Online Supplement to "Privacy Protection and Statistical Efficiency Trade-off for Federated Learning"

Haobo Qi, Feifei Wang, Hansheng Wang

Appendix A: Technical Proofs

Appendix A.1: Proof of Theorem 1

Recall that by iteratively applying the updating formula, we can obtain that

$$\widehat{\theta}_t = \left(\widehat{\Delta}_\alpha\right)^t \widehat{\theta}_0 + \left\{I - \left(\widehat{\Delta}_\alpha\right)^t\right\} \widehat{\theta}_{\text{ols}} + \mathcal{E}_t(\alpha),$$

where $\mathcal{E}_t(\alpha) = \alpha \sum_{s=1}^{t} \left(\widehat{\Delta}_\alpha\right)^{t-s} \xi_s$. If we further define $\mathcal{E}(\alpha) = \sum_{r=0}^{\infty} (\widehat{\Delta}_\alpha)^r \xi_{-r}$, then one can verify that

$$\begin{aligned} \left\|\widehat{\theta}_t - \widehat{\theta}_t^{\infty}\right\| &= \left\|\widehat{\Delta}_\alpha\right\|^t \left\|\widehat{\theta}_0 - \widehat{\theta}_{\text{ols}}\right\| + \left\|\mathcal{E}_t(\alpha) - \mathcal{E}_t^{\infty}(\alpha)\right\| \\ &\leq \left\|\widehat{\Delta}_\alpha\right\|^t \left\|\widehat{\theta}_0 - \widehat{\theta}_{\text{ols}}\right\| + \left\|\widehat{\Delta}_\alpha\right\|^t \left\|\mathcal{E}(\alpha)\right\|. \end{aligned} \tag{A.1}$$

According to inequality (A.1), as long as we can prove that there exists a sufficiently small $\eta > 0$ and positive constant $C > 0$ such that

$$P\left(\left\|\widehat{\Delta}_\alpha\right\| \leq 1 - \alpha\tau_{\min}\right) \geq 1 - 2p \exp\left(-\frac{CN\eta^2}{p}\right), \tag{A.2}$$

then the theorem result holds automatically. We then turn to prove (A.2). Note that $\left\|\widehat{\Delta}_\alpha\right\| = \left\|I - \alpha\widehat{\Sigma}_{xx}\right\| \leq \left\|I - \alpha\Sigma_{xx}\right\| + \alpha\left\|\widehat{\Sigma}_{xx} - \Sigma_{xx}\right\|$. Then by the assumptions, we know that there exist two positive constants $G_1, G_2 > 0$ such that $\|N^{-1}X_iX_i^\top - N^{-1}\Sigma_{xx}\| \leq G_1 p/N$ and $v(\widehat{\Sigma}_{xx}) = \|E\{(\widehat{\Sigma}_{xx} - \Sigma_{xx})(\widehat{\Sigma}_{xx} - \Sigma_{xx})\}\| \leq G_2 p/N$. By the results of Matrix Bernstein Inequality, we have for any $\eta > 0$,

$$P\left(\left\|\widehat{\Sigma}_{xx} - \Sigma_{xx}\right\| > \eta\right) \leq 2p \exp\left(\frac{-\eta^2/2}{v(\widehat{\Sigma}_{xx}) + G_1 p/(3\eta N)}\right).$$

By choosing a sufficiently small η, we should have $\tau_{\min} \leq \lambda_{\min}(\Sigma_{xx}) + \eta$. Then there exists a positive constant C, such that

$$P\Big(\big\|\widehat{\Delta}_\alpha\big\| \leq 1 - \alpha\tau_{\min}\Big) \geq P\left(\big\|\widehat{\Sigma}_{xx} - \Sigma_{xx}\big\| \leq \eta\right) \geq 1 - 2p\exp\left(-\frac{CN\eta^2}{p}\right).$$

This finishes the proof.

