Journal of Machine Learning Research 24 (2023) 1-46 Submitted 1/22; Revised 2/23; Published 3/23

Non-Asymptotic Guarantees for Robust Statistical Learning under Infinite Variance Assumption

Lihu Xu LIHUXU@UM.EDU.MO
Department of Mathematics
University of Macau, Taipa Macau, China
and
Zhuhai UM Science & Technology Research Institute, Zhuhai, China.

Fang Yao FYAO@MATH.PKU.EDU.CN
Department of Probability & Statistics; Center for Statistical Science
Peking University, Beijing, China.

Qiuran Yao YB97478@CONNECT.UM.EDU.MO
Department of Mathematics
University of Macau, Taipa Macau, China
and
Zhuhai UM Science & Technology Research Institute, Zhuhai, China.

Huiming Zhang [*] ZHANGHUIMING@BUAA.EDU.CN
Institute of Artificial Intelligence
Beihang University, Beijing, China
and
Zhuhai UM Science & Technology Research Institute, Zhuhai, China.
and
Department of Mathematics
University of Macau, Taipa Macau, China.

Editor: Ji Zhu

Abstract

There has been a surge of interest in developing robust estimators for models with heavy-tailed and bounded variance data in statistics and machine learning, while few works impose unbounded variance. This paper proposes two types of robust estimators, the ridge log-truncated M-estimator and the elastic net log-truncated M-estimator. The first estimator is applied to convex regressions such as quantile regression and generalized linear models, while the other one is applied to high dimensional non-convex learning problems such as regressions via deep neural networks. Simulations and real data analysis demonstrate the robustness of log-truncated estimations over standard estimations.

Keywords: data with infinite variance, excess risk bounds, robust ridge regressions, robust elastic net regressions, robust non-convex regressions, robust deep neural network (DNN) regressions.

[*] Corresponding authors: H. Zhang. The authors are in alphabetical order. This work was carried out when H. Zhang was a Postdoctoral Research Fellow at the University of Macau.

1. Introduction

1.1 Backgrounds

Robust statistics is a traditional topic that has been well studied since the pioneering work of Huber (1964) and Tukey (1960) in 1960s. Distributionally robust learning is nowadays revitalized and invigorating in statistical learning; see Nemirovskij and Yudin (1983) for median-of-means estimators, Baraud et al. (2017) for minimax types estimators, and Catoni (2012) for log-truncated estimators. For further details, we refer the reader to the note by Lerasle (2019) and the review paper by Lugosi and Mendelson (2019) for comprehensive introductions.

Many existing pieces of research on excess risk bounds for robust estimators heavily rely on one or more of the following three assumptions: (1) bounded loss functions (Bartlett and Mendelson, 2006; Yi et al., 2020); (2) bounded Lipschitz condition (Chinot et al., 2019; Shen et al., 2021b); (3) bounded data (Liu and Tao, 2014; Brownlees et al., 2015; Zhang and Zhou, 2018) or unbounded data with sub-Gaussian assumption (Xu et al., 2020; Lecué and Mendelson, 2013; Loh, 2017; Ostrovskii and Bach, 2021). However, there are a lot of statistical models which do not satisfy any of the above three assumptions, see for instance Zhang and Jia (2022) and Chi (2010).

When a distribution has no exponential moment, it is often called heavy-tailed; see Resnick (2007). However, many kinds of data, such as network, finance, and wealth distribution data, only have finite β-th moment with $\beta \in (1,2)$; see Peng and Qi (2017). Zhang and Zhou (2018) studied log-truncated M-estimator for least absolute deviation (LAD) regression under the assumption that the data have 2nd moment, while Chen et al. (2021a) extended their work to the data with β-th moment for $\beta \in (1,2)$.

Most of the minimization problems in machine learning have non-convex loss functions; see *mixture density estimation* in (Khamaru and Wainwright, 2019), *the mixture of two linear regressions* in (Klusowski et al., 2019), *truncated Cauchy non-negative matrix factorization* in (Guan et al., 2017), and regressions under deep neural networks (DNNs) in Fan et al. (2021). In practice, the dimension of DNN regressions is usually much larger than the dimension of input, and the computation costs of training large neural networks may be huge (Frankle and Carbin, 2018). To avoid training over-parameterized DNN, many works propose penalized DNN-based estimators for effectively learning the sparse DNN problems; see Wen et al. (2021); Ohn and Kim (2022).

1.2 Contributions

This paper proposes a log-truncated M-estimator for a large family of statistical regressions and establishes its excess risk bounds under the condition that the data have β-th moment with $\beta \in (1,2)$, our *contributions* are summarized as the following three aspects.

A general function $\lambda(x)$ and the associated ridge regression. We replace the function $x^2/2$ in Catoni's truncation defined by (4) by a new function $\lambda(x)$; the choice of the new function will play a crucial role in our robust learning problems for the data with infinite variance. Table 1 below lists the choices of $\lambda(x)$, which has been reported in literature. In this paper, we propose a new $\lambda(x)$, under certain conditions on λ and loss functions; we establish an error bound for the associated ridge regression, see Theorem 2 below. We allow the dimension p to increase with the sample size n. Because the parameter set Θ is bounded, it seems that our ridge regression is more natural and reasonable.

Ridge regressions with special $\lambda(x) = |x|^\beta/\beta$ and examples. Taking $\lambda(x) = |x|^\beta/\beta$ with $\beta \in (1, 2)$ and applying Theorem 2, we establish a new log-truncated robust estimator in Theorem 4, which not only extends the results about estimators of least absolute deviation (LAD) regressions in Zhang and Zhou (2018) and Chen et al. (2021a), but also covers many other convex loss examples such as robust quantile regressions (QR) and robust generalized linear models (GLMs). For GLMs, we obtain a general result for bounding excess risk and apply it to two typical classifications and count data models: logistic regression and negative binomial regression.

High dimensional non-convex regressions with elastic net and DNN. In the high dimension setting $p \gg n$, we propose a new robust elastic net estimator defined by (8) below and obtain the error bound of the excess risk. As applications, we study the non-convex regressions via DNN, and we apply our results to study several typical regression problems, such as LAD regression and logistic regression. In practice, the DNN regressions can be solved by some algorithms based on stochastic gradient descents (SGDs). Empirical studies, including Boston housing and MNIST datasets are performed well by the proposed robust DNN regression models. We stress that Theorems 4 and 7 below can be applied to many other non-convex regressions, e.g., robust two-component mixed linear regression and robust non-negative matrix factorization, in which one has to design specific algorithms rather than use SGDs.

1.3 Related works

Catoni (2012) put forward a logarithm truncation for mean regression with finite 2nd data and obtained an estimator whose confidence interval has a length comparable with that of the classical mean estimation with sub-Gaussian data. Since then, Catoni's idea has been extensively applied to study regressions and estimations with heavy-tailed data; see Fan et al. (2017); Sun et al. (2020); Wang et al. (2022); Sun (2021). Zhang and Zhou (2018) used Catoni's truncation technique to study a LAD regression for the data with 2nd moment and showed that the associated estimator is consistent in the measure of excess risk. By modifying the logarithm truncation of Catoni, Chen et al. (2021a) extended the results in Zhang and Zhou (2018) to the data with β-th moment for $\beta \in (1, 2)$. Due to the increasing applications whose data do not have 2nd moment, more and more generalizations of Catoni's truncation have been proposed, see Lam and Cheng (2021), and Lee et al. (2020). Xu et al. (2020) studied the excess risk bounds for learning with general non-convex truncated losses, in which $\lambda(x) = O(x^\beta)$ with $\beta = 1$ or $\beta \geq 2$. Under bounded input assumption, Shen et al. (2021c) studied non-asymptotic error bounds of DNN regression models with heavy-tailed error output having a finite β-th moment. For canonical GLMs, Zhu and Zhou (2021) required 4th moment condition on output to study the consistency property of their proposed robust estimators. For robust mean estimation, Minsker (2018), Lam and Cheng (2021) and Lee et al. (2020) also considered the extensions of ψ for different motivations (see the table below).

Both Zhang and Zhou (2018) and Chen et al. (2021a) assumed that the parameters to be estimated are located in a compact set, they did not consider adding a penalty on their

ERM problems. In this paper, we propose a robust ridge regression and a robust elastic net regression, and derive their excess risk bounds. They can be applied to classical statistical models such as QR and GLMs, and to high dimensional non-convex learning problems such as DNN regressions.

Robust and sparse estimation for DNN learning has recently drawn a lot of attentions. Taheri et al. (2021) studied the ℓ_1-regularized neural networks with the specific least square loss, while Wen et al. (2021) derived the risk bound for sparse DNNs regression by $L_{1,\infty}$-weight normalization under the bounded loss assumption. Under the heavy-tailed output, Lederer (2020); Shen et al. (2021a); Fan et al. (2022) studied risk bounds for robust DNN linear regressions by assuming that the input data is bounded or fixed if the loss is LAD or Huber or Cauchy type. In addition to heavy-tailed output, our work first attempts to study the heavy-tailed input setting systematically.

1.4 Notations and organizations

The following notations will be frequently used in the rest of this paper. Define the index set $[n] := \{1, 2, \cdots, n\}$ and let $\mathbb{N}$ be the non-negative integer set. The $\mathbb{R}_+$ denotes the set of positive real numbers. The r.v. is the shorthand for a random variable. Let $q \geq 1$ and $p \in \mathbb{N}$, for $\theta \in \mathbb{R}^p$, define $\|\theta\|_q := (\sum_{j=1}^{p} |\theta_j|^q)^{1/q}$. Define the unit ℓ_2-norm ball $B_2^p(r) := \{x \in \mathbb{R}^p : \|\theta\|_2 \leq r\}$ for $r > 0$, and the ℓ_0-norm ball $B_0^p(s) := \left\{x \in \mathbb{R}^d : \|\theta\|_0 \leq s\right\}$ for $s \in \mathbb{N} \cup \{0\}$. Let $\Theta \subset \mathbb{R}^p$, for an $\varepsilon > 0$, $\mathcal{N}(\Theta, \varepsilon) \subset \mathbb{R}^p$ is an *ε-net* of Θ if for all $x \in \Theta$, there is a $y \in \mathcal{N}(\Theta, \varepsilon)$ such that $\|y - x\|_2 \leq \varepsilon$. The *covering number* $N(\Theta, \varepsilon)$ is the smallest number of closed balls centered at Θ with radii ε whose union covers Θ.

For a probability measure μ and a measurable function f, let $\|f\|_{L^2(\mu)} := [\mathbb{E}_{X \sim \mu} f^2(X)]^{1/2}$ as long as the expectation is finite. Let $a > 0$ and denote by $L^2([0, a]^p)$ the square integrable function space with respect to domain $[0, a]^p$, and define the L^2-norm for a square integrable function g as $\|g\|_{[0,a]^p} := [\int_{[0,a]^p} g^2(x) \mathrm{d}x]^{1/2}$. Let $\lfloor x \rfloor$ be the largest integer strictly smaller than x. A function f is in the γ-Hölder function class with smoothness index $\gamma > 0$ if all partial derivatives of f up to order $\lfloor \gamma \rfloor$ exist and are bounded, and the γ-Hölder function space with domain $D \subset \mathbb{R}^p$ and radius $R > 0$ is defined as

$$\mathcal{C}^\gamma(D, R) = \left\{ f : D \to \mathbb{R} : \sum_{\alpha : \|\alpha\|_1 < \gamma} \|\partial^\alpha f\|_\infty + \sum_{\alpha : \|\alpha\|_1 = \lfloor \gamma \rfloor} \sup_{\substack{x, y \in D \\ x \neq y}} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{|x - y|_\infty^{\gamma - \lfloor \gamma \rfloor}} \leq R \right\}, \qquad (1)$$

where $\partial^\alpha = \partial^{\alpha_1} \ldots \partial^{\alpha_p}$ is multi-index notation with $\alpha = (\alpha_1, \ldots, \alpha_p) \in \mathbb{N}^p$. For two matrices A and B with compatible dimensions, denote $A \succ B$ if $B - A$ is positive definite. Let $\|A\|_F := \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2}$ be the Frobenius norm of a square matrix $A = (a_{ij}) \in \mathbb{R}^{m \times n}$.

The rest of the paper is organized as follows. Section 2 introduces the robust estimator based on the log-truncated loss function and provides three main theorems: Theorems 2,4 and 7, while Sections 3 and 4 give examples for Theorems 4 and 7, respectively. Section 5 includes simulations and real data analysis, which evaluate the effectiveness of the proposed log-truncated estimation for some regressions discussed in Section 3, and Section 6 provides some perspectives for future study.

2. Estimation with log-truncated loss and main results

2.1 Problem setup

We assume that $\{(X_i, Y_i)\}_{i=1}^n$ are a sequence of $\mathbb{R}^d \times \mathbb{R}$-valued independent identically distributed (i.i.d.) r.v.s and that each (X_i, Y_i) is a copy of r.v. (X, Y). Denote loss function by $l(y, x, \theta)$, where $y \in \mathbb{R}$ is the output variable, $x \in \mathbb{R}^d$ is the input variable, and $\theta \in \Theta$ with $\Theta \subset \mathbb{R}^p$ being the hypothesis space. Define

$$R_l(\theta) := \mathbb{E}[l(Y, X, \theta)],$$

the true parameter set $\Theta^* \subset \mathbb{R}^p$ is defined as the collection of all minimizers of $R_l(\cdot)$, i.e.,

$$\Theta^* := \left\{ \theta^* \in \operatorname*{arg\,min}_{\theta \in \Theta} R_l(\theta) \right\}. \tag{2}$$

Any $\theta^* \in \Theta^*$ is called true parameter, it may be not unique since our loss function $l(\cdot, \cdot, \theta)$ may be non-convex.

The estimator of θ^* based on the i.i.d. data $\{(X_i, Y_i)\}_{i=1}^n$ is obtained by the following *empirical risk minimization* (ERM, see Koltchinskii (2011)) problem:

$$\bar{\theta}_n \in \operatorname*{arg\,min}_{\theta \in \Theta} \hat{R}_l(\theta) \quad \text{with} \quad \hat{R}_l(\theta) := \frac{1}{n} \sum_{i=1}^n l(Y_i, X_i, \theta). \tag{3}$$

When the data are heavy tailed, the estimator $\bar{\theta}_n$ in (3) may not be a robust estimator, see for instance Ostrovskii and Bach (2021), Mathieu and Minsker (2021) and the references therein. Given an estimator $\hat{\theta}_n$, its excess risk bound is given by

$$R_l(\hat{\theta}_n) - \inf_{\theta \in \Theta} R_l(\theta).$$

In this paper, inspired by the log-truncation technique put forward by Catoni, i.e. (4) below, we shall study the statistical learning problems with heavy-tailed data by modifying the truncation function ψ in (4). In the sequel, we focus on learning problems under infinite variance data assumption.

2.2 Log-truncated loss and our estimators

Catoni (2012) put forward in his pioneering work a non-deceasing truncation function ψ such that

$$-\log\left(1 - x + \frac{x^2}{2}\right) \le \psi(x) \le \log\left(1 + x + \frac{x^2}{2}\right), \tag{4}$$

and obtained a robust mean estimator for i.i.d. data with finite variance. The truncation function $\psi(x)$ not only reduces the value of the exponential-scaled outliers but also largely retains the data fluctuation in an unbounded way, whereas the classical bounded truncated M-functions often lose the information of the data with large values. Catoni's truncation function has been modified in various ways to study many robust estimators/regressions, see Table 1 for a review.

We replace the function $x^2/2$ in (4) with a function $\lambda(x)$ satisfying the following conditions:

Table 1: Excess risk bound guarantees under various functions $\lambda(x)$ for the log-truncated losses and $\beta \in (1, 2)$.

