Estimating Stochastic Linear Combination of Non-linear Regressions Efficiently and Scalably ☆

Di Wang[a,*], Xiangyu Guo[a,**], Chaowen Guan[a], Shi Li[a], Jinhui Xu[a]

[a]Department of Computer Science and Engineering
State University of New York at Buffalo
338 Davis Hall, Buffalo, 14260

Abstract

Recently, many machine learning and statistical models such as non-linear regressions, the Single Index, Multi-index, Varying Coefficient Index Models and Two-layer Neural Networks can be reduced to or be seen as a special case of a new model which is called the *Stochastic Linear Combination of Non-linear Regressions* model. However, due to the high non-convexity of the problem, there is no previous work study how to estimate the model. In this paper, we provide the first study on how to estimate the model efficiently and scalably. Specifically, we first show that with some mild assumptions, if the variate vector x is multivariate Gaussian, then there is an algorithm whose output vectors have ℓ_2-norm estimation errors of $O(\sqrt{\frac{p}{n}})$ with high probability, where p is the dimension of x and n is the number of samples. The key idea of the proof is based on an observation motived by the Stein's lemma. Then we extend our result to the case where x is bounded and sub-Gaussian using the zero-bias transformation, which could be seen as a generalization of the classic Stein's lemma. We also show that with some additional assumptions there is an algorithm whose output vectors have ℓ_∞-norm estimation errors of $O(\frac{1}{\sqrt{p}} + \sqrt{\frac{p}{n}})$ with high probability. We also provide a concrete example to show that there exists some link function which satisfies the previous assumptions. Finally, for both Gaussian and sub-Gaussian cases we propose a faster sub-sampling based algorithm and show that when the sub-sample sizes are large enough then the estimation errors will not be sacrificed by too much. Experiments for both cases support our theoretical results. To the best of our knowledge, this is the first work that studies and provides theoretical guarantees for the stochastic linear combination of non-linear regressions model.

Keywords: linear regression, gradient descent, random projection

1. Introduction

Recently, many machine learning and statistical models can be reduced to or be seen as a special case of a new model which is called the *Stochastic Linear Combination of Non-linear Regressions* model, which can be defined as the followings.

Definition 1 (Stochastic Linear Combination of Non-linear Regressions). *Given variates $x \in \mathbb{R}^p$ and $z_1, \cdots, z_k \in \mathbb{R}$ such that $\mathbb{E}[x] = 0$ and z_i's for all $i \in [k]$ are i.i.d random variables independent of x with $\mathbb{E}[z_i] = 0$ and $Var(z_i) = 1$, the response y is given by*

$$y = \sum_{i=1}^{k} z_i f_i(\langle \beta_i^*, x \rangle) + \epsilon, \qquad (1)$$

where $\beta_1^*, \beta_2^*, \cdots, \beta_k^* \in \mathbb{R}^p$ are unknown parameters, f_i's for all $i \in [k]$ are known (but could be non-convex) link functions,

and ϵ is some random noise (from an unknown distribution) satisfying $\mathbb{E}[\epsilon] = 0$ and is independent of x and z_i's.

The goal is to estimate the parameters β_j^* for all $j \in [k]$ from n observations $(x_1, y_1, \{z_{1,i}\}_{i=1}^k)$, $(x_2, y_2, \{z_{2,i}\}_{i=1}^k), \cdots, (x_n, y_n, \{z_{n,i}\}_{i=1}^k)$.

This model has a close connection with many models in Statistics, Machine Learning, Signal Processing and Information Theory: (1) when $k = 1$, the model is reduced to the non-linear regression estimation problem which has been studied in [1; 2; 3; 4] and is related to compressed sensing and image recovery as well; (2) when $k = 1$ but the link function f_1 is unknown, it becomes the **Single Index Model**, which is one of the most fundamental models in statistics and has been studied for many years [5; 6; 7; 8; 9]; (3) when $k \geq 1$, z_i's are deterministic but f_i's are unknown, this model will be a special case of the **Multi-index Model** which has been studied in [10; 11; 12; 13]; (4) when $k \geq 1$, z_i's are stochastic but f_i's are unknown, it will be the **Varying Coefficient Index Model** which was introduced by [14] and has wide applications in economics and medical science [15]; (5) when all f_i's are the same, the model can be viewed as a **Two-layer Neural Network** with k hidden nodes and random hidden-output layer weights.

*Corresponding author
**The first two authors contributed equally

Email addresses: dwang45@buffalo.edu (Di Wang), xiangyug@buffalo.edu (Xiangyu Guo), chaowen@buffalo.edu (Chaowen Guan), shil@buffalo.edu (Shi Li), jinhui@buffalo.edu (Jinhui Xu)

To estimate the parameters in Model (1), the main challenge is that without the assumption that f_i's are convex or similarities between them, it is hard to establish an objective function that can be efficiently optimized using optimization methods such as (Stochastic) Gradient Descent. Thus, due to the high convexity and randomness, there is no previous study how to solve the model efficiently. Recently, some works including [8; 16; 13] studied and proposed efficient algorithms for the Single Index, Multi-index and Varying Coefficient Index models using Stein's Lemma. Their theoretical guarantees are measured in terms of $\|\beta_j - c\beta_j^*\|_2$, $j \in [k]$, where β_j is the estimator for β_j^* and c is a constant depending on many parameters in the models (such as f_i's, β_j^*'s and the distribution for x). However there is a common issue related to the constant c in these results: They did not provide a method to compute or even estimate c. Moreover, measuring the error by using the terms $\|\beta_j - c\beta_j^*\|_2$, $j \in [k]$ is quite meaningless due to the constant c. Thus, we wish to use other measurements and ideally, we hope to measure the error in terms of $\beta_j - \beta_j^*$ for all $j \in [k]$; that is, we do not introduce the constant c. The key question the paper tries to answer is:

Is there an efficient method whose output vectors β_1, $\beta_2, \cdots, \beta_k$ have small errors compared to $\beta_1^*, \beta_2^*, \cdots, \beta_k^*$?

In this paper, we answer the question in the affirmative under some mild assumptions on the model. Specifically, our contributions can be summarized as follows.

1. We first consider the case where x multivariate Gaussian. In this case, we show that there is a special structure for each β_j^*, $j \in [k]$: $\beta_j^* = c_j \beta_j^{ols}$, where c_j is a constant depending on the link function f_j and x, and β_j^{ols} is the Ordinary Lest Square estimator w.r.t yz_j and x, i.e., $\beta_j^{ols} = (\mathbb{E}[xx^T])^{-1}\mathbb{E}[z_jyx]$. Based on this key observation, we propose an algorithm which estimates c_j's and β_j^{ols}'s, and outputs $\{\beta_j\}_{j=1}^k$ satisfying $\|\beta_j - \beta_j^*\|_2 \leq O(\sqrt{\frac{p}{n}})$ for each $j \in [k]$ with high probability. Moreover, in order to make our algorithm faster, instead of using linear regression estimator to approximate β_j^{ols}, we use the sub-sampling covariance linear regression estimator [17]. We show that if the sub-sample size is large enough, the error bound is almost the same as in the previous ones.

2. We then extend our result to the case when x is (bounded) sub-Gaussian. The challenge is that the result for the Gaussian case depends on some properties of Gaussian distribution which are not satisfied in the sub-Gaussian case. To overcome this, we use the zero-bias transformation [18], which could be seen as a generalization of the Stein's lemma [19]. Particularly, we show that instead of the equality $\beta_j^* = c_j \beta_j^{ols}$, we have the ℓ_∞ norm estimation error $\|\beta_j^* - c_j\beta_j^{ols}\|_\infty \leq O(\frac{1}{\sqrt{p}})$ with some additional mild assumptions. Based on this and the same idea from the Gaussian case, we show that there exists an algorithm whose output vectors $\{\beta_j\}_{j=1}^k$ satisfy $\|\beta_j - \beta_j^*\|_\infty \leq O(\frac{1}{\sqrt{p}} + \sqrt{\frac{p}{n}})$ with high probability. Similarly, we also propose a sub-sampled version of our algorithm as in the Gaussian case.

3. While we provide some theoretical results in the previous parts, it is still unknown whether there exists any link function which satisfies these assumptions. To solve this problem, we consider the case where the link functions are sigmod function. And we show that with some assumptions on x, it indeed satisfies the previous assumptions.

4. At the end, we show the experimental results on both Gaussian and sub-Gaussian cases with single/mixed type of link functions, and these results support our theoretical results above. Specially, they show both of the effectiveness and scalability of our previous algorithms.

To the best of our knowledge, this is the first paper studying and providing the estimation error bound for Model (1) in both Gaussian and sub-Gaussian cases.

This paper is a substantially extended version of our previous work appeared in AAAI'20 [20]. The following are the main added contents. Firstly, we construct a concrete loss function which satisfies the assumptions in our theorems which has been studied in the conference version. Specifically, we show that under some assumptions on x, when the link function is the sigmoid function, then it satisfies the assumptions in Theorem 2 (see Theorem 7 for details). Secondly, we provide the proofs for all theorems and lemmas, and we believe the techniques can be used to other problems.

The rest of the paper is organized as follows. Section 2 introduces some related work. Section 3 gives some preliminaries on Sub-Gaussian random variable, necessary lemmas and assumptions throughout the paper. Section 4 describes our proposed algorithms for Gaussian case. Section 5 extends our algorithm to the Sub-Gaussian case. We provide all the proofs in Section 6. Finally, we experimentally study our methods in Section 7, and conclude them in Section 8.

2. Related Work

As we mentioned above, there is no previous work on Model (1) with guarantees on the ℓ_2 or ℓ_∞ norm of the errors $\beta_j - \beta_j^*$. Hence, below we compare with the results which are close to ours.

When the link functions f_j's are unknown, Model (1) is just the Varying Coefficient Index Model. [16] provided the first efficient algorithm for this model. Although they considered the high dimensional sparse case, their method requires the underlying distribution of x to be known, an unrealistic assumption for most applications. Moreover, their estimation errors are measured by the differences between β_j's and $c\beta_j^*$'s for an unknown c, while in our results we have fixed $c = 1$.

When the link functions f_j's are all the same, then our model can be reduced to the two-layer neural network with random hidden-output layer weights. Previous work on the convergence results all focused on the gradient descent type of methods such as those in [21; 22; 23]. However, our method is based on Stein's lemma and its generalization. Compared with the gradient descent type methods, our algorithm is non-interactive (that is, we do not need to update estimators in each iteration) and

parameter-free (that is, we don not need to tune the step-size). Moreover, our method can be extended to the case where the link functions f_j's are different.

