol{\theta}, F)$, we enable efficient aggregation of the information carried by the noisy MLE copies. It is important to emphasize that we do not assume any specific distribution for $(\boldsymbol{\theta}_m, F_m)$; rather, we are only defining the prior. The asymptotic properties discussed in Section 3.2 will hold with asymptotic normality of $\boldsymbol{\theta}_m$ and consistency of F_m only. For the model parameters $(\boldsymbol{\theta}, F)$, we impose an uninformative Jeffreys prior:

$$\pi(\boldsymbol{\theta}, F) \propto |F|^{-(p+1)/2}.$$

Our method estimates $(\boldsymbol{\theta}, F)$ using the Maximum a Posteriori (MAP) estimator:

$$(\widehat{\boldsymbol{\theta}}, \widehat{F}) = \underset{\boldsymbol{\theta}, F}{\operatorname{argmax}}\, \pi(\boldsymbol{\theta}, F|\boldsymbol{\theta}_1, F_1, \widetilde{\Theta}_{2:M}),$$

where $\pi(\boldsymbol{\theta}, F|\boldsymbol{\theta}_1, F_1, \widetilde{\Theta}_{2:M})$ is the marginal posterior density for $(\boldsymbol{\theta}, F)$. The full posterior density for $(\boldsymbol{\theta}, F, \boldsymbol{\theta}_{2:M}, F_{2:M})$ is provided in the Appendix, along with the intermediate marginal posterior density for $(\boldsymbol{\theta}, F, F_{2:M})$, obtained by integrating out the latent variables $\boldsymbol{\theta}_{2:M}$. We can further marginalize out $F_{2:M}$ analytically, leading to the marginal posterior density for $(\boldsymbol{\theta}, F)$. However, directly maximizing the marginal posterior with respect to F is challenging due to its requirement of positive definiteness.

To resolve this, we employ the Expectation-Maximization (EM) algorithm Dempster et al. (1977) on the intermediate marginal posterior. This approach facilitates the maximization of the final marginal posterior while avoiding explicit analytical marginalization and ensuring the positive definiteness of F. In the E-step of the EM algorithm, we need to compute the expectation of the log-posterior density with respect to the conditional distribution of the latent variables given the model parameters. Based on the intermediate posterior density, the conditional distribution of F_m given $(\boldsymbol{\theta}, F)$ is

$$F_m|\boldsymbol{\theta}, F \sim \mathcal{W}(\overline{F}_m(\boldsymbol{\theta}, F), n_m + K_m).$$

where

$$\begin{aligned} \overline{F}_m(\boldsymbol{\theta}, F)^{-1} = F^{-1} &+ \frac{K_m}{K_m + \psi_m}(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}_m)(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}_m)^T \\ &+ \psi_m^{-1} \sum_{k=1}^{K_m} (\widetilde{\boldsymbol{\theta}}_{mk} - \overline{\boldsymbol{\theta}}_m)(\widetilde{\boldsymbol{\theta}}_{mk} - \overline{\boldsymbol{\theta}}_m)^T \end{aligned} \tag{3}$$

with $\overline{\boldsymbol{\theta}}_m = \frac{1}{K_m}\sum_{k=1}^{K_m} \widetilde{\boldsymbol{\theta}}_{mk}$. This leads to the EM algorithm presented in Algorithm 1. The computational complexity is $O(p^3 M)$ per EM iteration, which can be reduced to $O(p^3)$ with parallel computing. The space complexity is $O(p^2 M)$. We refer to Algorithm 1 as the Information Assembler, or INFEMBLER for short.

Algorithm 1: INFEMBLER

Input: $n_{1:M}$, $\boldsymbol{\theta}_1$, F_1, $K_{2:M}$, $\psi_{2:M}$, and $\widetilde{\Theta}_{2:M}$
Output: $\widehat{\boldsymbol{\theta}}$, $\widehat{F}$, and $\widehat{F}_{2:M}$

$\widehat{F}_1 \leftarrow F_1$, $\overline{\boldsymbol{\theta}}_1 \leftarrow \boldsymbol{\theta}_1$, $K_1 \leftarrow 1$, $\psi_1 \leftarrow 0$;
Compute $\overline{\boldsymbol{\theta}}_m = \frac{1}{K_m}\sum_{k=1}^{K_m} \widetilde{\boldsymbol{\theta}}_{mk}$ for $m \geq 2$;
Initialize $\widehat{\boldsymbol{\theta}} \leftarrow \frac{1}{N}\sum_{m=1}^{M} n_m \overline{\boldsymbol{\theta}}_m$;
Initialize $\widehat{F} \leftarrow \frac{1}{n_1} F_1$;
repeat
  for $m \leftarrow 2$ **to** M **do**
    $\widehat{F}_m \leftarrow (n_m + K_m)\, \overline{F}_m(\widehat{\boldsymbol{\theta}}, \widehat{F})$; `// See (3)`
  $\widehat{F} \leftarrow \frac{1}{N+p+1}\sum_{m=1}^{M} \widehat{F}_m$;
  $\widehat{\boldsymbol{\theta}} \leftarrow \left(\sum_{m=1}^{M} \frac{K_m \widehat{F}_m}{K_m + \psi_m}\right)^{-1} \left(\sum_{m=1}^{M} \frac{K_m \widehat{F}_m \overline{\boldsymbol{\theta}}_m}{K_m + \psi_m}\right)$;
until *convergence*;

3 THEORETICAL PROPERTIES

Throughout this section, unless otherwise specified, D_0 refers to the entire dataset of a specific site and D_j, for $j \geq 1$, refers to a potential dataset at the same site, including D_0. It is important to note that these datasets should not be confused with the remote datasets indexed by m in Section 2. For each D_j, for

$j \geq 0$, the quantities $\boldsymbol{\theta}_j$, F_j, n_j, K_j, and ψ_j correspond to those associated with D_j. Two neighboring datasets are denoted as $D_1 \sim D_2$, which means D_2 is built by adding a data point to D_1 or vice versa.

