PROTECTING PRIVACY IN SIGNAL PROCESSING: ALGORITHMS AND COMPARISON FOR MEDIAN REGRESSION

E Chen[1] *Yang Cao*[2] *Yu Tang*[3] *Ying Miao*[4]

[1] Zhejiang Lab
[2] Hokkaido University
[3] Soochow University
[4] Tsukuba University

ABSTRACT

In signal processing, privacy protection plays a critically important role due to the handling of sensitive data from various sensors and devices, such as sound, images, and biometric features. Moreover, regression models are crucial in signal processing as they serve multiple purposes, including signal recovery, denoising, parameter estimation, prediction, and interpolation. By accurately fitting suitable mathematical models, we can enhance signal quality, extract valuable information, and gain profound insights into the behavior and characteristics of the signal. To address the impact of outliers and extreme values on estimation results, median regression has emerged as a reliable approach. In this context, we introduce three algorithms that satisfy differential privacy for median regression: DP-SGD, DP-Smooth and DP-ItSq. Among these algorithms, DP-Smooth achieves the highest accuracy, DP-ItSq demonstrates the fastest computation speed, and DP-SGD is employed as a control/reference for comparison.

***Index Terms*—** Differential Privacy, Median Regression, Signal Processing

1. INTRODUCTION

With the widespread availability of databases, there is a growing concern that personal privacy information may be exposed. As a result, there is an increasing demand for statistical analysis of these databases to prioritize individual privacy protection. As [1] described, differential privacy addresses the paradox of learning nothing about an individual while obtaining useful information about a population. Over the past few years, differential privacy has been investigated in machine learning [2] and has been applied in the real world [3]. Meanwhile, it is important to note that signal data often contains personal identities or sensitive personal information.

In the field of signal processing, signals are often contaminated by varying degrees of noise due to various factors such as environmental interference, noise during signal transmission and acquisition processes [4]. When the distribution of noise is unknown or when there are too many outliers, it can lead to signal distortion, and impact information extraction and accuracy. Extensive efforts have been dedicated to addressing this issue and enhancing recognition accuracy in noisy environments such as logistic regression [5], linear regression [6, 7, 8], kernel ridge regression [9].

However, in signal processing, there is often a significant amount of noise present, which necessitates the use of more robust fitting methods. Hence, we consider median regression [10] and minimize the objective function $L(\beta_0, \boldsymbol{\beta})$, which combines the sum of absolute residuals $F(\beta_0, \boldsymbol{\beta})$ with a ridge penalty term:

$$L(\beta_0, \boldsymbol{\beta}) = F(\beta_0, \boldsymbol{\beta}) + \frac{\lambda}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}. \tag{1}$$

The ridge penalty term is determined by the regularization parameter λ and the squared magnitude of the coefficients β.

In the field of privacy-preserving methods for empirical risk minimization, a general framework proposed in [11] requires the objective function to be doubly differentiable and strictly convex . However, this approach is not suitable for median regression. Hence, specialized privacy-preserving algorithms need to be designed for median regression.

2. BACKGROUND AND DEFINITIONS

Our objective is to generate a classifier β_0 and $\boldsymbol{\beta}$ while preserving the privacy of individual entries in the database $\mathbf{X}$ using differentially private algorithms.

In differential privacy, the most fundamental concept is that of neighboring datasets.

Definition 2.1. *Two datasets* $\mathbf{S}'$ *and* $\mathbf{S}$ *are neighboring if* $\mathbf{S}'$ *can be obtained from* $\mathbf{S}$ *by replacing one record. We denote this relationship as* $\mathbf{S} \triangledown \mathbf{S}'$.

Furthermore, a powerful statistical tool called Gaussian Differential Privacy (GDP) has been proposed [12] to analyze the privacy of algorithms, based on hypothesis testing [13]. Let H_0 represent the underlying dataset as $\mathbf{S}$, and H_1 represent the underlying dataset as $\mathbf{S}'$. By denoting U and V as the probabilities of H_0 and H_1 respectively, privacy protection can be defined using the Type I error (α) and Type II error (β) associated with a rejection rule ϕ.

Definition 2.2. *For any two probability distributions U and V on the same space Ω, the trade-off function $T(U, V) : [0, 1] \to [0, 1]$ is defined as*

$$T(U, V)(\alpha) = \inf\{\beta_\phi : \alpha_\phi \leq \alpha\}.$$

Definition 2.3 (μ-GDP)**.** *A mechanism M is said to satisfy μ-Gaussian Differential Privacy (μ-GDP) if it is G_μ-DP. That is, $T(M(\mathbf{S}), M(\mathbf{S}')) \geq G_\mu$ for all neighboring datasets $\mathbf{S}$ and $\mathbf{S}'$. Here $G_\mu = T(N(0, 1), N(\mu, 1))$ for $\mu \geq 0$. An explicit expression for G_μ reads $G_\mu(\alpha) = \Phi(\Phi^{-1}(1 - \alpha) - \mu)$, where $\Phi(\cdot)$ denotes the standard normal CDF.*

GDP precisely characterizes the Gaussian mechanism and can be fully described by the single mean parameter of a unit-variance Gaussian distribution, which facilitates the description and interpretation of privacy guarantees. The trade-off function is decreasing in μ, thus we can use the parameter "μ" to represent the privacy protection level of the following algorithms.

Next, we will introduce a randomized algorithm called the Gaussian mechanism, which is an effective method for privacy preservation. To begin with, we need to introduce the concept of l_1 sensitivity.

Definition 2.4 (l_1 sensitivity)**.** *The l_1 sensitivity of a function f that outputs a vector in $\mathbb{R}^k$ is:*

$$\Delta f = \max\nolimits_{\mathbf{S} \triangledown \mathbf{S}'} \|f(\mathbf{S}) - f(\mathbf{S}')\|_1.$$

The l_1 sensitivity of a function f captures the magnitude by which a single individual's data can change the function f in the worst case.

Definition 2.5. *Given any function f that outputs a vector in $\mathbb{R}^k$, the Gaussian mechanism is defined as:*

$$M(\mathbf{X}, f(\cdot), \mu) = f(\mathbf{S}) + \boldsymbol{Z},$$

where $\boldsymbol{Z}$ is drawn from the normal distribution $N(\mathbf{0}, (\frac{\Delta f}{\mu})^2 \mathbf{I}_k)$. The density function of the standard k dimensional normal distribution (centered at 0 with covariance matrix $\mathbf{I_k}$) is: $\varphi(\boldsymbol{x}) = (\frac{1}{\sqrt{2\pi}})^k e^{-\frac{\boldsymbol{x}'\boldsymbol{x}}{2}}$.