Our method is motivated by Stein's lemma [19] and its generalization, the zero-bias transformation. Several previous studies have used Stein's Lemma in various machine learning problems. For example, [24; 25] used it to accelerate some optimization procedures, [26] applied it to Bayesian inference, [27] appied it to estimate smooth Generalized Linear Model in differential privacy model and [8; 3; 16; 28] used it and its generalizations in the Single Index, Multi-index, Varying Coefficient Index and Generative models, respectively. The zero-bias transformation has also been used in [29] for estimating the Generalized Linear Model. However, due to the difference between the models, these algorithms cannot be applied to our problem.

3. Preliminaries

In this section, we review some necessary definitions and lemmas.

Definition 2 (Sub-Gaussian). *For a given constant κ, a random variable $x \in \mathbb{R}$ is said to be sub-Gaussian if it satisfies $\sup_{m \geq 1} \frac{1}{\sqrt{m}} \mathbb{E}[|x|^m]^{\frac{1}{m}} \leq \kappa$. The smallest such κ is the sub-Gaussian norm of x and it is denoted by $\|x\|_{\psi_2}$.*

Similarly, a random vector $x \in \mathbb{R}^p$ is called a sub-Gaussian vector if there exists a constant κ such that $\sup_{v \in S^{p-1}} \|\langle x, v \rangle\|_{\psi_2} \leq \kappa$, where S^{p-1} is the set of all p-dimensional unit vector.

In order to extend our results to the sub-Gaussian case, we will use the zero-bias transformation which is proposed by [18]. It is a generalization of the classic Stein's lemma in [19].

Definition 3. *Let z be a random variable with mean 0 and variance σ^2. Then there exists a random variable z^* such that for all differentiable functions f we have $\mathbb{E}[zf(z)] = \sigma^2 \mathbb{E}[f'(z^*)]$. The distribution of z^* is said to be the z-zero-bias distribution.*

The standard Gaussian distribution is the unique distribution whose zero-bias distribution is itself. This is just the basic **Stein's lemma**.

Lemma 1. *[17] Assume that $\mathbb{E}[x] = 0, \mathbb{E}[x_i x_i^T] = \Sigma \in \mathbb{R}^{p \times p}$, and $\Sigma^{-\frac{1}{2}} x$ and y are sub-Gaussian with norms κ_x and γ respectively. If $n \geq \Omega(\gamma \kappa_x p)$, then with probability at least $1 - 3 \exp(-p)$ we have*

$$\|\Sigma^{\frac{1}{2}}(\tilde{\beta}^{ols} - \beta^{ols})\|_2 \leq C_1 \kappa_x \gamma \sqrt{\frac{p}{n}}, \tag{2}$$

where $\beta^{ols} = \Sigma^{-1} \mathbb{E}[yx]$ is the OLS estimator w.r.t y and x, $\tilde{\beta}^{ols} = (X^T X)^{-1} X^T Y$ is the empirical one, and $C_1 > 0$ is some universal constant.

Lemma 2. *[29] Let $B^\delta(\tilde{\beta})$ denote the ball centered around $\tilde{\beta}$ with radius δ. For $i = 1, 2, \cdots, n$, let $x_i \in \mathbb{R}^p$ be i.i.d random vectors with a covariance matrix Σ. Given a function g that*

is uniformly bounded by L and G-Lipschitz, with probability at least $1 - \exp(-p)$ we have

$$\sup_{\beta \in B^\delta(\tilde{\beta})} \left| \frac{1}{n} \sum_{i=1}^n g(\langle x_i, \beta \rangle) - \mathbb{E}[g(\langle x, \beta \rangle)] \right| \leq 2(G(\|\tilde{\beta}\|_2 + \delta)\|\Sigma\|_2 + L) \sqrt{\frac{p}{n}}.$$

Assumption 1. *We assume that for each $j \in [k]$, the random variable yz_j is sub-Gaussian with its sub-Gaussian norm $\|yz_j\|_{\psi_2} = \gamma$.*

Note that this assumption holds if y is bounded and z_j is sub-Gaussian or z_j is bounded and y is sub-Gaussian.

Assumption 2. *We assume that there exist constants $G, L > 0$ such that for each $j \in [k]$, f_j' is G-Lipschitz and bounded by L. Also for $j \in [k]$, we let $\mathbb{E}[f_j'(\langle x, \beta_j^* \rangle)] \neq 0$.*

Notations. For a positive semi-definite matrix $M \in \mathbb{R}^{p \times p}$, we define the M-norm for a vector w as $\|w\|_M^2 = w^T M w$. Also we will denote $B_M^\delta(\tilde{\beta})$ as the ball around $\tilde{\beta}$ with radius δ under M-norm, i.e., $B_M^\delta(\tilde{\beta}) = \{\beta : \|M^{\frac{1}{2}}(\beta - \tilde{\beta})\|_2 \leq \delta\}$. $\lambda_{\min}(M)$ is the minimal singular value of the matrix M. For a semi positive definite matrix $M \in \mathbb{R}^{p \times p}$, let its SVD be $M = U^T \Sigma U$, where $\Sigma = \text{diag}(\lambda_1, \cdots, \lambda_p)$, then $M^{\frac{1}{2}}$ is defined as $M^{\frac{1}{2}} = U^T \Sigma^{\frac{1}{2}} U$ with $\Sigma^{\frac{1}{2}} = \text{diag}(\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_p})$.

4. Gaussian Case

In this section we consider the case where x is sampled from some multivariate Gaussian distribution, then we will extend our idea to the sub-Gaussian distribution case in next section.

Our algorithm is based on the following key observation using some properties of the multivariate Gaussian distribution.

Theorem 1. *Consider Model (1) in Definition 1 under Assumptions 1 and 2. Moreover, assume that the observations $\{x_i\}_{i=1}^n$ are i.i.d sampled from $\mathcal{N}(0, \Sigma)$. Then each $\beta_j^*, j \in [k]$ can be written as*

$$\beta_j^* = c_j \times \beta_j^{ols}, \tag{3}$$

where $\beta_j^{ols} = \Sigma^{-1} \mathbb{E}[z_j yx]$ and c_j is the root of the function $l_j(c) - 1$ where

$$l_j(c) = c \mathbb{E}[f_j'(\langle x, \beta_j^{ols} \rangle c)]. \tag{4}$$

From Theorem 1 we can see that, in order to estimate β_j^*, it is sufficient to estimate the terms $\beta_j^{ols} = \Sigma^{-1} \mathbb{E}[z_j yx]$ and c_j. If we denote $z_j y$ as the response and x as the variate, then the term β_j^{ols} is just the **Ordinary Least Square (OLS)** estimator. Thus we can use its empirical form $\tilde{\beta}_j^{ols} = (\sum_{i=1}^n x_i^T x_i)^{-1} \sum_{i=1}^n z_{i,j} y_i x_i = (X^T X)^{-1} X^T Y_j$ as an estimator, where $X = [x_1^T; x_2^T; \cdots; x_n^T] \in \mathbb{R}^{n \times d}$ is the data matrix and $Y_j = [z_{1,j} y_1, \cdots, z_{n,j} y_n]^T$ is the corresponding response vector.

After getting the estimator of β_j^{ols}, denoted by $\tilde{\beta}_j^{ols}$, we use it to approximate c_j. That is we find the root $\hat{c}_j$ of the empirical version of $l_j(c) - 1$, i.e., $\hat{l}_j(c) - 1$, where

$$\hat{l}_j(c) = \frac{c}{n} \sum_{i=1}^n [f_j'(\langle x_i, \tilde{\beta}_j^{ols} \rangle c)].$$

Note that there are numerous methods available to find a root of a function, such as Newton's root-finding method with quadratic convergence and Halley's method with cubic convergence. We also note that this step only cost $O(n)$ per-iteration. After that, we could estimate each β_j^* by $\hat{\beta}_j^{nlr} = \hat{c}_j \tilde{\beta}_j^{ols}$. In total, we have Algorithm 1.

Algorithm 1 SLS: Scaled Least Squared Estimators

Input: Data $\{(x_i, y_i, \{z_{i,j}\}_{j=1}^k)\}_{i=1}^n$, link functions $\{f_j\}_{j\in[k]}$.

1: **Option 1:** Let $X = [x_1, x_2, \cdots, x_n]^T \in \mathbb{R}^{n\times p}$ and compute the $\hat{\Sigma}^{-1} = (X^T X)^{-1}$.
2: **Option 2:** Construct a sub-sampling based OLS estimator, that is let $S \subset [n]$ be a random subset and take $\hat{\Sigma}_S^{-1} = \frac{|S|}{n}(X_S^T X_S)^{-1}$, where $X_S \in \mathbb{R}^{|S|\times p}$ is the data matrix constrained on indices of S.
3: **for** $j = 1, 2 \cdots, k$ **do**
4:     Let $Y_i = [z_{1,j}y_1, \cdots, z_{n,j}y_n]^T \in \mathbb{R}^n$. For **Option 1**, Compute the ordinary least squares estimator $\tilde{\beta}_j^{ols} = (\Sigma)^{-1}X^T Y_j$. For **Option 2**, take $\tilde{\beta}_j^{ols} = (\hat{\Sigma}_S)^{-1}X^T Y_j$.
5:     Denote $\tilde{y}_j = X\tilde{\beta}_j^{ols}$. Then use the Newton's root-finding method to the function $\frac{c}{n}\sum_{i=1}^n[f_j'(\tilde{y}_{j,i}c)] - 1$, denote the root as $\hat{c}_j$:
6:     **for** $t = 1, 2, \cdots$ until convergence **do**
7:         $c = c - \frac{c\frac{1}{n}\sum_{i=1}^n f_j'(c\tilde{y}_{j,i})-1}{\frac{1}{n}\sum_{i=1}^n\{f_j'(c\tilde{y}_{j,i})+c\tilde{y}_{j,i}f_j^{(2)}(c\tilde{y}_{j,i})\}}$.
8:     **end for**
9:     $\hat{\beta}_j^{nlr} = \hat{c}_j \cdot \tilde{\beta}_j^{ols}$.
10: **end for**
11:
12: **return** $(\hat{\beta}_j^{nlr})_{j\in[k]}$

The following theorem shows that the converge rate of the estimation error for each $\|\hat{\beta}_j^{nlr} - \beta_j^*\|_2$ is $O(\sqrt{\frac{p}{n}})$ under some additional mild assumptions on link functions $\{f_j\}_{j=1}^k$.