3.1 Differential Privacy of Noisy MLE Copies

A mechanism f is considered differentially private if the output $f(D)$, generated based on the dataset D, obscures the presence or absence of any specific individual within D. Specifically, for any two neighboring datasets D_1 and D_2, DP requires that the distributions of $f(D_1)$ and $f(D_2)$ are sufficiently similar, to prevent an observer from inferring information about the differing data point between D_1 and D_2. There are several definitions of differential privacy. In this work, we adopt the (ϵ, δ)-DP, defined as follows.

Definition 1 ((ϵ, δ)-Differential Privacy)**.** *A randomized algorithm f is said to be (ϵ, δ)-differentially private if, for all $B \subseteq \text{Range}(f)$ and for all neighboring datasets D_1 and D_2, the following holds:*

$$\Pr[f(D_1) \in B] \leq e^{\epsilon} \Pr[f(D_2) \in B] + \delta. \tag{4}$$

In our setting, the noisy MLE copies are treated as the output of a randomized mechanism, designed to protect the remote datasets from any recipient of these copies. Note that the communication is one-shot and one-way, meaning that remote sites only transmit information without receiving any input. Therefore, remote sites are not vulnerable to active adversarial attacks.

The standard Gaussian mechanism Dwork and Roth (2014) adds noise proportional to the L_2 sensitivity. However, while it assumes independent and identically sized noise for each individual entry, our mechanism introduces correlated noise. Therefore, we require a modified version of L_2 sensitivity, defined as follows.

Definition 2 (Modified Local L_2 sensitivity)**.**

$$S(D_1) = \max_{D_2 \sim D_1} \|(F_1/n_1)^{1/2}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\|_2.$$

If we choose ψ_1 based on the global sensitivity, $S = \max_{D_1} S(D_1)$, it will introduce a substantial amount of noise, resulting in reduced utility. On the other hand, selecting ψ_1 based on the local sensitivity allows for very small noise, but DP will not be guaranteed, as the noise level depends on the dataset D_1. The β-smooth sensitivity, proposed by Nissim et al. (2007), offers a good compromise by guaranteeing DP while maintaining a noise level significantly lower than that of the global sensitivity.

Definition 3 (β-smooth sensitivity, Nissim et al. (2007))**.**

$$S_\beta(D_0) = \max_{D_1} S(D_1) e^{-\beta d(D_0, D_1)},$$

where $d(\cdot, \cdot)$ refers to the distance between two datasets.

Unfortunately, the β-smooth sensitivity is not directly applicable in our case, as it works only for noise with data-independent correlation, whereas our mechanism introduces data-dependent correlation. Therefore, we also propose a modified version of the β-smooth sensitivity that accommodates noise with data-dependent correlation as follows.

Consider a graph $\mathcal{G}$ over all datasets where there is a directed edge between neighboring datasets. The distance of a directed edge $D_1 \to D_2$ is adjusted as follows.

$$d_\beta(D_1, D_2) = \beta - \frac{1}{2} \log \|(F_1/n_1)(F_2/n_2)^{-1}\|_2.$$

Then, the modified β-smooth sensitivity is defined as a maximum over all paths from the dataset D_0.

Definition 4 (Modified β-smooth sensitivity)**.**

$$S^*(D_0) = \max_{j \geq 0 : D_0 \to \cdots \to D_j} S(D_j) e^{-\sum_{l=1}^{j} d_\beta(D_{l-1}, D_l)}.$$

The modified β-smooth sensitivity serves as an upper bound for the modified sensitivity and satisfies the smoothness property, as stated in Lemma 1. Notably, the smoothness property accounts for the sensitivity of Fisher information, and therefore the sensitivity of distant datasets is less discounted compared to the original β-smooth sensitivity.

Lemma 1 (β-smooth upper bound)**.** *$S^*(D)$ is a β-smooth upper bound of $S(D)$ in the sense that*

$$S^*(D_0) \geq S(D_0),$$
$$S^*(D_1) \leq S^*(D_0) e^{\beta} \|(F_0/n_0)(F_1/n_1)^{-1}\|_2^{-1/2},$$

for all neighboring datasets D_0 and D_1.

All proofs are provided in Appendix B. We are now ready to calibrate the noise level of our proposed mechanism. Let $\alpha = \frac{\epsilon}{2\sqrt{2\log(2.5/\delta)}}$ and $\beta = \frac{\epsilon}{4\sqrt{pK\log(2/\delta)} + 4\log(2/\delta)}$. Theorem 1 implies that, if we choose

$$\psi(D) = nKS^*(D)^2/\alpha^2, \tag{5}$$

where $S^*(D)$ is the modified β-smooth sensitivity, then the proposed noising mechanism satisfies (ϵ, δ)-DP.

Theorem 1. *Suppose $\widetilde{\boldsymbol{\theta}}^0_{1:K}$ and $\widetilde{\boldsymbol{\theta}}^1_{1:K}$ are the noisy copies of MLEs obtained from two neighboring datasets*

D_0 and D_1, respectively. Assume ψ_0 and ψ_1 are chosen according to (5). Then, for any Borel set $B \subset \mathbb{R}^{p\times K}$, it follows that

$$\Pr(\widetilde{\boldsymbol{\theta}}^1_{1:K} \in B) \le e^{\epsilon} \Pr(\widetilde{\boldsymbol{\theta}}^0_{1:K} \in B) + \delta.$$

3.2 Asymptotic Properties

INFEMBLER exhibits excellent asymptotic properties under very mild regularity conditions.

Assumption 1. *All MLEs are independent and asymptotically normal;*

$$\begin{bmatrix} \sqrt{n_1}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0) \\ \vdots \\ \sqrt{n_M}(\boldsymbol{\theta}_M - \boldsymbol{\theta}_0) \end{bmatrix} \to_d \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathcal{I}_1^{-1} & & 0 \\ & \ddots & \\ 0 & & \mathcal{I}_M^{-1} \end{bmatrix}\right),$$

as $N \to \infty$ where $\mathcal{I}_m$ are symmetric and positive definite.