Higher order functions $\lambda(x)$	References	Moment conditions
0	Xu et al. (2020)	$\mathbb{E}\,\lvert X\rvert^2 < \infty$
$\frac{1}{2}\lvert x\rvert^2$	Catoni (2012)	$\mathbb{E}\,\lvert X\rvert^2 < \infty$
$\left(\frac{\beta-1}{\beta} \vee \sqrt{\frac{2-\beta}{\beta}}\right) \lvert x\rvert^\beta$	Minsker (2018)	$\mathbb{E}\,\lvert X\rvert^\beta < \infty$
$\frac{1}{\beta}\lvert x\rvert^\beta$	Chen et al. (2021a)	$\mathbb{E}\,\lvert X\rvert^\beta < \infty$
$\left[2(\frac{2-\beta}{\beta-1})^{1-2/(1+\varepsilon)} + (\frac{2-\beta}{\beta-1})^{2-2/\beta}\right]^{-\beta/2} \lvert x\rvert^\beta$	Lee et al. (2020)	$\mathbb{E}\,\lvert X\rvert^\beta < \infty$
$\beta^{-\beta/2}(2-\beta)^{1-\beta/2}(\beta-1)^{(\beta-1)}\lvert x\rvert^\beta$	Lam and Cheng (2021)	$\mathbb{E}\,\lvert X\rvert^\beta < \infty$
$\sum_{k=2}^m \frac{x^k}{k!}$, $(m \geq 2)$	Xu et al. (2020)	$\sum_{k=2}^m \frac{\mathbb{E}\lvert X\rvert^k}{k!} < \infty$
the function $\lambda(x)$ in (C.1)	this paper	$\mathbb{E}[\lambda(H_{Y,X})] < \infty$

- (C.1) The function $\lambda(x) : \mathbb{R}_+ \to \mathbb{R}_+$ is a continuous non-decreasing function such that $\lim_{x\to\infty} \frac{\lambda(x)}{x} = \infty$. Moreover, there exist some $c_2 > 0$ and a function $f : \mathbb{R}_+ \to \mathbb{R}_+$ such that
  - (C.1.1) $\lambda(tx) \leq f(t)\lambda(x)$ for all $t, x \in \mathbb{R}_+$, where $\lim_{t\to 0^+} f(t)/t = 0$;
  - (C.1.2) $\lambda(x+y) \leq c_2[\lambda(x) + \lambda(y)]$ for all $x, y \in \mathbb{R}_+$.

We further replace ψ in (4) with ψ_λ which satisfies:

$$-\log\left[1 - x + \lambda(|x|)\right] \leq \psi_\lambda(x) \leq \log\left[1 + x + \lambda(|x|)\right], \quad \forall x \in \mathbb{R}. \tag{5}$$

We assume that our parameter space Θ satisfies

- (C.2): The parameter space $\Theta \subseteq \mathbb{R}^p$ is convex and there exists some $r_n \in (0, \infty)$, which may depend on the size n of the observed data, such that $\|\theta\|_2 \leq r_n$, $\forall \theta \in \Theta$.

The condition (C.2) naturally induces a ridge penalty for the ERM problem (3) as the following:

$$\hat{\theta}_n \in \underset{\theta\in\Theta}{\arg\min}\{\hat{R}_{\psi_\lambda,l,\alpha}(\theta) + \rho\|\theta\|_2^2\} \quad \text{with} \qquad \hat{R}_{\psi_\lambda,l,\alpha}(\theta) := \frac{1}{n\alpha}\sum_{i=1}^n \psi_\lambda[\alpha l(Y_i, X_i, \theta)], \tag{6}$$

where $\alpha > 0$ is a *robustification parameter* to be tuned, and $\rho > 0$ is a *penalty parameter* for ℓ_2-regularization.

As we shall see below, the estimator defined by (6) can only work for the case of $p < n$ with $p = o(n/\log n)$ and thus rules out the high dimensional learning problems with $p > n$. To solve this problem, we assume an s_n-sparsity condition:

$$\Theta^s := \{\theta \in \Theta : \ \|\theta\|_0 \leq s_n\}, \tag{7}$$

and introduce an elastic net (Zou and Hastie, 2005) as follows:

$$\hat{\theta}_n \in \underset{\theta\in\Theta}{\arg\min}\{\frac{1}{n\alpha}\sum_{i=1}^{n}\psi_\lambda[\alpha l(Y_i, X_i, \theta)] + \rho\|\theta\|_2^2 + \gamma\|\theta\|_1\}, \tag{8}$$

where $\rho > 0$ and $\gamma > 0$ are both tuning parameter to be chosen later. Note that the elastic net adopts the idea of lasso, using ℓ_1 penalty to approximately solve optimization problem with ℓ_0 constraint.

The α will be chosen according to the sample size n and tend to 0 as $n \to \infty$. In order to obtain the optimal α as deriving excess risk bounds, we assume the following conditions for further analysis:

- (C.3) **Local Lipschitz condition**: $\exists$ *locally Lipschitz constant* $H_{y,x}$ s.t. $l(y, x, \cdot)$ satisfies
$$|l(y, x, \theta_2) - l(y, x, \theta_1)| \leq H_{y,x}\|\theta_2 - \theta_1\|_2,\ \theta_1, \theta_2 \in \Theta.$$

- (C.4) **Moment condition**: $\mathbb{E}[\lambda(H_{Y,X})] < \infty$.

- (C.5) **The existence of risk function**:
$$R_{\lambda\circ l}(\Theta) := \sup_{\theta\in\Theta} R_{\lambda\circ l}(\theta) < \infty,$$
where $R_{\lambda\circ l}(\theta) := \mathbb{E}\{\lambda[l(Y, X, \theta)]\}, \theta \in \Theta$.

Remark 1 *The assumptions (C.1)-(C.5) hold for a large class of examples, including classical regressions and high dimensional non-convex regressions via DNN; see concrete examples in the following sections.*

(C.4) is essentially an assumption on the moments of X and Y. For instance, as $\lambda(x) = |x|^\beta/\beta$ with $\beta \in (1, 2)$, (C.4) implies that $H_{Y,X}$ has β-th moment.

As the data satisfy the condition $\mathbb{E}\{\lambda[l(Y, X, 0)]\} < \infty$, which is true for all the examples in this paper, by (C.2), (C.3) and (C.4), we immediately see that (C.5) holds.

2.3 Main results

In this subsection, we state our main results, Theorems 2, 4 and 7 below, the first theorem is a general result for the ridge regression (6) under the assumptions (C.1)-(C.5), while the other two provide the error bounds of excess risks of the regressions (6) and (8) as $\lambda(x) = |x|^\beta/\beta,\ \beta \in (1, 2)$.

Theorem 2 *Let $\hat{\theta}_n$ be defined by* (6)*. For $\delta \in (0, 1/2)$ and $\kappa > 0$, under (C.1)-(C.5), we have with probability at least $1 - 2\delta$*

$$\begin{aligned} &R_l(\hat{\theta}_n) - \inf_{\theta\in\Theta} R_l(\theta) \\ &\leq 2\kappa\{\mathbb{E}H_{Y,X} + \frac{c_2 f(\alpha\kappa)}{\alpha\kappa}\mathbb{E}[\lambda(H_{Y,X})]\} + \frac{(c_2+1)f(\alpha)}{\alpha}R_{\lambda\circ l}(\Theta) + \frac{1}{n\alpha}\log\frac{N(\Theta,\kappa)}{\delta^2} + \rho\|\Theta^*\|_2^2, \end{aligned}$$

where c_2 is a constant in (C.1.1); $f(t)$ is the function in (C.1.2); $\|\Theta^*\|_2 := \inf_{\theta^*\in\Theta^*}\|\theta^*\|_2$.

In particular, choose $\kappa = 1/n$ and tune α accordingly, then with probability at least $1-2\delta$

$$R_l(\hat{\theta}_n) - \inf_{\theta\in\Theta} R_l(\theta) \le \frac{2\mathbb{E}H_{Y,X}}{n} + \frac{c_2\mathbb{E}[\lambda(H_{Y,X})]}{\alpha} f(\frac{\alpha}{n}) + \frac{2}{n\alpha}\log\frac{N(\Theta, n^{-1})}{\delta^2} + \rho\|\Theta^*\|_2^2, \quad (9)$$

where $\alpha = f^{-1}\left(\frac{1}{n(c_2+1)}[R_{\lambda\circ l}(\Theta)]^{-1}\log\frac{N(\Theta,n^{-1})}{\delta^2}\right)$.

Remark 3 *The terms $\frac{(c_2+1)f(\alpha)}{\alpha}R_{\lambda\circ l}(\Theta)$ and $\frac{1}{n\alpha}\log\frac{N(\Theta,\kappa)}{\delta^2}$ in Theorem 2 can be viewed as variance and bias respectively. We choose the tuning parameter α by setting $\frac{(c_2+1)f(\alpha)}{\alpha}R_{\lambda\circ l}(\Theta) = \frac{1}{n\alpha}\log\frac{N(\Theta,\kappa)}{\delta^2}$. Note that Zhang and Zhou (2018) and Chen et al. (2021a) chose their α without this delicate consideration.*

Under infinite variance assumption, by Theorem 2 we can derive our *second main result*, in which we need the condition $p < n$ but allow p to grow with n.

Theorem 4 *Set $\lambda(x) = |x|^\beta/\beta$, $\beta \in (1,2)$, $\alpha = \frac{1}{n^{1/\beta}}\left[\frac{C_{\delta,n,r}(p)}{(2^{\beta-1}+1)R_{\lambda\circ l}(\Theta)}\right]^{1/\beta}$ in Theorem 2 with $C_{\delta,n,r}(p) := \log(\delta^{-2}) + p\log(1+2nr_n)$, and assume $\mathbb{E}H_{Y,X}^\beta < \infty$. Then, with probability at least $1-2\delta$, one has*

$$\begin{aligned} & R_l(\hat{\theta}_n) - \inf_{\theta\in\Theta} R_l(\theta) \\ \le\ & \frac{2\mathbb{E}H_{Y,X}}{n} + C_{\beta,R_{\lambda\circ l}}\left[\frac{C_{\delta,n,r}(p)}{n}\right]^{\frac{\beta-1}{\beta}} + \rho\|\Theta^*\|_2^2 = O\left[\frac{1}{n} + \left(\frac{p\log(nr_n)}{n}\right)^{\frac{\beta-1}{\beta}}\right] + \rho\|\Theta^*\|_2^2, \end{aligned}$$

where $C_{\beta,R_{\lambda\circ l}} := \left[2(2^{\beta-1}+1)R_{\lambda\circ l}(\Theta) + \frac{2^{\beta-1}}{\beta n^\beta}\mathbb{E}H_{Y,X}^\beta\right]/[(2^{\beta-1}+1)R_{\lambda\circ l}(\Theta)]^{\frac{\beta-1}{\beta}}$. Additionally, if we replace $\mathbb{E}H_{Y,X}^\beta < \infty$ with following increasing moment conditions

$$\mathbb{E}H_{Y,X} = q_n, \quad \mathbb{E}H_{Y,X}^\beta = z_{n,\beta}, \quad (10)$$

where $\{q_n\}$ and $\{z_{n,\beta}\}$ are positive sequence of n, then, the excess risk has a convergence rate

$$R_l(\hat{\theta}_n) - \inf\nolimits_{\theta\in\Theta} R_l(\theta) = O\left(\frac{q_n}{n} + \left(1 + \frac{z_{n,\beta}}{n^\beta}\right)\left(\frac{p\log(nr_n)}{n}\right)^{\frac{\beta-1}{\beta}} + \rho\|\Theta^*\|_2^2\right).$$

Remark 5 *For condition (10), taking $q_n = o(n)$ and $z_{n,\beta} = O(n^\beta)$, one has*

$$\mathbb{E}H_{Y,X} = o(n), \mathbb{E}H_{Y,X}^\beta = O(n^\beta). \quad (11)$$

Then, the consistency of excess risk is valid under $\rho = o(1)$, i.e.

$$R_l(\hat{\theta}_n) - \inf\nolimits_{\theta\in\Theta} R_l(\theta) = o_p(1).$$

For condition (11), we provide two examples. In quantile regressions, $H_{y,x} \propto \|x\|_2 \le \sqrt{d}\|x\|_\infty$; as long as $d = o(n^2)$ and $\mathbb{E}\|X\|_\infty^\beta < \infty$, we have $\mathbb{E}H_{Y,X} \propto \mathbb{E}[\sqrt{d}\|X\|_\infty] = o(n)$ and $\mathbb{E}H_{Y,X}^\beta \propto \mathbb{E}[\sqrt{d}\|X\|_\infty]^\beta = O(n^\beta)$ in condition (11). Similarly, negative binomial loss satisfies condition (11) if $d = o(n^2)$, $\mathbb{E}Y^{2\beta} < \infty$ and $\mathbb{E}\|X\|_\infty^{2\beta} < \infty$, by $H_{y,x} \propto y\,\|x\|_2$.

Remark 6 *Theorem 1 in Zhang and Zhou (2018) focused on the excess risk bound of robust LAD regression under $\mathbb{E}\|X\|^2 < \infty$. As a special case of (6), while Theorem 4.1 in Chen et al. (2021a) considered the excess risk bound of robust LAD regression that allows infinite variance of input using $\lambda(x) = |x|^\beta/\beta$ with $\beta \in (1, 2)$. Theorem 4 extends these two results to a large class of loss functions, which include many other regressions.*

Our *third main result* is the following theorem about the estimator of elastic net defined by (8) under the s_n-sparsity condition (7).

Theorem 7 *Let Θ^* be defined by (2) and let $\hat{\theta}_n$ be given by (8) with $\lambda(x) = |x|^\beta/\beta$ with $\beta \in (1, 2)$. If*

$\alpha = \frac{1}{n^{1/\beta}} \left(\frac{\log(\delta^{-2}/\sqrt{2es_n}) + s_n \log[(1+2nr_n)ep/s_n]}{(2^{\beta-1}+1)R_{\lambda\circ l}(\Theta)} \right)^{1/\beta}$ *with (C.3), $R_{\lambda\circ l}(\Theta) < \infty$ and $\mathbb{E}H^\beta_{Y,X} < \infty$,*

then with probability at least $1 - 2\delta$ one has

$$R_l(\hat{\theta}_n) - \inf_{\theta\in\Theta} R_l(\theta) \leq \frac{2\mathbb{E}H_{Y,X}}{n} + \frac{C_{\beta,R_{\lambda\circ l}}}{n^{\frac{\beta-1}{\beta}}} \left(\log(\frac{\delta^{-2}}{2es_n}) + s_n \log\left[(1 + 2nr_n)\frac{ep}{s_n}\right] \right)^{\frac{\beta-1}{\beta}} + \|\Theta^*\|_{\rho,\gamma},$$

where $C_{\beta,R_{\lambda\circ l}}$ is a constant given in Theorem 4, and $\|\Theta^*\|_{\rho,\gamma} := \inf_{\theta^*\in\Theta^*}(\rho\|\theta^*\|_2^2 + \gamma\|\theta^*\|_1)$.

Remark 8 *Suppose that Θ^* is bounded, Theorem 7 implies a rate $O((s_n \log(nr_n p/s_n)/n)^{(\beta-1)/\beta})$ excess risk bound if*

$$\rho \vee \gamma \lesssim (s_n \log(nr_n p/s_n)/n)^{(\beta-1)/\beta},$$

and it works for the high-dimension setting $p \gg n$. Moreover, put $(s_n \log(nr_n p/s_n)/n)^{(\beta-1)/\beta} = o(1)$, which implies the consistency of excess risk if $r_n = s_n(np)^{-1}e^{o(n/s_n)}$.

3. Examples for Theorem 4 ($p < n$)

This section provides examples of several robust regressions, which include quantile regression and GLMs. We assume that the data in this section has the finite β-th moment with $\beta \in (1, 2)$. In the all the models in the section, the dimension of the input X equals that of the parameter θ, i.e. $d = p$.

3.1 Robust quantile regressions

Consider

$$Y_i = X_i^\top \theta^* + \epsilon_i, (i = 1, \dots, n), \tag{12}$$

where $X_i = (X_{i1}, \dots, X_{ip})^\top$ is the i-th stochastic design point in $\mathbb{R}^p$, and random errors ϵ_i's are i.i.d. and satisfy $P(\epsilon_i < 0|X_i) = \tau$ for $0 < \tau < 1$. The unknown regression coefficient $\theta^* = (\theta_1^*, \dots, \theta_p^*)^\top$ may depend on τ, but we suppress such dependence for the notational simplicity. The conditional distribution of Y given x is $F(y|x) = P(Y \leq y|x)$ and the τth conditional quantile of Y given x is $Q_{y|x}(\tau) = \inf\{t : F(t|x) \geq \tau\}$. The problem of interest is to estimate the unknown slope coefficient θ^* by regressing the conditional quantile function

$$Q_{Y_i|X_i}(\tau) = X_i^\top \theta^*, (i = 1, \dots, n).$$

Recall that the loss function of quantile regression is

$$l(y,x,\theta)=\rho_\tau(y-x^\top\theta) \text{ with } \rho_\tau(u)=u[\tau-I(u<0)],$$

see more details in Koenker and Bassett Jr (1978).