Theorem 2. *Consider Option 1 in Algorithm 1. Under the Assumptions 1, 2 and the assumptions in Theorem 1, for each $j \in [k]$ we define the function $\ell_j(c,\beta) = c\mathbb{E}[f_j'(\langle x,\beta\rangle c)]$ and its empirical counter part as*

$$\hat{\ell}_j(c,\beta) = \frac{c}{n}\sum_{i=1}^n f_j'(\langle x_i,\beta\rangle c).$$

Assume that there exist some constants $\eta, \bar{c}_j$ such that $\ell_j(\bar{c}_j, \beta_j^{ols}) > 1 + \eta$. Then there exists $c_j > 0$ satisfying the equation $1 = \ell_j(c_j, \beta_j^{ols})$ for each $j \in [k]$.

Further, assume that n is sufficiently large:

$$n \geq \Omega(p\|\Sigma\|_2\|\beta_j^*\|_2^2)$$

Then, with probability at least $1 - k\exp(-p)$ there exist constants $\hat{c}_j \in (0, \bar{c}_j)$ satisfying the equations

$$1 = \frac{\hat{c}_j}{n}\sum_{i=1}^n f_j'(\langle x_i,\tilde{\beta}_j^{ols}\rangle\hat{c}_j).$$

Moreover, if for all $j \in [k]$ the derivative of $z \mapsto \ell_j(z, \beta_j^{ols})$ is bounded below in absolute value (does not change sign) by $M > 0$ in the interval $z \in [0, c_j]$. Then with probability at least $1 - 4k\exp(-p)$ the outputs $\{\hat{\beta}_j^{nlr}\}_{j=1}^k$ satisfy for each $j \in [k]$

$$\|\hat{\beta}_j^{nlr} - \beta_j^*\|_2 \leq O(\max\{1, \|\beta_j^*\|_2\}\lambda_{\min}^{-\frac{1}{2}}(\Sigma)\sqrt{\frac{p}{n}}), \qquad (5)$$

where $G, L, \gamma, M, c_j, \eta$ are assumed to be $\Theta(1)$ and thus omitted in the Big-O and Ω notations (see Appendix for the explicit forms).

Note that in Theorem 2 the link functions f_j are not required to be convex. Hence this is quite useful in non-convex learning models.

Time Complexity Analysis. Under Option 1 of Algorithm 1, we can see that the first step takes $O(np^2 + p^3)$ time, calculating $\tilde{\beta}_j^{ols}$ for all $j \in [k]$ takes $O(k(np + p^2))$ time and each iteration of finding $\hat{c}_j$ takes $O(n)$ time. Thus, if k, the number of link functions f_j, is a constant, then the total time complexity is $O(np^2 + p^3 + nT)$, where T is the number of iterations for finding c_j.

However, the term np^2 is prohibitive in the large scale setting where n, p are huge (see [30; 31] for details). To further reduce the time complexity, we propose another estimator based on sub-sampling.

Note that the term $O(np^2)$ comes from calculating the empirical covariance matrix $X^T X$. Thus, to reduce the time complexity, instead of calculating the covariance via the whole dataset, we here use the sub-sampled covariance matrix. More precisely, we first randomly sample a set of indices $S \subset [n]$ whose size $|S|$ will be specified later. Then we calculate $\frac{|S|}{n}(X_S^T X_S)^{-1}$ to estimate $(X^T X)^{-1}$, where $X_S \in \mathbb{R}^{|S|\times p}$ is the data matrix constrained on indices of S. We can see that the time complexity in this case will only be $O(|S|p^2 + p^3)$. The following lemma, which is given by [17; 24] shows the convergence rate of the OLS estimator based on the sub-sampled covariance matrix. This is a generalization of Lemma 1.

Lemma 3. *Under the same assumptions as in Lemma 1, if $|S| \geq \Omega(\gamma\kappa_x p)$, then with probability at least $1 - 3\exp(-p)$ the sub-sampled covariance OLS estimator $\tilde{\beta}^{ols} = \frac{|S|}{n}(X_S^T X_S)^{-1}X^T Y$ satisfies*

$$\|\tilde{\beta}^{ols} - \beta^{ols}\|_2 \leq C_2\kappa_x\gamma\sqrt{\frac{p}{|S|}}.$$

We have the following approximation error based the sub-sampled covariance OLS estimator:

Theorem 3. *Under the same assumptions as in Theorem 2, in Algorithm 1 if we use Option 2 with $|S| \geq \Omega(\gamma\kappa_x p)$, then with probability at least $1 - 4k\exp(-p)$ the outputs $\{\hat{\beta}_j^{nlr}\}_{j=1}^k$ satisfy for each $j \in [k]$*

$$\|\hat{\beta}_j^{nlr} - \beta_j^*\|_2 \leq O(\max\{1, \|\beta_j^*\|_2\}\lambda_{\min}^{-\frac{1}{2}}(\Sigma)\sqrt{\frac{p}{|S|}}).$$

Moreover, it is also possible to accelerate the algorithm using the sub-sampling method in the step 5 (finding the root) and we can see the estimation error will be the same as in Theorem 3 (by the proof of Theorem 3). Due to the space limit, we omit it here.

5. Extension to Sub-Gaussian Case

Note that Theorem 2 is only suitable for the case when x is Gaussian. This is due to the requirements on some properties of the Gaussian distribution in the proof of Theorem 1. In this section we will first extend Theorem 1 to the sub-Gaussian case.

Remember that the proof of Theorem 1 is based on the classic Stein's lemma. Thus, in order to generalize to sub-Gaussian case, we will use the zero-bias transformation in Definition 3 since it is a generalization of the Stein's lemma. With some additional assumptions, we can get the following theorem.

Theorem 4. *Let $x_1, \cdots, x_n \in \mathbb{R}^p$ be i.i.d realizations of a random vector x which is sub-Gaussian with zero mean, whose covariance matrix Σ has $\Sigma^{\frac{1}{2}}$ being diagonally dominant [1], and whose distribution is supported on a ℓ_2-norm ball of radius r. Let $v = \Sigma^{-\frac{1}{2}}x$ be the whitened random vector of x with sub-Gaussian norm $\|v\|_{\psi_2} = \kappa_x$. If for all $j \in [k]$, each v has constant first and second conditional moments (i.e., $\forall s \in [p]$ and $\tilde{\beta}_j = \Sigma^{\frac{1}{2}}\beta_j^*$, $\mathbb{E}[v_s| \sum_{t \neq s} \tilde{\beta}_j v_t]$ and $\mathbb{E}[v_s^2| \sum_{t \neq s} \tilde{\beta}_j v_t]$ are deterministic) and the link functions f_j' satisfy Assumption 2. Then for $c_j = \frac{1}{\mathbb{E}[f_j'(\langle x, \beta_j^* \rangle)]}$, the following holds for the model in (1) for all $j \in [k]$*

$$\|\frac{1}{c_j} \cdot \beta_j^* - \beta_j^{ols}\|_\infty \leq 16Grk\kappa_x^3 \sqrt{\rho_2}\rho_\infty \frac{\|\beta_j^*\|_\infty^2}{\sqrt{p}}, \qquad (6)$$

where ρ_q for $q = \{2, \infty\}$ is the conditional number of Σ in ℓ_q norm and $\beta_j^{ols} = \Sigma^{-1}\mathbb{E}[xyz_j]$ is the OLS vector w.r.t yz_j and x.

Remark 1. *Note that compared with the equality relationship between β_j^* and $c_j\beta^{ols}$ in Theorem 1, in Theorem 4 we only has the ℓ_∞ norm of their difference. Also, here we need more assumptions on the distribution of x, and these assumptions ensure that the estimation error decays in the rate of $O(\frac{1}{\sqrt{p}})$.*

Theorem 4 indicates that we can use the same idea as in the Gaussian case to estimate each β_j^*. Note that the forms of c_j in Theorem 1 and 4 are different. In Theorem 1 each c_j is based on β_j^{ols}, while in Theorem 4 it is based on β_j^*. However, due to the closeness of β_j^* and β_j^{ols} in (6), we can still use $\frac{1}{\mathbb{E}[f_j'(\langle x, \beta_j^{ols} \rangle \tilde{c}_j)]}$ to approximate c_j, where $\tilde{c}_j$ is the root of $c\mathbb{E}[f_j'(\langle x_i, \beta_j^{ols} \rangle c)] - 1$. Because of this similarity, we can still use Algorithm 1 for the sub-Gaussian case under the assumptions in Theorem 4, and we can get the following estimation error:

---

[1] A square matrix is said to be diagonally dominant if, for every row of the matrix, the magnitude of the diagonal entry in a row is larger than or equal to the sum of the magnitudes of all the other (non-diagonal) entries in that row.

Theorem 5. *Consider Option 1 in Algorithm 1. Under Assumptions 1, 2 and the assumptions in Theorem 4, for each $j \in [k]$, if we define the function $\ell_j(c, \beta) = c\mathbb{E}[f_j'(\langle x, \beta \rangle c)]$ and its empirical counter part as*

$$\hat{\ell}_j(c, \beta) = \frac{c}{n} \sum_{i=1}^n f_j'(\langle x_i, \beta \rangle c).$$

Assume that there exist some constants $\eta, \bar{c}_j$ such that $\ell_j(\bar{c}_j, \beta_j^{ols}) > 1 + \eta$. Then there exists $\tilde{c}_j > 0$ satisfying the equation $1 = \ell_j(\tilde{c}_j, \beta_j^{ols})$ for each $j \in [k]$.