Assumption 2. *All Fisher information matrices are consistent, i.e., $\max_{m\ge 1} \|F_m/n_m - \mathcal{I}_m\|_2 = o_p(1)$ as $N \to \infty$.*

Assumption 3. *For $m \ge 1$, $\lambda_{\min}(\mathcal{I}_m) > t$ and $\lambda_{\max}(\mathcal{I}_m) < T$ for some constants $0 < t < T < \infty$, where $\lambda_{\min}$ and $\lambda_{\max}$ refer to the smallest and largest eigenvalues, respectively.*

All assumptions are mild under trivial situations. Assumptions 1 and 2 capture the typical asymptotic behaviors of MLE and empirical Fisher information. Note that we allow for heterogeneous Fisher information, $\mathcal{I}_m$, in Assumptions 1 and 2.

Theorem 2. *Suppose Assumptions 1-3 hold. Assume $N \to \infty$, $n_m/N \to \alpha_m > 0$, $K_m/n_m \to \gamma_m > 0$, and $\psi_m/K_m \to 0$ for all m. Then, it follows that (i) $\widehat{F} \to_p \mathcal{I}_0$ and (ii) $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to_d \mathcal{N}(\mathbf{0}, \mathcal{I}_0^{-1}\mathcal{I}_*\mathcal{I}_0^{-1})$ where $\mathcal{I}_0$ is the zero of*

$$g(\mathcal{I}) = \alpha_1\mathcal{I}_1 + \sum_{m\ge 2} \alpha_m(1+\gamma_m)(\mathcal{I}^{-1} + \gamma_m\mathcal{I}_m^{-1})^{-1} - \mathcal{I},$$

and

$$\mathcal{I}_* = \alpha_1\mathcal{I}_1 + \sum_{m\ge 2} \alpha_m G_m \mathcal{I}_m^{-1} G_m$$

with $G_m = (1+\gamma_m)(\mathcal{I}_0^{-1} + \gamma_m\mathcal{I}_m^{-1})^{-1}$ for $m \ge 2$.

Note that the function $g(\mathcal{I})$ is concave and therefore the zero of $g(\mathcal{I})$ is unique. There are several remarks that result from Theorem 2.

Remark 1. *If $\max_{m\ge 2}\alpha_m \to 0$ or $\max_{m\ge 2}\gamma_m \to 0$, then $\mathcal{I}_0 \to \mathcal{I}_1$ and $\mathcal{I}_* \to \mathcal{I}_1$. As a result, the asymptotic variance of $\widehat{\boldsymbol{\theta}}$ converges to $\mathcal{I}_1^{-1}$. This implies that if the sample size at the central site is significantly larger than those at the remote sites, or if the number of noisy copies is insufficiently large, the asymptotic behavior of $\widehat{\boldsymbol{\theta}}$ is primarily determined by the central site. This accounts for the remarkable adaptivity and robustness of INFEMBLER.*

Remark 2. *More importantly, it can be shown that, if $\gamma_m > 0$ for some $m \ge 2$, the asymptotic variance of INFEMBLER is strictly smaller than that of the non-DP average estimator weighted by α_m given by $\sum_m \alpha_m\mathcal{I}_m^{-1}$. This implies that INFEMBLER is asymptotically more efficient than the AVG approach (1).*

Remark 3. *If $\gamma_m \to \infty$ for all $m \ge 2$, then $\mathcal{I}_0 \to \sum_{m\ge 1}\alpha_m\mathcal{I}_m$ and $\mathcal{I}_* \to \sum_{m\ge 1}\alpha_m\mathcal{I}_m$ resulting in an asymptotic variance of $(\sum_{m\ge 1}\alpha_m\mathcal{I}_m)^{-1}$. This suggests that if the number of noisy MLE copies is sufficiently large relative to the sample sizes of the remote sites, the asymptotic behavior of $\widehat{\boldsymbol{\theta}}$ closely approximates that of the optimal estimator.*

It remains to be a question whether the proposed privatization mechanism is admissible for the asymptotic property stated in Theorem 2 or the stronger property discussed in Remark 3. The answer to both is yes, and a bit tedious arguments are provided in Appendix B.3. We conclude with a statement on the statistical efficiency of INFEMBLER.

Theorem 3. *Suppose Assumptions 1-3 hold. Assume $N \to \infty$, $n_m/N \to \alpha_m > 0$, $K_m/n_m \to \infty$, and $\psi_m/K_m \to 0$ for all m. Then, it follows that (i) $\widehat{F} \to_p \mathcal{I}_0$ and (ii) $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to_d \mathcal{N}(\mathbf{0}, \mathcal{I}_0^{-1})$ where $\mathcal{I}_0 = \sum_{m\ge 1}\alpha_m\mathcal{I}_m$.*

4 EXPERIMENTS

4.1 Baselines

In our experiments, we compare INFEMBLER against five existing methods: DIP (Bi and Shen, 2022), AVG (as defined in (1)), DP-SGD (Abadi et al., 2016), Site1 (the MLE at the central site), and OPT (the optimal estimator that pools all datasets). For all (ϵ, δ)-DP approaches, we set $\delta = 1/n$. We implemented AVG, DP-SGD, Site1, INFEMBLER, and OPT in R, while the author of Bi and Shen (2022) kindly provided the R code for DIP.

4.2 Simulation

We conduct simulations to compare INFEMBLER with baseline methods across three different scenarios: (i) increasing the sample size at each site while keeping the number of sites fixed, (ii) increasing the number of sites while maintaining a fixed sample size per site, and (iii) increasing the privacy budget while