The Gaussian mechanism adds noise to the true query function f and produces a noisy output. Indeed, GDP provides a privacy bound that is lossless, meaning that it preserves privacy without any degradation or loss of privacy guarantees. We present the following two necessary lemmas, and the detailed proofs can be found in the article [12].

Lemma 2.1. *The Gaussian mechanism is μ-GDP.*

Lemma 2.2 (μ-GDP composition)**.** *The k-fold composition (applying the same algorithm to a dataset k times) of μ_0-GDP mechanisms is $\sqrt{k}\mu_0$-GDP.*

3. ALGORITHMS

In this section, we put forward three privacy preserving algorithms, so called DP-SGD, DP-Smooth and DP-ItSq, for median regression and calculate their privacy parameters respectively.

3.1. Algorithm 1: DP-SGD

Inspired by [11], we apply a similar approach to minimize the non-differentiable objective function $L(\beta_0, \boldsymbol{\beta})$, utilizing its directional derivatives. As an illustration, if $\boldsymbol{e}_k$ represents the coordinate direction of β_k variation, it gives rise to two directional derivatives: $d_{\boldsymbol{e}_k^+} L(\beta_0, \boldsymbol{\beta})$ and $d_{\boldsymbol{e}_k^-} L(\beta_0, \boldsymbol{\beta})$.

We present a theorem regarding privacy, while demonstrating conclusions about utility through simulation results. Following the approach in [14], we utilize a greedy coordinate descent algorithm to update the direction of parameter β_k. This update is based on the equation $Dir^{(k)} = \min\{d_{\boldsymbol{e}_k^+} L(\beta_0, \boldsymbol{\beta}), d_{\boldsymbol{e}_k^-} L(\beta_0, \boldsymbol{\beta})\}$. We terminate the update of β_k if both coordinate directional derivatives are nonnegative.

Theorem 3.1. *Given a set of n samples $\mathbf{X}_1, \ldots, \mathbf{X}_n$ over $\mathbb{R}^d$ with labels $Y_1, \ldots, Y_n$, where for each i ($1 \leq i \leq n$), $\|\mathbf{X}_i\|_1 \leq 1$ and $|Y_i| \leq B$, the output of DP-SGD preserves μ-GDP.*

Algorithm 1 : DP-SGD

```
Input: Privacy parameters $\mu$, design matrix $\mathbf{X}$, response vector $\mathbf{Y}$,
    regularization parameter $\lambda$, learning rate $\eta$, privacy parameter
    within one iteration $\mu_0$ and the number of iterations $N_0$
Output: $\hat{\boldsymbol{\omega}} \in \mathbb{R}^{d+1}$.
Initialize the algorithm with a vector $(\hat{\beta}_0(0), \hat{\beta}^{\mathrm{T}}(0))^{\mathrm{T}}$
for t = $0, 1, \cdots, N_0 - 1$ do
    Subsampling: take a uniformly random subsample $I_t \subseteq [n]$
    with batch size $m$.
    for $i \in I_t$ do
        Compute gradient:
        $\boldsymbol{g}^i(t) = Dir^i(t+1)$
        Clip to norm 1:
        $\tilde{\boldsymbol{g}}^i(t) = \boldsymbol{g}^i(t) / \max(1, \|\boldsymbol{g}^i(t)\|_2)$
    end for
    $\hat{\boldsymbol{\beta}}(t + 1) = \hat{\boldsymbol{\beta}}(t) - \eta(\frac{1}{m}\sum_{i \in [n]} \tilde{\boldsymbol{g}}^i(t) + \boldsymbol{U}(t))$,
    $U(t) \sim N(\mathbf{0}, \frac{4\mu_0^2}{m^2}\boldsymbol{I}_d)$. $\mu_0$ is the value that satisfies
    $\mu = \sqrt{2}c\sqrt{e^{\mu_0^2}\Phi(1.5\mu_0) + 3\Phi(-0.5\mu_0) - 2}$ and $c =$
    $m\sqrt{N_0}/n$.
    $\hat{\beta}_0(t+1) = \frac{1}{m}\sum_{i=1}^m (Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}(t+1))$.
end for
return $\hat{\boldsymbol{\omega}} = (\hat{\beta}_0(N_0), \hat{\boldsymbol{\beta}}(N_0)^T)^T$
```

3.2. Algorithm 2: DP-Smooth

The finite smoothing method [15] is an important tool to solve non-differentiable problem. In addition, the solution of smooth function can estimate the solution of the original function well. This idea is applied in DP-Smooth by an analogous technique. Let γ be a non-negative parameter which indicates the degree of approximation. Define

$$\rho_\gamma(x) = \begin{cases} x^2/(2\gamma), & \text{if } |x| \le \gamma, \\ |x| - \frac{1}{2}\gamma, & \text{if } |x| > \gamma. \end{cases} \tag{2}$$

Then the nondifferentiable function $L(\beta_0, \boldsymbol{\beta})$ is approximated by $L_\gamma(\beta_0, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \rho_\gamma(r_i(\beta_0, \boldsymbol{\beta})) + \frac{\lambda}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}$, where $r_i(\beta_0, \boldsymbol{\beta})$ is the residual of i-th observation.

Algorithm 2 : DP-Smooth

```
Input: Privacy parameter $\mu$, design matrix $\mathbf{X}$, response vector $\mathbf{Y}$, regularization parameter $\lambda$ and approximation parameter $\gamma$
Output: $\boldsymbol{\omega}^* \in \mathbb{R}^{d+1}$
Generate a random vector $\boldsymbol{b}$ from $d+1$ dimensional normal distribution with mean $\mathbf{0}$ and covariance $\frac{16}{\mu^2}\mathbf{I}_{d+1}$
Compute $(\beta_0^*, \boldsymbol{\beta}^{*\mathrm{T}})^{\mathrm{T}} = \mathbf{argmin}_{\beta_0,\boldsymbol{\beta}} L_\gamma(\beta_0, \boldsymbol{\beta}) + \frac{\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\omega}}{n} + \frac{\beta_0^2}{\sqrt{n}}$, where $\boldsymbol{\omega} = (\beta_0, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ is a $d+1$ dimensional column vector.
return $\boldsymbol{\omega}^* = (\beta_0^*, \boldsymbol{\beta}^{*\mathrm{T}})^{\mathrm{T}}$
```

This algorithm bears a striking resemblance to the smoothing median regression convex program [16], resulting in similar running times as smoothing regression. In fact, ω^* can be obtained by using the interior point method. By following a similar proof as presented in [17], we can demonstrate that DP-Smooth preserves privacy and holds utility.