Further, assume that n is sufficiently large:

$$n \geq \Omega(\|\Sigma\|_2 p^2 \rho_2 \rho_\infty^2 \|\beta_j^*\|_\infty^2 \max\{1, \|\beta_j^*\|_\infty^2\}).$$

Then, with probability at least $1 - k\exp(-p)$ there exist constants $\hat{c}_j \in (0, \bar{c}_j)$ satisfying the equations

$$1 = \frac{\hat{c}_j}{n} \sum_{i=1}^n f_j'(\langle x_i, \tilde{\beta}_j^{ols} \rangle \hat{c}_j).$$

Moreover, if for all $j \in [k]$, the derivative of $z \mapsto \ell_j(z, \beta_j^{ols})$ is bounded below in absolute value (does not change sign) by $M > 0$ in the interval $z \in [0, \max\{\bar{c}_j, c_j\}]$. Then with probability at least $1 - 4k\exp(-p)$ the outputs $\{\hat{\beta}_j^{nlr}\}_{j=1}^k$ satisfy for each $j \in [k]$

$$\|\hat{\beta}_j^{nlr} - \beta_j^*\|_\infty \leq O(\sqrt{\rho_2}\rho_\infty \lambda_{\min}^{-\frac{1}{2}}(\Sigma) \sqrt{\frac{p}{n}} \|\beta_j^*\|_\infty$$
$$\times \max\{1, \|\beta_j^*\|_\infty\} + \rho_2 \rho_\infty^2 \frac{\max\{\|\beta_j^*\|_\infty^2, 1\}\|\beta_j^*\|_\infty^2}{\sqrt{p}}),$$

where $\eta, G, L, \gamma, M, \bar{c}_j, r, \kappa_x, c_j$ are assumed to be $\Theta(1)$ and thus omitted in the Big-O and Ω notations (see Appendix for the explicit forms).

Remark 2. *Compared with the converge rate in the ℓ_2-norm of $O(\sqrt{\frac{p}{n}})$ in Theorem 2, Theorem 5 shows that for the sub-Gaussian case, the converge rate of the estimation error is $O(\frac{1}{\sqrt{p}} + \sqrt{\frac{p}{n}})$ in the ℓ_∞-norm (if we omit other terms). This is due to the estimation error in Theorem 4. Moreover, compared with the assumptions of link functions in Theorem 2, there are additional assumptions in Theorem 5.*

In order to reduce the time complexity and make the algorithm faster, we can also use the sub-sampled covariance OLS estimator. This is the same as that in the Gaussian case.

Theorem 6. *Under the same assumptions as in Theorem 5, if we use Option 2 in Algorithm 1, then with probability at least $1 - 4k\exp(-p)$, the outputs $\{\hat{\beta}_j^{nlr}\}_{j=1}^k$ satisfy for each $j \in [k]$*

$$\|\hat{\beta}_j^{nlr} - \beta_j^*\|_\infty \leq O(\sqrt{\rho_2}\rho_\infty \lambda_{\min}^{-\frac{1}{2}}(\Sigma) \sqrt{\frac{p}{|S|}} \|\beta_j^*\|_\infty$$
$$\times \max\{1, \|\beta_j^*\|_\infty\} + \rho_2 \rho_\infty^2 \frac{\max\{\|\beta_j^*\|_\infty^2, 1\}\|\beta_j^*\|_\infty^2}{\sqrt{p}}).$$

An unsatisfactory issue of all above theorems is that they need quite a few assumptions/conditions. Although almost all of them commonly appear in some related works, the assumptions on functions $\ell_j(\cdot, \beta_j^{ols})$ seem to be a little strange. They are introduced for ensuring that the functions $\ell_j(\cdot, \beta_j^{ols}) - 1$ and $\hat{\ell}_j(\cdot, \tilde{\beta}_j^{ols}) - 1$ have roots and $\hat{c}_j$ is close to c_j for large enough n. The following theorem shows that the sigmoid link function indeed satisfies the assumptions in Theorem 2 when x is some Gaussian and $\|\beta_j^{ols}\|_2 = O(\sqrt{p})$.

Theorem 7. *Consider the case where the link function $f_j(z) = \frac{1}{1+e^{-z}}$ and $x \sim \mathcal{N}(0, \frac{1}{p}I_p)$ and $\|\beta_j^{ols}\|_2 = \frac{\sqrt{p}}{20}$. Then when $\bar{c}_j = 6$ and $\eta = 0.22$, $\ell_j(\bar{c}_j, \beta_j^{ols}) > 1 + \eta$. Moreover, $\ell'(z, \beta^{ols})$ is bounded by constant $M = 0.19$ on $[0, \bar{c}_j]$ from below.*

As we will see later, our experimental results show that the algorithm actually performs quite well even though some of the assumptions are not satisfied, such as the model with logistic and cubic link functions.

To deal with the privacy issue, recently, [27] studied the problem of estimating smooth Generalized Linear Model and non-linear regression problem in a model which is called *Non-interactive Local Differential Privacy with public unlabeled data*. And they proposed several methods based on the Stein's lemma and zero-bias transformation. As we mentioned above, non-linear regression is one special case of our model, thus, we can easily get an algorithm which could estimate our problem in the the model of Non-interactive Local Differential Privacy with public unlabeled data.

6. Omitted Proofs

Proof of Theorem 1. Fix $j \in [k]$, we have

$$\mathbb{E}[z_j y x] = \mathbb{E}[z_j(\sum_{i=1}^{k} z_i f_i(\langle x, \beta_i^* \rangle) + \epsilon)x]$$
$$= \mathbb{E}[z_j^2 f_j(\langle x, \beta_j^* \rangle)x] = \mathbb{E}[f_j(\langle x, \beta_j^* \rangle)x], \quad (7)$$

where Eq. (7) is due to the assumption of $\{z_j\}_{j=1}^{k}$ are i.i.d and $\text{Var}(z_j) = 1$.

Now, denote by $\phi(x|\Sigma)$ the multivariate normal density with mean 0 and covariance matrix Σ. We recall the well-known property of Gaussian density $\frac{d\phi(x|\Sigma)}{dx} = -\Sigma^{-1}x\phi(x|\Sigma)$ (**this is just the Stein's lemma**). Using this and the integration we have

$$\mathbb{E}[f_j(\langle x, \beta_j^* \rangle)x] = \int x f_j(\langle x, \beta_j^* \rangle)\phi(x|\Sigma)dx$$
$$= \int -f_j(\langle x, \beta_j^* \rangle)d\phi(x|\Sigma) = \Sigma\beta_j^*\mathbb{E}[f_j'(\langle x, \beta_j^* \rangle)].$$

Thus we have

$$\frac{1}{\mathbb{E}[f_j'(\langle x, \beta_j^* \rangle)]}\Sigma^{-1}\mathbb{E}[z_j y x] = \beta_j^*. \quad (8)$$

$\square$

Proof of Theorem 4. Now we fix $j \in [k]$, by the assumptions in the model we have

$$\mathbb{E}[z_j y x] = \mathbb{E}[z_j(\sum_{i=1}^{k} z_i f_i(\langle x, \beta_i^* \rangle) + \epsilon)x]$$
$$= \mathbb{E}[z_j^2 f_j(\langle x, \beta_j^* \rangle)x] = \mathbb{E}[f_j(\langle x, \beta_j^* \rangle)x]$$
$$= \Sigma^{\frac{1}{2}}\mathbb{E}[v f_j(\langle v, \tilde{\beta}_j \rangle)].$$

Thus we have

$$\beta_j^{ols} = \Sigma^{-\frac{1}{2}}\mathbb{E}[v f_j'(\langle v, \tilde{\beta}_j \rangle)]. \quad (9)$$

For convenience we will omit subscript j. Now we define the partial sum $V_{-i} = \langle v, \tilde{\beta} \rangle - v_i\tilde{\beta}_i$ for $i \in [p]$. We will focus on the i-th entry of the above expectation given in (9). Denote the zero-bias-transformation of v_i conditioned on V_{-i} by v_i^*, we have

$$\mathbb{E}[v_i f_j(\langle v, \tilde{\beta}_j \rangle)] = \mathbb{E}[\mathbb{E}[v_i f_j(v_i\tilde{\beta}_i + V_{-i})|V_{-i}]]$$
$$= \tilde{\beta}_i\mathbb{E}[f_j'(v_i^*\tilde{\beta}_i + V_{-i})] = \tilde{\beta}_i\mathbb{E}[f_j'((v_i^* - v_i)\tilde{\beta}_i + \langle \tilde{\beta}, v \rangle)].$$

Combining with the above equation, we have

$$\beta^{ols} = \Sigma^{-\frac{1}{2}}D\tilde{\beta} = \Sigma^{-\frac{1}{2}}D\Sigma^{\frac{1}{2}}\beta^*, \quad (10)$$

where D is a diagonal matrix where the the entry $D_{ii} = \mathbb{E}[f'((v_i^* - v_i)\tilde{\beta}_i + \langle \tilde{\beta}, v \rangle)]$.

By the Lipschitz property of f we have

$$|D_{ii} - \frac{1}{c_j}| \leq |\mathbb{E}[f'((v_i^* - v_i)\tilde{\beta}_i + \langle \tilde{\beta}, v \rangle)] - \mathbb{E}[f'(\langle \tilde{\beta}, v \rangle)]|$$
$$\leq G\mathbb{E}[|(v_i^* - v_i)\tilde{\beta}_i|]. \quad (11)$$

Now we will bound the term of $\mathbb{E}[|(v_i^* - v_i)|]$, by the same method as in [24] we can get

$$\mathbb{E}[|(v_i^* - v_i)|] \leq \frac{3}{2}\mathbb{E}[|v_i|^3] \leq 8\kappa_x^3 \quad (12)$$

where the second inequality comes from $\frac{1}{\sqrt{3}}\mathbb{E}[|v_i|^3]^{\frac{1}{3}} \leq \|v\|_{\psi_2} \leq \kappa_x$.

Thus, in total we have $\max_{i\in[p]}|D_{ii} - \frac{1}{c_j}| \leq 8G\kappa_x^3\|\Sigma^{\frac{1}{2}}\beta_j^*\|_\infty$ and

$$\|\beta_j^{ols} - \frac{1}{c_j}\beta_j^*\|_\infty = \|(\Sigma^{-\frac{1}{2}}(D - \frac{1}{c_j}I)\Sigma^{\frac{1}{2}}\|_\infty)\|\beta_j^*\|_\infty$$
$$\leq \max_{i\in[p]}|D_{ii} - \frac{1}{c_j}|\|\Sigma^{\frac{1}{2}}\|_\infty\|\Sigma^{-\frac{1}{2}}\|_\infty\|\beta_j^*\|_\infty$$
$$\leq 8G\kappa_x^3\rho_\infty\|\Sigma^{\frac{1}{2}}\|_\infty\|\beta_j^*\|_\infty^2.$$

Due to the diagonal dominance property we have

$$\|\Sigma^{\frac{1}{2}}\|_\infty = \max_i \sum_{j=1}^{p}|\Sigma_{ij}^{\frac{1}{2}}| \leq 2\max\Sigma_{ii}^{\frac{1}{2}} \leq 2\|\Sigma\|_2^{\frac{1}{2}}.$$

Since we have $\|x\|_2 \leq r$, we write

$$r^2 \geq \mathbb{E}[\|x\|_2^2] = \text{Trace}(\Sigma) \geq p\|\Sigma \geq \frac{p\|\Sigma\|_2}{\rho_2}.$$

Thus we have $\|\Sigma^{\frac{1}{2}}\|_\infty \leq 2r\sqrt{\frac{\rho_2}{p}}$. $\square$

Since Theorem 5 is more complex than Theorem 2, we will proof Theorem 5 first. Before that, we need to show the following lemma.