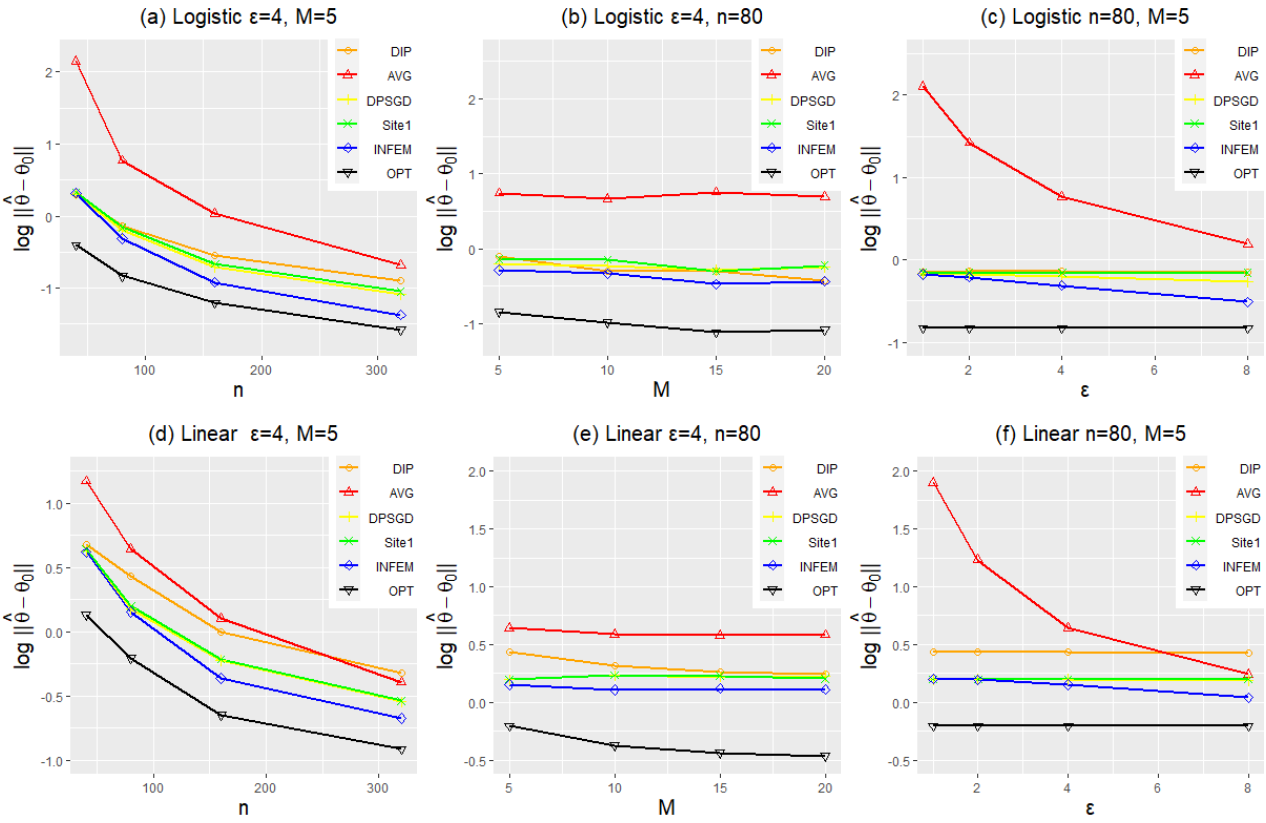

Figure 1: Simulation study results for logistic regressions (upper panels) and linear regressions (lower panels). Columns correspond to three scenarios: (i) increasing sample size per site, (ii) increasing the number of sites, and (iii) increasing the privacy budget. The dimension is fixed at $p = 8$. Reported values are averages of log-transformed L_2 distances between the estimated and true coefficients across all methods.

keeping both the number of sites and the sample size per site constant.

Each dataset consists of $p = 8$ predictors, with the $n \times p$ design matrix X^m of site m having entries x_{ij}^m for $1 \leq i \leq n$ and $1 \leq j \leq p$, where $x_{ij}^m \sim \mathcal{N}(0, 1/m)$. The true coefficients are set as $\boldsymbol{\theta}_0 = (0.5, 0.5, 0.5, 0.5, 0, 0, 0, 0)^T$, and the response variables are sampled from $y_i^m \sim \mathcal{N}(\boldsymbol{\theta}_0^T \mathbf{x}_i^m, 2^2)$ for linear regression and from $y_i^m \sim \text{Bernoulli}(q_i^m)$ with $q_i^m = (1 + e^{-\boldsymbol{\theta}_0^T \mathbf{x}_i^m})^{-1}$ for logistic regression. We generate 100 random datasets for each scenario and compute the L_2-distance between the estimates and the true coefficients. The averages of the log-distances are reported in Figure 1.

Logistic Regression. Figures 1(a)-(c) illustrate the performance of all methods for logistic regression in three different scenarios. Notably, INFEMBLER consistently outperforms all competitors except OPT. In contrast, AVG struggles to effectively de-noise the DP noise. Among all methods, only INFEMBLER and DP-SGD demonstrate better performance than Site1, advocating the benefits of FL. Given that DP-SGD reports the best result from multiple trials with varying clipping sizes and numbers of iterations, these simulation results further highlight the efficiency and robustness of INFEMBLER.

Linear Regression. Figures 1(d)-(f) illustrate the performance of all methods for linear regression in three different scenarios. Similarly to logistic regression, INFEMBLER consistently outperforms all methods except OPT. AVG continues to struggle with effectively de-noising the DP noise. Again, it is remarkable that INFEMBLER guarantees improved performance over Site1 in all cases, despite having no tuning parameter.

Sensitivity to K. INFEMBLER performs robustly provided that K_m is sufficiently large to effectively capture uncertainty information. We include additional sensitivity analyses in Table 1, demonstrating that the method is largely insensitive to variations in the tuning parameter K_m. Please note that other parameters (e.g., ψ_m and β) are derived directly from

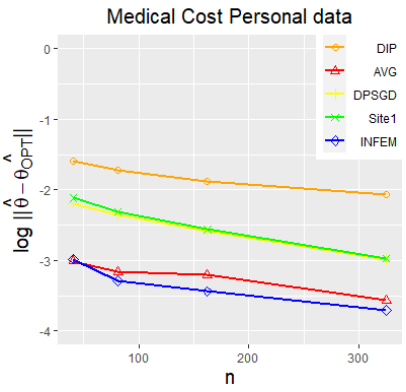

Figure 2: Results of linear regression on the Medical Cost Personal dataset. OPT is treated as the gold standard. The total sample size N is fixed, while the per-site sample size n (and thus the number of sites M) varies. The privacy budget is set to $\epsilon = 4$.