Appendix A.2: Proof of Theorem 2

We study the conditional bias and conditional variance of the noise-added FGD estimator one by one. First, according to the updating formula (2), we can calculate that $E(\widehat{\theta}_T - \widehat{\theta}_{\text{ols}}|\mathcal{D}) = \big(\widehat{\Delta}_\alpha\big)^T\big(\widehat{\theta}_0 - \widehat{\theta}_{\text{ols}}\big)$. Note that $\|\widehat{\Delta}_\alpha\| \leq (1 - \alpha\tau_{\min})$ and $\|\widehat{\theta}_0 - \widehat{\theta}_{\text{ols}}\| = O_p(\sqrt{p})$. We then have $E(\widehat{\theta}_T - \widehat{\theta}_{\text{ols}}|\mathcal{D}) = O_p\{\sqrt{p}\exp(-\alpha T)\}$. Thus the conditional bias is negligible if and only if $\exp(\alpha T)/\sqrt{N} \to \infty$, which implies $(\alpha T)/\log N \to \infty$. This implies the theorem conclusion (a).

We then focus on the conditional covariance. Under the theorem conditions, we know that $\|\widehat{\Delta}_\alpha\| < 1$. Then one can verified that $\text{cov}(\widehat{\theta}_T|\mathcal{D}) = \text{cov}\{\mathcal{E}_T(\alpha)\} \preceq \text{cov}\{\mathcal{E}_T^\infty(\alpha)\}$, where

$$\text{cov}\{\mathcal{E}_T^\infty(\alpha)\} = \alpha^2\left(\sum_{k=1}^K \widehat{\pi}_k^2\sigma_k^2\right)\left(I - \widehat{\Delta}_\alpha^2\right)^{-1} = \frac{\alpha c_{\text{MA}}^2 C^2 KT\log(1/\delta)}{\varepsilon^2 N^2}\widehat{\Sigma}_{xx}^{-1}\big(2I - \alpha\widehat{\Sigma}_{xx}\big)^{-1}.$$

Note that $\text{tr}\{\text{cov}(\widehat{\theta}_{\text{ols}})\} = O(p/N)$, then the noise-added FGD estimator can become as efficient as the OLS estimator $\widehat{\theta}_{\text{ols}}$ only if $\text{tr}\{\text{cov}(\widehat{\theta}_T|\mathcal{D})\} = o_p(p/N)$. It can be easily verified that

$$\begin{aligned}\text{tr}\{\text{cov}(\widehat{\theta}_T^\infty)\} &\leq c_{\text{MA}}^2 C^2(\alpha KT)\log(1/\delta)\,\text{tr}\left\{\widehat{\Sigma}_{xx}^{-1}\big(2I - \alpha\widehat{\Sigma}_{xx}\big)^{-1}\right\}/(\varepsilon^2 N^2) \\ &= O_p\left\{\frac{p}{N}\cdot\frac{\alpha T}{n}\cdot\frac{\log(1/\delta)}{\varepsilon^2}\right\}.\end{aligned}$$

Consequently, for any given (ε, δ), we know that $\text{tr}\{\text{cov}(\widehat{\theta}_T^\infty)\}$ can be negligible if $\alpha T/n \to 0$. This finishes the proof.

Appendix A.3: Proof of Theorem 3

We study the conditional bias and conditional variance of the averaged estimator $\bar{\theta}_T$ one by one. First, according to the updating formula in Algorithm 3, we can calculate

$$E(\bar{\theta}_T - \widehat{\theta}_{\text{ols}}|\mathcal{D}) = T^{-1}\sum_{t=1}^{T}\big(\widehat{\Delta}_\alpha\big)^t\big(\widehat{\theta}_0 - \widehat{\theta}_{\text{ols}}\big).$$

Note that $\|\widehat{\Delta}_\alpha\| \le (1-\alpha\tau_{\min})$. We then have $\|E(\widehat{\theta}_T - \widehat{\theta}_{\text{ols}}|\mathcal{D})\| \le \|\widehat{\theta}_0 - \widehat{\theta}_{\text{ols}}\|/(\alpha T\tau_{\min})$. Thus the conditional bias is negligible if $(\alpha T)/\sqrt{N} \to \infty$, which implies the theorem conclusion (a). Next, we consider the conditional covariance. As we discussed in Appendix A.2, given $\|\widehat{\Delta}_\alpha\| < 1$, we should have $\text{cov}(\widehat{\theta}_t|\mathcal{D}) = \text{cov}\{\mathcal{E}_t(\alpha)|\mathcal{D}\} \preceq \text{cov}\{\mathcal{E}_t^\infty(\alpha)|\mathcal{D}\}$, where

$$\text{cov}\{\mathcal{E}_t^\infty(\alpha)|\mathcal{D}\} = \frac{\alpha c_{\text{MA}}^2 C^2 K T\log(1/\delta)}{\varepsilon^2 N^2}\widehat{\Sigma}_{xx}^{-1}\big(2I - \alpha\widehat{\Sigma}_{xx}\big)^{-1}. \tag{A.3}$$