Lemma 4. *Under Assumption 2, if we define the function $\ell_j(c,\beta) = c\mathbb{E}[f'_j(\langle x,\beta\rangle c)]$ and its empirical counter part as*

$$\hat{\ell}_j(c,\beta) = \frac{c}{n}\sum_{i=1}^{n} f'_j(\langle x_i,\beta\rangle c).$$

Assume that each x_i are i.i.d sub-Gaussian with $\|x_i\|_{\psi_2} \leq \kappa_x$ and there exist some constant $\eta, \bar{c}_j$ such that $\ell_j(\bar{c}_j, \beta_j^{ols}) > 1 + \eta$. Then there exists $\tilde{c}_j > 0$ satisfying the equation $1 = \ell_j(\tilde{c}_j, \beta_j^{ols})$. Further, assume that for sufficiently large n such that

$$n \geq \Omega\left(\frac{p\|\Sigma\|_2\|\beta_j^{ols}\|_2^2 G^2 L^2 \bar{c}_j^4 \kappa_x^4 \gamma^2}{\eta^2}\right).$$

Then, with probability at least $1 - k\exp(-p)$ there exist constants $\hat{c}_j \in (0, \bar{c}_j)$ satisfying the equations

$$1 = \frac{\hat{c}_j}{n}\sum_{i=1}^{n} f'_j(\langle x_i, \tilde{\beta}_j^{ols}\rangle \hat{c}_j).$$

Moreover, if for all $j \in [k]$ the derivative of $z \mapsto \ell_j(z, \beta_j^{ols})$ is bounded below in absolute value (does not change sign) by $M > 0$ in the interval $z \in [0, \bar{c}_j]$. Then with probability at least $1 - 4k\exp(-p)$ for all $j \in [k]$ we have

$$|\hat{c}_j - \tilde{c}_j| \leq O\left(M^{-1}GL\bar{c}_j^2\kappa_x^2\gamma\|\Sigma\|_2^{1/2}\|\beta_j^{ols}\|_2\sqrt{\frac{p}{n}}\right).$$

Proof of Lemma 4. We first proof the existence of $\tilde{c}_j$. By the definition we know that $\ell_j(0, \beta_j^{ols}) = 0$ and $\ell_j(\bar{c}_j, \beta_j^{ols}) > 1$, since ℓ_j is continuous, there must exists a $\tilde{c}_j$ which satisfies $\ell_j(\tilde{c}_j, \beta_j^{ols}) = 1$.

Secondly, we will proof the existence of $\hat{c}_j \in (0, \bar{c}_j)$. Denote $v_i = \Sigma^{-\frac{1}{2}}x_i$ as the whitened random variable. Note that by Assumption 1 we know that z_jy is sub-Gaussian with norm $\|z_jy\|_{\psi_2} \leq \gamma$, thus Lemma 1 holds with γ. We denote the upper bound in Lemma 1 as δ, that is $\delta = C_1\kappa_x\gamma\sqrt{\frac{p}{n}}$ and suppose the event in Lemma 1 holds. Then we have

$$\tilde{\beta}_j^{ols} \in B_\Sigma^\delta(\beta_j^{ols}) = \{\beta : \|\Sigma^{\frac{1}{2}}(\beta - \beta_j^{ols})\|_2 \leq \delta\}.$$

Thus we have for all $c \in (0, \bar{c}_j]$

$$|\hat{\ell}_j(c, \tilde{\beta}_j^{ols}) - \ell_j(c, \tilde{\beta}_j^{ols})| \tag{13}$$

$$\leq \sup_{\beta \in B_\Sigma^\delta(\beta_j^{ols})} |\hat{\ell}_j(c, \beta) - \ell_j(c, \beta)|$$

$$\leq \sup_{c \in (0, \bar{c}_j]} \sup_{\beta \in B_\Sigma^\delta(\beta_j^{ols})} |\hat{\ell}_j(c, \beta) - \ell_j(c, \beta)|$$

$$\leq \bar{c}_j \sup_{c \in (0, \bar{c}_j]} \sup_{\beta \in B_\Sigma^\delta(\beta_j^{ols})} |\frac{1}{n}\sum_{i=1}^{n} f'_j(\langle x_i, \beta\rangle c) - \mathbb{E}[f'_j(\langle x, \beta\rangle c)]|$$

$$= \bar{c}_j \sup_{c \in (0, \bar{c}_j]} \sup_{\beta \in B_\Sigma^\delta(\beta_j^{ols})} |\frac{1}{n}\sum_{i=1}^{n} f'_j(\langle v_i, \Sigma^{\frac{1}{2}}\beta\rangle c) - \mathbb{E}[f'_j(\langle v, \Sigma^{\frac{1}{2}}\beta\rangle c)]|$$

$$= \bar{c}_j \sup_{c \in (0, \bar{c}_j]} \sup_{\beta \in B^\delta(\bar{\beta}_j^{ols})} |\frac{1}{n}\sum_{i=1}^{n} f'_j(\langle v_i, \beta\rangle c) - \mathbb{E}f'_j(\langle w, \beta\rangle c)|$$

$$= \bar{c}_j \sup_{\beta \in B^{\bar{c}_j\delta}(\bar{\beta}_j^{ols})} |\frac{1}{n}\sum_{i=1}^{n} f'_j(\langle v_i, \beta\rangle) - \mathbb{E}[f'_j(\langle w, \beta\rangle)]|. \tag{14}$$

Where the second equality is due to that $\beta \in B_\Sigma^\delta(\beta_j^{ols})$ is equivalent to $\Sigma^{\frac{1}{2}}\beta \in B^\delta(\bar{\beta}_j^{ols})$.

Thus by Lemma 2 we have with probability at least $1 - \exp(-p)$ (we have $n \geq \Omega(p\kappa_x^2\gamma^2\bar{c}_j^2)$)

$$\sup_{\beta \in B^{\bar{c}_j\delta}(\bar{\beta}_j^{ols})} |\frac{1}{n}\sum_{i=1}^{n} f'_j(\langle v_i, \beta\rangle) - \mathbb{E}[f'_j(\langle w, \beta\rangle)]|$$

$$\leq 2(G(\|\bar{\beta}_j^{ols}\| + \bar{c}_j\delta)\|I\|_2 + L)\sqrt{\frac{p}{n}}$$

$$= O(GL\kappa_x\gamma\|\Sigma\|_2^{1/2}\|\beta_j^{ols}\|_2\sqrt{\frac{p}{n}}). \tag{15}$$

Moreover by the G-Lipschitz property of f'_j we have for any β_1, β_2 and $c \in [0, \bar{c}_j]$,

$$|\ell_j(c, \beta_1) - \ell_j(c, \beta_2)| \leq G\bar{c}_j^2\mathbb{E}[\langle v, \Sigma^{\frac{1}{2}}(\beta_1 - \beta_2)\rangle]$$

$$\leq G\bar{c}_j^2\kappa_x\|\Sigma^{\frac{1}{2}}(\beta_1 - \beta_2)\|_2\mathbb{E}[\|v\|_2]$$

$$\leq O(G\bar{c}_j^2\kappa_x\|\Sigma^{\frac{1}{2}}(\beta_1 - \beta_2)\|_2). \tag{16}$$

Where the last inequality is due the definition of κ_x.

Take $\beta_1 = \tilde{\beta}_j^{ols}$ and $\beta_2 = \beta_j^{ols}$ and by Lemma 1 we have

$$|\ell_j(c, \tilde{\beta}_j^{ols}) - \ell_j(c, \beta_j^{ols})| \leq O(G\bar{c}_j^2\kappa_x^2\gamma\sqrt{\frac{p}{n}}). \tag{17}$$

Combining with (14), (15), (16) and (17), in total we have for all $c \in (0, \bar{c}_j]$

$$|\hat{\ell}_j(c, \tilde{\beta}_j^{ols}) - \ell_j(c, \beta_j^{ols})| \leq O(GL\bar{c}_j^2\kappa_x^2\gamma\|\Sigma\|_2^{1/2}\|\beta_j^{ols}\|_2\sqrt{\frac{p}{n}}). \tag{18}$$

Thus, when $c = \bar{c}_j$ we have

$$\hat{\ell}_j(\bar{c}_j, \tilde{\beta}_j^{ols}) \geq 1 + \eta - O(GL\bar{c}_j^2\kappa_x^2\gamma\|\Sigma\|_2^{1/2}\|\beta_j^{ols}\|_2\sqrt{\frac{p}{n}}). \tag{19}$$

7

Thus when $n \geq \Omega(\frac{p\|\Sigma\|_2\|\beta_j^{ols}\|_2^2 G^2 L^2 \bar{c}_j^4 \kappa_x^4 \gamma^2}{\eta^2})$. The left term of (19) is greater than 1, also by the definition we have $\hat{\ell}_j(0, \tilde{\beta}_j^{ols}) = 0$. Thus, there exists a $\hat{c}_j$ such that $\hat{\ell}_j(\hat{c}_j, \tilde{\beta}_j^{ols}) = 1$.