K_m and the privacy budget.

Table 1: Sensitivity of INFEMBLER to variations in K values. Reported values are averages of log-transformed L_2 distances between the estimated and true coefficients across all methods.

K	Logistic (INFEM)	Linear (INFEM)
40	-0.918	-0.334
80	-0.900	-0.327
160	-0.982	-0.326
320	-0.927	-0.334
640	-0.939	-0.341

Sensitivity to p. We have conducted additional analyses for dimensions exceeding $p = 8$ with $M = 8$ and $n = 160$, and the results are shown in Table 2. INFEMBLER outperforms in higher-dimensional settings as well.

Table 2: Performance comparison of different methods on Logistic and Linear models when $M = 8$ and $n = 160$ for various p.

Logistic	DIP	AVG	DPSGD	INFEM	OPT
$p = 8$	-0.545	0.030	-0.706	-0.977	-1.205
$p = 16$	-0.185	0.595	-0.178	-0.435	-0.869
$p = 32$	NA	1.650	0.406	0.320	-0.519
$p = 64$	NA	4.152	1.767	1.642	-0.100
Linear	DIP	AVG	DPSGD	INFEM	OPT
$p = 8$	-0.025	0.101	-0.225	-0.354	-0.647
$p = 16$	0.373	0.730	0.254	0.160	-0.250
$p = 32$	NA	1.433	0.688	0.649	0.100
$p = 64$	NA	2.365	1.173	1.172	0.462

Running Time. We conducted our experiments using R as the experimental platform and an Intel Xeon Platinum 8358 2.60GHz CPU for each experimental setting. The computational time required varied depending on the data size. Table 3 presents the average running times over 100 repetitions of logistic and linear regressions when $M = 8$ and $n = 160$ for $p = 8, 16, 32, 64$.

Table 3: Average running times in seconds for the logistic and linear regressions in simulation study when $M = 8$ and $n = 160$ with various p.

Logistic	DIP	AVG	DPSGD	INFEM	OPT
$p = 8$	$3.2e{-3}$	$1.9e{-3}$	$1.1e{-3}$	$1.6e{-2}$	$5.0e{-3}$
$p = 16$	$5.3e{-1}$	$7.0e{-3}$	$8.9e{-3}$	$2.3e{-2}$	$5.5e{-3}$
$p = 32$	NA	$7.8e{-3}$	$1.3e{-2}$	$5.2e{-2}$	$8.6e{-3}$
$p = 64$	NA	$1.5e{-2}$	$2.6e{-2}$	$3.6e{-1}$	$2.1e{-2}$
Linear	DIP	AVG	DPSGD	INFEM	OPT
$p = 8$	$1.6e{-3}$	$6.1e{-5}$	$6.0e{-5}$	$8.3e{-3}$	$7.4e{-7}$
$p = 16$	$5.5e{-1}$	$2.9e{-4}$	$3.3e{-3}$	$1.5e{-2}$	$8.2e{-5}$
$p = 32$	NA	$2.9e{-4}$	$3.6e{-3}$	$4.3e{-2}$	$8.6e{-5}$
$p = 64$	NA	$3.8e{-4}$	$5.4e{-3}$	$1.6e{-1}$	$1.2e{-4}$

4.3 Real Data Analysis

We compare INFEMBLER with the baseline methods using real datasets. Given the absence of a gold standard in real data, we take the coefficients obtained by OPT as the reference coefficients for our comparisons.

4.3.1 Medical Cost Personal data

For the Medical Cost Personal dataset (Moro et al., 2014), which includes 1,338 individuals with medical insurance and various attributes such as age, sex, and medical costs, we fit linear regression models to predict individual medical costs based on other attributes.

To simulate the federated learning environment, we randomly partition the dataset into multiple subsets, corresponding to different numbers of sites (4, 8, 12, and 16). Each method is then applied to these subsets, and the estimated coefficients are compared to those of OPT. This process is repeated 100 times with different random splits, and the performance of all methods is evaluated based on the average log L_2-distance from OPT.

Figure 2 demonstrates that INFEMBLER consistently outperforms all competitors across all considered numbers of sites. Unlike in the simulation study, AVG performs comparably to INFEMBLER, while DP-SGD remains on par with Site1.

4.3.2 Georgia Coverdell Acute Stroke Registry

Next, we apply linear regression to analyze data from the Georgia Coverdell Acute Stroke Registry

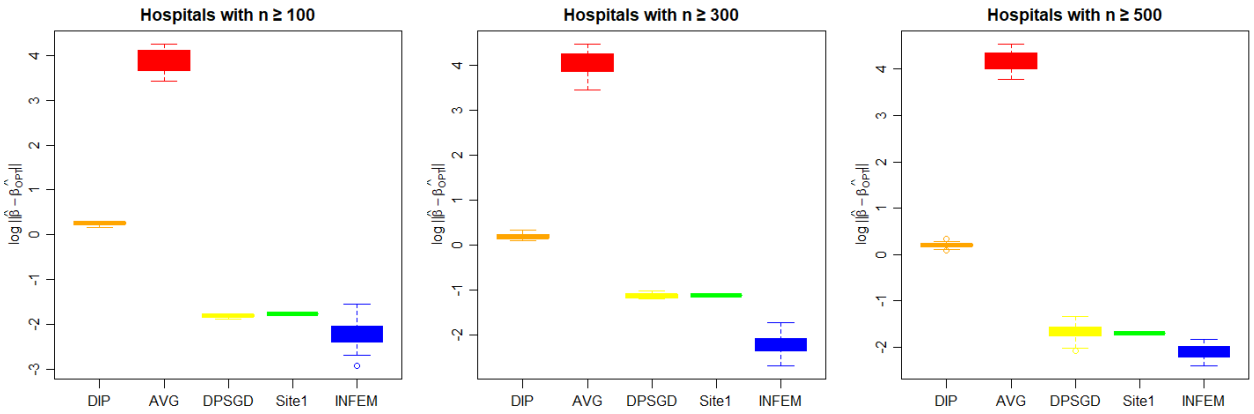

Figure 3: Box plots showing the performance variation of all methods on the Georgia Coverdell Acute Stroke Registry dataset. The non-DP approach Site1 exhibits no variation. OPT is treated as the gold standard, and the privacy budget is set to $\epsilon = 4$.