Theorem 3.2. *Given a set of n samples $\mathbf{X}_1, \ldots, \mathbf{X}_n$ over $\mathbb{R}^d$, with labels $Y_1, \ldots, Y_n$, where for each i, $\|\mathbf{X}_i\|_1 \le 1$ and $|Y_i| \le B$. If for each i, $r_i(\mu^*, \beta^*) \ne \gamma$, then the output of DP-Smooth would preserve μ-GDP.*

Theorem 3.3. *Given an l_1 regression problem with regularization parameter λ, let $\boldsymbol{\omega}_1$ be the classifier that minimizes $L_\gamma(\beta_0, \boldsymbol{\beta}) + \frac{\beta_0^2}{\sqrt{n}}$, and $\boldsymbol{\omega}_2$ be the classifier output by DP-smooth respectively. Then, with probability $1 - \alpha$, $\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|_1 \le \frac{16(d+1)t_{1-\alpha}(d+1)}{n\mu^2 \min(\lambda, \frac{2}{\sqrt{n}})}$. Here, $t_{1-\alpha}(d+1)$ represents the $1-\alpha$ quantile of the t-distribution with $d+1$ degrees of freedom.*

Note that when the sample size n is large enough, $\boldsymbol{\omega}_2$ approximates $\boldsymbol{\omega}_1$ closely. Furthermore, $\boldsymbol{\omega}_1$ is also close to the true parameter in $\text{argmin}_{\boldsymbol{\omega}}\ L_\gamma(\boldsymbol{\omega})$. However, it should be noted that this is simply a theoretical lower bound on accuracy. In practical simulations, the value of lambda can be set to 0.

3.3. Algorithm 3: DP-ItSq

The DP-ItSq technique, which stands for **d**ifferentially **p**rivate **it**erative least **sq**uares regression, merges the methodologies of least absolute deviations regression and least squares regression. By doing so, it effectively converts a complex problem involving least absolute deviations into a simple weighted least squares regression [18]. For the $(t+1)$-th iteration, we set w_i as $\frac{1}{|r(t)_i|+e}$, with $r(t)_i$ being the residual of the i-th sample at the t-th iteration. This leads to the iterative process expressed as:

$$I(t+1) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{|r(t)_i| + e} r^2(t+1)_i + \frac{\lambda}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}. \tag{3}$$

If $|r(t)_i - r(t+1)_i| \approx 0, i = 1, 2, \ldots, n$, $I(t+1)$ is close to $L(\beta_0, \boldsymbol{\beta})$. In practice, we set e as a small positive value. According to the relationship between constrained optimization and regularization terms, there exists a mutual correspondence between v and λ, where $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} \le v$.

Algorithm 3 : DP-ItSq

```
Input: privacy parameter $\mu$, design matrix $\mathbf{X}$, response vector $\mathbf{Y}$, regularization parameter $\lambda$, tolerance parameter $\tau$ and the number of iteration $N_0$
Output: $\hat{\boldsymbol{\omega}} \in \mathbb{R}^{d+1}$
Partition the dataset into $N_0$ disjoint subsets, numbered from 0 to $N_0 - 1$.
Initialize the algorithm with $\hat{\beta}_0(0)$ and $\hat{\boldsymbol{\beta}}(0)$
$(\hat{\beta}_0(1), \hat{\boldsymbol{\beta}}^{\mathrm{T}}(1))^{\mathrm{T}} = \text{argmin}_{\beta_0,\boldsymbol{\beta}} I(1)$
for $t = 1, \cdots, N_0 - 1$ do
    Estimate the parameters using the $t$-th subset.
    while $|\hat{\beta}_0(t) - \hat{\beta}_0(t-1)| > \tau$ or $\|\hat{\boldsymbol{\beta}}(t) - \hat{\boldsymbol{\beta}}(t-1)\|_1 > \tau$ do
        $(\hat{\beta}_0(t+1), \hat{\boldsymbol{\beta}}^{\mathrm{T}}(t+1))^{\mathrm{T}} = \text{argmin}_{\beta_0,\boldsymbol{\beta}} I(t+1)$
    end while
end for
return $\hat{\boldsymbol{\omega}} := (\hat{\beta}_0(N_0), \hat{\boldsymbol{\beta}}^{\mathrm{T}}(N_0))^{\mathrm{T}} + \boldsymbol{U}$, where $\boldsymbol{U}$ is a $d+1$ dimensional normal random variable with mean $\mathbf{0}$ and covariance $\frac{c^2}{\mu^2}\mathbf{I}_{d+1}$, where $c = \frac{12(\sqrt{dv}+B)}{n\min(\frac{2}{2(\sqrt{dv}+B)+e},\lambda)e}$.
```

Theorem 3.4. *Given a set of n samples $\mathbf{X}_1, \ldots, \mathbf{X}_n$ over $\mathbb{R}^d$, with labels $Y_1, \ldots, Y_n$, where for each i ($1 \le i \le n$), $\|\mathbf{X}_i\|_1 \le 1, |Y_i| \le B$, the output of DP-ItSq preserves μ-GDP.*

Theorem 3.5. *Given an l_1 regression problem with a full-rank covariance matrix $\mathbf{X}$ and regularization parameter λ, let $\boldsymbol{\omega}_1$ be the unique minimizer of $L_e(\beta_0, \boldsymbol{\beta})$ and $\boldsymbol{\omega}_2 = \boldsymbol{\omega}_1 + \boldsymbol{U}$, where $\boldsymbol{U}$ is a $d+1$ dimensional normal random variable with mean $\mathbf{0}$ and covariance $\frac{c^2}{\mu^2}\mathbf{I}_{d+1}$, where c is $\frac{12(\sqrt{dv}+B)}{\min(\frac{2}{2(\sqrt{dv}+B)+e},\lambda)ne}$. Then, with probability $1-\alpha$, $\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|_1 \le \frac{c}{\mu}\sqrt{d+1}t_{1-\alpha}(d+1)$.*

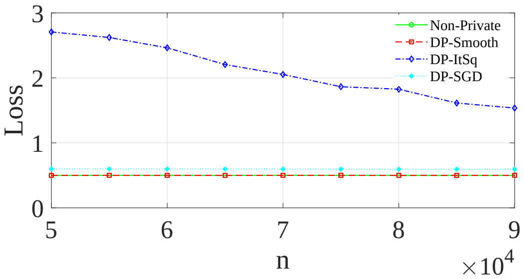

Fig. 1. The utility of three algorithms for various sample size.

4. EXPERIMENTAL RESULTS

Synthetic dataset We constructed the synthetic dataset by sampling $\boldsymbol{x}_i \in \mathbb{R}^3$, for $i \in [n]$, independently from a multivariate normal distribution with mean $\bar{\boldsymbol{x}} = (0.2, 0.6, 0.3)$ and covariance $\Sigma = \mathbf{I}_3$. Each $\boldsymbol{x}_i$ is associated with a corresponding y_i that is generated as $y_i = 0.2 - 3x_{i,1} + 0.5x_{i,2} - x_{i,3} + u_i$, where u_i is sampled from a standard normal distribution $N(0, 0.5)$. In practice, we set the hyperparameters in three algorithms as listed in Table 1. Figure 1 demonstrates that DP-Smooth achieves the highest accuracy across different values of n. The comparison results are not surprising due to the fact that the derivative of DP-SGD only captures sign information, and DP-ItSq performs well for large n.