Thirdly, we will show the approximation error of $\hat{c}_j$ w.r.t $\tilde{c}_j$. Using the Taylor's series expansion of $c \mapsto \ell_j(c, \beta_j^{ols})$ around $\tilde{c}_j$ and the assumption on the derivative of ℓ with respect to its first argument, we obtain

$$
\begin{aligned}
M|\hat{c}_j - \tilde{c}_j| &\leq |\ell_j(\hat{c}_j, \beta_j^{ols}) - \ell_j(\tilde{c}_j, \beta_j^{ols})| \\
&= |\ell_j(\hat{c}_j, \beta_j^{ols}) - \ell_j(\hat{c}_j, \tilde{\beta}_j^{ols})| + |\ell_j(\hat{c}_j, \tilde{\beta}_j^{ols}) - 1| \\
&= |\ell_j(\hat{c}_j, \beta_j^{ols}) - \ell_j(\hat{c}_j, \tilde{\beta}_j^{ols})| + |\ell_j(\hat{c}_j, \tilde{\beta}_j^{ols}) - \hat{\ell}_j(\hat{c}_j, \tilde{\beta}_j^{ols})| \\
&\leq O(GL\bar{c}_j^2 \kappa_x^2 \gamma \|\Sigma\|_2^{1/2} \|\beta_j^{ols}\|_2 \sqrt{\frac{p}{n}}).
\end{aligned}
$$

Where the last inequality is due to (18) and (14). $\qquad\square$

Proof of Theorem 5 . We have for each $\hat{\beta}_j^{nlr}$

$$
\begin{aligned}
\|\hat{\beta}_j^{nlr} - \beta_j^*\|_\infty &\leq \|\hat{c}_j \tilde{\beta}_j^{ols} - \tilde{c}_j \beta_j^{ols}\|_\infty + \|\tilde{c}_j \beta_j^{ols} - \beta_j^*\|_\infty \\
&\leq \|\hat{c}_j \tilde{\beta}_j^{ols} - \tilde{c}_j \beta_j^{ols}\|_\infty + \|\tilde{c}_j \beta_j^{ols} - c_j \beta_j^{ols}\|_\infty + \|c_j \beta_j^{ols} - \beta_j^*\|_\infty.
\end{aligned}
\tag{20}
$$

We first bound the term of $|\tilde{c}_j - c_j|$. By definition we have $c_j \mathbb{E}[f_j'(\langle x, \beta_j^* \rangle)] = 1$ and $\tilde{c}_j \mathbb{E}[f_j'(\langle x, \beta_j^{ols} \rangle \tilde{c}_j)] = 1$, we get

$$
\begin{aligned}
|\ell_j(\tilde{c}_j, \beta_j^{ols}) - \ell_j(c_j, \beta_j^{ols})| &= |1 - \ell_j(c_j, \beta_j^{ols})| \\
&= |c_j \mathbb{E}[f_j'(\langle x, \beta_j^* \rangle)] - c_j \mathbb{E}[f_j'(\langle x, \beta_j^{ols} \rangle c_j)]| \\
&\leq G|c_j| \mathbb{E}[\langle x, \beta_j^* - c_j \beta_j^{ols} \rangle] \\
&\leq Gc_j \|\beta_j^* - c_j \beta_j^{ols}\|_\infty \mathbb{E}\|x\|_1 \\
&\leq Gc_j \kappa_x \|\beta_j^* - c_j \beta_j^{ols}\|_\infty.
\end{aligned}
$$

Thus, by the assumption of the bounded deviation of $\ell(c, \beta_j^{ols})$ on $\max\{\bar{c}_j, c_j\}$ we have

$$
M|\tilde{c}_j - c_j| \leq |\ell_j(\tilde{c}_j, \beta_j^{ols}) - \ell_j(c_j, \beta_j^{ols})| \leq Gc_j \kappa_x \|\beta_j^* - c_j \beta_j^{ols}\|_\infty.
$$

Thus by Theorem 4 we have

$$
|\tilde{c}_j - c_j| \leq O(M^{-1} rG^2 c_j^2 \kappa_x^4 \sqrt{\rho_2} \rho_\infty \frac{\|\beta_j^*\|_\infty^2}{\sqrt{p}}).
\tag{21}
$$

Thus for the second term of (20) we have

$$
\|\tilde{c}_j \beta_j^{ols} - c_j \beta_j^{ols}\|_\infty \leq O(M^{-1} rG^2 c_j^2 \kappa_x^4 \sqrt{\rho_2} \rho_\infty \frac{\|\beta_j^{ols}\|_\infty \|\beta_j^*\|_\infty^2}{\sqrt{p}}).
\tag{22}
$$

For the first term of (20) we have

$$
\begin{aligned}
\|\hat{c}_j \tilde{\beta}_j^{ols} - \tilde{c}_j \beta_j^{ols}\|_\infty &\leq |\hat{c}_j| \|\tilde{\beta}_j^{ols} - \beta_j^{ols}\|_\infty + |\hat{c}_j - \tilde{c}_j| \|\beta_j^{ols}\|_\infty. \\
&\leq O((\tilde{c}_j + M^{-1} GL\bar{c}_j^2 \kappa_x^3 \gamma^2 \|\Sigma\|_2^{1/2} \|\lambda_{\min}(\Sigma)^{-\frac{1}{2}} \beta_j^{ols}\|_2 \frac{p}{n}) \\
&\quad + M^{-1} GL\bar{c}_j^2 \kappa_x^2 \gamma \|\Sigma\|_2^{1/2} \|\beta_j^{ols}\|_2 \sqrt{\frac{p}{n}} \|\beta_j^{ols}\|_\infty) \\
&\leq O(\bar{c}_j \kappa_x \gamma \lambda_{\min}^{-\frac{1}{2}}(\Sigma) \sqrt{\frac{p}{n}} \max\{1, \|\beta_j^{ols}\|_\infty\}).
\end{aligned}
\tag{23}
$$

By Theorem 4 we have the third term of (20) is bounded by

$$
\|c_j \beta_j^{ols} - \beta_j^*\|_\infty \leq O(c_j Gr\kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|\beta_j^*\|_\infty^2}{\sqrt{p}}).
\tag{24}
$$

By (22), (23) and (24) we have

$$
\begin{aligned}
\|\hat{\beta}_j^{nlr} - \beta_j^*\|_\infty &\leq O(M^{-1} rG^2 c_j^2 \kappa_x^4 \sqrt{\rho_2} \rho_\infty \frac{\|\beta_j^{ols}\|_\infty \|\beta_j^*\|_\infty^2}{\sqrt{p}} + \\
&\quad \bar{c}_j \kappa_x \gamma \lambda_{\min}^{-\frac{1}{2}}(\Sigma) \sqrt{\frac{p}{n}} \max\{1, \|\beta_j^{ols}\|_\infty\} + c_j Gr\kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|\beta_j^*\|_\infty^2}{\sqrt{p}}).
\end{aligned}
\tag{25}
$$

Take this into the previous equation and take $G, L, M, r, \kappa_x, \gamma, C, c_j, \bar{c}_j = \Theta(1)$ we can get

$$
\begin{aligned}
\|\hat{\beta}_j^{nlr} - \beta_j^*\|_\infty &\leq O(\sqrt{\rho_2} \rho_\infty \frac{\|\beta_j^*\|_\infty^2 \max\{1, \|\beta_j^{ols}\|_\infty\}}{\sqrt{p}} \\
&\quad + \lambda_{\min}^{-\frac{1}{2}}(\Sigma) \sqrt{\frac{p}{n}} \max\{1, \|\beta_j^{ols}\|_\infty\})
\end{aligned}
\tag{26}
$$

By Theorem 4 we have

$$
\|\beta_j^{ols}\|_\infty \leq O(\|\frac{1}{c_j} \beta_j^*\|_\infty + Gr\kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|\beta_j^*\|_\infty^2}{\sqrt{p}}).
$$

Thus we have

$$
\begin{aligned}
\|\hat{\beta}_j^{nlr} - \beta_j^*\|_\infty &\leq O(\max\{1, \sqrt{\rho_2} \rho_\infty\} \lambda_{\min}^{-\frac{1}{2}}(\Sigma) \sqrt{\frac{p}{n}} \|\beta_j^*\|_\infty \max\{1, \|\beta_j^*\|_\infty\} \\
&\quad + \rho_2 \rho_\infty^2 \frac{\max\{\|\beta_j^*\|_\infty^2, 1\} \|\beta_j^*\|_\infty^2}{\sqrt{p}}).
\end{aligned}
$$

Note that the equality holds when

$$
n \geq \Omega(\frac{p\|\Sigma\|_2 \|\beta_j^{ols}\|_2^2 G^2 L^2 \bar{c}_j^4 \kappa_x^4 \gamma^2}{\eta^2}).
$$

Which will holds when

$$
\begin{aligned}
n &\geq \Omega(\frac{\|\Sigma\|_2 p^2 G^4 L^2 \bar{c}_j^4 \kappa_x^{10} \max\{\rho_2 \rho_\infty^2, 1\} r^2 \gamma^2 \|\beta_j^*\|_\infty^2 \max\{1, \|\beta_j^*\|_\infty^2\}}{\eta^2} \\
&\quad \max\{1, \frac{1}{c_j^2}\}) \geq \Omega(\frac{p^2 \|\Sigma\|_2 \|\beta_j^{ols}\|_\infty^2 G^2 L^2 \bar{c}_j^4 \kappa_x^4 \gamma^2}{\eta^2})
\end{aligned}
\tag{27}
$$

$\qquad\square$

Proof of Theorem 2. Since x is Gaussian, Lemma 1 holds with $\kappa_x = O(1)$. The proof of Theorem 2 is almost the same as in the proof of Theorem 4.