One can verify that $\text{cov}(\bar{\theta}_T|\mathcal{D}) = T^{-2}\,\text{cov}\big(\sum_{t=1}^T \widehat{\theta}_t|\mathcal{D}\big) = Q_1 + Q_2$, where

$$Q_1 = T^{-2}\sum_{t=1}^{T}\text{cov}\big(\widehat{\theta}_t|\mathcal{D}\big) \quad \text{and} \quad Q_2 = T^{-2}\sum_{s\neq t}\text{cov}\big(\widehat{\theta}_t, \widehat{\theta}_s|\mathcal{D}\big)$$

Based on the previous analysis, we know that

$$\text{tr}(Q_1) = \frac{\alpha c_{\text{MA}}^2 C^2 K\log(1/\delta)}{\varepsilon^2 N^2}\widehat{\Sigma}_{xx}^{-1}\big(2I - \alpha\widehat{\Sigma}_{xx}\big)^{-1} = O_p\left\{\frac{p}{N}\cdot\frac{\alpha}{n}\cdot\frac{\log(1/\delta)}{\varepsilon^2}\right\}.$$

We then turn to consider Q_2. A simple calculation reveals that $\text{cov}\big(\widehat{\theta}_t, \widehat{\theta}_s|\mathcal{D}\big) = \alpha^2\sigma_\xi^2\widehat{\Delta}_\alpha^{|t-s|}$. Then we have $Q_2 = \alpha^2\sigma_\xi^2 T^{-2}Q_3$ with $Q_3 = \sum_{t=1}^T(T+1-t)\widehat{\Delta}_\alpha^{2t}$. One can verify that $(I - \widehat{\Delta}_\alpha^2)Q_3 = T\widehat{\Delta}_\alpha^2 - \widehat{\Delta}_\alpha^2(I + \widehat{\Delta}_\alpha^2 + \cdots + \widehat{\Delta}_\alpha^{2T-2}) \preceq TI_p$, and thus $Q_3 \preceq T(I - \widehat{\Delta}_\alpha^2)^{-1} = \alpha^{-1}T\widehat{\Sigma}_{xx}^{-1}(2I - \alpha\widehat{\Sigma}_{xx})^{-1}$. Combine all the results, we have

$$\begin{aligned}\text{cov}(\bar{\theta}_T|\mathcal{D}) = Q_1 + Q_2 &\preceq 2\frac{\alpha c_{\text{MA}}^2 C^2 K\log(1/\delta)}{\varepsilon^2 N^2}\widehat{\Sigma}_{xx}^{-1}\big(2I - \alpha\widehat{\Sigma}_{xx}\big)^{-1} \\ &= O_p\left\{\frac{p}{N}\cdot\frac{\alpha}{n}\cdot\frac{\log(1/\delta)}{\varepsilon^2}\right\}\end{aligned}$$

Note that $\text{tr}\{\text{cov}(\widehat{\theta}_{\text{ols}})\} = O(p/N)$, then the averaged estimator $\bar{\theta}_T$ can become as efficient as the OLS estimator $\widehat{\theta}_{\text{ols}}$ only if $\text{tr}\{\text{cov}(\bar{\theta}_T|\mathcal{D})\} = o_p(p/N)$. For any given $(\varepsilon,\, \delta)$, we know

that $\mathrm{tr}\{\mathrm{cov}(\bar{\theta}_T)\}$ can be negligible if $\alpha/n \to 0$. This finishes the proof.

Appendix A.4: Proof of Theorem 4

Recall the equation (2.3) as

$$\widetilde{\theta}_T - \widehat{\theta}_T = \alpha \sum_{t=1}^{T} \left(\prod_{s=t+1}^{T} \widehat{\Delta}_\alpha^s \right) \xi_t. \tag{A.4}$$