Table 1. Hyperparameters set of algorithms

Hyperparameters	Value	Description
μ	1	Privacy budget
λ	0.02	Regularized parameter
γ	0.05	Smoothing parameter in Alg 2
e	0.05	Correcting parameter in Alg 3
v	1	Constraint parameter in Alg 3
B	3	Bound of the response variable
τ	10^{-6}	Tolerance parameter in Alg 3
p	0.02	Sample ratio $p = m/n$ in Alg 1
N_0	100, 957	The number of iterations in Alg 1, 3
η	0.01	Learning rate in Alg 1
μ_0	0.5	Privacy budgets per iteration in Alg 1

Carbon nanotubes UCI dataset We obtained the data from a dataset investigating atomic coordinates in carbon nanotubes [19]. Although this dataset does not include sensitive information, we utilize it to assess the performance of differential privacy algorithms on diverse real datasets. Once we removed all points that do not lie within the interval $[0, 1] \times [0, 1]$, the resulting dataset comprises a total of $10, 721$ datapoints. All parameters are set the same as Table 1, except for μ. Figure 2 demonstrates that as the privacy budget increases, DP-ItSq converges rapidly to a vicinity of the optimal solution, while DP-SGD shows no significant change. DP-Smooth exhibits the best fitting performance at the cost of higher computational complexity.

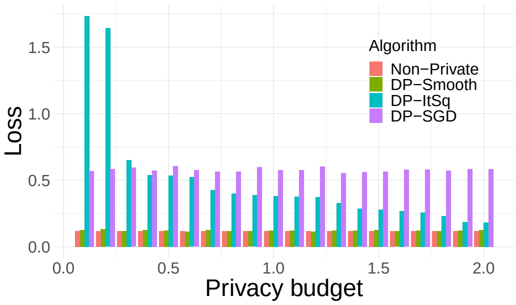

Fig. 2. Performance of three algorithms on carbon nanotubes dataset for various privacy budget μ.

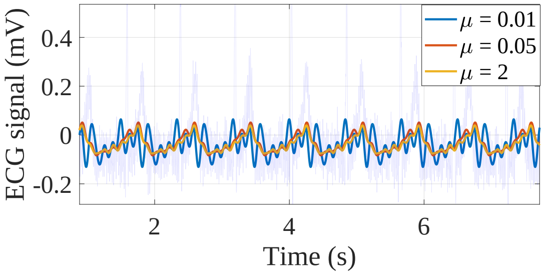

Fig. 3. Comparing fitting accuracy of ECG signal with DP-Smooth under various privacy budgets

Electrocardiogram signal ECG signals, which represent the electrical activity of the heart, are frequently affected by substantial noise and outliers. In this study, we utilize data from the ”ECG-ID” database, a part of the PhysioNet biological signal bank [20], which consists of 310 ECG recordings obtained from 90 individuals. Considering the findings from previous analysis, we employ DP-Smooth for fitting purposes. Given that the ECG signal is a periodic sequence, we employ a 7-th order Fourier series expansion [21] with respect to the independent variable ”Time”. That is, $Y = a_0 + \sum_{i=1}^{7}(a_n cos(i\omega t) + b_n sin(i\omega t))$, where $\omega = \frac{2\pi}{T}$, with period $T = \frac{5}{6}$s. Figure 3 indicates that the smaller the degree of privacy protection, the better the fitting effect. This can be observed from the consistency of peak values between the original sequence and the fitted sequence.

5. CONCLUSION

Privacy protection plays a crucial role in signal processing and median regression is a reliable approach to handle outliers in signal recovery and denoising. In this context, three differential privacy algorithms, namely DP-Smooth, DP-ItSq, and DP-SGD, are introduced as solutions for median regression. Among these algorithms, DP-Smooth achieves the highest accuracy, while DP-ItSq demonstrates the fastest computation speed.

6. REFERENCES

[1] Cynthia Dwork, Aaron Roth, et al., "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[2] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[3] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, 2014, pp. 1054–1067.

[4] Alan T Welford, "Signal, noise, performance, and age," *Human Factors*, vol. 23, no. 1, pp. 97–109, 1981.

[5] Ø Birkenes, Tomoko Matsui, Kunio Tanabe, Sabato Marco Siniscalchi, Tor André Myrvoll, and Magne Hallstein Johnsen, "Penalized logistic regression with hmm log-likelihood regressors for speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1440–1454, 2010.

[6] Jen-Tzung Chien, "Linear regression based bayesian predictive classification for speech recognition," *IEEE transactions on speech and audio processing*, vol. 11, no. 1, pp. 70–79, 2003.

[7] Xiaodong Cui and Abeer Alwan, "Noise robust speech recognition using feature compensation based on polynomial regression of utterance snr," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1161–1172, 2005.

[8] T Tony Cai, Yichen Wang, and Linjun Zhang, "The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy," *The Annals of Statistics*, vol. 49, no. 5, pp. 2825–2850, 2021.

[9] Jie Chen, Lingfei Wu, Kartik Audhkhasi, Brian Kingsbury, and Bhuvana Ramabhadrari, "Efficient one-vs-one kernel ridge regression for speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2454–2458.

[10] Keming Yu, Zudi Lu, and Julian Stander, "Quantile regression: applications and current research areas," *Journal of the Royal Statistical Society Series D: The Statistician*, vol. 52, no. 3, pp. 331–350, 2003.

[11] Di Wang, Minwei Ye, and Jinhui Xu, "Differentially private empirical risk minimization revisited: Faster and more general," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[12] Jinshuo Dong, Aaron Roth, and Weijie J Su, "Gaussian differential privacy," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 84, no. 1, pp. 3–37, 2022.

[13] Peter Kairouz, Sewoong Oh, and Pramod Viswanath, "The composition theorem for differential privacy," in *International conference on machine learning*. PMLR, 2015, pp. 1376–1385.

[14] Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani, "Pathwise coordinate optimization," *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 302 – 332, 2007.

[15] Gianni Di Pillo, Luigi Grippo, and Stefano Lucidi, "A smooth method for the finite minimax problem," *Mathematical Programming*, vol. 60, no. 1-3, pp. 187–214, 1993.

[16] Marcelo Fernandes, Emmanuel Guerre, and Eduardo Horta, "Smoothing quantile regressions," *Journal of Business & Economic Statistics*, vol. 39, no. 1, pp. 338–357, 2021.