By Theorem 1 we have for each $\hat{\beta}_j^{nlr}$

$$
\|\hat{\beta}_j^{nlr} - \beta_j^*\|_2 = \|\hat{c}_j \tilde{\beta}_j^{ols} - c_j \beta_j^{ols}\|_2
\tag{28}
$$

$$
\leq \hat{c}_j \|\tilde{\beta}_j^{ols} - \beta_j^{ols}\|_2 + |\hat{c}_j - c_j| \|\beta_j^{ols}\|_2.
\tag{29}
$$

8

For the first term of (29) we have

$$\hat{c}_j \|\tilde{\beta}_j^{ols} - \beta_j^{ols}\|_2 \le (c_j + M^{-1}GL\bar{c}_j^2\gamma^2\|\Sigma\|_2^{1/2}\|\beta_j^{ols}\|_2\frac{p}{n})\lambda_{\min}^{-\frac{1}{2}}(\Sigma)$$

$$\le O(c_j\gamma\lambda_{\min}^{-\frac{1}{2}}(\Sigma)\sqrt{\frac{p}{n}}). \qquad (30)$$

For the second term of (29) we have

$$|\hat{c}_j - c_j|\|\beta_j^{ols}\|_2 \le O(M^{-1}GL\bar{c}_j^2\gamma\|\Sigma\|_2^{1/2}\|\beta_j^{ols}\|_2\sqrt{\frac{p}{n}}\frac{1}{c_j}\|\beta_j^*\|_2). \qquad (31)$$

In total we have

$$\|\hat{\beta}_j^{nlr} - \beta_j^*\|_2 \le O(c_j\gamma\lambda_{\min}^{-\frac{1}{2}}(\Sigma)\sqrt{\frac{p}{n}}$$

$$+ M^{-1}GLc_j^2\gamma\|\Sigma\|_2^{1/2}\|\beta_j^{ols}\|_2\sqrt{\frac{p}{n}}\frac{1}{c_j}\|\beta_j^*\|_2) \qquad (32)$$

If we assume $\eta, c_j, \gamma, C, M, G, L = \Theta(1)$ we can get

$$\|\hat{\beta}_j^{nlr} - \beta_j^*\|_2 \le O(\max\{1, \|\beta_j^*\|_2\}\lambda_{\min}^{-\frac{1}{2}}\sqrt{\frac{p}{n}}). \qquad (33)$$

The sample show satisfies

$$n \ge \Omega(\frac{p\|\Sigma\|_2\|\beta_j^{ols}\|_2^2G^2L^2\bar{c}_j^4\gamma^2}{\eta^2}) = \Omega(\frac{p\|\Sigma\|_2\|\beta_j^*\|_2^2G^2L^2\bar{c}_j^4\gamma^2}{\eta^2}\frac{1}{c_j^2})$$

$\square$

Proofs of Theorem 3 and 6. The proofs of Theorem 3 and 6 are almost the same as them of Theorem 2 and 5, respectively. Instead of using Lemma 1 here we will use Lemma 3. For convenience we omit them here. $\square$

Proof of Theorem 7 . The proof is motivated by [29]. For the function $f(z) = \frac{1}{1+e^{-z}}$ we have

$$f'(z) = \frac{e^z}{(1+e^z)^2}, f^{(3)}(z) = \frac{e^z(1-4e^z+e^{2z})}{(1+e^z)^4}.$$

It is notable that both of $f'(z)$ and $f^{(3)}$ are even functions. We will use the local convexity for $z \ge 0$ around $z = 2.5$, thus we can get

$$f'(z) \ge f'(2.5) + f^{(2)}(2.5)(z-2.5) = a - bz, \qquad (34)$$

where $a = f'(2.5) - 2.5f^{(2)}(2.5)$ and $b = -f^{(2)}(2.5)$. Denote $W = \mathcal{N}(0,1)$ and ϕ as the density of the standard Gaussian distribution, by the assumption we have

$$\ell(z,\beta^{ols}) = z\mathbb{E}[f'(\langle x, \beta^{ols}\rangle)z] = z\mathbb{E}[f'(\frac{Wz}{20})]$$

$$= 2z\int_0^\infty f'(\frac{wz}{20})\phi(w)dw$$

$$\ge 2z\int_0^{\frac{20a}{bz}}(a - \frac{bwz}{20})\phi(w)dw$$

$$= 2z\{a\Phi(\frac{20a}{bz}) - \frac{a}{2} - \frac{bz}{20\sqrt{2\pi}}(1-\exp(-\frac{200a^2}{b^2z^2}))\}.$$

Take $z = 6$ and $\eta = 0.22$ we have $\ell(z,\beta^{ols}) > 1 + \eta$. Next we will prove that $\ell'(z,\beta^{ols})$ is lower bounded by 0.19.

By the Stein's lemma (the special case of Definition 3 with z is Gaussian) we can get

$$\ell'(z,\beta^{ols}) = \mathbb{E}[f'(\frac{Wz}{20})] + (\frac{z}{20})^2\mathbb{E}[f^{(3)}(\frac{Wz}{20})]$$

$$\ge 2\{a\Phi(\frac{20a}{bz}) - \frac{a}{2} - \frac{bz}{20\sqrt{2\pi}}(1-\exp(-\frac{200a^2}{b^2z^2}))\}$$

$$- \frac{9}{100}\max|f^{(3)}| \ge 0.19.$$

$\square$

7. Experiments

We conduct experiments on three types of link functions: polynomial, sigmoid, and logistic function, as well as an arbitrary mix of them. Formally, the polynomial link functions include $f(x) = x, x^3, x^5$; the sigmoid link function is defined as $f(x) = (1 + e^{-x})^{-1}$; the logistic link function refers to $f(x) = \log(1 + e^{-x})$. Due to the statistical setting we focused on in the paper, we will perform our algorithm on the synthetic data, and the same experimental setting has been used in the previous work such as [16; 13; 24].

Experimental setting. We sample all coefficient $z_{i,j}$ and noise ϵ i.i.d. from standard Gaussian distribution $N(0,1)$ across each experiment. Each β_j^* is generated by sampling from $N(1, 16\mathbb{I}_d)$. We consider two distributions for generating x: Gaussian and Uniform distribution (corresponds to thr sub-Gaussian case). To satisfy the requirement of Theorem 7, in both cases the standard variance is scaled by $1/p$ and this is also used in [24], where p is the data dimension. Thus, in the Gaussian case, $x \sim N(0, \frac{1}{p}\mathbb{I}_p)$, while in the sub-Gaussian case x is sampled from a uniform distribution, i.e., $x \sim U([-1/p, 1/p])^p$. Finally, given the list $F = [f_1, \ldots, f_k]$ of link functions, the response y is computed via (1). It is notable that the experimental results with different number of link functions k are incomparable since when k is changed Model (1) will also be changed.

These experiments are divided into two parts, examining how the sample size n and the size of the sub-sample set S affect the algorithm performance. In the first part we vary n from $100\,000$ to $500\,000$ with fixed $p = 20$ and $|S| = n$, while in the second part we vary $|S|$ from $0.01n$ to n, with fixed $n = 500\,000$ and $p = 20$. In each part we test the algorithm against various data distribution/link function combinations. For each experiment, in order to support our theoretical analysis, we will use the (maximal) relative error as the measurement, that is when x is Gaussian we use $\max_{i\in[k]}\frac{\|\beta_i - \beta_i^*\|_2}{\|\beta_i^*\|_2}$ and when x is sub-Gaussian we will use $\max_{i\in[k]}\frac{\|\beta_i - \beta_i^*\|_\infty}{\|\beta_i^*\|_\infty}$. For each experiment we repeat 20 times and take the average as the final output.

Experiment results. Each of Figure 1-3 illustrates the result for a single type of link function. We can see that the relative error decreases steadily as the sample size n grows which is due to the $O(\frac{1}{\sqrt{n}})$ converge rate as our theorem states. Also, we can see that the size of S doesn't affect the final relative error much

if $|S|$ is large enough, *i.e., in all cases, choosing large enough $|S| \geq 0.2n$ is sufficient to achieve a relative error roughly the same as when $|S| = n$,* which also has been mentioned in our theorems theoretically.

We further investigate the case when F contains different types of link functions. In Figure 4a, we let F contain polynomials with different degrees (x, x^3, x^5), and there are roughly $\frac{k}{3}$ functions for each degree. Similarly, in Figure 4b we also mix polynomial links with the other two types of links, i.e., logistic link and log-exponential link, and there are roughly $\frac{k}{3}$ functions for each type of link function. Our algorithm achieves similar performance as in the previous settings.
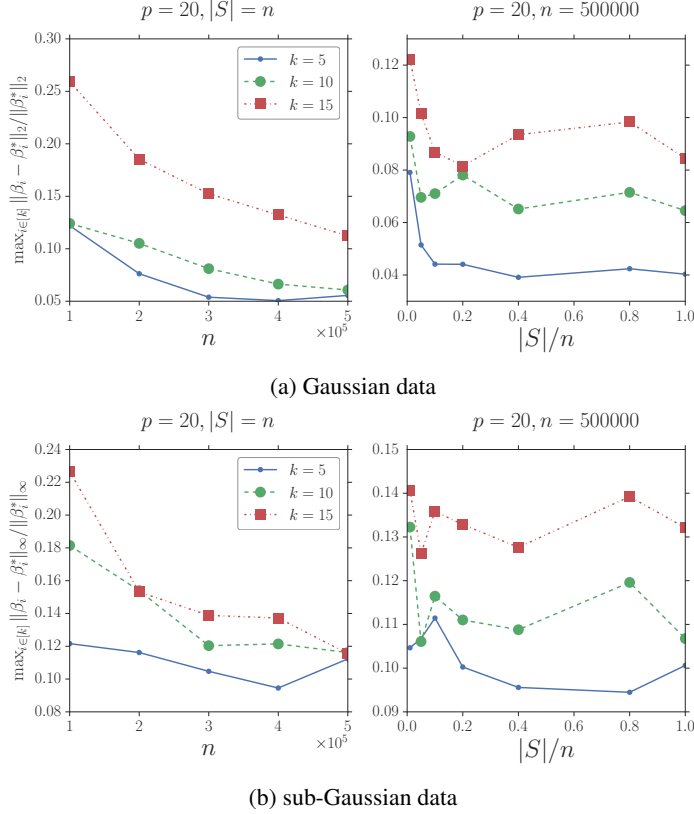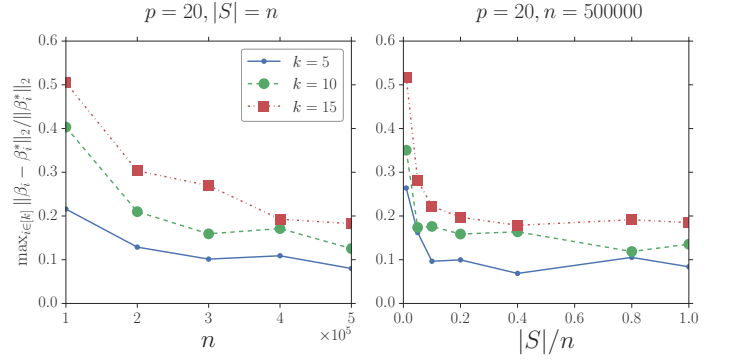

(a) Gaussian data

(b) sub-Gaussian data

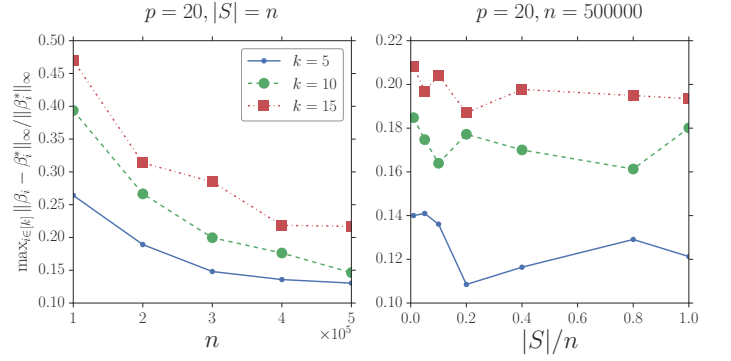Figure 1: Single type of link function $f(x) = x^3$

8. Conclusion

We studied a new model called *stochastic linear combination of non-linear regressions* and provided the first estimation error bounds for both Gaussian and bounded sub-Gaussian cases. Our algorithm is based on Stein's lemma and its generalization, the zero-bias transformation. Moreover, we used the subsampling of the covariance matrix to accelerate our algorithm. Finally, we conducted experiments whose results support our theoretical analysis.