First, by the assumptions, we know that there exist two positive constants $\tau_{\min}$ and $\tau_{\max}$ such that $(1-\alpha\tau_{\max})I_p \preceq (\widehat{\Delta}_\alpha^{(t)}) \preceq (1-\alpha\tau_{\min})I_p$. Consequently, we should have

$$\begin{aligned} & E\Big\{ \big\|\widehat{\theta}_t - \widehat{\theta}\big\|^2 \,|\mathcal{D}\Big\} = E\Big\{ \big\|\widehat{\theta}_t - \widetilde{\theta}_t + \widetilde{\theta}_t - \widehat{\theta}\big\|^2 \,|\mathcal{D}\Big\} \\ \leq \;& 2E\Big\{ \big\|\widehat{\theta}_t - \widetilde{\theta}_t\big\|^2 |\mathcal{D}\Big\} + 2\big\|\widetilde{\theta}_t - \widehat{\theta}\big\|^2. \end{aligned}$$

We then derive them one by one. First, according to the updating formula of the gradient descent algorithm, we have

$$\begin{aligned} \big\|\widetilde{\theta}_t - \widehat{\theta}\big\|^2 &= \big\|\widetilde{\theta}_{t-1} - \alpha\dot{\mathcal{L}}(\widetilde{\theta}_{t-1}) - \widehat{\theta}\big\|^2 \\ &= \big\|\widetilde{\theta}_{t-1} - \widehat{\theta}\big\|^2 - 2\alpha < \dot{\mathcal{L}}(\widetilde{\theta}_{t-1}), \widetilde{\theta}_{t-1} - \widehat{\theta} > + \alpha^2 \big\|\dot{\mathcal{L}}(\widetilde{\theta}_{t-1})\big\|^2 \\ &\leq \left(1 - \alpha\frac{2\tau_{\max}\tau_{\min}}{\tau_{\max}+\tau_{\min}}\right) \big\|\widetilde{\theta}_{t-1} - \widehat{\theta}\big\|^2 + \alpha\left(\alpha - \frac{2}{\tau_{\min}+\tau_{\max}}\right) \big\|\dot{\mathcal{L}}(\widetilde{\theta}_{t-1})\big\|^2 \\ &\leq \left(1 - \alpha\frac{2\tau_{\max}\tau_{\min}}{\tau_{\max}+\tau_{\min}}\right)^t \big\|\widehat{\theta}_0 - \widehat{\theta}\big\|^2. \end{aligned}$$

Next, we consider $E\Big\{ \big\|\widehat{\theta}_t - \widetilde{\theta}_t\big\|^2 |\mathcal{D}\Big\}$. According to equation (A.4), let $q = (1-\alpha\tau_{\min})$,

we then can derive that

$$
\begin{aligned}
& E\Big\{\big\|\widehat{\theta}_t-\widetilde{\theta}_t\big\|^2|\mathcal{D}\Big\}=\alpha^2E\left\{\Big\|\sum_{t=1}^{T}\left(\prod_{s=t+1}^{T}\widehat{\Delta}_\alpha^s\right)\xi_t\Big\|^2\Big|\mathcal{D}\right\} \\
\leq\ & \alpha^2\sum_{t=1}^{T}q^{2(T-t)}E\big\|\xi_t\big\|^2+2\alpha^2\sum_{t=1}^{T}\sum_{s=t+1}^{T}q^{2T-(t+s)}E\Big\{\big(\max_t\big\|\xi_t\big\|\big)^2\Big\} \\
\leq\ & \frac{\alpha^2}{1-q^2}E\big\|\xi_t\big\|^2+2\alpha^2\frac{q}{(1-q)^2}E\Big\{\big(\max_t\big\|\xi_t\big\|\big)^2\Big\} \\
=\ & c_1\frac{\alpha c_{\mathrm{MA}}^2C^2KT\log(1/\delta)p}{\varepsilon^2N^2}+c_2\frac{\alpha c_{\mathrm{MA}}^2C^2K\log(T)T\log(1/\delta)p}{\varepsilon^2N^2} \\
=\ & O_p\left(\frac{p}{N}\cdot\frac{\alpha T\log(T)}{n}\cdot\frac{\log(1/\delta)}{\varepsilon^2}\right).
\end{aligned}
$$

Consequently, if we want $\big\|\widetilde{\theta}_t-\widehat{\theta}\big\|^2=o_p(p/N)$, we should have $\alpha T/\log N\to\infty$. On the other hand, if we want $E\Big\{\big\|\widehat{\theta}_t-\widetilde{\theta}_t\big\|^2|\mathcal{D}\Big\}=o_p(p/N)$, we should have $\alpha T\log(T)/n\to 0$.