[17] Kamalika Chaudhuri and Claire Monteleoni, "Privacy-preserving logistic regression," *Advances in neural information processing systems*, vol. 21, 2008.

[18] EJ Schlossmacher, "An iterative technique for absolute deviations curve fitting," *Journal of the American Statistical Association*, vol. 68, no. 344, pp. 857–859, 1973.

[19] Mehmet Acı and Mutlu Avcı, "Artificial neural network approach for atomic coordinate prediction of carbon nanotubes," *Applied Physics A*, vol. 122, pp. 1–14, 2016.

[20] Tatiana S Lugovaya, "Biometric human identification based on electrocardiogram," *Master's thesis, Faculty of Computing Technologies and Informatics, Electrotechnical University 'LETI', Saint-Petersburg, Russian Federation*, 2005.

[21] Gerald B Folland, *Fourier analysis and its applications*, vol. 4, American Mathematical Soc., 2009.

[22] David R Hunter and Kenneth Lange, "A tutorial on mm algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

7. APPENDIX

7.1. Details of DP-SGD

The computation of directional gradients is crucial in this context. Specifically, the directional derivative of β_k can be computed as follows.

$$\begin{aligned} d_{\boldsymbol{e}_k^+} L(\beta_0, \boldsymbol{\beta}) &= lim_{\tau \to 0^+} \frac{L(\beta_0, \boldsymbol{\beta} + \tau \boldsymbol{e}_k) - L(\beta_0, \boldsymbol{\beta})}{\tau} \\ &= d_{\boldsymbol{e}_k^+} F(\beta_0, \boldsymbol{\beta}) + \lambda \beta_k, \end{aligned}$$

and

$$\begin{aligned} d_{\boldsymbol{e}_k^-} L(\beta_0, \boldsymbol{\beta}) &= lim_{\tau \to 0^-} \frac{L(\beta_0, \boldsymbol{\beta} + \tau \boldsymbol{e}_k) - L(\beta_0, \boldsymbol{\beta})}{\tau} \\ &= d_{\boldsymbol{e}_k^-} F(\beta_0, \boldsymbol{\beta}) + \lambda \beta_k. \end{aligned}$$

In l_1 regression, the coordinate direction derivatives are

$$d_{\boldsymbol{e}_k^+} F(\beta_0, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \begin{array}{rl} -x_{ik}, & r_i(\beta_0, \boldsymbol{\beta}) < 0, \\ x_{ik}, & r_i(\beta_0, \boldsymbol{\beta}) > 0, \\ |x_{ik}|, & r_i(\beta_0, \boldsymbol{\beta}) = 0, \end{array} \right. \tag{4}$$

and

$$d_{\boldsymbol{e}_k^-} F(\beta_0, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \begin{array}{rl} -x_{ik}, & r_i(\beta_0, \boldsymbol{\beta}) < 0, \\ x_{ik}, & r_i(\beta_0, \boldsymbol{\beta}) > 0, \\ -|x_{ik}|, & r_i(\beta_0, \boldsymbol{\beta}) = 0. \end{array} \right. \tag{5}$$

Furthermore, the sub-sampling property of GDP is provided by Lemma 7.1, and a detailed proof of this lemma can be found in [12].

Lemma 7.1. *If M is f-DP on X^m, then the subsampled mechanism $M(Sample_m)$ is $C_p(f)$-DP on X^n, where the sampling ratio $p = m/n$. Especially, if $m\sqrt{N_0}/n = c$ and $f = G_{\mu_0}$, then $C_p(G_{\mu_0})^{\otimes N_0}$ is asymptotically μ-GDP with*

$$\mu = \sqrt{2}c\sqrt{e^{\mu_0^2}\Phi(1.5\mu_0) + 3\Phi(-0.5\mu_0) - 2}.$$

Proof of Theorem 3.1

Proof. For $(x, y) \in (\mathbf{X}(t), \mathbf{Y}(t))$, it suffices to prove the privacy guarantee for the t-th iteration of the algorithm and use μ-GDP composition to obtain full privacy bound. At the t-th iteration, the algorithm first updates the non-sparse estimate of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}(t+1) = \hat{\boldsymbol{\beta}}(t) - \eta \left(\frac{1}{m} \sum_{i \in [n]} \tilde{\boldsymbol{g}}^i(t) + \boldsymbol{U}(t) \right),$$

where $\boldsymbol{U}(t) \sim N\left(\mathbf{0}, \frac{4\sigma^2}{m^2}\boldsymbol{I}_d\right)$. For convenience, assume the first $m-1$ records are the same, then the sensitivity of $\frac{1}{m}\sum_{i\in[m]} \tilde{\boldsymbol{g}}^i(t) \le 2/m$.

Combined with Lemma 2.1, $\hat{\beta}(t+1)$ satisfies μ_0-GDP. In addition, since $\hat{\mu} = \frac{1}{n_0}\sum_{i=1}^{n_0}(Y_i - \mathbf{X}_i\hat{\beta})$, it is differentially private by post-processing. By using Lemma 7.1, DP-SGD is asymptotically μ-GDP, where

$$\mu = \sqrt{2}c\sqrt{e^{\mu_0^2}\Phi(1.5\mu_0) + 3\Phi(-0.5\mu_0) - 2}.$$

□

7.2. Details of DP-Smooth

Since the absolute value function is not differentiable at the cuspidal point, a smooth method for minimizing function (1) is considered. Denote $F_\gamma(\beta_0, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \rho_\gamma(r_i(\beta_0, \boldsymbol{\beta}))$, and $L_\gamma(\beta_0, \boldsymbol{\beta}) = F_\gamma(\beta_0, \boldsymbol{\beta}) + \frac{\lambda}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}$. The sign vector

$$\mathbf{S}_\gamma(\beta_0, \boldsymbol{\beta}) = (s_1(\beta_0, \boldsymbol{\beta}), \ldots, s_n(\beta_0, \boldsymbol{\beta}))^{\mathrm{T}}$$

is given by

$$s_i(\beta_0, \boldsymbol{\beta}) = \left\{ \begin{array}{rl} -1, & \text{if} \quad r_i(\beta_0, \boldsymbol{\beta}) < -\gamma, \\ 0, & \text{if} \quad -\gamma \le r_i(\beta_0, \boldsymbol{\beta}) \le \gamma, \\ 1, & \text{if} \quad r_i(\beta_0, \boldsymbol{\beta}) > \gamma. \end{array} \right. \tag{6}$$