Under the i.i.d. data $\{(X_i,Y_i)\}_{i=1}^n$, we study the log-truncated estimator $\hat{\theta}_n$ for the quantile loss:

$$\hat{\theta}_n := \operatorname*{arg\,min}_{\theta\in\Theta} \hat{R}_{\psi_\lambda,\rho_\tau,\alpha}(\theta), \tag{13}$$

where $\hat{R}_{\psi_\lambda,\rho_\tau,\alpha}(\theta) := \frac{1}{n\alpha}\sum_{i=1}^n \psi_\lambda[\alpha\rho_\tau(Y_i-X_i^\top\theta)]$ and $\lambda(x)=|x|^\beta/\beta,\ \beta\in(1,2)$. The *tuning parameter* α will be specified. The true parameter θ^* is defined as the minimizer

$$\theta^* := \operatorname*{arg\,min}_{\theta\in\Theta} R_{\rho_\tau}(\theta), \tag{14}$$

where $R_{\rho_\tau}(\theta) := \mathbb{E}[\rho_\tau(Y-X^\top\theta)]$ for $\theta\in\Theta$. Besides (C.2), we further assume

- (Q.1): $\mathbb{E}\,\|X\|_2^\beta<\infty$.
- (Q.2): $R_{\lambda\circ\rho_\tau}(\Theta) := \sup_{\theta\in\Theta} R_{\lambda\circ\rho_\tau}(\theta)<\infty$ with $R_{\lambda\circ\rho_\tau}(\theta) := \mathbb{E}[\lambda(\rho_\tau(Y-X^\top\theta))]$.

Corollary 9 *Let $\tau\in(0,1)$, $\delta\in(0,1/2)$. Define $\hat{\theta}_n$ by* (13)*, and θ^* is given by* (14)*. Under (C.2), (Q.1) and (Q.2), if we put $\alpha=\frac{1}{n^{1/\beta}}\left[\frac{C_{\delta,n,r}(p)}{(2^{\beta-1}+1)R_{\lambda\circ\rho_\tau}(\Theta)}\right]^{1/\beta}$. Then, with probability at least $1-2\delta$ one has*

$$R_l(\hat{\theta}_n)-R_l(\theta^*)\le \frac{2l_\tau\mathbb{E}\|X\|_2}{n}+C_{\beta,R_{\lambda\circ\rho_\tau}}\left[\frac{C_{\delta,n,r}(p)}{n}\right]^{\frac{\beta-1}{\beta}}+\rho\|\theta^*\|_2^2,$$

where $C_{\beta,R_{\lambda\circ\rho_\tau}} := \left[2(2^{\beta-1}+1)R_{\lambda\circ\rho_\tau}(\Theta)+\frac{2^{\beta-1}l_\tau^\beta}{\beta n^\beta}\mathbb{E}\,\|X\|_2^\beta\right]/[(2^{\beta-1}+1)R_{\lambda\circ\rho_\tau}(\Theta)]^{\frac{\beta-1}{\beta}}$ and $l_\tau :=$ $\max\{1+\tau,2-\tau\}$.

(Koenker, 2005, Section 4.1.2) stressed that 2nd moment condition of the input is required to show the consistency for the ERM estimator $\bar{\theta}_n := \arg\min_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^n\rho_\tau(Y_i-X_i^\top\theta)$, so our new method is essential in quantile regression for the data with β-th moment $(1<\beta<2)$.

3.2 Robust generalized linear models (GLMs)

We consider the general loss function of GLMs (McCullagh and Nelder, 1989) as below. In this part, we assume that $\{Y_i\}_{i=1}^n$ satisfy some moment conditions rather than put specific conditions on its distribution (as in the classical GLMs).

Let $u(\cdot)$ be a known link function. Consider the quasi-GLMs loss function:

$$\hat{R}_l(\theta) := \frac{1}{n}\sum_{i=1}^n l(Y_i,X_i^\top\theta),\ \theta\in\mathbb{R}^p \tag{15}$$

where the loss function is given by $l(y,x^\top\theta) := k(x^\top\theta)-yu(x^\top\theta)$ with $k(t) := b(u(t))$ for a known function $b(\cdot)$. If $u(t)=t$, we say the quasi-GLMs has the *canonical link*.

Let α be a tuning parameter to be specified, the log-truncated robust estimator $\hat{\theta}_n$ for quasi-GLMs is

$$\hat{\theta}_n := \mathop{\arg\min}_{\theta \in \Theta} \hat{R}_{\psi_\lambda, l, \alpha}(\theta), \tag{16}$$

where $\hat{R}_{\psi_\lambda, l, \alpha}(\theta) := \frac{1}{n\alpha} \sum_{i=1}^n \psi_\lambda[\alpha(k(X_i^\top \theta) - Y_i u(X_i^\top \theta))]$ with $\psi_\lambda(x) := \operatorname{sign}(x) \log(1 + |x| + \beta^{-1}|x|^\beta)$. We assume the GLMs-related conditions:

- (G.1): Assume that $u(\cdot)$ is continuous differentiable and $\dot{u}(\cdot) \geq 0$, there exist a positive constant A and a positive function $g_A(\cdot)$:
$$0 \leq \dot{u}(x^\top \theta) \leq g_A(x), \text{ for } \|\theta\|_2 \leq A.$$
- (G.2): Given a function $b(\cdot)$ such that $\ddot{b}(\cdot) > 0$, suppose that $k(\cdot)$ is continuous differentiable and $k(\cdot) \geq 0$, there exist a positive function $h_A(\cdot)$:
$$0 < \dot{k}(x^\top \theta) \leq h_A(x), \text{ for } \|\theta\|_2 \leq A.$$
- (G.3): Mixed moment condition: $\mathbb{E}|[|Y| g_A(X) + h_A(X)] \left\| X \right\|_2 |^\beta < \infty$.
- (G.4): Assume $\sigma_R := \sup_{\theta \in \Theta} \mathbb{E}[k(X^\top \theta) - Y u(X^\top \theta)]^\beta / \beta < \infty$.

The conditions (G.1), (G.2) and (G.3) imply (C.4), while (G.4) implies (C.5); see Remark 20 in Appendix A.6 for more discussions for (G.1) and (G.2). Theorem 4 is applicable to obtain the following result.

Corollary 10 *Let* $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}[l(y, x, \theta)]$ *with loss* $l(y, x, \theta) := k(x^\top \theta) - y u(x^\top \theta)$ *defined in* (15)*, and* $\hat{\theta}_n$ *is given by* (16)*. Under (C.2), (G.1)-(G.4), if* $\alpha = \frac{1}{n^{1/\beta}} \left[\frac{C_{\delta, r}(p)}{(2^{\beta-1}+1)\sigma_R} \right]^{1/\beta}$*, then with probability at least* $1 - 2\delta$

$$R_l(\hat{\theta}_n) - R_l(\theta^*) \leq \mathbb{E}\{[|Y| g_{r_n}(X) + h_{r_n}(X)] \left\| X \right\|_2\} \frac{2}{n} + C_{\beta, \sigma_R} \left[\frac{C_{\delta, r}(p)}{n} \right]^{\frac{\beta-1}{\beta}} + \rho \|\theta^*\|_2^2, \tag{17}$$

where $C_{\beta, \sigma_R} := \left[2(2^{\beta-1} + 1)\sigma_R + \frac{2^{\beta-1}}{\beta n^\beta} \mathbb{E}|[|Y| g_{r_n}(X) + h_{r_n}(X)] \left\| X \right\|_2 |^\beta \right] / [(2^{\beta-1} + 1)\sigma_R]^{\frac{\beta-1}{\beta}}$.

Corollary 10 can be applied to the following two examples, robust logistic regression and robust negative binomial regression; see Appendix A.6 for derivations.

Example 1 (Robust logistic regression) *The output in logistic regression can take only two values: "0, 1". Formally, let* Y_i*'s* $\in \{0, 1\}$ *be the random outputs and* θ^* *be a* $p \times 1$ *vector of unknown regression coefficients belonging to a compact subset of* $\mathbb{R}^p$*. Given* n *random input* X_i*'s* $\in \mathbb{R}^{n \times p}$*, the logistic regression assumes* $P(Y_i = 1 | X_i; \theta^*) := \frac{e^{X_i^\top \theta^*}}{1 + e^{X_i^\top \theta^*}}$*. The empirical loss function of logistic regression is*

$$\hat{R}_l(\theta) = \tfrac{-1}{n} \textstyle\sum_{i=1}^n [Y_i X_i^\top \theta - \log(1 + e^{X_i^\top \theta})].$$

Note that $H_{y,x} = 2 \left\| x \right\|_2$ *in Corollary 10 under logistic loss. To obtain the finite excess risk* (17)*, the robust logistic regression requires the moment condition*

$$\mathbb{E}\|X\|_2^{\beta} < \infty.$$

For modeling count data regressions, the classical Poisson regression as the canonical link GLMs has equal dispersion assumption (i.e. $\mathbb{E}(Y|X) = \operatorname{Var}(Y|X)$), which has little practical motivation. Nevertheless, it motivates us to study the more flexible count data regressions, as shown below.

Example 2 (Robust negative binomial regression) *As a generalization of Poisson regression, negative binomial regression (NBR) relaxes the equadispersion assumption to the quadratic relationship between the mean and variance of the responses. NBR assumes that the overdispersed responses $\{Y_i\}_{i=1}^n$ are modelled by two-parameter negative binomial distribution with the connection of covariates: $P(Y_i = y|\theta, \mu_i) = \frac{\Gamma(\eta+y)}{\Gamma(\eta)y!}(\frac{\mu_i}{\eta+\mu_i})^y(\frac{\eta}{\eta+\mu_i})^{\eta}$ with $\log \mu_i = X_i^{\top}\theta$, where $\eta > 0$ is the known dispersion parameter, which can be estimated previously. One has $\mathbb{E}(Y_i|X_i) = \mu_i \leq \operatorname{Var}(Y_i|X_i) = \mu_i + \mu_i^2/\eta$. The NBR empirical loss function is*

$$\hat{R}_l(\theta) = \frac{-1}{n}\sum_{i=1}^{n}\{Y_i[X_i^{\top}\theta - \log(\eta + e^{X_i^{\top}\theta})] - \eta\log(\eta + e^{X_i^{\top}\theta})\},$$

see Zhang and Jia (2022) for details. In Corollary 10, NBR loss has $H_{y,x} = (y+\eta)\,\|x\|_2$. Note that there are no assumptions for the distribution of output, and it only requires the moment conditions

$$\mathbb{E}\|XY\|_2^{\beta} < \infty \text{ and } \mathbb{E}\|X\|_2^{\beta} < \infty$$

to guarantee the excess risk bound (17).

4. Examples for Theorem 7 ($p > n$): non-convex regressions via DNN

In many statistical learning problems, loss functions are non-convex, whereby the associated ERMs have multiple local minima; see Guan et al. (2017); Chen et al. (2021b); Klusowski et al. (2019). Regressions via DNN is a large family of highly non-convex learning problems due to the multiple compositions of activation functions. In this section, we shall apply Theorem 7 to study high dimensional non-convex regressions via DNN.

We consider the DNN function class as follows:

$$\mathcal{NN}(N,L) := \left\{f_\theta(x) = W_L\sigma_L\left(W_{L-1}\sigma_{L-1}\left(\ldots W_1\sigma_1\left(W_0x\right)\right)\right) \mid \theta := (W_0,\ldots,W_L)\right\}, \quad (18)$$

where $W_j \in \mathbb{R}^{N_j \times N_{j+1}}$ for $j = 0, 1, \ldots, L-1$ with $N_0 = d$. Here L represents the depth of this class of DNNs, each activation function $\sigma_j : \mathbb{R}^{N_j} \to \mathbb{R}^{N_j}$, $j = 1, 2, \ldots, L$, and θ is the vectorized parameter consisting of weighted matrices with the width $N = \max\{N_1, \ldots, N_L\}$; see Fan et al. (2021) for details.

For i.i.d. observations $\{(X_i, Y_i)\}_{i=1}^n$ and a given loss function $l(\cdot,\cdot)$, the risk function is

$$R_l(f) = \mathbb{E}[l(Y, f(X))], \qquad \text{for a function } \ f : \mathbb{R}^d \to \mathbb{R}.$$

In general, the true function f^* belongs to a certain function family and is defined by (Fan et al., 2021)

$$f^* \in \operatorname*{argmin}_f R_l(f). \quad (19)$$

From DNN function class (18) with parameter space $\Theta \subset \mathbb{R}^{\sum_{l=0}^{L} N_{l+1}N_l}$, we define $\theta^*_{\mathcal{N}}$ as

$$\theta^*_{\mathcal{N}} \in \operatorname*{argmin}_{\theta\in\Theta} R_l(f_\theta) \text{ for } f_\theta \in \mathcal{NN}(N, L). \tag{20}$$

Note that $p = \sum_{l=0}^{L} N_{l+1}N_l$ and $d = N_0$ in this case. Denote $\Theta^*_{\mathcal{N}} := \{\theta^*_{\mathcal{N}} \in \arg\min_{\theta\in\Theta} R_l(f_\theta)\}$ for $f_\theta \in \mathcal{NN}(N, L)$.

Now we fit the regression problem (20) into the framework of the elastic net regression (8), whose corresponding form is as follows:

$$\hat{\theta}_n \in \operatorname*{arg\,min}_{\theta\in\Theta} \left\{ \frac{1}{n\alpha} \sum_{i=1}^{n} \psi_\lambda\big(\alpha l(Y_i, f_\theta(X_i))\big) + \rho\|\theta\|_2^2 + \gamma\|\theta\|_1 \right\}, \tag{21}$$

where ρ and $\gamma > 0$ are *penalty parameters.* We have the following assumptions:

- (D.1): Assume that the $l(\cdot,\cdot)$ satisfies Lipschitz condition with a *Lipschitz constant* $D_{x,y}$:

  $$|l(y, f_{\theta_2}(x)) - l(y, f_{\theta_1}(x))| \le D_{x,y}|f_{\theta_2}(x) - f_{\theta_1}(x)| \text{ for } \theta_1,\ \theta_2 \in \Theta,$$

  where the DNN function is indexed by the s_n*-sparse parameter space*

  $$\Theta := \{\theta := (W_1, \ldots, W_L) : \|\theta\|_2 \le r_n,\ \|\theta\|_0 \le s_n\} \subseteq \mathbb{R}^p, \tag{22}$$

  where r_n, s_n are both allowed to increase with the size n of the observed data. Further assume that there exists some $W > 0$ so that $\max_{0\le j\le L} \sigma_{\max}(W_j) \le W$, where $\sigma_{\max}(W_j)$ is the largest singular value of W_j.

- (D.2): We assume that the true function f^* belongs to the Hölder function class $f^* \in \mathcal{C}^\gamma([0, a_n]^d, B)$ with smoothness index γ, where B is a positive constant (see definition of $\mathcal{C}^\gamma([0, a_n]^d, B)$ in (1)), and $\{a_n\}$ is a sequence of n.

- (D.3): For a fixed d, we assume that: $\mathbb{E}\|X\|_\infty \le b$ for $X \in \mathbb{R}^d$ and $b > 0$. Moreover, $\mathbb{E}D^2_{X,Y} < \infty$.

Remark 11 *The Lipschitz condition (D.1), together with Proposition 6 in Taheri et al. (2021), immediately implies (C.3). In real-world applications, the input data is usually transformed into an interval* $[0, a_n]^d$*, this motivates the assumption (D.2).*

Theorem 12 *Assume that (D.1)-(D.3) hold and that* $\mathbb{E}|\|X\|_2 D_{X,Y}|^\beta < \infty, \beta \in (1, \infty)$*. Let* f^* *be defined by* (19)*, and let* $\hat{\theta}_n$ *be given by* (21) *with* $\lambda(x) = |x|^\beta/\beta$ *and* $\beta \in (1, 2)$*. For a* $\delta \in (0, 1/2)$*, if we choose*

$$\alpha = \tfrac{1}{n^{1/\beta}} \left(\tfrac{\log(\delta^{-2}/\sqrt{2es_n}) + s_n \log[(1+2nr_n)ep/s_n]}{(2^{\beta-1}+1)R_{\lambda\circ l}(\Theta)} \right)^{1/\beta},$$

then with probability at least $1 - 2\delta$ *we have*

$$R_l(f_{\hat{\theta}_n}) - R_l(f^*) \le E_1 + E_2 + E_3 + E_4, \tag{23}$$

for any $\|f - f^*\|_\infty \leq F < \infty$ *with* $f \in \mathcal{N}\mathcal{N}(N, L)$, *where*

$$E_1 := \frac{4W^L\sqrt{L}}{n}\mathbb{E}[\|X\|_2|D_{X,Y}|],\ E_2 := \inf_{\theta^*_{\mathcal{N}} \in \Theta^*_{\mathcal{N}}}(\rho\|\theta^*_{\mathcal{N}}\|_2^2 + \gamma\|\theta^*_{\mathcal{N}}\|_1),$$

$$E_3 := 2\sqrt{\mathbb{E}D^2_{X,Y}}b^{\frac{\gamma}{2\gamma+1}}\left[\frac{(2B+1)\left(1+d^2+\gamma^2\right)6^d MF^{2\gamma}}{2^m} + \frac{3^\gamma BF^{2\gamma}}{N^{\gamma/d}}\right]^{\frac{1}{2\gamma+1}},$$

$$E_4 := \frac{F_{\beta,L,W}(R_{\lambda \circ l})}{n^{(\beta-1)/\beta}}\left(\log(\frac{\delta^{-2}}{\sqrt{2es_n}}) + s_n \log\left[(1+2nr_n)\frac{ep}{s_n}\right]\right)^{(\beta-1)/\beta},$$

with $F_{\beta,L,W}(R_{\lambda\circ l}) := \left[2(2^{\beta-1}+1)R_{\lambda\circ l}(\Theta) + \frac{(4W^L\sqrt{L})^\beta}{2\beta n^\beta}\mathbb{E}|\|X\|_2 D_{X,Y}|^\beta\right]/[(2^{\beta-1}+1)R_{\lambda\circ l}(\Theta)]^{(\beta-1)/\beta}$, *the integer* $m \geq 1$ *and* $M \geq (\gamma+1)^d \vee (B+1)e^d$, *and* $L \leq 8 + (m+5)\left(1 + \lceil\log_2(d \vee \gamma)\rceil\right)$.