Appendix B: Additional Simulation Results

Appendix B.1: Effect of Number of Clients K

We consider the effect of K. In general, the experimental setting is similar to that in Section 3.1, except that we fix $\varepsilon=0.8$ and let K vary among $\{5,15,30\}$. The MSE values of the baseline FGD estimator without DP, the NA-FGD estimator, and the ANA-FGD estimator are evaluated by a total of $B=100$ times replications for each parameter combination. The MSE results under the sparse case and dense case are reported in Figures B.1 and B.2, respectively. By the two figures, we find that the statistical efficiency of all three estimators improves as the number of clients K increases. Moreover, we can still find that the ANA-FGD estimator exhibits better statistical properties compared to the NA-FGD estimator.

Appendix B.2: Laplacian Noise Mechanism

Since our theories are established based on the Gaussian noise mechanism, it is of great interest to investigate whether our method remains valid for the Laplacian mechanism. To

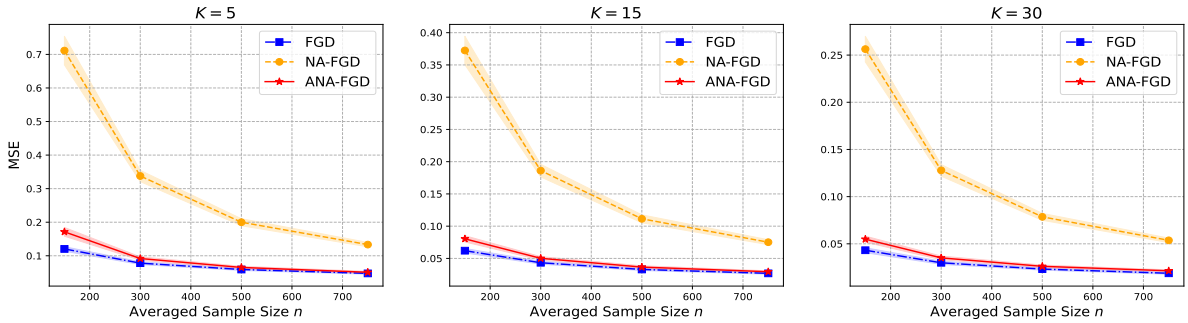

Figure B.1: The MSE results of the three estimators under different number of clients K with the sparse true parameter. The lines represent the MSE values and the shaded areas represent the standard error bounds.

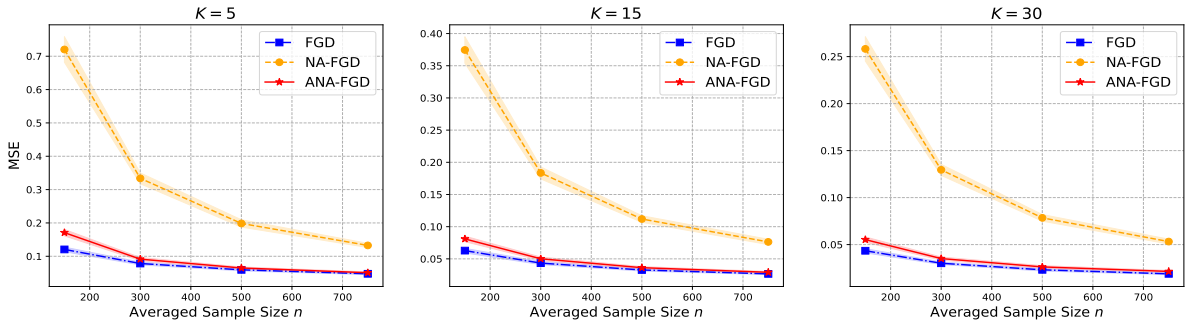

Figure B.2: The MSE results of the three estimators under different number of clients K with the dense true parameter. The lines represent the MSE values and the shaded areas represent the standard error bounds.

this end, we keep the setting similar to that in Section 3.1, except that the noises added to the gradients follow a Laplace distribution with noise level proportional to the ℓ_1 sensitivity of the local gradients. Therefore, we should clip the ℓ_1 norm of the local gradients instead of clipping the ℓ_2 norm of the local gradients for the Gaussian mechanism. The MSE results of the three estimators under the sparse case and dense case are reported in Figures B.3 and B.4, respectively. As shown by these two figures, we find the statistical properties of the two DP estimators become closer to the baseline FGD estimator as the privacy budget ε increases. In addition, the ANA-FGD estimator performs much better than the NA-FGD estimator, and nearly matches the baseline FGD estimator as the sample size increases. Last, compared with the MSE results with the Gaussian noise mechanism (Figures 1 and 2), we find the statistical efficiencies of ANA-FGD and NA-FGD estimators with the Laplacian noise mechanism behave worse.