Let $w_i(\beta_0, \boldsymbol{\beta}) = 1 - s_i^2(\beta_0, \boldsymbol{\beta})$, then

$$\begin{aligned} \rho_\gamma(r_i(\beta_0, \boldsymbol{\beta})) &= \frac{1}{2\gamma} w_i(\beta_0, \boldsymbol{\beta}) r_i^2(\beta_0, \boldsymbol{\beta}) \\ &\quad + s_i(\beta_0, \boldsymbol{\beta}) \left[r_i(\beta_0, \boldsymbol{\beta}) - \frac{1}{2}\gamma s_i(\beta_0, \boldsymbol{\beta}) \right]. \end{aligned}$$

Denote $\mathbf{W}_\gamma(\beta_0, \boldsymbol{\beta})$ as the diagonal $n \times n$ matrix whose diagonal elements are $w_i(\beta_0, \boldsymbol{\beta})$. So $\mathbf{W}_\gamma(\beta_0, \boldsymbol{\beta})$ has value 1 in those diagonal elements related to small residuals and 0 elsewhere. For $\beta_0 \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^d$, the derivation of $F_\gamma(\beta_0, \boldsymbol{\beta})$ is

$$\frac{\partial F_\gamma(\beta_0, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{n}\mathbf{X}^{\mathrm{T}} \left[\frac{1}{\gamma}\mathbf{W}_\gamma(\beta_0, \boldsymbol{\beta}) r(\beta_0, \boldsymbol{\beta}) + \mathbf{S}_\gamma(\beta_0, \boldsymbol{\beta}) \right],$$

and

$$\frac{\partial F_\gamma(\beta_0, \boldsymbol{\beta})}{\partial \beta_0} = \frac{1}{n}\mathbf{1}^{\mathrm{T}} \left[\frac{1}{\gamma}\mathbf{W}_\gamma(\beta_0, \boldsymbol{\beta}) r(\beta_0, \boldsymbol{\beta}) + \mathbf{S}_\gamma(\beta_0, \boldsymbol{\beta}) \right].$$

Notice that $L_\gamma(\beta_0, \boldsymbol{\beta})$ is strictly convex and differentiable. By intuition, a minimizer of $L(\beta_0, \boldsymbol{\beta})$ is close to a minimizer of $L_\gamma(\beta_0, \boldsymbol{\beta})$ when γ is close to zero. Furthermore, the l_1 solution can be detected when $\gamma > 0$ is small enough [15]. That is, it is not necessary to let γ converge to zero in order to find a minimizer of $L_\gamma(\beta_0, \boldsymbol{\beta})$. This observation is essential for the efficiency and the numerical stability of DP-Smooth.

Proof of Theorem 3.2

Proof. Let $\boldsymbol{a}_1$ and $\mathbf{a}_2$ be two row vectors over $\mathbb{R}^d$ with l_1 norm at most 1 and $y_1, y_2 \in [-B, B]$. Consider the two inputs

D_1 and D_2 where D_2 is obtained from D_1 by replacing one record $(\boldsymbol{a}_1, y_1)$ into $(\boldsymbol{a}_2, y_2)$. For convenience, assume the first $n-1$ records are the same. For any output $\boldsymbol{\omega}^*$ by Algorithm 1, there is a unique value of $\boldsymbol{b}$ that maps the input to the output. This uniqueness holds, because both the regularization function and the loss function are differentiable everywhere, and the objective function is strictly convex. Denote $\tilde{\boldsymbol{a}}_1$ as $(1, \boldsymbol{a}_1)$ and $\tilde{\boldsymbol{a}}_2$ as $(1, \boldsymbol{a}_2)$, w_n and w'_n are corresponding coefficients. In addition, let $\mathbf{X}$ and $\mathbf{X}'$ represent samples about independent variables of D_1 and D_2, and $\mathbf{Y}$ and $\mathbf{Y}'$ represent their corresponding response vectors, respectively. Denote $\tilde{\mathbf{X}} = (\mathbf{1}, \mathbf{X})$ and $\tilde{\mathbf{X}}' = (\mathbf{1}, \mathbf{X}')$, where $\mathbf{1}$ is an n dimensional column vector whose elements are 1. Let the values of $d+1$ dimensional vector $\boldsymbol{b}$ for D_1 and D_2 respectively, be $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$. Since $\boldsymbol{\omega}^*$ is the value that minimizes both the optimization problems, the derivative of both optimization functions at $\boldsymbol{\omega}^*$ is 0. This implies that for every $\boldsymbol{b}_1$ in the first case, there exists a $\boldsymbol{b}_2$ in the second case such that:

$$\begin{aligned} & \boldsymbol{b}_1 + \tilde{\mathbf{X}}^{\mathrm{T}}(1/\gamma \mathbf{W}_\gamma(\mu^*, \boldsymbol{\beta}^*)(\tilde{\mathbf{X}}\boldsymbol{\omega}^* - \mathbf{Y}) + \mathbf{S}_\gamma(\mu^*, \boldsymbol{\beta}^*)) \\ = \quad & \boldsymbol{b}_2 + \tilde{\mathbf{X}'}^{\mathrm{T}}(1/\gamma \mathbf{W}'_\gamma(\mu^*, \boldsymbol{\beta}^*)(\tilde{\mathbf{X}}'\boldsymbol{\omega}^* - \mathbf{Y}') + \mathbf{S}'_\gamma(\mu^*, \boldsymbol{\beta}^*)), \end{aligned}$$

which indicates that:

$$\begin{aligned} & \boldsymbol{b}_1 + \tilde{\boldsymbol{a}}_1^{\mathrm{T}}(1/\gamma w_n(\mu^*, \boldsymbol{\beta}^*)(\mu^* + \boldsymbol{a}_1\boldsymbol{\beta}^* - y_1) + s_n(\mu^*, \boldsymbol{\beta}^*)) \\ = \quad & \boldsymbol{b}_2 + \tilde{\boldsymbol{a}}_2^{\mathrm{T}}(1/\gamma w'_n(\mu^*, \boldsymbol{\beta}^*)(\mu^* + \boldsymbol{a}_2\boldsymbol{\beta}^* - y_2) + s'_n(\mu^*, \boldsymbol{\beta}^*)). \end{aligned}$$

According to the definitions of $\mathbf{W}_\gamma(\mu^*, \boldsymbol{\beta}^*)$ and $\mathbf{S}_\gamma(\mu^*, \boldsymbol{\beta}^*)$, it is clear that

$$-1 \leq 1/\gamma w_n(\mu^*, \boldsymbol{\beta}^*) * (\mu^* + \boldsymbol{a}_1\boldsymbol{\beta}^* - y_1) + s_n(\mu^*, \boldsymbol{\beta}^*) \leq 1$$

and

$$-1 \leq 1/\gamma w'_n(\mu^*, \boldsymbol{\beta}^*) * (\mu^* + \boldsymbol{a}_2\boldsymbol{\beta}^* - y_2) + s'_n(\mu^*, \boldsymbol{\beta}^*) \leq 1.$$

Since $\|\tilde{\boldsymbol{a}}_1\|_1 \leq 2$ and $\|\tilde{\boldsymbol{a}}_2\|_1 \leq 2$, we have $\|\boldsymbol{b}_1 - \boldsymbol{b}_2\|_1 \leq 4$ and $-4 \leq \|\boldsymbol{b}_1\|_1 - \|\boldsymbol{b}_2\|_1 \leq 4$, which implies that DP-smooth approximately satisfies μ-GDP. □

In theory, for each i, the probability of $\boldsymbol{\omega}^*$ satisfies $|r_i(\mu^*, \boldsymbol{\beta}^*)| \neq \gamma$ is 1.