Remark 13 E_1, E_2, E_3, E_4 *can be interpreted as the bias, the penalization error, the error between the true function* f^* *and the DNN, and the statistical error, respectively. For* E_3, *we require sparsity* $s \leq 141(d+\gamma+1)^{3+d}M(m+6)$ *in* (22).

The following corollary gives an upper bound for the depth and a lower bound for the width of a DNN designed to realize the regression (21).

Corollary 14 *Under the setting in Theorem 12, if* depth-sample *and* width-sample *of DNNs, which may increase with* n, *satisfy the following conditions:*

$$L_n \lesssim \frac{\log n + \log(s_n \log(nr_np/s_n))}{\log W}, \quad N_n \gtrsim b^d\left[\frac{n}{s_n\log(nr_np/s_n)}\right]^{\frac{d(2\gamma+1)}{\gamma}\cdot\frac{\beta-1}{\beta}} \tag{24}$$

with $W > 1$, *and* order of tuning parameters $\rho \vee \gamma \lesssim (s_n \log(nr_np/s_n)/n)^{(\beta-1)/\beta}$, *then*

$$R_l(f_{\hat{\theta}_n}) - R_l(f^*) \leq C_\delta((s\log(nr_np/s_n)/n)^{(\beta-1)/\beta})$$

with probability at least $1 - 2\delta$, *for a certain constant* $C_\delta > 0$.

We finish this section by giving three concrete examples of robust DNN regressions, and the models therein will be used in simulations or real data studies.

Example 3 (Robust DNN LAD regression) *Suppose that i.i.d. observations* $\{(Y_i, X_i) \sim (Y, X)\}_{i=1}^n \in \mathbb{R} \times \mathbb{R}^d$ *satisfy*

$$Y_i = f^*(X_i) + e_i, \quad \mathbb{E}(e_i|X_i) = 0,\ i = 1, 2, \cdots, n,$$

where $\{e_i \sim e\}_{i=1}^n$ *are i.i.d. noise. Similar to QR with* $\tau = 0.5$, *the robust DNN LAD regression problem* (21) *has loss function* $l(x, y, \theta) = |y - f_\theta(x)|$ *with* $f_\theta \in \mathcal{N}\mathcal{N}(N, L)$. *Thus* $R_{\lambda\circ l}(\theta) := \beta^{-1}\mathbb{E}|Y - f_\theta(X)|^\beta = \mathbb{E}|e|^\beta/\beta$ *and we have* $R_{\lambda\circ l}(\theta) < \infty$ *in Theorem 12 if*

$$\mathbb{E}|e|^\beta < \infty.$$

The LAD regression loss has Lipschitz constant $D_{x,y} = 1$ *and thus* $H_{y,x} = 2W^L\sqrt{L}\|x\|_2$ *in Theorem 12 also requires*

$$\mathbb{E}\|X\|_2^\beta < \infty. \tag{25}$$

Recently, Padilla et al. (2022); Shen et al. (2021a) studied the DNN quantile regression with fixed inputs, and their estimators are only robust for output. Their setting can not deal with the robustness of the random input with heavy-tail condition (25).

Example 4 (Robust DNN logistic regression) *Assume that i.i.d. observations* $\{(Y_i, X_i) \sim (Y, X)\}_{i=1}^n \in \{0, 1\} \times \mathbb{R}^d$ *satisfy*

$$P(Y_i = 1|X_i) := \frac{e^{f^*(X_i)}}{1 + e^{f^*(X_i)}}, \quad P(Y_i = 0|X_i) = 1 - P(Y_i = 1|X_i). \tag{26}$$

The robust DNN logistic regression problem (21) *has loss function* $l(x, y, \theta) = -[yf_\theta(x) - \log(1+e^{f_\theta(x)})]$ *with* $f_\theta \in \mathcal{NN}(N, L)$*. Theorem 12 requires* $R_{\lambda \circ l}(\Theta) := \beta^{-1}\sup_{\theta\in\Theta}\mathbb{E}[Y f_\theta(X) - \log(1 + e^{f_\theta(X)})]^\beta \lesssim \sup_{\theta\in\Theta}\mathbb{E}|f_\theta(X)|^\beta < \infty$*. For logistic loss, it gives* $D_{x,y} = y + 1 \leq 2$ *and* $H_{y,x} = 4W^L\sqrt{L}\|x\|_2$*. The moment conditions for robust DNN logistic regression are*

$$\mathbb{E}\,\|X\|_2^\beta < \infty \text{ and } \sup_{\theta\in\Theta}\mathbb{E}|f_\theta(X)|^\beta < \infty.$$

Example 5 (Robust DNN NBR) *Let* $\eta > 0$ *be the known dispersion parameter. Suppose that i.i.d. observations* $\{(Y_i, X_i) \sim (Y, X)\}_{i=1}^n \in \mathbb{Z} \times \mathbb{R}^d$ *satisfy*

$$P(Y_i = y|X_i) = \frac{\Gamma(\eta + y)}{\Gamma(\eta)y!}(\frac{\mu_i}{\eta + \mu_i})^y(\frac{\eta}{\eta + \mu_i})^\eta, \text{ with } \log\mu_i = f^*(X_i).$$

The robust DNN NBR problem (21) *has loss function* $l(x, y, \theta) = -y[f_\theta(x) - \log(\eta+e^{f_\theta(x)})] - \eta\log(\eta + e^{f_\theta(x)})$*. If* $\sup_{\theta\in\Theta}\mathbb{E}|Y f_\theta(X)|^\beta < \infty$*, then Theorem 12 gives*

$$\begin{aligned} R_{\lambda\circ l}(\Theta) :&= \beta^{-1}\sup_{\theta\in\Theta}\mathbb{E}\{Y[f_\theta(X) - \log(\eta + e^{f_\theta(X)})] - \eta\log(\eta + e^{f_\theta(X)})\}^\beta \\ &\lesssim \sup_{\theta\in\Theta}\mathbb{E}|Y f_\theta(X)|^\beta + \sup_{\theta\in\Theta}\mathbb{E}|f_\theta(X)|^\beta < \infty. \end{aligned}$$

For NBR loss, we get $D_{x,y} = y+\eta$ *and* $H_{y,x} = 2W^L\sqrt{L}\|x\|_2(y+\eta)$*. (D.3) needs* $\mathbb{E}(Y+\eta)^2 < \infty$*. In summary, the required moment conditions are*

$$\mathbb{E}(Y + \eta)^2 < \infty, \ \mathbb{E}[\|X\|_2(Y + \eta)]^\beta < \infty \text{ and } \sup_{\theta\in\Theta}\mathbb{E}|Y f_\theta(X)|^\beta < \infty.$$

Note that if $L = 0$ and $\theta = W_0 \in \mathbb{R}^d$, Theorem 12 cannot work since the proof of excess risk bound requires $L \geq 1$. In this degenerate DNN regression, it is just the common robust ERM problem with elastic net penalty (8). Under s_n-sparse parameter space, we obtain the excess risk bound in Theorem 7 for this special and important parametric regressions when $d > n$.

5. Simulation and real data studies

5.1 Simulations on normal regression models

In this part, by stochastic gradient descent (SGD) algorithms, we illustrate the effectiveness of regressions based on log-truncated ERM by the numerical experiments of ordinary logistic regression and negative binomial regression. The elastic net DNN regressions are optimized by the Adam algorithm (SGD-based algorithm, Kingma and Ba (2015)), which is an extension of SGD. Moreover, the Adam algorithm is more computationally efficient than SGD under a large number of parameters, and it has few memory requirements.

5.1.1 SGD

1. **SGD for our estimation**

   Let us consider a regularized optimization with a given penalty function $\Omega(\theta)$:

   $$\hat{\theta}_n(\alpha,\rho) := \mathop{\arg\min}_{\theta\in\Theta}\{\hat{R}_{\psi_\lambda,l,\alpha}(\theta)+\Omega(\theta)\}, \tag{27}$$

   where $\hat{R}_{\psi_\lambda,l,\alpha}(\theta) := \frac{1}{n\alpha}\sum_{i=1}^n \psi_\lambda[\alpha l(Y_i,X_i,\theta)]$ with $l(y,x,\theta)$ being some specific losses, and $\alpha > 0$ is another tuning parameter to be chosen.

   - For $\Omega(\theta) = \rho\|\theta\|_2^2$, this is a ℓ_2-regularization, where $\rho > 0$ is the penalty parameter.
   - For $\Omega(\theta) = \rho\|\theta\|_2^2 + \gamma\|\theta\|_1$, this is an elastic net regularization, where $\rho,\gamma > 0$ are two penalty parameters.

   In practice, this optimization problem is solved by stochastic gradient descent (SGD) as the following:

   $$\theta_{t+1} = \theta_t - \frac{r_t}{\alpha}\nabla_\theta\{\psi_\lambda[\alpha l(Y_{i_t},X_{i_t},\theta_t)]+\Omega(\theta_t)\}, t=0,1,2,\cdots, \tag{28}$$

   where i_t denotes a random sampled index, $\{r_t\}$ is the learning rate. For ℓ_2-regularization, we employ five-fold cross validation (CV) method to find the optimal parameter pair (α,ρ) in a certain effective subset of $\mathbb{R}_+^2$. For the elastic net regularized DNN model, we select the optimal parameters (α,β,γ) by evaluating the performances of their corresponding training models on validation data set whose size is 1/5 of the size of the training set.

2. **SGD for the comparative estimations**

   For the standard ridge regression without truncation, the corresponding optimization problem is

   $$\hat{\theta}_n(\rho) := \mathop{\arg\min}_{\theta\in\Theta}\{\frac{1}{n}\sum_{i=1}^n l(Y_i,X_i,\theta)+\Omega(\theta)\}, \tag{29}$$

   where ρ is the penalty parameter for regularization, this optimization problem can be solved by the following SGD:

   $$\theta_{t+1} = \theta_t - r_t\nabla_\theta\{l(Y_{i_t},X_{i_t},\theta_t)+\Omega(\theta_t)\}, t=0,1,2,\cdots. \tag{30}$$

   We also consider the Cauchy log-truncated function in Table 1, which has the form $\phi_\alpha(x) = \alpha\log(1+\frac{x}{\alpha})$. Similarly, the estimator $\hat{\theta}_n^C$ is solved by

   $$\hat{\theta}_n^C := \mathop{\arg\min}_{\theta\in\Theta}\{\frac{1}{n}\sum_{i=1}^n \phi_\alpha\big(l(Y_i,X_i,\theta)\big)+\Omega(\theta)\}.$$

   The corresponding SGD iterations are

   $$\theta_{t+1} = \theta_t - r_t\nabla_\theta\{\phi_\alpha(Y_{i_t},X_{i_t},\theta_t)+\Omega(\theta_t)\}, t=0,1,2,\cdots.$$

   In both standard ridge and Cauchy log-truncated regressions, we also take five-fold CV to find the optimal parameters $\rho\in\mathbb{R}_+$ and $(\alpha,\rho)\in\mathbb{R}_+^2$.

5.1.2 Numerical experiments

1. **Simulation study**

   For $\mathbb{R}^d$-valued covariates $\{X_i\}_{i=1}^n$, each X_i can be written as

$$X_i = X_i' + \xi_i,$$

where $\{X_i'\}_{i=1}^n$ are i.i.d. $\mathbb{R}^d$-valued random vectors with normal distribution $N(\mathbf{0}, \mathbf{Q}(\varsigma))$. Here the covariance matrix $\mathbf{Q}(\varsigma)$ is an identity matrix ($\varsigma = 0$) or a Toeplitz matrix ($\varsigma = 0.5$) which is formed

$$\mathbf{Q}(\varsigma) = \begin{bmatrix} 1 & \varsigma & \varsigma^2 & \cdots & \cdots & \varsigma^{d-1} \\ \varsigma & 1 & \varsigma & \ddots & & \vdots \\ \varsigma^2 & \varsigma & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \varsigma & \varsigma^2 \\ \vdots & & \ddots & \varsigma & 1 & \varsigma \\ \varsigma^{d-1} & \cdots & \cdots & \varsigma^2 & \varsigma & 1 \end{bmatrix}.$$

$\{\xi_i\}_{i=1}^n$ are i.i.d. $\mathbb{R}^d$-valued random noisy vectors, which satisfy one of the following conditions:

(i) **Pareto noise** (heavy tail): the noise $\xi_i := (\xi_{i1}, ..., \xi_{id})^\top$ whose entries $\{\xi_{ij}\}$ are independently drawn from Pareto distribution with scale parameter 1 and shape parameter $\tau \in \{1.6, 1.8, 2.01, 4.01, 6.01\}$.

(ii) **Uniform noise** (outlier): the noise $\xi_i = Z\xi_i'$ where $\xi_i' := (\xi_{i1}', ..., \xi_{id}')^\top$ are independently drawn from uniform distribution $U(-2, 2)$ in logistic regression and negative binomial regression, and $Z := \text{diag}(\zeta_1, ..., \zeta_d)$ where $\{\zeta_j\}_{j=1}^d$ are i.i.d. Bernoulli r.v.s with probability $\pi \in (0, 1)$ taking 1. We will choose $\pi \in \{0.3, 0.5, 0.8\}$.

For each fixed pair (d, n), the true value $\theta = (\theta_1, ..., \theta_d)^\top \in \mathbb{R}^d$ is designed by drawing each θ_j from $U(0, 1)$ independently. We compute the ℓ_2-estimation error for each estimator $\hat{\theta}_n$, i.e., $\|\hat{\theta}_n - \theta\|_2$. We choose the high order log-truncated function as (5) with $\lambda(x) = |x|^\beta/\beta$. When $\{\xi_i\}_{i=1}^n$ are Pareto noises, we choose $\beta = 1.5$ as $\tau \in \{1.6, 1.8\}$ and $\beta = 2.0$ as $\tau \in \{2.01, 4.01, 6.01\}$. When $\{\xi_i\}_{i=1}^n$ are uniform noises, we always choose $\beta = 2.0$.

We will conduct experiments for the following three cases:

$$(d, n) = (100, 200), (200, 500), (1000, 1000).$$

Tables 2 and 3 present the comparison results of average ℓ_2-estimation errors and standard errors (in bracket) for predicted logistic regression coefficients with 100 replications. It is obvious that the log-truncated estimators perform much better than standard regression estimators under two noise settings. Our proposed log-truncated estimators based on high-order functions have smaller estimation errors than Cauchy

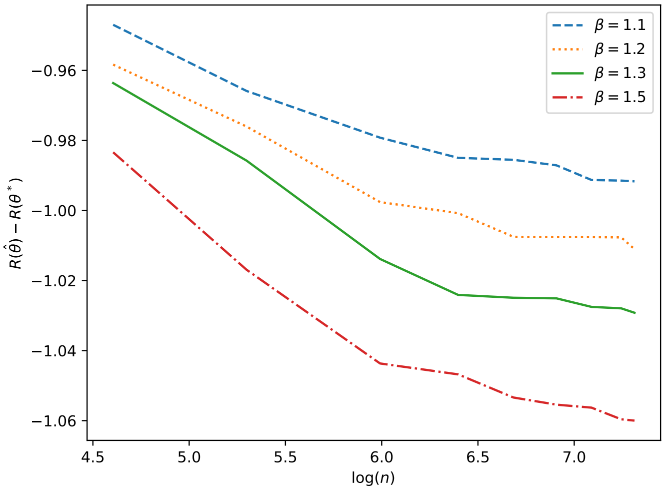

Figure 1: Plot of $R(\hat{\theta}_n) - R(\theta^*)$ for different β over various n based on Pareto noise model with $\tau = 1.6$, $d = 100$ and n in $[100, 1500]$.

log-truncated estimators. It reveals that the log-truncated regression with a high-order function is more flexible than the 1-order log-truncation in coping with the contaminated or heavy-tailed data. We obtain similar results from simulations on NBR, which are displayed in Appendix A.10.