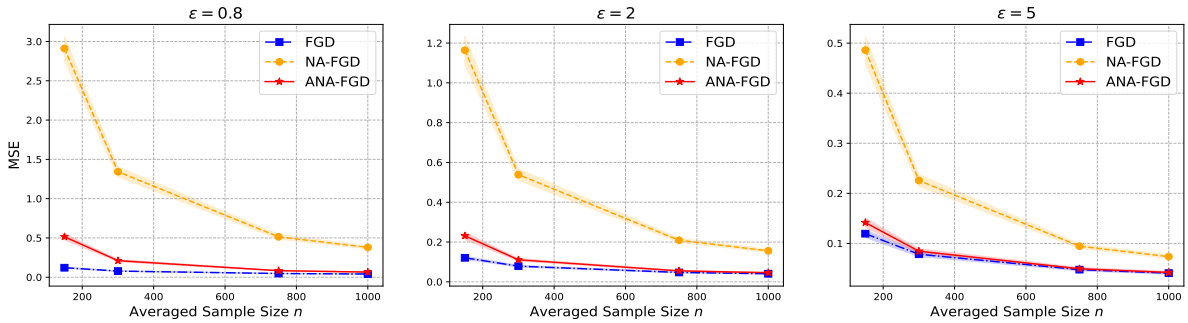

Figure B.3: The MSE results of the three estimators under three different privacy budgets with Laplacian mechanism and the sparse true parameter. The lines represent the MSE values and the shaded areas represent the standard error bounds.

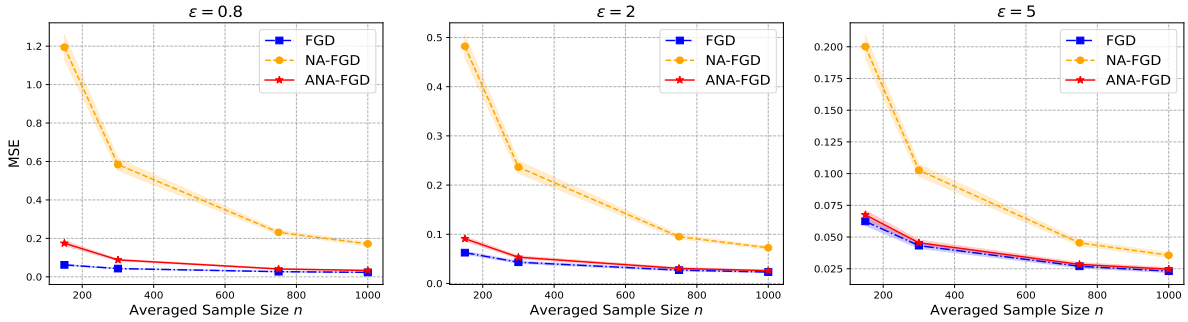

Figure B.4: The MSE results of the three estimators under three different privacy budgets with Laplacian mechanism and dense true parameter. The lines represent the MSE values and the shaded areas represent the standard error bounds.

Appendix B.3: Diminishing Learning Rates

We investigate the performance of using the scheduling strategy with diminishing learning rates, since it is a commonly used strategy to ensure the convergence for FGD algorithms. Specifically, we fix the feature dimension $p = 200$ and the number of clients $K = 15$. Suppose the local sample size $n_k = n$ for all k, and let n vary among $\{150, 300, 500, 750\}$. We further fix $\varepsilon = 0.8$ and $\delta = 10^{-5}$. We consider a polynomial decay strategy, i.e., $\alpha_t = \alpha_0 t^{-\gamma}$ for a positive constant $\alpha_0 = 1$ and $\gamma \in \{0.2, 0.5, 0.8\}$. The MSE results of the three estimators under three different γs and the sparse or dense cases are reported in Figures B.5 and B.6, respectively. As shown by Figures B.5 and B.6, we find that as the tuning parameter γ increases, the statistical efficiency of the NA-FGD estimator is improved. According to the classic stochastic gradient descent literature (Chen et al., 2020), we know that an appropriate choice of γ should lie in $(1/2, 1]$. We find that our simulation results match this theoretical

claim. In addition, for all choices of γ, the ANA-FGD estimator can achieve better statistical efficiency than the NA-FGD estimator.