Lemma 7.2. *[17] Let $G(\boldsymbol{\omega})$ and $g(\boldsymbol{\omega})$ be two convex functions, which are continuous and differentiable at all points. If $\boldsymbol{\omega}_1 = argmin_{\boldsymbol{\omega}} G(\boldsymbol{\omega})$ and $\boldsymbol{\omega}_2 = argmin_{\boldsymbol{\omega}}\ G(\boldsymbol{\omega}) + g(\boldsymbol{\omega})$, then $\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|_1 \leq \frac{g_1}{G_2}$. Here, $g_1 = \max_{\boldsymbol{\omega}} \|\nabla g(\boldsymbol{\omega})\|_1$ and $G_2 = \min_{\boldsymbol{v}} \min_{\boldsymbol{\omega}} \boldsymbol{v}^{\mathrm{T}} \nabla^2 G(\boldsymbol{\omega})\boldsymbol{v}$, for any unit vector $\boldsymbol{v}$.*

The main idea of the proof is to examine the gradient and the Hessian of the functions G and g around $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$. For fixed $\boldsymbol{b}$, let $\boldsymbol{\omega} = (\beta_0, \boldsymbol{\beta})$, $G(\boldsymbol{\omega}) = L_\gamma(\beta_0, \boldsymbol{\beta})$ and $g(\boldsymbol{\omega}) = \frac{\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\omega}}{n}$ + $\frac{\beta_0^2}{\sqrt{n}}$, we can obtain accuracy of DP-Smooth. Notice that $\boldsymbol{b}$ is a random variable, it is necessary to provide its property of convergence.

Lemma 7.3. *If $\boldsymbol{b}$ is a random variable drawn from $d+1$ dimensional normal distribution with mean $\mathbf{0}$ and covariance $\mathbf{I}_{d+1}$, then $\|\boldsymbol{b}\|_2$ obeys t distribution with parameter $d+1$.*

Proof. Since $\mathbf{b} = (b_1, \cdots, b_{d+1})' \sim N(\mathbf{0}, \mathbf{I}_{d+1})$, then for $1 \leq i < j \leq d+1$, b_i and b_j are independent and identically distributed as $N(0,1)$. Hence, $\|\mathbf{b}\|_2^2$ is the sum of $d+1$ independent standard normal variables, which follows a $\chi^2(d+1)$ distribution. By the definition of t distribution, the result is obtained. □

Proof of Theorem 3.3

Proof. According to Lemma 7.2, we take $G(\boldsymbol{\omega}) = L_\gamma(\beta_0, \boldsymbol{\beta}) + \frac{\beta_0^2}{\sqrt{n}}$ and $g(\boldsymbol{\omega}) = \frac{\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\omega}}{n}$. Because $F_\gamma(\beta_0, \boldsymbol{\beta})$ is a convex function, if we define the second derivative of $F_\gamma(\beta_0, \boldsymbol{\beta})$ to be 0 at nondifferentiable points, then the Hessian matrix of $F_\gamma(\beta_0, \boldsymbol{\beta})$ is positive semi-definite. Notice that

$$\nabla^2\left(\frac{\beta_0^2}{\sqrt{n}}\right) = \begin{bmatrix} \frac{2}{\sqrt{n}} & \mathbf{0}_{1\times d} \\ \mathbf{0}_{d\times 1} & \mathbf{0}_{d\times d} \end{bmatrix}_{(d+1)\times(d+1)},$$

and

$$\nabla^2\left(\frac{\lambda}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\right) = \begin{bmatrix} 0 & 0 \\ 0 & \lambda \mathbf{I}_{d\times d} \end{bmatrix}_{(d+1)\times(d+1)}.$$

Hence, for any unit vector $\boldsymbol{v}$, $G_2 \geq \min(\lambda, \frac{2}{\sqrt{n}})$ and $g_1 = \frac{\|\boldsymbol{b}\|_1}{n}$, $\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|_1 \leq \frac{\|\boldsymbol{b}\|_1}{n \min(\lambda, \frac{2}{\sqrt{n}})} \leq \frac{(d+1)\|\boldsymbol{b}\|_2}{n \min(\lambda, \frac{2}{\sqrt{n}})}$. then the theorem is obtained. □

7.3. Details of DP-ItSq

For $e > 0$, define a perturbation of $L(\beta_0, \boldsymbol{\beta})$ as

$$L_e(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} |r_i(\beta_0, \boldsymbol{\beta})| - \frac{e}{2}\log(e + |r_i(\beta_0, \boldsymbol{\beta})|) + \frac{\lambda}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}.$$

[22] proved that iterative least squares algorithm without adding noise is a special case of Majorization-Minimization (MM) algorithm for the objective function $L_e(\beta_0, \boldsymbol{\beta})$ and obtained the following convergence results.

Proposition 7.1. *For linear median regression with a full-rank covariate matrix $\mathbf{X}$, the iterative least squares algorithm without adding noise converges to the unique minimizer of $L_e(\beta_0, \boldsymbol{\beta})$.*

Proposition 7.2. *If $(\hat{\beta}_{0,e}, \hat{\boldsymbol{\beta}}_e)$ minimizes $L_e(\beta_0, \boldsymbol{\beta})$, then any limit point of $(\hat{\beta}_{0,e}, \hat{\boldsymbol{\beta}}_e)$ as e tends to 0 minimizes $L(\beta_0, \boldsymbol{\beta})$. If $L(\beta_0, \boldsymbol{\beta})$ has a unique minimizer $(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}})$, then $\lim_{e\to 0}(\hat{\beta}_{0,e}, \hat{\boldsymbol{\beta}}_e)$ = $(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}})$.*

Proposition 7.1 and 7.2 show that $(\hat{\beta}_0(N_0), \hat{\boldsymbol{\beta}}^{\mathrm{T}}(N_0))^{\mathrm{T}}$ can estimate true value well. Notice that DP-ItSq is simply adding normal noise to true parameters, hence, the accuracy of output in DP-ItSq is given directly.