We also use the Pareto noisy model with $\tau = 1.6$ to illustrate the rate $O\big((\frac{d\log(n)}{n})^{\frac{\beta-1}{\beta}}\big)$ of excess risk bound $R(\hat{\theta}_n) - R(\theta^*)$ for different β and sample size n in Theorem 4, if r_n is constant. Figure 1 demonstrates that the numerical excess risk bound linearly decreases with the increase of n, and β is closer to 1, the excess risk bound is larger.

2. **Tuning Parameter selection**

Tuning parameter selection is a crucial step in the experiments, correct parameter can enhance the accuracy of the prediction of a model. It is interesting for us to explore the way to select the optimal tuning parameter in our proposed model. A simple and user-friendly tuning parameter selection method is the grid search. For example, for the regularized optimization problem (27), we can use the binary search to find an effective interval of the parameter pair (α, ρ), and use the grid search to select the optimal $(\hat{\alpha}, \hat{\rho})$ which minimizes the ℓ_2-estimation error $\|\hat{\theta}(\alpha, \rho) - \theta\|_2$. However, the ground truth of the regression coefficients θ are generally unknown and good fitting model on the training data can not say the model exactly works well. Thus, cross-validation is a popular technique to select the optimal tuning parameter. In our experiments, we use five-fold CV to select the optimal $(\hat{\alpha}, \hat{\rho})$ optimizing $\frac{1}{n}\sum_{j=1}^{5}\sum_{i\in K_j}|\hat{f}^{-j}_{\alpha,\rho}(X_i) - Y_i|$ in an effective subset of $\mathbb{R}^2_+$, where K_j is the validation data set in the five-fold CV. Figure 2 plots the heat map of the criterion values of grid search and five-fold CV under the Pareto noise model as $d = 200, n = 500$. The number in the upper half of the cell is the criterion of grid search, i.e., $\|\hat{\theta}(\alpha, \rho) - \theta\|_2$. The number in the lower half of the

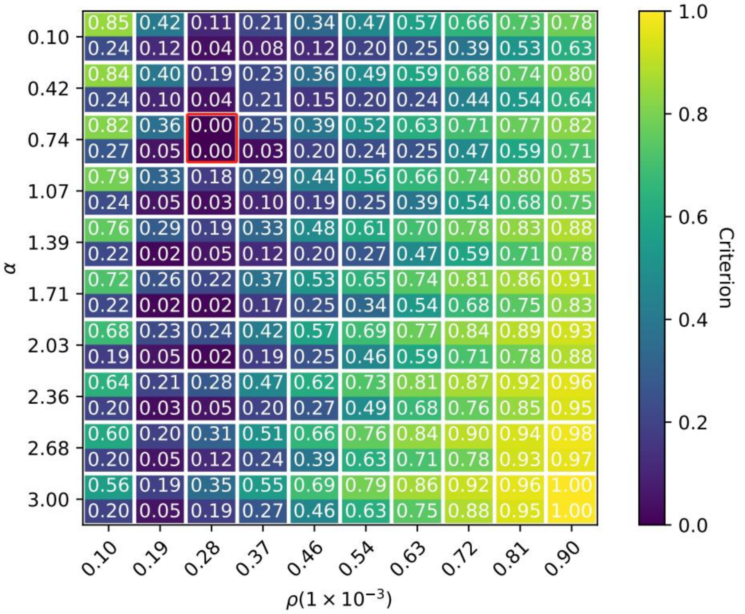

Figure 2: Comparisons of the tuning parameter selection by grid search and five-fold CV under the Pareto noise model as $d = 200, n = 500$. For each cell, the number in the upper half of the cell is the criterion of grid search, i.e., $\|\hat{\theta}(\alpha, \rho) - \theta\|_2$. The number in the lower half of the cell is the criterion of five-fold CV, i.e., $\frac{1}{n} \sum_{j=1}^{5} \sum_{i \in K_j} |\hat{f}_{\alpha,\rho}^{-j}(X_i) - Y_i|$. All the values are normalized in $(0, 1)$.

cell is the criterion of five-fold CV, i.e., $\frac{1}{n}\sum_{j=1}^{5}\sum_{i\in K_j}|\hat{f}_{\alpha,\rho}^{-j}(X_i)-Y_i|$. All the values are normalized in $(0,1)$. We find that the selected optimal parameters based on the ground truth are consistent with the ones based on the mean absolute values of the response.

5.2 Simulations on robust deep regression models

We also compare our proposed robust elastic net penalized DNN logistic regression and DNN NBR in Examples 4 and 5 of Section 3 with their non-truncated elastic net penalized DNN versions. The non-parametric function $f^*(x)$ in (21) satisfies

$$f^*(X_i)=f^*(X_i'+\xi_i),\ i=1,...,n, \tag{31}$$

where $\{X_i'\}_{i=1}^n$ are i.i.d. with law $N(\mathbf{0},\mathbf{I}_{d\times d})$. Here $\{\xi_i\}_{i=1}^n\in\mathbb{R}^d$ also follow Pareto or uniform distribution in the above noise setting. Differently, in the uniform noise setting, $\xi_i':=(\xi_{i1}',...,\xi_{id}')^\top$ are independently drawn from uniform distribution $U(2,10)$. Two types of the real function f^* are considered:

(1) **A complex function**: f^* is a complex function (referred to (Ohn and Kim, 2022)):

$$f^*(x)=0.8\exp\left(0.03x_1+x_2^2-\sqrt{x_3+5}\right)-\cot\left(\frac{1}{0.01+|x_4-2x_5+x_6|}\right),$$

where $d=6$. And we put $n=200$.

(2) **DNN**: taking into account the higher dimensional input data, we generate a real two-layers deep neural network by Pytorch as the function f^*:

The function $f^*(x)$: $\mathbb{R}^d\to\mathbb{R}$ is generated by a ReLu activated fully connected two-layers deep neural network with network width $(d,0.6d,0.4d,1)$ by Pytorch. The real weights of the two-layers DNN are drawn from $U(-1,1)$ independently. We set $(d,n)=(100,1000)$.

The elastic net penalized robust DNN logistic regression is trained by the ERM problem (21) with the ReLU activated 2-layers DNN model and network width $(d,0.6d,0.4d,1)$. We use the Adam optimization algorithm in PyTorch as implement with $n/4$ batch size in each case. The same network configurations and optimization algorithms are used to train the non-truncated elastic net penalized DNN regressions. For the elastic net based robust DNN logistic regression, we compute the accuracy of predictors $\{\hat{Y}_i\}_{i=1}^n$ for inputs $\{Y_i\}_{i=1}^n\in\{0,1\}$, defined as

$$\text{Accuracy}:=\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(\hat{Y}_i=Y_i)\times 100\%.$$

Table 4 shows the average accuracy for robust DNN logistic regression under Pareto and uniform noise settings with 100 replications. Different values of β are selected for different noises. The results reveal that the proposed elastic net penalized robust DNN logistic regression has higher flexibility and stronger robustness than the non-truncated elastic net-based DNN logistic regression for fitting the contaminated or heavy-tail data. The elastic net penalized robust DNN NBR results are demonstrated in Appendix A.10.

Table 2: Comparison of average ℓ_2-estimation error for logistic regression on Pareto noise model.

ℓ2-estimation error for logistic regression							
			$\varsigma = 0$			$\varsigma = 0.5$	
				$d = 100, n = 200$			
Pareto	β	High-order	Cauchy	Non-truncation	High-order	Cauchy	Non-truncation
1.60	1.5	2.9463(0.0110)	3.1587(0.0167)	3.5344(0.1532)	2.9279(0.1401)	3.0992(0.1490)	3.5128(0.0306)
1.80	1.5	2.9922(0.0811)	3.2056(0.0260)	3.6353(0.0753)	2.9241(0.0664)	3.0896(0.1063)	3.5103(0.0241)
2.01	2.0	2.9335(0.0217)	3.2157(0.0416)	3.6312(0.0715)	2.7939(0.0906)	2.9930(0.1713)	3.5097(0.0425)
4.01	2.0	2.9210(0.0184)	3.1100(0.0144)	3.5047(0.0803)	2.8205(0.0151)	2.9485(0.0275)	3.5084(0.0162)
6.01	2.0	2.8361(0.0153)	3.0689(0.0458)	3.4677(0.1294)	2.8538(0.1114)	2.8994(0.1178)	3.5092(0.0410)
				$d = 200, n = 500$			
Pareto	β	High-order	Cauchy	Non-truncation	High-order	Cauchy	Non-truncation
1.60	1.5	3.8489(0.0378)	4.0187(0.1169)	4.4003 (0.0203)	4.0401(0.1436)	4.0698(0.1761)	4.5583(0.1816)
1.80	1.5	3.8551(0.0225)	4.0001(0.1370)	4.4332(0.1774)	3.9112(0.2145)	4.1223(0.1159)	4.5953(0.0784)
2.01	2.0	3.8271(0.0667)	3.9885(0.0727)	4.2360(0.0605)	4.0839(0.1136)	4.1373(0.2145)	4.6304(0.1079)
4.01	2.0	3.9291(0.0194)	4.0467(0.1243)	4.4708(0.1353)	3.9896(0.0853)	4.0554(0.1133)	4.5967(0.1821)
6.01	2.0	3.9502(0.0724)	4.0155(0.1278)	4.3148(0.1211)	3.9835(0.0880)	4.0297(0.1287)	4.5948(0.1424)
				$d = 1000, n = 1000$			
Pareto	β	High-order	Cauchy	Non-truncation	High-order	Cauchy	Non-truncation
1.60	1.5	10.2445(0.0274)	10.4860(0.0925)	11.2177(0.0302)	10.1364(0.0633)	10.4548(0.0382)	10.2014(0.1289)
1.80	1.5	10.2147(0.0372)	10.4757(0.1345)	11.1923(0.0221)	10.0405(0.0857)	10.4431(0.0118)	4.5953(0.0784)
2.01	2.0	10.2983(0.0174)	10.6056(0.0327)	10.9366(0.0202)	10.0366(0.0951)	10.4361(0.0318)	11.4579(0.0854)
4.01	2.0	10.2070(0.0206)	10.4250(0.1389)	11.7001(0.0307)	9.9773(0.0584)	10.4361(0.0318)	11.5536(0.1378)
6.01	2.0	10.2290(0.0208)	10.2782(0.0216)	11.1155(0.0433)	10.0486(0.0893)	10.4462(0.0052)	11.4853(0.0799)

Table 3: Comparison of average ℓ_2-estimation error for logistic regression on Uniform noise model.

ℓ_2-estimation error for logistic regression							
			$\varsigma = 0$			$\varsigma = 0.5$	
				$d = 100, n = 200$			
Uniform	β	High-order	Cauchy	Non-truncation	High-order	Cauchy	Non-truncation
0.3	2.0	2.9615(0.0958)	3.0826(0.1750)	3.9509(0.0060)	2.8887(0.0685)	2.9750(0.1000)	3.3121(0.0938)
0.5	2.0	2.9567(0.0141)	3.0464(0.1687)	3.9428(0.0024)	2.9549(0.1425)	3.0426(0.1195)	3.3372(0.1275)
0.8	2.0	3.0271(0.1597)	3.0896(0.1546)	3.9447(0.0081)	2.9613(0.0184)	3.0237(0.1622)	3.3591(0.0957)
				$d = 200, n = 500$			
Uniform	β	High-order	Cauchy	Non-truncation	High-order	Cauchy	Non-truncation
0.3	2.0	4.3418(0.0154)	4.9339(0.0452)	5.4107(0.0018)	4.0917(0.1150)	4.2971(0.0507)	4.8207(0.1894)
0.5	2.0	4.4815(0.0745)	5.0787(0.0687)	5.4075(0.0149)	4.1469(0.1168)	4.3671(0.1057)	4.8559(0.0574)
0.8	2.0	4.4230(0.0312)	5.0416(0.1101)	5.3869(0.0046)	4.1265(0.0615)	4.4216(0.0935)	4.8762(0.1569)
				$d = 1000, n = 1000$			
Uniform	β	High-order	Cauchy	Non-truncation	High-order	Cauchy	Non-truncation
0.3	2.0	10.6409(0.0149)	10.6697(0.0314)	10.7218(0.0176)	10.4514(0.0753)	10.7567(0.1724)	11.2571(0.0977)
0.5	2.0	10.6744(0.1574)	10.7760(0.1275)	10.7279(0.0128)	10.4686(0.0857)	10.6281(0.0694)	11.2776(0.1124)
0.8	2.0	10.7316(0.0715)	10.8551(0.0957)	10.7961(0.0196)	10.4816(0.0965)	10.7613(0.2020)	11.4005(0.0542)

Table 4: Comparison of average accuracy for DNN logistic regression under two noise settings.

Accuracy (%) for DNN logistic regression					
		$d = 6$, $n = 200$ (Complex function)		$d = 100$, $n = 1000$ (DNN)	
β	Pareto (τ)	High-order	Non-truncation	High-order	Non-truncation
1.5	1.60	86.06(0.04)	82.46(0.02)	81.90(0.01)	74.33(0.01)
1.5	1.80	85.60(0.02)	82.25(0.02)	80.04(0.01)	77.97(0.01)
2.0	2.01	86.57(0.04)	81.07(0.02)	83.49(0.01)	80.62(0.01)
2.0	4.01	85.93(0.04)	83.98(0.02)	93.36(0.02)	80.13(0.01)
2.0	6.01	87.90(0.01)	85.06(0.05)	95.75(0.02)	84.54(0.01)
β	Uniform (π)	High-order	Non-truncation	High-order	Non-truncation
2.0	0.3	84.00(0.02)	81.33(0.02)	89.05(0.01)	88.28(0.01)
2.0	0.5	82.64(0.02)	80.26(0.03)	88.28(0.01)	87.86(0.01)
2.0	0.8	82.04(0.01)	80.47(0.02)	88.61(0.01)	87.81(0.01)

5.3 Real data analysis

5.3.1 Boston housing dataset

We use the Boston housing dataset provided by the python library Scikit-Learn to learn the log-truncated standard and deep LAD models. Boston housing dataset contains $n = 506$ cases, and each case includes 14 variables. We aim to predict Median Value (MEDV) of Owner-Occupied Housing Units as output, by the remaining 13 variables as input. To this end, we randomly split the dataset into two groups, one as the training set and the other as the testing set, and train a l_2-regularized standard LAD regression in (27) and a 3-layers elastic net penalized DNN LAD regression model separately.

We denote the variable MEDV by $Y \in \mathbb{R}$ and the other 13 variables by $X \in \mathbb{R}^{13}$, so the Boston housing dataset can be represented as $\{(Y_i, X_i)\}_{i=1}^{506}$. In our experiment, $n_1 = 339$ samples are randomly selected for training and validation, and the remaining $n_2 = 167$ samples are the testing set, denoted by $\{(Y_{tr,1}, X_{tr,1}), ..., (Y_{tr,n_1}, X_{tr,n_1})\}$ and $\{(Y_{te,1}, X_{te,1}), ..., (Y_{te,n_2}, X_{te,n_2})\}$ respectively. We use 4/5 of the training samples to train the log-truncated standard LAD model (DNN model), then we select the optimal parameters on the remaining 1/5 of the training set, and feed the testing set $\{X_{te,1}, ..., X_{te,n_2}\}$ into the trained model to get a prediction set $\{\hat{Y}_1, ..., \hat{Y}_{n_2}\}$. To assess the obtained model, we compute the absolute average errors (MAEs) of the prediction, which is defined as

$$\mathrm{MAE} := \frac{1}{n_2} \sum_{i=1}^{n_2} |\hat{Y}_i - Y_{te,i}|.$$

The function λ in (5) is chosen as $\lambda(x) = |x|^\beta/\beta$. To select an appropriate β, we consider ten values of $\beta \in (1, 2]$ with $\beta = 1.1, 1.2, ..., 1$; see Table 5. The corresponding normal LAD regressions without truncation are also considered, which are trained in the same way. Table 5 indicates that the log-truncated ℓ_2-regularized standard LAD regression has smaller prediction errors than its non-truncated version. The truncated deep LAD model outperforms the non-truncated deep LAD model in all settings. In particular, as $\beta = 1.8$, the prediction errors are smallest for both truncated ℓ_2-regularized standard LAD model and truncated deep LAD model. Thus, we choose $\beta = 1.8$ for these two models.