There are still many open problems: 1) In the paper, we only focused on the low dimensional case (where $n \gg p$). However, it is unknown whether we can extend our methods to the high dimensional sparse case. 2) In the paper we assume x is either
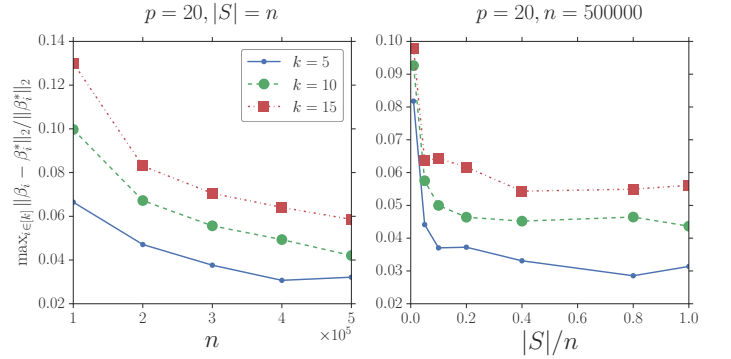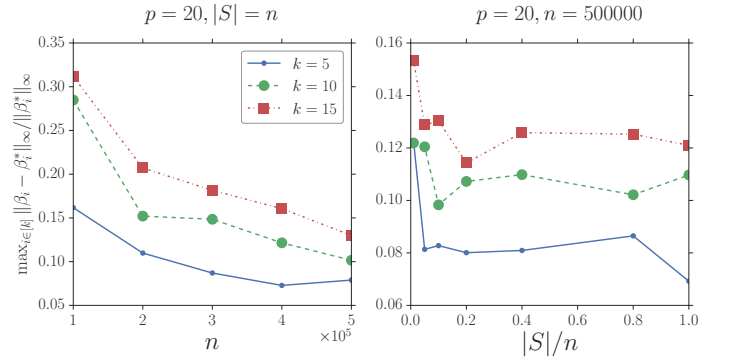

(a) Gaussian data

(b) sub-Gaussian data

Figure 2: Single type of link function $f(x) = (1 + e^{-x})^{-1}$

(a) Gaussian data

(b) sub-Gaussian data

Figure 3: Single type of link function $f(x) = \log(1 + e^{-x})$

(a) $f(x) = x, x^3, x^5$, sub-Gaussian data

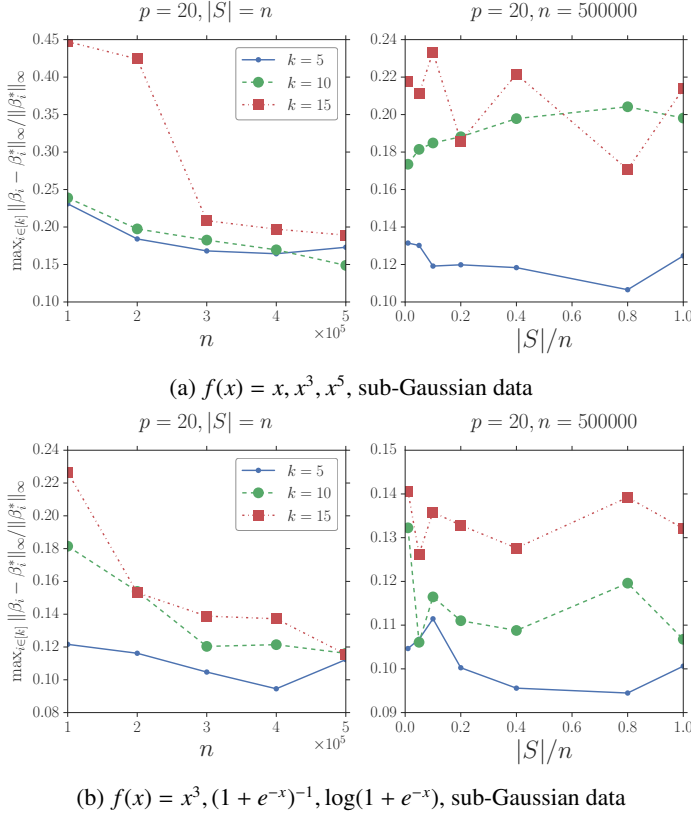(b) $f(x) = x^3, (1 + e^{-x})^{-1}, \log(1 + e^{-x})$, sub-Gaussian data

Figure 4: Mixed different type of link functions

Gaussian or Sub-Gaussian, we do not know its behaviors in the Sub-Exponential distribution case. 3) Where we can further accelerate our methods is still unknown. We will leave these problems as future work.

References

[1] K. Zhang, Z. Yang, Z. Wang, Nonlinear structured signal estimation in high dimensions via iterative hard thresholding, in: International Conference on Artificial Intelligence and Statistics, 2018, pp. 258–268.

[2] A. Beck, Y. C. Eldar, Sparse signal recovery from nonlinear measurements, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 5464–5468.

[3] Z. Yang, Z. Wang, H. Liu, Y. Eldar, T. Zhang, Sparse nonlinear regression: parameter estimation under nonconvexity, in: International Conference on Machine Learning, 2016, pp. 2472–2481.

[4] R. D. Cook, H. Lee, Dimension reduction in binary response regression, Journal of the American Statistical Association 94 (448) (1999) 1187–1200.

[5] H. Ichimura, Semiparametric least squares (sls) and weighted sls estimation of single-index models, Journal of Econometrics 58 (1-2) (1993) 71–120.

[6] J. L. Horowitz, Semiparametric and nonparametric methods in econometrics, Vol. 12, Springer, 2009.

[7] S. M. Kakade, V. Kanade, O. Shamir, A. Kalai, Efficient learning of generalized linear and single index models with isotonic regression, in: Advances in Neural Information Processing Systems, 2011, pp. 927–935.

[8] Z. Yang, K. Balasubramanian, H. Liu, High-dimensional non-gaussian single index models via thresholded score function estimation, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 3851–3860.

[9] P. Radchenko, High dimensional single index models, Journal of Multivariate Analysis 139 (2015) 266–282.

[10] K.-C. Li, Sliced inverse regression for dimension reduction, Journal of the American Statistical Association 86 (414) (1991) 316–327.

[11] K.-C. Li, On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma, Journal of the American Statistical Association 87 (420) (1992) 1025–1039.

[12] K.-C. Li, N. Duan, et al., Regression analysis under link violation, The Annals of Statistics 17 (3) (1989) 1009–1052.

[13] Z. Yang, K. Balasubramanian, Z. Wang, H. Liu, Learning non-gaussian multi-index model via second-order steins method, Advances in Neural Information Processing Systems 30 (2017) 6097–6106.

[14] S. Ma, P. X.-K. Song, Varying index coefficient models, Journal of the American Statistical Association 110 (509) (2015) 341–356.

[15] J. Fan, W. Zhang, Statistical methods with varying coefficient models, Statistics and its Interface 1 (1) (2008) 179.

[16] S. Na, Z. Yang, Z. Wang, M. Kolar, High-dimensional index volatility models via stein's identity, arXiv preprint arXiv:1811.10790.

[17] P. Dhillon, Y. Lu, D. P. Foster, L. Ungar, New subsampling algorithms for fast least squares regression, in: Advances in neural information processing systems, 2013, pp. 360–368.

[18] L. Goldstein, G. Reinert, et al., Stein's method and the zero bias transformation with application to simple random sampling, The Annals of Applied Probability 7 (4) (1997) 935–952.

[19] D. R. Brillinger, A generalized linear model with" gaussian" regressor variables, A Festschrift For Erich L. Lehmann (1982) 97.

[20] D. Wang, X. Guo, C. Guan, S. Li, J. Xu, Scalable estimating stochastic linear combination of non-linear regressions, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York City, New York, USA, February 7-12, 2020., 2020.

[21] X. Zhang, Y. Yu, L. Wang, Q. Gu, Learning one-hidden-layer relu networks via gradient descent, in: The 22nd International Conference on Artificial Intelligence and Statistics, 2019, pp. 1524–1534.

[22] D. Zou, Y. Cao, D. Zhou, Q. Gu, Stochastic gradient descent optimizes over-parameterized deep relu networks, arXiv preprint arXiv:1811.08888.

[23] A. Nitanda, T. Suzuki, Refined generalization analysis of gradient descent for over-parameterized two-layer neural networks with smooth activations on classification problems, arXiv preprint arXiv:1905.09870.

[24] M. A. Erdogdu, L. H. Dicker, M. Bayati, Scaled least squares estimator for glms in large-scale problems, in: Advances in Neural Information Processing Systems, 2016, pp. 3324–3332.

[25] M. A. Erdogdu, Newton-stein method: an optimization method for glms via stein's lemma, The Journal of Machine Learning Research 17 (1) (2016) 7565–7616.

[26] Q. Liu, D. Wang, Stein variational gradient descent: A general purpose bayesian inference algorithm, in: Advances In Neural Information Processing Systems, 2016, pp. 2378–2386.

[27] D. Wang, H. Zhang, M. Gaboardi, J. Xu, Estimating smooth glm in non-interactive local differential privacy model with public unlabeled data, arXiv preprint arXiv:1910.00482.

[28] X. Wei, Z. Yang, Z. Wang, On the statistical rate of nonlinear recovery in generative models with heavy-tailed data, in: International Conference on Machine Learning, 2019, pp. 6697–6706.

[29] M. A. Erdogdu, M. Bayati, L. H. Dicker, Scalable approximations for generalized linear problems, The Journal of Machine Learning Research 20 (1) (2019) 231–275.

[30] D. Wang, J. Xu, Large scale constrained linear regression revisited: Faster algorithms via preconditioning, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[31] D. Wang, J. Xu, Faster constrained linear regression via two-step preconditioning, Neurocomputing 364 (2019) 280–296.