(GCASR), which was collected from multiple hospitals in Georgia, USA, covering nearly 80% of patients admitted for acute stroke between 2005 and 2013. The task is to predict each patient's CT arrival time, using 11 selected attributes.

In this experiment, we take a different perspective by analyzing the distribution of outputs from all methods, given the inherent randomness of DP methods. We repeat the analysis 100 times using the same dataset and visualize the range of performance measures through box plots in Figure 3. To neutralize the impact of sample size, we consider subsets of hospitals with varying case thresholds (100 or more, 300 or more, and 500 or more), designating the hospital with the median number of cases as the central site in each scenario. As observed in Figure 3, INFEMBLER significantly outperforms all other competitors across all cases.

5 CONCLUSION

We conclude that INFEMBLER is an efficient and privacy preserving method for federated statistical learning. It outperforms existing methods in representative statistical models, showcasing its practical utility, robustness, and broader applicability. We have rigorously established the differential privacy (DP) property of our noising mechanism and examined the asymptotic properties of INFEMBLER. Both the theoretical analyses and extensive experiments strongly confirm INFEMBLER's significantly improved performance over existing methods.

Lastly, we note that INFEMBLER offers communication efficiency by requiring only a single communication between the central site and each remote site. This feature not only helps maintain a low level of DP noise but also eliminates the need for a dedicated communication environment, allowing for manual communication among users, which greatly increases the usability of INFEMBLER.

Acknowledgements

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.IITP-2026-RS-2024-00360227, Leading Generative AI Human Resources Development, No.RS-2025-25442824, AI Star Fellowship Program(Ulsan National Institute of Science and Technology), & No.RS-2020-II201336, Artificial Intelligence Graduate School Program(UNIST)).

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, page 308–318.

Agarwal, N., Suresh, A. T., Yu, F. X. X., Kumar, S., and McMahan, B. (2018). cpsgd: Communication-efficient and differentially-private distributed sgd. *Advances in Neural Information Processing Systems*, 31.

Andrew, G., Thakkar, O., McMahan, B., and Ramaswamy, S. (2021). Differentially private learning with adaptive clipping. In *Advances in Neural Information Processing Systems*, volume 34, pages 17455–17466.

Avella-Medina, M. (2021). Privacy-preserving parametric inference: A case for robust statistics.

Journal of the American Statistical Association, 116(534):969–983.

Bi, X. and Shen, X. (2022). Distribution-invariant differential privacy. *Journal of Econometrics*. In Press.

Cai, T. T., Wang, Y., and Zhang, L. (2021). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850.

Chang, C., Bu, Z., and Long, Q. (2022). Cedar: Communication efficient distributed analysis for regressions. *Biometrics*, 79(3):2357–2369.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3).

Chen, X., Wu, S. Z., and Hong, M. (2020). Understanding gradient clipping in private SGD: A geometric perspective. In *Advances in Neural Information Processing Systems*, volume 33, pages 13773–13782.

Chen, Y., Dong, G., Han, J., Pei, J., Wah, B. W., and Wang, J. (2006). Regression cubes with lossless compression and aggregation. *IEEE Transactions on Knowledge and Data Engineering*, 18(12):1585–1599.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Duan, R., Ning, Y., and Chen, Y. (2022). Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika*, 109(1):67–83.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

Foulds, J., Geumlek, J., Welling, M., and Chaudhuri, K. (2016). On the theory and practice of privacy-preserving bayesian data analysis. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, page 192–201, Arlington, Virginia, USA. AUAI Press.

Heikkilä, M., Lagerspetz, E., Kaski, S., Shimizu, K., Tarkoma, S., and Honkela, A. (2017). Differentially private bayesian learning on distributed data. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jordan, M. I., Lee, J. D., and Yang, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338.

Lu, C.-L., Wang, S., Ji, Z., Wu, Y., Xiong, L., Jiang, X., and Ohno-Machado, L. (2015). WebDISCO: a web service for distributed cox model learning without patient-level data sharing. *Journal of the American Medical Informatics Association*, 22(6):1212–1219.

Moro, S., Cortez, P., and Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31.

Nissim, K., Raskhodnikova, S., and Smith, A. (2007). Smooth sensitivity and sampling in private data analysis. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, STOC '07, page 75–84, New York, NY, USA. Association for Computing Machinery.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wasserman, L. and Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389.

Wu, Y., Jiang, X., Kim, J., and Ohno-Machado, L. (2012). Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. *Journal of the American Medical Informatics Association*, 19(5):758–764.

Zhang, Y., Duchi, J. C., and Wainwright, M. J. (2013). Communication-efficient algorithms for statistical optimization. *J. Mach. Learn. Res.*, 14(1):3321–3363.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Private, Efficient, and Robust Federated Statistical Learning: Supplementary Materials

A Algorithm for INFEMBLER

A.1 Posterior Densities

The log joint posterior density for $(\boldsymbol{\theta}, F, \boldsymbol{\theta}_{2:M}, F_{2:M})$ is given by

$$\log \pi(\boldsymbol{\theta}, F, \boldsymbol{\theta}_{2:M}, F_{2:M} | \boldsymbol{\theta}_1, F_1, \widetilde{\Theta}_{2:M}) = C + \sum_{m=2}^{M} \frac{n_m + k_m - p}{2} \log |F_m| - \frac{N+p+1}{2} \log |F| - \frac{1}{2} \operatorname{tr} F^{-1} \sum_{m=1}^{M} F_m$$
$$- \frac{1}{2} \sum_{m=1}^{M} (\boldsymbol{\theta}_m - \boldsymbol{\theta})^T F_m (\boldsymbol{\theta}_m - \boldsymbol{\theta}) - \sum_{m=2}^{M} \frac{\sum_{k=1}^{K_m} (\widetilde{\boldsymbol{\theta}}_{mk} - \boldsymbol{\theta}_m)^T F_m (\widetilde{\boldsymbol{\theta}}_{mk} - \boldsymbol{\theta}_m)}{2\psi_m}.$$