Table 5: Comparison of MAEs on Boston housing dataset.

	Standard LAD regression		Deep LAD regression	
β	Truncation	Non-truncation	Truncation	Non-truncation
1.1	6.2538	6.4843	6.2822	6.3186
1.2	6.2551	6.4843	6.2203	6.3186
1.3	6.2563	6.4843	6.2150	6.3186
1.4	6.2578	6.4843	6.1898	6.3186
1.5	6.2592	6.4843	6.1810	6.3186
1.6	6.2611	6.4843	6.2096	6.3186
1.7	6.2631	6.4843	6.1620	6.3186
1.8	**6.2448**	6.4843	**6.0628**	6.3186
1.9	6.2676	6.4843	6.1598	6.3186
2.0	6.2705	6.4843	6.1711	6.3186

The optimal β in Table 5 can be roughly interpreted by Theorem 4, from which we can see that there is a trade-off between the constant $\mathbb{E}H_{Y,X}^{\beta}$ and the excess risk convergence rate $\frac{1}{n^{(\beta-1)/\beta}}$ if β varies from 1.1 to 2.0.

5.3.2 MNIST database

We use a handwritten digits database MNIST to learn a 6-layers elastic net penalized DNN LAD model, and compared it with the non-truncated model, which is learned in the same way. The activation function in (18) is ReLU. The two DNN models with elastic-net regularization are learned with Adam optimization algorithm. MNIST database contains 70000 28×28 grayscale images of the 10 digits. In experiments, we treat the digit images as the input variables X and their corresponding labels as output $Y \in \{0, 1, 2, \cdots, 9\}$. We randomly split the 70000 images into three groups: the validation set (10000 images), denoted by $\{(X_{va,i}, Y_{va,i})\}_{i=1}^{10000}$; training set (50000 images), denoted by $\{(X_{tr,i}, Y_{tr,i})\}_{i=1}^{50000}$ and testing set (10000 images), denoted by and $\{(X_{te,i}, Y_{te,i})\}_{i=1}^{10000}$ respectively. Firstly, we normalize the 28×28 pixels of each image in range $(-1, 1)$ and train several candidate DNN models using the parameters $(\alpha, \rho, \gamma) \in \mathbb{R}_+^3$. Then, we select the optimal parameters by computing the classification accuracy of their corresponding trained DNN models on the validation set, that is:

$$\text{Accuracy}(\alpha, \rho, \gamma) := \frac{1}{10000} \sum_{i=1}^{10000} \mathbf{1}(\hat{Y}_{va,i}(\alpha, \rho, \gamma) = Y_{va,i}) \times 100\%, \tag{32}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function, $\{\hat{Y}_{va,i}\}_{i=1}^{10000}$ are the predictors from 10000 images in validation set. The batch size is 64. Next, we feed the 10000 testing images into the trained DNN model which is corresponded to the optimal parameters and compute their classification accuracy.

Three types of noises are added on the 28×28 pixels of the original data: Gaussian noise $N(0.5, 2)$, uniform noise $U(2, 5)$ and Pareto noise with shape parameter $\beta = 2.01$. For each batch, we randomly generate $20\%, 50\%$ and 80% noises to contaminate the data. The truncation function is also chosen as (5) with $\lambda(x) = |x|^\beta/\beta$. We select the optimal values of β on $(1, 2]$ with step 0.1 for the three classes of noisy settings. We repeat the experiment 100 times and record the average accuracy, see Table 6 (the values in the bracket are standard errors of accuracy). The results in Table 6 show that the classification results of the truncated DNN LAD model are better than the standard DNN LAD model. Especially under strong noise disturbance, the truncated DNN LAD model is more robust than the standard DNN LAD model.

We can see from the tables that the optimal index β does not change according to the proportions of noises, and this is a significant advantage of our truncation regression models.

Table 6: Comparison of classification accuracy on MNIST dataset

Accuracy (%)			
Gaussian Noise $N(0.5, 2)$			
β	Noise proportion	Truncation	Non-truncation
2.0	20%	86.85 (0.01)	75.71 (0.01)
2.0	50%	84.39 (0.02)	74.27 (0.03)
2.0	80%	76.06 (0.03)	72.96 (0.03)
Uniform Noise $U(2, 5)$			
β	Noise proportion	Truncation	Non-truncation
1.5	20%	96.56 (0.01)	95.91 (0.09)
1.5	50%	93.87 (0.01)	92.46 (0.01)
1.5	80%	93.45 (0.03)	91.65 (0.01)
Pareto Noise $\beta = 2.01$			
β	Noise proportion	Truncation	Non-truncation
2.0	20%	88.75 (0.02)	87.85 (0.01)
2.0	50%	76.03 (0.03)	74.74 (0.01)
2.0	80%	65.12 (0.02)	60.96 (0.04)

6. Discussion and future study

Our proposed robust elastic net estimators, in practice, need to be obtained by SGD-based algorithms, which are deserved to be studied in future research.

From the simulations and real data analysis above, we can see that selecting the β is a crucial issue for prediction. The tail index estimation has been intensively studied in extreme-value statistics; see Fedotenkov (2020) for a review. We leave the research of estimating the index β to future study.

As mentioned above, Theorem 4 or Theorem 7 can be applied to study other robust statistical learning models, in which one may have to design specific algorithms according to the concrete problems at hand. We conclude this paper with the following two examples, robust two-component mixed linear regression and robust non-negative matrix factorization, whose algorithms differ from SGD. Here we only roughly address their theoretical results and leave detailed study to future work.

Robust two-component mixed linear regression. One challenging non-convex problem is the mixture of two linear regressions. Suppose $\{(Y_i, X_i) \sim (Y, X)\}_{i=1}^n$ are $\mathbb{R} \times \mathbb{R}^d$-valued i.i.d. random variables. With probability π, (X, Y) has conditional density function $p(y, x^\top \eta_0),\ \eta_0 \in \mathbb{R}^d$ for $Y = y|X = x$, and with probability $1 - \pi$, (X, Y) has conditional density function $p(y, x^\top \eta_1),\ \eta_1 \in \mathbb{R}^d$ for $Y = y|X = x$, where η_0 and η_1 are unknown coefficients. Given the input x, the negative log-likelihood function of the output y is

$$l(y, x; \pi, \eta_0, \eta_1) = -\log[\pi p(y, x^\top \eta_0) + (1 - \pi) p(y, x^\top \eta_1)], \tag{33}$$

where $\pi \in (0, 1)$ is an unknown mixing probability. Mei et al. (2018) studied Gaussian mixture models, while Khamaru and Wainwright (2019) considered mixture density estimation under sub-exponential condition.

In order to fit this example to our theory, we write $\theta := (\pi, \eta_0, \eta_1)$. Under the following moment conditions: $\mathbb{E}(\|X\|_2 |Y|)^\beta < \infty$, $\mathbb{E}\|X\|_2^{2\beta} < \infty$, by Theorem 4 with $p = 2d + 1$, we can obtain an excess risk in the order of $O(((2d + 1)\log(nr_n)/n)^{(\beta-1)/\beta})$ if the regularization error $\rho\|\theta^*\|_2^2 \lesssim (\frac{(2d+1)\log(nr_n)}{n})^{(\beta-1)/\beta}$; see Appendix A.9 for details.

Robust non-negative matrix factorization. Given n vector samples $\{X_i\}_{i=1}^n \in \mathbb{R}^p$ arranged in a nonnegative matrix $\mathrm{S} := [X_1, \ldots, X_n] \in \mathbb{R}_+^{d \times n}$ and positive integer $z \geq 1$, the log-truncated non-negative matrix factorization (NMF) considers the decomposition for S into a product of two non-negative matrices:

$$\mathrm{S} = \mathrm{BC} + \mathrm{R},$$

where $\mathrm{B} \in \mathbb{R}_+^{d \times z}$ is the given basis, $\mathrm{C} = [h_1, \ldots, h_n] \in \mathbb{R}_+^{z \times n}$ is the coefficients, and $\mathrm{R} \in \mathbb{R}^{d \times n}$ is the random error matrix.

By minimizing the ℓ_2-distance between their product and the original data matrix, the ordinary NMF decomposes a data matrix into the product of two lower dimensional non-negative factor matrices B and C. When the original data matrix is corrupted by heavy-tailed outliers that seriously violate the second-moment assumption (Guan et al., 2017). We assume the element of R_{ij} in R has only β-th moment, i.e. $\sup_{i \in [n], j \in [d]} \mathbb{E}|R_{ij}|^\beta < \infty$ with $\beta \in (1, 2)$, it is of interest to study the log-truncated NMF and the robust algorithm by using the following non-convex optimization problem:

$$\hat{\mathrm{C}} = \mathop{\arg\min}_{\mathrm{C} \in \mathbb{R}_+^{z \times n}} \frac{1}{np\alpha} \sum_{i \in [n], j \in [d]} \psi_\lambda \left[\alpha\,(\mathrm{S} - \mathrm{BC})_{ij}^2\right]$$

where $\lambda(x) = \beta^{-1}|x|^\beta$, and the α is tuning parameter implicitly determined by the random error matrix. We expect that $\hat{\mathrm{C}}$ is able to learn a subspace on a dataset through the original data matrix that is contaminated by a heavy-tailed noise matrix.

Acknowledgments

F. Yao is supported by the National Key R&D Program of China (No. 2020YFE0204200, 2022YFA1003801), the National Natural Science Foundation of China (No. 12292981, 11931001), the LMAM and the LMEQF. L. Xu is supported in part by the National Natural Science Foundation of China (No. 12071499), Macao S.A.R grant FDCT 0090/2019/A2 and University of Macau grant MYRG2018-00133-FST. H. Zhang is supported in part by the National Natural Science Foundation of China (No. 12101630) and the University of Macau under UM Macao Talent Programme (UMMTP-2020-01).

Appendix A. Proofs and additional results

The appendix includes the proofs of the lemmas, corollaries, and theorems in the main body, and additional remarks and simulation results is also given.

A.1 Useful lemmas of covering number bounds

The following covering number bound of ℓ_2-ball is in Corollary 4.2.13 in Vershynin (2018).

Lemma 15 (Covering number bound of the ℓ_2-ball) *For any $\kappa > 0$, the covering number of the p-dimensional unit ℓ_2-ball $B_2^p(1)$ satisfy*

$$\left(\tfrac{1}{\kappa}\right)^p \le N\left(B_2^p(1), \kappa\right) \le \left(\tfrac{2}{\kappa}+1\right)^p.$$

The next lemma is modified from Vershynin (2009). Here, we provide a sharper covering number bound, while (2) in Vershynin (2009) contains a unknown universal constant.

Lemma 16 (Covering number bound of the s-sparse ℓ_2-ball) *For any $\kappa > 0$, the covering number of the p-dimensional unit s-sparse ℓ_2-ball satisfy*

$$N\left(B_2^p(1) \cap B_0^p(s), \kappa\right) < \tfrac{1}{\sqrt{2es}} \left(\tfrac{(\kappa+2)ep}{\kappa s}\right)^s.$$

Proof To get the result, we consider a union bound over k-dimensional subspaces of $B_2^p(1)$ by using the upper bound in Lemma 15 to bound the covering number of s-sparse ℓ_2-ball,

$$N\left(B_2^p(1) \cap B_0^p(s), \kappa\right) \le \begin{pmatrix} p \\ s \end{pmatrix} N\left(B_2^s, \kappa\right) < \tfrac{(ep/s)^s}{\sqrt{2es}} \left(\tfrac{2}{\kappa}+1\right)^s = \tfrac{1}{\sqrt{2es}} \left(\tfrac{(\kappa+2)ep}{\kappa s}\right)^s,$$

where the last inequality is by Stirling's approximation $\begin{pmatrix} p \\ s \end{pmatrix} < \frac{(ep/s)^s}{\sqrt{2es}}$ for $1 \le s \le p$.

■

A.2 The proof of Theorem 2

Proof By the definition of $\hat{\theta}_n$, we have for all $\theta^* \in \Theta^*$

$$\begin{aligned}\hat{R}_{\psi_\lambda,l,\alpha}(\hat{\theta}_n) + \rho\|\hat{\theta}_n\|_2^2 &= \frac{1}{n\alpha}\sum_{i=1}^n \psi_\lambda[\alpha l(Y_i, X_i, \hat{\theta}_n)] + \rho\|\hat{\theta}_n\|_2^2 \\ &\le \frac{1}{n\alpha}\sum_{i=1}^n \psi_\lambda[\alpha l(Y_i, X_i, \theta^*)] + \rho\|\theta^*\|_2^2 =: \hat{R}_{\psi_\lambda,l,\alpha}(\theta^*) + \rho\|\theta^*\|_2^2 \qquad (34)\end{aligned}$$

which yields $\hat{R}_{\psi_\lambda,l,\alpha}(\hat{\theta}_n) - \hat{R}_{\psi_\lambda,l,\alpha}(\theta^*) \le \rho(\|\theta^*\|_2^2 - \|\hat{\theta}_n\|_2^2) \le \rho\|\theta^*\|_2^2$ and thus

$$\begin{aligned}R_l(\hat{\theta}_n) - R_l(\theta^*) &= [R_l(\hat{\theta}_n) - \hat{R}_{\psi_\lambda,l,\alpha}(\hat{\theta}_n)] + [\hat{R}_{\psi_\lambda,l,\alpha}(\hat{\theta}_n) - \hat{R}_{\psi_\lambda,l,\alpha}(\theta^*)] + [\hat{R}_{\psi_\lambda,l,\alpha}(\theta^*) - R_l(\theta^*)] \\ &\le [\hat{R}_{\psi_\lambda,l,\alpha}(\theta^*) - R_l(\theta^*)] + [R_l(\hat{\theta}_n) - \hat{R}_{\psi_\lambda,l,\alpha}(\hat{\theta}_n)] + \rho\|\theta^*\|_2^2. \qquad (35)\end{aligned}$$

Under the two high provability events $\mathcal{E}_1(l, \lambda, \theta^*)$ in Lemma 17 and $\mathcal{E}_2(l, \lambda, \hat{\theta}_n)$ in Lemma 18 below, we have by inequality (35)

$$\begin{aligned}&R_l(\hat{\theta}_n) - R_l(\theta^*) \le \left(\frac{f(\alpha)}{\alpha}R_{\lambda\circ l}(\theta^*) + \frac{1}{n\alpha}\log\frac{1}{\delta}\right) \\ &+ 2\kappa\mathbb{E}H_{Y,X} + \frac{c_2 f(\alpha)}{\alpha}\sup_{\theta\in\Theta}R_{\lambda\circ l}(\theta) + \frac{c_2 f(\alpha\kappa)}{\alpha}\mathbb{E}[\lambda(H_{Y,X})] + \frac{1}{n\alpha}\log\frac{N(\Theta,\kappa)}{\delta} + \rho\|\theta^*\|_2^2 \\ &= 2\kappa\mathbb{E}H_{Y,X} + \frac{c_2 f(\alpha\kappa)}{\alpha}\mathbb{E}[\lambda(H_{Y,X})] + \frac{(c_2+1)f(\alpha)}{\alpha}R_{\lambda\circ l}(\Theta) + \frac{1}{n\alpha}\log\frac{N(\Theta,\kappa)}{\delta^2} + \rho\|\theta^*\|_2^2, \qquad (36)\end{aligned}$$

with probability at least $1 - 2\delta$.

For the last two terms in (36), we put $\frac{(c_2+1)f(\alpha)}{\alpha}R_{\lambda\circ l}(\Theta) = \frac{1}{n\alpha}\log\frac{N(\Theta,\kappa)}{\delta^2}$, i.e. the variance term equals to the bias term. Then we get $\alpha = f^{-1}\left(\frac{1}{n(c_2+1)}R_{\lambda\circ l}^{-1}(\Theta)\log\frac{N(\Theta,\kappa)}{\delta^2}\right)$. So (36) implies

$$R_l(\hat{\theta}_n) \le R_l(\theta^*) + 2\kappa\mathbb{E}H_{Y,X} + \frac{c_2 f(\alpha\kappa)}{\alpha}\mathbb{E}[\lambda(H_{Y,X})] + \frac{2}{n\alpha}\log\frac{N(\Theta,\kappa)}{\delta^2} + \rho\|\theta^*\|_2^2.$$

Let $\kappa = \frac{1}{n}$ and take infimum over for each $\theta^* \in \Theta^*$, we obtain

$$R_l(\hat{\theta}_n) - \inf_{\theta\in\Theta}R_l(\theta) \le \frac{2\mathbb{E}H_{Y,X}}{n} + \frac{c_2\mathbb{E}[\lambda(H_{Y,X})]}{\alpha}f(\frac{\alpha}{n}) + \frac{2}{n\alpha}\log\frac{N(\Theta,n^{-1})}{\delta^2} + \rho\|\Theta^*\|_2^2,$$

with probability at least $1 - 2\delta$.