Proof of Theorem 3.4

Proof. Denote $\hat{\boldsymbol{\omega}} = (\hat{\beta}_0(N_0), \hat{\boldsymbol{\beta}}^{\mathrm{T}}(N_0))^{\mathrm{T}}$ and the l_1 sensitivity of $\hat{\boldsymbol{\omega}}$ as $s(\hat{\boldsymbol{\omega}})$. Let $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ be two row vectors over $\mathbb{R}^d$ with l_1 norm at most 1 and $y_1, y_2 \in [-B, B]$. Consider the two inputs D_1 and D_2 where D_2 is obtained from D_1 by changing one record $(\boldsymbol{a}_1, y_1)$ into $(\boldsymbol{a}_2, y_2)$. For convenience, assume the first $n-1$ records are the same and denote $\tilde{\boldsymbol{a}}_1$ as $(1, \boldsymbol{a}_1)$ and $\tilde{\boldsymbol{a}}_2$ as $(1, \boldsymbol{a}_2)$. According to Lemma 7.2, let $G(\hat{\boldsymbol{\omega}}) = I(N_0)$ and $g(\hat{\boldsymbol{\omega}}) = \frac{1}{n} w_n (\tilde{\boldsymbol{a}}_2 \hat{\boldsymbol{\omega}} - y_2)^2 - \frac{1}{n} w_n (\tilde{\boldsymbol{a}}_1 \hat{\boldsymbol{\omega}} - y_1)^2$. Similar to the proof in Theorem 3.2, we can achieve that

$$\|\nabla g(\hat{\boldsymbol{\omega}})\|_1 = \left\| \left. \frac{\partial g(\boldsymbol{\omega})}{\partial \beta_0} \right|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}} \right\|_1 + \left\| \left. \frac{\partial g(\boldsymbol{\omega})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}} \right\|_1,$$

where

$$\left. \frac{\partial g(\boldsymbol{\omega})}{\partial \beta_0} \right|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}} = \frac{2w_n}{n} [(\boldsymbol{a}_2 - \boldsymbol{a}_1)\hat{\boldsymbol{\beta}}(N_0) - (y_2 - y_1)],$$

$$\begin{aligned} \left. \tfrac{\partial g(\boldsymbol{\omega})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}} &= \tfrac{2w_n}{n} [(\hat{\beta}_0(N_0) + \boldsymbol{a}_2 \hat{\boldsymbol{\beta}}(N_0) - y_2)\boldsymbol{a}_2^{\mathrm{T}} \\ &\quad -(\hat{\beta}_0(N_0) + \boldsymbol{a}_1 \hat{\boldsymbol{\beta}}(N_0) - y_1)\boldsymbol{a}_1^{\mathrm{T}}]. \end{aligned}$$

According to the triangle inequality, we can achieve that

$$\begin{aligned} \left\| \left. \tfrac{\partial g(\boldsymbol{\omega})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}} \right\|_1 &\leq \tfrac{2}{n} |w_n| (|\hat{\beta}_0(N_0)| + |\boldsymbol{a}_2 \hat{\boldsymbol{\beta}}(N_0)| + |y_2| \\ &\quad + |\hat{\beta}_0(N_0)| + |\boldsymbol{a}_1 \hat{\boldsymbol{\beta}}(N_0)| + |y_1|). \end{aligned}$$

Notice that $(\hat{\beta}_0(N_0), \hat{\boldsymbol{\beta}}^{\mathrm{T}}(N_0))^{\mathrm{T}} = \mathrm{argmin}_{\beta_0, \boldsymbol{\beta}} I(N_0)$, then $\frac{\partial I(N_0)}{\partial \beta_0} = 0$ at $\beta_0 = \hat{\beta}_0(N_0)$, that is,

$$\begin{aligned} & \textstyle\sum_{i=1}^n w_i (\hat{\beta}_0(N_0) + \mathbf{X}_i \hat{\boldsymbol{\beta}}(N_0) - Y_i) = 0 \\ \Longleftrightarrow \quad & \hat{\beta}_0(N_0) = \tfrac{-\sum_{i=1}^n w_i (\mathbf{X}_i \hat{\beta}(N_0) - Y_i)}{\sum_{i=1}^n w_i}. \end{aligned}$$

Since $0 < w_i \leq 1/e$ $(i = 1, \ldots, n)$ and $\|\hat{\beta}(N_0)\|_1 \leq \sqrt{d}\|\hat{\beta}(N_0)\|_2 \leq \sqrt{dv}$, then we have $|\hat{\beta}_0(N_0)| \leq \sqrt{dv} + B$ and we obtain the following inequlity:

$$\left\| \left. \frac{\partial g(\boldsymbol{\omega})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}} \right\|_1 \leq \frac{8(\sqrt{dv} + B)}{ne}.$$

Similarly, it can be verified that

$$\left\| \left. \frac{\partial g(\boldsymbol{\omega})}{\partial \beta_0} \right|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}} \right\|_1 \leq \frac{4(\sqrt{dv} + B)}{ne}.$$

Therefore,

$$g_1 = \max_{\hat{\boldsymbol{\omega}}} \|\nabla g(\hat{\boldsymbol{\omega}})\|_1 \leq \frac{12(\sqrt{dv} + B)}{ne}.$$

Notice that above inequalities are still true in t-th $(t \geq 2)$ iteration and hence $\frac{1}{2(\sqrt{dv}+B)+e} \leq w_i \leq \frac{1}{e}$. In addition, denote $F_e(\boldsymbol{\omega}) = \frac{1}{n} \sum_{i=1}^n w_i r_i^2(\beta_0, \boldsymbol{\beta})$. It can be checked that $F_e(\boldsymbol{\omega})$ is convex and

$$\frac{\partial F_e^2(\boldsymbol{\omega})}{\partial \beta_0^2} = \frac{2}{n} \sum_{i=1}^n w_i \geq \frac{2}{2(\sqrt{dv} + B) + e},$$

$$\frac{\partial^2 (\frac{\lambda}{2} \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \lambda \mathbf{I},$$

where $\mathbf{I}$ is an identity matrix with size $d \times d$, then $G_2 \geq \min(\frac{2}{2(\sqrt{dv}+B)+e}, \lambda)$ and $s(\hat{\boldsymbol{\omega}}) \leq \frac{12(\sqrt{dv}+B)}{n \min(\frac{2}{2(\sqrt{dv}+B)+e}, \lambda) e}$.

According to Lemma 2.1, the result is obtained. □

Proof of Theorem 3.5

Proof. By using Lemma 7.3, $\frac{\mu}{c}\|\mathbf{U}\|_2 \sim t(d+1)$. Then with probability $1-\alpha$, $\|\mathbf{U}\|_1 \leq \sqrt{d+1}\|\mathbf{U}\|_2 \leq \sqrt{d+1} t_{1-\alpha}(d+1)$, then the theorem is obtained. □