If we marginalize out $\boldsymbol{\theta}_{2:M}$, we obtain

$$\begin{aligned} \log \pi(\boldsymbol{\theta}, F, F_{2:M} | \boldsymbol{\theta}_1, F_1, \widetilde{\Theta}_{2:M}) &= C + \sum_{m=2}^{M} \frac{n_m + K_m - p - 1}{2} \log |F_m| - \frac{N+p+1}{2} \log |F| \\ &- \frac{1}{2} (\boldsymbol{\theta}_1 - \boldsymbol{\theta})^T F_1 (\boldsymbol{\theta}_1 - \boldsymbol{\theta}) - \frac{1}{2} \operatorname{tr} F^{-1} F_1 - \frac{1}{2} \sum_{m=2}^{M} \operatorname{tr} F_m \overline{F}_m(\boldsymbol{\theta}, F)^{-1}, \end{aligned} \tag{6}$$

where $\overline{F}_m(\boldsymbol{\theta}, F)^{-1}$ is given in (3).

A.2 EM Algorithm

A.2.1 E-step

Based on (6), the conditional distribution of F_m given $(\boldsymbol{\theta}, F)$ is given as

$$F_m | \boldsymbol{\theta}, F \sim \mathcal{W}(\overline{F}_m(\boldsymbol{\theta}, F), n_m + K_m), \quad m \geq 2.$$

Therefore, the E-step goes as follows.

$$\widehat{F}_m \leftarrow (n_m + K_m) \overline{F}_m(\widehat{\boldsymbol{\theta}}, \widehat{F}), \quad m \geq 2.$$

A.2.2 M-step

Based on (6), M-step updates are as follows.

$$\begin{aligned} \widehat{F} &\leftarrow \frac{1}{N+p+1} \sum_{m=1}^{M} \widehat{F}_m, \\ \widehat{\boldsymbol{\theta}} &\leftarrow \left(\sum_{m=1}^{M} \frac{K_m \widehat{F}_m}{K_m + \psi_m} \right)^{-1} \left(\sum_{m=1}^{M} \frac{K_m \widehat{F}_m \overline{\boldsymbol{\theta}}_m}{K_m + \psi_m} \right), \end{aligned} \tag{7}$$

where $\widehat{F}_1 = F_1$, $\overline{\boldsymbol{\theta}}_1 = \boldsymbol{\theta}_1$, $K_1 = 1$, and $\psi_1 = 0$.

B Proofs

B.1 Differential Privacy

Proof of Lemma 1. The first claim is trivial considering the length $j = 0$ path. Suppose $D_1 \to D_2 \to \cdots \to D_j$ is the path with which $S^*(D_1)$ is attained. Then, since $D_0 \to D_1 \to \cdots \to D_j$ is a path from D_0 and to D_j, we have

$$\begin{aligned} S^*(D_0) &\geq S(D_j)e^{-\sum_{l=1}^{j} d(D_{l-1},D_l)} \\ &= S(D_j)e^{-\sum_{l=2}^{j} d(D_{l-1},D_l)}e^{-d(D_0,D_1)} \\ &= S^*(D_1)e^{-\beta}\|(F_0/n_0)(F_1/n_1)^{-1}\|_2^{1/2}. \end{aligned}$$

□

Before we prove Theorem 1, we present a dilation property of Gaussian distribution which is slightly different than the one presented in Nissim et al. (2007).

Lemma 2 (Dilation Property of Gaussian Distribution)**.** *Suppose $\mathbf{z} \sim \mathcal{N}(0, I)$ with dimension p and let $\epsilon > 0$ and $\delta \in (0, 1)$. Then, for any Borel set $B \subset \mathbb{R}^p$, it follows that*

$$\Pr(\mathbf{z} \in B) \leq e^{\epsilon/2} \Pr(\mathbf{z} \in e^{\lambda} B) + \delta/2,$$

if $|\lambda| < \frac{\epsilon}{4\sqrt{p \log(2/\delta)} + 4 \log(2/\delta)}$.

Proof of Lemma 2. Let $\phi(\cdot)$ is the density function of the standard Gaussian distribution with dimension p. Assume $\lambda > 0$ and we require

$$P\left(\frac{\phi(e^{-\lambda}\mathbf{z})e^{-\lambda p}}{\phi(\mathbf{z})} > e^{\epsilon/2}\right) = P\left(\frac{1 - e^{-2\lambda}}{2}\mathbf{z}^T\mathbf{z} - \lambda p > \frac{\epsilon}{2}\right) \leq P(Y - p > \epsilon/(2\lambda)) \leq \delta/2,$$

where $Y \sim \chi^2_p$, since $1 - e^{-2\lambda} \leq 2\lambda$. By Lemma 3.3 in Laurent and Massart (2000), it suffices to have

$$\lambda \leq \frac{\epsilon}{4\sqrt{p \log(2/\delta)} + 4 \log(2/\delta)}.$$

Assume $\lambda < 0$ and we require

$$P\left(\frac{\phi(e^{-\lambda}\mathbf{z})e^{-\lambda p}}{\phi(\mathbf{z})} > e^{\epsilon/2}\right) = P\left(\frac{e^{-2\lambda} - 1}{2}\mathbf{z}^T\mathbf{z} + \lambda p < -\frac{\epsilon}{2}\right) \leq P(Y - p < \epsilon/(2\lambda)) \leq \delta/2,$$

since $e^{-2\lambda} - 1 \geq -2\lambda$. By Lemma 3.3 in Laurent and Massart (2000), it suffices to have

$$\lambda \geq -\frac{\epsilon}{4\sqrt{p \log(2/\delta)}}.$$

□

Proof of Theorem 1. It follows, from the property of the standard Gaussian mechanism, that

$$\Pr(\widetilde{\boldsymbol{\theta}}^1_{1:K} \in B) = \Pr(\mathbf{z}_{1:K} \in (F_1/\psi_1)^{1/2}(B - \boldsymbol{\mu}_1\mathbf{1}^T)) \leq e^{\epsilon/2} \Pr(\mathbf{z}_{1:K} \in (F_1/\psi_1)^{1/2}(B - \boldsymbol{\mu}_0\mathbf{1}^T)) + \delta/2,$$

because, due to the definition of the modified sensitivity and Lemma 1, the sliding distance is sufficiently small:

$$\psi_1^{-1}K(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T F_1(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \leq \psi_1^{-1}Kn_1S^*(D_1)^2 = \alpha^2.$$