■

Lemma 17 (Concentration error bound) *For general loss function $l(\cdot,\cdot,\cdot)$, under (C.5), we have for all $\theta^* \in \Theta^*$*

$$\mathbb{P}\{\mathcal{E}_1(l, \lambda, \theta^*)\} \ge 1 - \delta,$$

where $\mathcal{E}_1(l, \lambda, \theta^*) := \{\hat{R}_{\psi_\lambda,l,\alpha}(\theta^*) - R_l(\theta^*) \le \frac{f(\alpha)}{\alpha}R_{\lambda\circ l}(\theta^*) + \frac{1}{n\alpha}\log\frac{1}{\delta}\}$.

Proof The proof is based on bounding the exponential moment of $n\alpha\hat{R}_{\psi_\lambda,l,\alpha}(\theta^*)$ from Markov's inequality. Applying the upper truncated function in (5), $1+x \le e^x$, one has

$$\begin{aligned}
\mathbb{E}e^{n\alpha\hat{R}_{\psi_\alpha\circ l}(\theta^*)} &= \mathbb{E}e^{\sum\limits_{i=1}^{n}\psi[\alpha l(Y_i,X_i,\theta^*)]} \le \mathbb{E}\{\prod_{i=1}^{n}[1+\alpha l(Y_i,X_i,\theta^*)+\lambda[\alpha l(Y_i,X_i,\theta^*)]]\} \\
\text{[By independence]} &= \prod_{i=1}^{n}\{\mathbb{E}[1+\alpha l(Y_i,X_i,\theta^*)+\lambda[\alpha l(Y_i,X_i,\theta^*)]]\} \\
&\le e^{\alpha\sum_{i=1}^{n}\mathbb{E}l(Y_i,X_i,\theta^*)+\sum_{i=1}^{n}\mathbb{E}\{\lambda[\alpha l(Y_i,X_i,\theta^*)]\}} \le e^{n[\alpha R_l(\theta^*)+f(\alpha)R_{\lambda\circ l}(\theta^*)]}.
\end{aligned}$$

Via Markov's inequality with the exponential transform, it gives

$$\begin{aligned}
&\mathbb{P}\{\mathcal{E}_1^c(l,\lambda,\theta^*)\} = \mathbb{P}\{\hat{R}_{\psi_\lambda,l,\alpha}(\theta^*) > [R_l(\theta^*)+\frac{f(\alpha)}{\alpha}R_{\lambda\circ l}(\theta^*)]+\frac{\log(1/\delta)}{n\alpha}\} \\
&= \mathbb{P}\{e^{n\alpha\hat{R}_{\psi_\lambda,l,\alpha}(\theta^*)} > e^{n\alpha[R_l(\theta^*)+\frac{f(\alpha)}{\alpha}R_{\lambda\circ l}(\theta^*)]+\log(1/\delta)}\} \le \frac{\mathbb{E}e^{n\alpha\hat{R}_{\psi_\lambda,l,\alpha}(\theta^*)}}{e^{n\alpha[R_l(\theta^*)+\frac{f(\alpha)}{\alpha}R_{\lambda\circ l}(\theta^*)]+\log(1/\delta)}} \le \delta.
\end{aligned}$$

■

Lemma 18 (Generalization error bound) *For any $\kappa > 0$, under (C.1)-(C.5), one has*

$$\mathbb{P}\{\mathcal{E}_2(l,\lambda,\hat{\theta}_n)\} \ge 1-\delta,$$

where $\mathcal{E}_2(l,\lambda,\hat{\theta}_n) := \{R_l(\hat{\theta}_n)-\hat{R}_{\psi_\lambda,l,\alpha}(\hat{\theta}_n) \le 2\kappa\mathbb{E}H_{Y,X}+\frac{c_2f(\alpha)}{\alpha}R_{\lambda\circ l}(\Theta)+\frac{c_2f(\alpha\kappa)}{\alpha}\mathbb{E}[\lambda(H_{Y,X})]+\frac{1}{n\alpha}\log\frac{N(\Theta,\kappa)}{\delta}\}$.

Proof Let $\mathcal{N}(\Theta,\kappa)$ be an κ-net of Θ and denote its covering number as $N(\Theta,\varepsilon)$. For each $\hat{\theta}_n \in \Theta$, the definition of κ-net implies that there exists a $\tilde{\theta} \in \mathcal{N}(\Theta,\kappa)$ satisfying $\|\hat{\theta}_n-\tilde{\theta}\|_2 \le \kappa$, by Lipschitz condition (C.3), we obtain

$$l(Y_i,X_i,\hat{\theta}_n) \ge l(Y_i,X_i,\tilde{\theta})-\kappa H_{Y_i,X_i},\ i=1,2,\cdots,n.$$

Since $\psi_\lambda(\cdot)$ is non-decreasing and the last inequality gives

$$\hat{R}_{\psi_\lambda,l,\alpha}(\hat{\theta}_n) = \frac{1}{n\alpha}\sum_{i=1}^{n}\psi_\lambda[\alpha l(Y_i,X_i,\hat{\theta}_n)] \ge \frac{1}{n\alpha}\sum_{i=1}^{n}\psi_\lambda[\alpha l(Y_i,X_i,\tilde{\theta})-\kappa\alpha H_{Y_i,X_i}]. \tag{37}$$

In below, we continue to derive the lower bound in (37) by applying covering number techniques. From the lower bound in (5), we could estimate the exponential moment bound:

$$\begin{aligned}
&\mathbb{E}e^{-\sum_{i=1}^{n}\psi_\lambda[\alpha l(Y_i,X_i,\tilde\theta)-\alpha\kappa H_{Y_i,X_i}]}\\
&\le \mathbb{E}\prod_{i=1}^{n}\{1-\alpha\mathbb{E}[l(Y_i,X_i,\tilde\theta)]+\alpha\kappa H_{Y_i,X_i}+\lambda[\alpha(l(Y_i,X_i,\tilde\theta)-\kappa H_{Y_i,X_i})]\}\\
&\le \prod_{i=1}^{n}\left\{1-\alpha\mathbb{E}[l(Y_i,X_i,\tilde\theta)]+\alpha\kappa\mathbb{E}H_{Y_i,X_i}+\mathbb{E}\{\lambda[\alpha(l(Y_i,X_i,\tilde\theta)-\kappa H_{Y_i,X_i})]\}\right\}\\
[(\text{C.1.2})]\ &\le \prod_{i=1}^{n}\left\{1-\alpha\mathbb{E}[l(Y_i,X_i,\tilde\theta)]+\alpha\kappa\mathbb{E}H_{Y_i,X_i}+c_2\mathbb{E}\{\lambda[\alpha(l(Y_i,X_i,\tilde\theta))]+\lambda[\alpha\kappa H_{Y_i,X_i}]\}\right\}\\
[(\text{C.1.1})]\ &\le e^{n\alpha\{-R_l(\tilde\theta)+\kappa\mathbb{E}H_{Y,X}+\frac{c_2f(\alpha)}{\alpha}R_{\lambda\circ l}(\tilde\theta)+\frac{c_2f(\alpha\kappa)}{\alpha}\mathbb{E}[\lambda(H_{Y,X})]\}},
\end{aligned}$$

where the last inequality stems from $1+x\le e^x$. By the last exponential moment bound, Markov's inequality shows that, for a fixed $\tilde\theta\in\mathcal{N}(\Theta,\kappa)$,

$$\begin{aligned}
&\mathbb{P}\left\{\frac{-1}{n\alpha}\sum_{i=1}^{n}\psi_\lambda(\alpha l(Y_i,X_i,\tilde\theta)-\kappa\alpha H_{Y_i,X_i})>-R_l(\tilde\theta)+\frac{c_2f(\alpha)}{\alpha}R_{\lambda\circ l}(\tilde\theta)\right.\\
&\left.+\kappa\mathbb{E}H_{Y,X}+\frac{c_2f(\alpha\kappa)}{\alpha}\mathbb{E}[\lambda(H_{Y,X})]+\frac{\log(1/s)}{n\alpha}\right\}\\
&\le\frac{e^{n\alpha\{-R_l(\tilde\theta)+\kappa\mathbb{E}H_{Y,X}+\frac{c_2f(\alpha)}{\alpha}R_{\lambda\circ l}(\tilde\theta)+\frac{c_2f(\alpha\kappa)}{\alpha}\mathbb{E}[\lambda(H_{Y,X})]\}}}{e^{n\alpha\{-R_l(\tilde\theta)+\kappa\mathbb{E}H_{Y,X}+\frac{c_2f(\alpha)}{\alpha}R_{\lambda\circ l}(\tilde\theta)+\frac{c_2f(\alpha\kappa)}{\alpha}\mathbb{E}[\lambda(H_{Y,X})]\}+\log(1/s)}}=s\in(0,1).
\end{aligned}\tag{38}$$

The set $\mathcal{N}(\Theta,\kappa)$ has $N(\Theta,\kappa)$ elements. From the single bound (38) and putting $s=\delta/N(\Theta,\kappa)$, we have for all $\tilde\theta\in\mathcal{N}(\Theta,\kappa)$

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{\tilde\theta\in\mathcal{N}(\Theta,\kappa)}\left\{\frac{-1}{n\alpha}\sum_{i=1}^{n}\psi_\lambda(\alpha l(Y_i,X_i,\tilde\theta)-\kappa\alpha H_{Y,X})>-R_l(\tilde\theta)+\frac{c_2f(\alpha)}{\alpha}R_{\lambda\circ l}(\tilde\theta)\right.\right.\\
&\left.\left.+\kappa\mathbb{E}H_{Y,X}+\frac{c_2f(\alpha\kappa)}{\alpha}\mathbb{E}[\lambda(H_{Y,X})]+\frac{\log(1/s)}{n\alpha}\right\}\right)\\
&\le N(\Theta,\kappa)\cdot\mathbb{P}\left\{\frac{-1}{n\alpha}\sum_{i=1}^{n}\psi_\lambda(\alpha l(Y_i,X_i,\tilde\theta)-\kappa\alpha H_{Y,X})\ge-R_l(\tilde\theta)+\frac{c_2f(\alpha)}{\alpha}R_{\lambda\circ l}(\tilde\theta)\right.\\
&\left.+\kappa\mathbb{E}H_{Y,X}+\frac{c_2f(\alpha\kappa)}{\alpha}\mathbb{E}[\lambda(H_{Y,X})]+\frac{\log(1/s)}{n\alpha}\right\}\le N(\Theta,\kappa)s=:\delta.
\end{aligned}\tag{39}$$

Then the complementary set in (39) hold with probability at least $1-\delta$. Inequalities (37) and (39) give the following lower bound for all $\tilde{\theta} \in \mathcal{N}(\Theta, \kappa)$

$$\begin{aligned}
&\hat{R}_{\psi_\lambda, l, \alpha}(\hat{\theta}) \geq \frac{1}{n\alpha} \sum_{i=1}^{n} \psi_\lambda[\alpha l(Y_i, X_i, \tilde{\theta}) - \kappa\alpha H_{Y,X}] \\
&\geq R_l(\tilde{\theta}) - \left\{ \kappa \mathbb{E} H_{Y,X} + \frac{c_2 f(\alpha)}{\alpha} R_{\lambda \circ l}(\tilde{\theta}) + \frac{c_2 f(\alpha\kappa)}{\alpha} \mathbb{E}[\lambda(H_{Y,X})] + \frac{1}{n\alpha} \log \frac{N(\Theta, \kappa)}{\delta} \right\} \\
&\geq R_l(\tilde{\theta}) - \left\{ \kappa \mathbb{E} H_{Y,X} + \frac{c_2 f(\alpha)}{\alpha} R_{\lambda \circ l}(\Theta) + \frac{c_2 f(\alpha\kappa)}{\alpha} \mathbb{E}[\lambda(H_{Y,X})] + \frac{1}{n\alpha} \log \frac{N(\Theta, \kappa)}{\delta} \right\} \qquad (40)
\end{aligned}$$

with probability at least $1-\delta$.

It remains to find the lower bound for $R_l(\tilde{\theta})$ in (40) by the error bound of $|R_l(\hat{\theta}_n) - R_l(\tilde{\theta})|$. To this end, the (C.2) implies

$$R_l(\hat{\theta}_n) - R_l(\tilde{\theta}) \leq \mathbb{E}[H_{Y,X} \|\hat{\theta}_n - \tilde{\theta}\|_2] \leq \kappa \mathbb{E} H_{Y,X} \quad \hat{\theta}, \tilde{\theta} \in \Theta,$$

which gives $R_l^n(\tilde{\theta}) \geq R_l(\hat{\theta}_n) - \kappa \mathbb{E} H_{Y,X}$. Thus, (40) has a further lower bound:

$$\begin{aligned}
&\hat{R}_{\psi_\lambda, l, \alpha}(\hat{\theta}_n) \\
&\geq R_l(\hat{\theta}_n) - \{2\kappa \mathbb{E} H_{Y,X} + \frac{c_2 f(\alpha)}{\alpha} R_{\lambda \circ l}(\Theta) + \frac{c_2 f(\alpha\kappa)}{\alpha} \mathbb{E}[\lambda(H_{Y,X})] + \frac{1}{n\alpha} \log \frac{N(\Theta, \kappa)}{\delta} \}
\end{aligned}$$

with probability at least $1-\delta$. Then, we conclude Lemma 18. ■

A.3 The proof of Theorem 4

Proof Let $\kappa = 1/n$ in (36), we get with probability at least $1-2\delta$

$$\begin{aligned}
&R_l(\hat{\theta}_n) - \inf_{\theta \in \Theta} R_l(\theta) \\
&\leq \frac{2}{n} \mathbb{E} H_{Y,X} + \frac{c_2 f(\alpha/n)}{\alpha} \mathbb{E}[\lambda(H_{Y,X})] + \frac{(c_2+1) f(\alpha)}{\alpha} R_{\lambda \circ l}(\Theta) + \frac{1}{n\alpha} \log \frac{N(\Theta, 1/n)}{\delta^2} + \rho \|\Theta^*\|_2^2.
\end{aligned}$$

The $\lambda(x) = |x|^\beta/\beta,\ \beta \in (1,2)$ satisfies weak triangle inequality and homogeneous inequality

$$|x+y|^\beta/\beta \leq 2^{\beta-1}[|x|^\beta/\beta + |y|^\beta/\beta], \quad |tx|^\beta/\beta \leq |t|^\beta \cdot |x|^\beta/\beta \qquad (41)$$

by $|a+b|^\beta \leq 2^{\beta-1}(|a|^\beta + |b|^\beta)$ for $\beta > 1$. Thus, we have $c_2 = 2^{\beta-1}$ and $f(t) = t^\beta$ for $t > 0$ in (C.1).

Using the variance-bias tradeoff, i.e. the variance term equals to the bias term, put $\frac{(c_2+1)f(\alpha)}{\alpha} R_{\lambda \circ l}(\Theta) = \frac{1}{n\alpha} \log \frac{N(\Theta,\kappa)}{\delta^2}$. Note that $f^{-1}(t) = t^{1/\beta}$ for $t > 0$. So we have

$$\alpha = f^{-1}\left(\tfrac{1}{n(c_2+1)} R_{\lambda \circ l}^{-1}(\Theta) \log \tfrac{N(\Theta,\kappa)}{\delta^2} \right) = \tfrac{1}{n^{1/\beta}} \left(\tfrac{\log[N(\Theta,\kappa)/\delta^2]}{(2^{\beta-1}+1) R_{\lambda \circ l}(\Theta)} \right)^{1/\beta}.$$

Observe that $f(t)/t = t^{\beta-1}$ for $t > 0$. Then

$$\begin{aligned}
\frac{(c_2+1)f(\alpha)}{\alpha}R_{\lambda\circ l}(\Theta)+\frac{1}{n\alpha}\log\frac{N(\Theta,1/n)}{\delta^2} &= \frac{2(c_2+1)f(\alpha)}{\alpha}R_{\lambda\circ l}(\Theta)\\
=\frac{2(2^{\beta-1}+1)}{n^{(\beta-1)/\beta}}\left(\frac{\log[N(\Theta,1/n)/\delta^2]}{(2^{\beta-1}+1)R_{\lambda\circ l}(\Theta)}\right)^{(\beta-1)/\beta} & R_{\lambda\circ l}(\Theta)\\
=\frac{2(2^{\beta-1}+1)}{n^{(\beta-1)/\beta}}\left(\frac{\log[N(\Theta,1/n)/\delta^2]}{2^{\beta-1}+1}\right)^{(\beta-1)/\beta} & [R_{\lambda\circ l}(\Theta)]^{\beta^{-1}}\\
\leq\frac{2(2^{\beta-1}+1)}{n^{(\beta-1)/\beta}}\left(\frac{\log(\delta^{-2})+p\log\left(1+2r/\kappa\right)}{2^{\beta-1}+1}\right)^{(\beta-1)/\beta} & [R_{\lambda\circ l}(\Theta)]^{\beta^{-1}},
\end{aligned}$$

where the last inequality is by

$$\log\frac{N(\Theta,\kappa)}{\delta^2}\leq\log\frac{N(B_2^p(r_n),\kappa)}{\delta^2}\leq\log(\frac{1}{\delta^2})+p\log\left(1+\frac{2r_n}{\kappa}\right) \tag{42}$$

from Lemma 15 and (C.2).