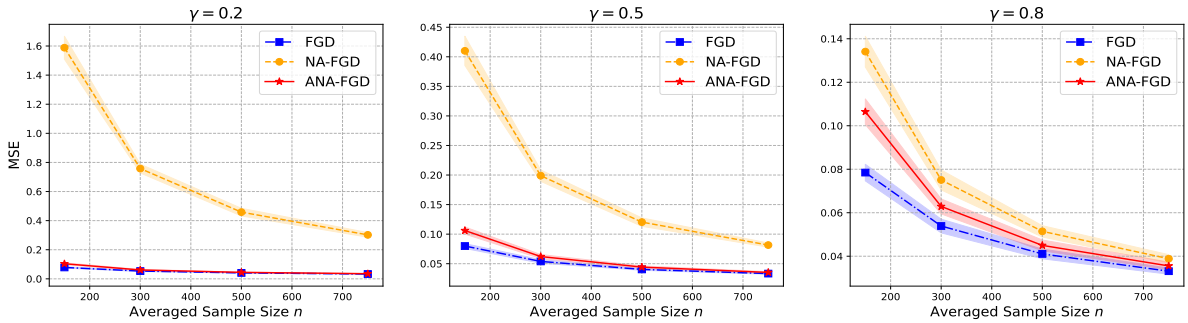

Figure B.5: The MSE results of the three estimators under three different diminishing learning rate settings and the sparse true parameter. The lines represent the MSE values and the shaded areas represent the standard error bounds.

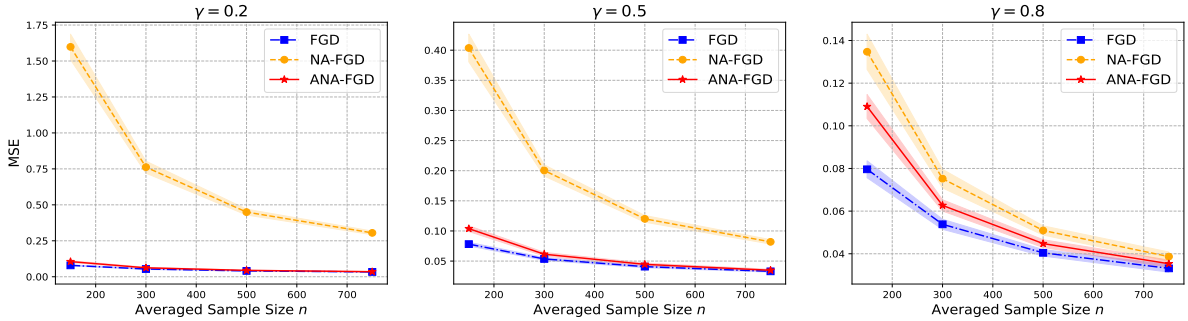

Figure B.6: The MSE results of the three estimators under three different diminishing learning rate settings and the dense true parameter. The lines represent the MSE values and the shaded areas represent the standard error bounds.

Appendix B.4: Heterogeneous Local Sample Sizes

We consider the scenario of heterogeneous local sample sizes, where n_k differs across clients. Under the experimental settings used in Section 3.1, we further assume that n_ks are generated from a normal distribution with a mean n and a standard deviation Δ. Here Δ represents the discrepancy level of the local sample sizes, which varies among $\{\sqrt{n}, 3\sqrt{n}, 5\sqrt{n}\}$. We then calculate the mean squared error (MSE) of the four different estimators. For a reliable evaluation, the experiment is replicated for a total of $B = 100$ times for each parameter combination. The MSE results under the sparse case are reported in Figure B.7. By

Figure B.7, we can find that the statistical efficiency of the three DP estimators (ANA-FGD, Opacus, and NbAFL) becomes worse when the discrepancy level of the local sample sizes increases. However, we can draw a similar conclusion that the performance of the ANA-FGD estimator surpasses that of Opacus and NbAFL estimators even under the heterogeneous scenario.