On the other hand, we have

$$\Pr(\mathbf{z}_{1:K} \in (F_1/\psi_1)^{1/2}(B - \boldsymbol{\mu}_0\mathbf{1}^T)) \leq e^{\epsilon/2} \Pr(\mathbf{z}_{1:K} \in (F_0/\psi_0)^{1/2}(B - \boldsymbol{\mu}_0\mathbf{1}^T)) + \delta/2 = e^{\epsilon/2} \Pr(\widetilde{\boldsymbol{\theta}}^0_{1:K} \in B) + \delta/2,$$

by Lemma 2, since Lemma 1 implies

$$e^{-\beta}I \leq (\psi_1/\psi_0)^{1/2}F_0^{1/2}F_1^{-1/2} = (S^*(D_1)/S^*(D_0))(F_0/n_0)^{1/2}(F_1/n_1)^{-1/2} \leq e^{\beta}I.$$

These two results combined yield (ϵ, δ)-DP following the proof of the basic composition rule of (ϵ, δ)-DP. □

B.2 Asymptotic Properties

Proof of Theorem 2. First note that Algorithm 1 implies that $\widehat{F}$ satisfies the equation

$$g_N(F) = \frac{F_1 + \sum_{m>1}(n_m + K_m)\overline{F}_m(\widehat{\boldsymbol{\theta}}, F)}{N + p + 1} - F = 0.$$

Equation (7) implies that $\widehat{\boldsymbol{\theta}}$ is a weighted average of $\overline{\boldsymbol{\theta}}_{1:M}$. Therefore, we have $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 \leq \max_{m\geq 1} \|\overline{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0\|_2$ and therefore

$$\begin{aligned}\max_{m\geq 1} \|\overline{\boldsymbol{\theta}}_m - \widehat{\boldsymbol{\theta}}\|_2 &\leq \max_{m\geq 1} \|\overline{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0\|_2 + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 \\ &\leq 2 \max_{m\geq 1} \|\overline{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0\|_2 = o_p(1).\end{aligned}$$

Since $K_m/n_m \to \gamma_m$, we have

$$\psi_m^{-1} \sum_{k=1}^{K} (\widetilde{\boldsymbol{\theta}}_{mk} - \overline{\boldsymbol{\theta}}_m)(\widetilde{\boldsymbol{\theta}}_{mk} - \overline{\boldsymbol{\theta}}_m)^T \to_p \gamma_m \mathcal{I}_m^{-1},$$

for all $m \geq 2$, and thus we have

$$\max_{m\geq 2} \|\overline{F}_m(\widehat{\boldsymbol{\theta}}, F)^{-1} - (F^{-1} + \gamma_m \mathcal{I}_m^{-1})\| = o_p(1)$$

and

$$\max_{m\geq 2} \|\widehat{F}_m/n_m - (1+\gamma_m)(F^{-1} + \gamma_m \mathcal{I}_m^{-1})^{-1}\| = o_p(1),$$

which implies $g_N(F) \to_p g(F)$ for any F. Then, by the standard arguments for the consistency of Z-estimators (Theorem 5.9 in van der Vaart (1998)), the first claim follows.

Note that

$$\mathrm{Var}(\overline{\boldsymbol{\theta}}_m | F_m) = \mathrm{Var}(\boldsymbol{\theta}_m) + \frac{\psi_m}{K_m} F_m^{-1}, \quad m \geq 2.$$

According to (7) of Algorithm 1, since $\psi_m/K_m \to 0$, it follows from the Slutsky Theorem that

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to_d \mathcal{N}(\mathbf{0}, \mathcal{I}_0^{-1} \mathcal{I}_* \mathcal{I}_0^{-1}).$$

Hence, the second claim has been proved. □

Proof of Theorem 3. The proof is almost identical to the proof of Theorem 2. Since $K_m/n_m \to \infty$, it follows that

$$\max_{m\geq 2} \|\widehat{F}_m/n_m - \mathcal{I}_m\| = o_p(1),$$

which immediately implies the first claim follows. The second claim also immediately follows by the first claim and the Slutsky Theorem. □

B.3 Admissibility

For simplicity, assume that the parameter space and the data domain are bounded, and that the Hessian of the log-likelihood is both bounded and Lipschitz continuous. Consider a dataset D_0 and let D_j denote any dataset within distance j from D_0. Take $\delta = 1/n_0$. As the local sample size n_0 increases, we typically have $F_0 = O_p(n_0)$ and $\boldsymbol{\theta}_0 - \boldsymbol{\theta}_1 = O_p(n_0^{-1})$, leading to $S(D_j) = O_p((n_0 - j)^{-1})$. Assuming $K_0 = n_0$, we note that $\beta \approx (n_0 \log n_0)^{-1/2}$ and $\|(F_0/n_0)(F_1/n_1)^{-1}\|_2 = O_p(n_0^{-1})$, which yields the quantity in Definition 4 $-\sum_{l=1}^{j} d(D_{l-1}, D_l) = O_p(j(n_0 \log n_0)^{-1/2})$. Thus, we obtain $S^*(D_0) = O_p(n_0^{-1})$, resulting in $\psi_0 = O_p(\log n_0)$. Since $\psi_0/K_0 \to 0$, the proposed noising mechanism is admissible for Theorem 2. If we instead choose $K_0 = n_0 \log n_0$, the same analysis leads to $\psi_0 = O_p((\log n_0)^2)$. Since we have $K_0/n_0 \to \infty$ and $\psi_0/K_0 \to 0$, the proposed noising mechanism is also admissible for Theorem 3.