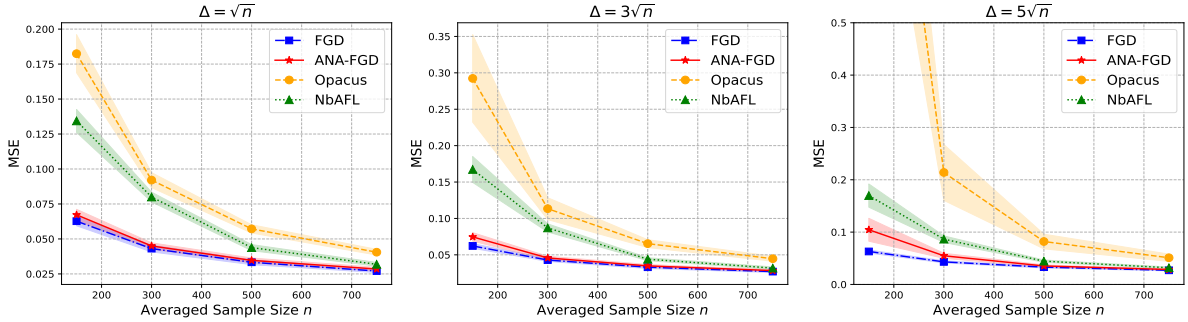

Figure B.7: The MSE results of the four estimators under heterogeneous local sample sizes with three different values of the discrepancy level Δ and the sparse true parameter. The lines represent the MSE values and the shaded areas represent the standard error bounds.

Appendix C: Additional Real Data Analysis

Appendix C.1: The Summary of Variables in NIO Data

See Table C.1 for the summary of variables in the NIO dataset.

Appendix C.2: The Calculation of Noise Levels

Our theoretical analysis provides a quantitative result regarding the trade-off relationship between statistical efficiency and privacy protection, which can further offer practical guidance on setting appropriate noise levels. Specifically, according to our theory, the required noise level σ_k is determined by three adjustable factors. They are, respectively, (i) the privacy budget ε, (ii) the failure probability δ, and (iii) the number of iterations T of the algorithm. In this real application, the privacy budget ε and the failure probability δ are two hyperparameters in differential privacy, which should be decided by the user itself subjectively. ε controls the level of privacy protection: a lower ε corresponds to stronger

Table C.1: The Summary of Variables in the NIO Dataset

Type	Variables	Description
Response	PostCount	The total number of posts published by each user.
Basic Feature	Gender	The gender of each user. Male: 64.40%, Female: 22.51%, Unknown: 13.09%.
	Identity	The identity of each user. Not owner: 57.54%, Owner: 32.77%, Co-owner: 7.13%, Pre-owner: 2.56%.
	RegisterTime	The registration time of each user, from 2017 to 2021.
	Location	The location of each user, like Beijing, Shanghai, Hefei, Guangdong, etc.
	NIOStaff	Whether the user is NIO staff or not. Yes: 3.16%, No: 96.84%.
Social Feature	FollowingCount	The number of followees of each user.
	FollowerCount	The number of follower of each user.
	ActivityCount	The number of joined activities of each user.
	RecommendCount	The number of posts selected and recommended by the community system to the front page.
	PostContent	The latest twenty posts published by each user

privacy. Typically, users prefer a relatively low ε to enhance privacy. For presentation purposes, we consider $\varepsilon \in \{0.8, 2, 5\}$ to represent different levels of privacy protection. As for the failure probability δ, a small value for δ is often preferred to minimize the probability of privacy leakage. Thus in this application, we set $\delta = 10^{-5}$ to ensure a very low probability of failure. In terms of the number of iterations T, according to our Theorem 3, it should satisfy the theorem conditions: (a) $\alpha T/\sqrt{N} \to \infty$ and (b) $\alpha/n \to 0$. For the given averaged local sample size n, one can first choose a sufficiently small learning rate α such that condition (b) is satisfied. Then practically one can set for example $T = cN^{3/4}/\alpha$ for some positive constant c to meet condition (a). Here we consider $\alpha = 0.01$ and $c = 0.25$ for illustration purpose. Then the iteration number T can be computed accordingly. Once ε, δ and T are specified, the noise level can be determined as $\sigma_k = c_{MA} C\sqrt{T\log(1/\delta)}/(\varepsilon n_k)$ according to Theorem 3. Here n_k is the local sample size of client k, c_{MA} is a positive constant, which is calculated as 0.473 according to Abadi et al. (2016), and C is the gradient clipping norm, which is set to be 1.