Thus, $\frac{f(\alpha/n)}{\alpha}=\alpha^{\beta-1}n^{-\beta}$ and $\lambda(x)=|x|^{\beta}/\beta$ show

$$\begin{aligned}
\frac{c_2 f(\alpha/n)}{\alpha}\mathbb{E}[\lambda(H_{Y,X})] &= \frac{2^{\beta-1}/\beta}{n^{1-\beta^{-1}+\beta}}\left(\frac{\log[N(\Theta,1/n)/\delta^2]}{(2^{\beta-1}+1)R_{\lambda\circ l}(\Theta)}\right)^{(\beta-1)/\beta}\mathbb{E}H_{Y,X}^{\beta}\\
&\leq\frac{2^{\beta-1}/\beta}{n^{1-\beta^{-1}+\beta}}\left(\frac{\log(\delta^{-2})+p\log\left(1+2rn\right)}{2^{\beta-1}+1}\right)^{(\beta-1)/\beta}\frac{\mathbb{E}H_{Y,X}^{\beta}}{[R_{\lambda\circ l}(\Theta)]^{(\beta-1)/\beta}},
\end{aligned}$$

where the last inequality is by (42). Then we have

$$\begin{aligned}
R_l(\hat{\theta}_n)-\inf_{\theta\in\Theta}R_l(\theta) &\leq \frac{2}{n}\mathbb{E}H_{Y,X}+\frac{1}{n^{\frac{\beta-1}{\beta}}}\left(\frac{\log(\delta^{-2})+p\log\left(1+2r_n n\right)}{2^{\beta-1}+1}\right)^{\frac{\beta-1}{\beta}}\\
&\quad\cdot\left[\frac{2^{\beta-1}\mathbb{E}H_{Y,X}^{\beta}}{\beta n^{\beta}[R_{\lambda\circ l}(\Theta)]^{\frac{\beta-1}{\beta}}}+2(2^{\beta-1}+1)[R_{\lambda\circ l}(\Theta)]^{\beta^{-1}}\right]\\
&\leq\frac{2\mathbb{E}H_{Y,X}}{n}+C_{\beta,R_{\lambda\circ l}}\left[\frac{C_{\delta,n,r}(p)}{n}\right]^{\frac{\beta-1}{\beta}}+\rho\|\Theta^*\|_2^2
\end{aligned} \tag{43}$$

with probability at least $1-2\delta$.

Additionally, if we consider the conditions $\mathbb{E}H_{Y,X}=q_n$ and $\mathbb{E}H_{Y,X}^{\beta}=z_{n,\beta}$. By (43), the excess risk has a convergence rate

$$\begin{aligned}
R_l(\hat{\theta}_n)-\inf_{\theta\in\Theta}R_l(\theta) &\leq \frac{2q_n}{n}+\frac{1}{n^{\frac{\beta-1}{\beta}}}\left(\frac{\log(\delta^{-2})+p\log\left(1+2r_n n\right)}{2^{\beta-1}+1}\right)^{\frac{\beta-1}{\beta}}\\
&\quad\cdot\left[\frac{2^{\beta-1}z_{n,\beta}}{\beta n^{\beta}[R_{\lambda\circ l}(\Theta)]^{\frac{\beta-1}{\beta}}}+2(2^{\beta-1}+1)[R_{\lambda\circ l}(\Theta)]^{\beta^{-1}}\right]\\
&=O_p\left(\frac{q_n}{n}+\left(1+\frac{z_{n,\beta}}{n^{\beta}}\right)\left(\frac{p\log(nr_n)}{n}\right)^{\frac{\beta-1}{\beta}}+\rho\|\Theta^*\|_2^2\right).
\end{aligned}$$

Details for Remark 5. NBR loss has $H_{Y,X} \propto Y \left\|X\right\|_2 \le \sqrt{d}Y\|X\|_\infty$, and we need $\mathbb{E}H_{Y,X} \propto \mathbb{E}[\sqrt{d}Y\|X\|_\infty] \le \sqrt{d}[\mathbb{E}Y^2]^{1/2}[\mathbb{E}\|X\|_\infty^2]^{1/2} = o(n)$ and

$$\mathbb{E}H_{Y,X}^\beta \propto \mathbb{E}[\sqrt{d}Y\|X\|_\infty]^\beta \le d^{\beta/2}[\mathbb{E}Y^{2\beta}]^{1/2}[\mathbb{E}\|X\|_\infty^{2\beta}]^{1/2} = O(n^\beta)$$

under conditions $n > d = o(n^2)$, $\mathbb{E}Y^{2\beta} < \infty$ and $\mathbb{E}\|X\|_\infty^{2\beta} < \infty$. ■

A.4 The proof of Theorem 7

Proof Similar to the inequality (36) in the proof of Theorem 2, with ridge penalty replaced by elastic net penalty, we have for all $\theta^* \in \Theta^*$

$$\begin{aligned} R_l(\hat{\theta}_n) - R_l(\theta^*) \le{}& 2\kappa\mathbb{E}H_{Y,X} + \frac{c_2 f(\alpha\kappa)}{\alpha}\mathbb{E}[\lambda(H_{Y,X})] \\ &+ \frac{(c_2+1)f(\alpha)}{\alpha}R_{\lambda\circ l}(\Theta) + \frac{1}{n\alpha}\log\frac{N(\Theta,\kappa)}{\delta^2} + \rho\|\theta^*\|_2^2 + \gamma\|\theta^*\|_1, \end{aligned} \tag{44}$$

with probability at least $1-2\delta$.

For the term $\frac{1}{n\alpha}\log\frac{N(\Theta,\kappa)}{\delta^2}$ in (44) with $\kappa = 1/n$, Lemma 16 implies

$$\log N\left(\Theta, 1/n\right) \le \log N\left(B_2^p(r_n) \cap B_0^p(s_n), 1/n\right) \le \log\left(\frac{1}{\sqrt{2es_n}}\right) + s_n\log\left[\frac{e(1+2r_nn)p}{s_n}\right].$$

Let $f(t) = t^\beta$ and we have $c_2 = 2^{\beta-1}$, which shows that

$$\begin{aligned} R_l(\hat{\theta}_n) - R_l(\theta^*) \le{}& \rho\|\theta^*\|_2^2 + \gamma\|\theta^*\|_1 + \frac{2\mathbb{E}H_{Y,X}}{n} + \frac{c_2 f(\alpha\kappa)}{\alpha}\mathbb{E}[\lambda(H_{Y,X})] \\ &+ \frac{(c_2+1)f(\alpha)}{\alpha}R_{\lambda\circ l}(\Theta) + \frac{1}{n\alpha}\left\{s_n\log\left[e(1+2r_nn)\frac{p}{s_n}\right] + \log\left(\frac{\delta^{-2}}{\sqrt{2es_n}}\right)\right\} \end{aligned} \tag{45}$$

with probability at least $1-\delta$.

For the last two terms in (45), we put

$$\alpha^{\beta-1}(c_2+1)R_{\lambda\circ l}(\Theta) = \tfrac{1}{n\alpha}\left\{s_n\log\left[e(1+2r_nn)\tfrac{p}{s_n}\right] + \log\left(\tfrac{\delta^{-2}}{\sqrt{2es_n}}\right)\right\}.$$

Then we obtain $\alpha = \frac{1}{n^{1/\beta}}\left(\frac{\log(\delta^{-2}/\sqrt{2es_n})+s_n\log[(1+2r_nn)ep/s_n]}{(2^{\beta-1}+1)R_{\lambda\circ l}(\Theta)}\right)^{1/\beta}$. Moreover, in (44)

$$\begin{aligned} \frac{c_2 f(\alpha/n)}{\alpha}\mathbb{E}[\lambda(H_{Y,X})] &= \frac{2^{\beta-1}/\beta}{n^{1-\beta^{-1}+\beta}}\left(\frac{\log[N(\Theta,1/n)/\delta^2]}{(2^{\beta-1}+1)R_{\lambda\circ l}(\Theta)}\right)^{(\beta-1)/\beta}\mathbb{E}H_{Y,X}^\beta \\ &\le \frac{2^{\beta-1}/\beta}{n^{1-\beta^{-1}+\beta}}\left(\log(\frac{\delta^{-2}}{2es_n}) + s_n\log\left[(1+2r_nn)\frac{ep}{s_n}\right]\right)^{(\beta-1)/\beta}\frac{\mathbb{E}H_{Y,X}^\beta}{[R_{\lambda\circ l}(\Theta)]^{(\beta-1)/\beta}}. \end{aligned}$$

By the above terms, with probability at least $1-2\delta$, one has

$$R_l(\hat{\theta}_n) - \inf_{\theta\in\Theta}R_l(\theta) \le \frac{2\mathbb{E}H_{Y,X}}{n} + \frac{C_{\beta,R_{\lambda\circ l}}}{n^{\frac{\beta-1}{\beta}}}\left(\log(\frac{\delta^{-2}}{2es_n}) + s_n\log\left[(1+2r_nn)\frac{ep}{s_n}\right]\right)^{\frac{\beta-1}{\beta}} + \|\Theta^*\|_{\rho,\gamma},$$

where $C_{\beta,R_{\lambda\circ l}}$ is a constant given in Theorem 4, and $\|\Theta^*\|_{\rho,\gamma} := \inf_{\theta^*\in\Theta^*}(\rho\|\theta^*\|_2^2+\gamma\|\theta^*\|_1)$. ■

A.5 The proof of Corollary 9

We use Knight's identity (Knight, 1998) to obtain the expression of $H_{y,x}$ for QR.

$$\rho_\tau(u-v)-\rho_\tau(u)=-v[\tau-1(u<0)]+\int_0^v[1(u\le s)-1(u\le 0)]ds,$$

which provides a Taylor-like expansion for non-smooth function. On the one hand, we have

$$\begin{aligned}\rho_\tau(y-x^\top\eta_1)-\rho_\tau(y-x^\top\eta_2)&=x^\top(\eta_2-\eta_1)[\tau-1(y-x^\top\eta_2)]\\&\quad+\int_0^{x^\top(\eta_1-\eta_2)}[1(y-x^\top\eta_2\le s)-1(y-x^\top\eta_2\le 0)]ds\\&\le|x^\top(\eta_2-\eta_1)|\cdot|\tau-1(y-x^\top\eta_2)|\\&\quad+\left|\int_0^{x^\top(\eta_1-\eta_2)}[1(y-x^\top\eta_2\le s)-1(y-x^\top\eta_2\le 0)]ds\right|\\&\le\max\{\tau,1-\tau\}|x^\top(\eta_2-\eta_1)|+|x^\top(\eta_2-\eta_1)|\\&\le\max\{1+\tau,2-\tau\}\|\eta_2-\eta_1\|_2\|x\|_2.\end{aligned}$$

Let $l_\tau:=\max\{1+\tau,2-\tau\}$. Hence, $H_{y,x}=l_\tau\left\|x\right\|_2$.

A.6 Remarks and proofs for GLMs

Remark 19 *The* (15) *is originally derived from negative log-likelihood functions of exponential family. The exponential family contains many sub-exponential and sub-Gaussian distributions such as binomial, Poisson, negative binomial, Normal, Gamma distributions (McCullagh and Nelder, 1989). Let $\nu(\cdot)$ be some dominated measure and $b(\cdot)$ be a function. Consider a Y follows the distribution of the natural exponential families P_η indexed by parameter η*

$$P_\eta(dy)=c(y)\exp\{y\eta-b(\eta)\}\nu(dy),\tag{46}$$

where the function $c(y)>0$ is free of $\eta\in\Xi:=\{\eta:\int c(y)\exp\{y\eta\}\nu(dy)<\infty\}$.

Remark 20 *For GLMs, the link function $u(x)$ is canonical, i.e. $u(x)=x$, whence we can choose $g_A(\cdot)\equiv 1$ in (G1). Note that in this case $k(t)=b(t)$, a choice of $h_A(\cdot)$ in condition (G2) is derived by $\ddot b(t)>0$*

$$\dot k(x^\top\theta)\le\dot b(r_n\left\|x\right\|_2)=:h_{r_n}(x),\ \textit{for}\ \|\theta\|_2\le r_n\ \textit{due to}\ |x^\top\theta|\le r_n\left\|x\right\|_2.$$

For GLM with non-canonical link function $u(x)$, we first choose $g_A(\cdot)$ in condition (G1) by $\dot u(x^\top\theta)\le\dot u(r_n\left\|x\right\|_2)=:g_{r_n}(x)$, for $\|\theta\|_2\le r_n$ due to $|x^\top\theta|\le r_n\left\|x\right\|_2$. Under the (C.2) and (G.1), it implies (G.2) with $A=r_n$ and $h_{r_n}(x)=g_{r_n}(x)\dot b(u(r_n\left\|x\right\|_2))$ by the following inequality

$$\dot k(x^\top\theta)=\dot u(x^\top\theta)\dot b(u(x^\top\theta))\le g_{r_n}(x)\dot b(u(r_n\left\|x\right\|_2)):=h_{r_n}(x),\ \textit{for}\ \|\theta\|_2\le r_n.$$

Suppose the input $\{X_i\}_{i=1}^n$ is i.i.d. drawn from X, and X is bounded (see Yang et al. (2021)). Under the (C.2), the $k(X^\top\theta)$ and $u(X^\top\theta)$ are also bounded, then (G.4) is true under the finite second moments of output

$$\mathbb{E}[k(X^\top\theta)-Yu(X^\top\theta)]^2\le 2\mathbb{E}[k(X^\top\theta)]^2+2\mathbb{E}[Yu(X^\top\theta)]^2\le C_1+C_2\mathbb{E}Y^2<\infty$$

for $\|\theta\|_2 \leq r_n$, where C_1 and C_2 are some positive constants. Then $\sigma_R < \infty$ in (G.4).

From (46), one can formally derive the quasi-GLMs loss in (15). Indeed, given $\{X_i\}_{i=1}^n$, the conditional likelihood function of $\{Y_i\}_{i=1}^n$ is the product of n terms in (46) with $\eta_i := u(X_i^\top \theta)$, and the average negative log-likelihood function is

$$\hat{R}_l(\theta) := \tfrac{-1}{n} \textstyle\sum_{i=1}^n [Y_i u(X_i^\top \theta) - b(u(X_i^\top \theta))] = \tfrac{1}{n} \textstyle\sum_{i=1}^n l(Y_i, X_i^\top \theta),\ \theta \in \mathbb{R}^p.$$

The (C.2) can be obtained by a first-order Taylor expansion of $l(y, x, \cdot)$ as the following

$$l(y, x, \eta_2) = l(y, x, \eta_1) + (\eta_2 - \eta_1)^\top \dot{l}[y, x, (t\eta_2 + (1-t)\eta_1)]; \eta_1, \eta_2 \in \Theta,\ \exists\ t \in (0, 1),$$

where $\dot{l}(y, x, \cdot)$ is a (sub-)gradient, we can choose a $H_{y,x}$ satisfying

$$H_{y,x} \geq \sup_{\eta_1,\eta_2 \in \Theta} \|\dot{l}[y, x, (t\eta_2 + (1-t)\eta_1)]\|_2. \tag{47}$$

Fix a $\eta \in \Theta$, we compute the gradient for the loss function in (15)

$$\dot{l}(y, x, \eta) := \nabla_\eta l(y, x^\top \eta) = [-y\dot{u}(x^\top \eta) + \dot{k}(x^\top \eta)]x^\top.$$

From (47), $H_{y,x}$ in Theorem 4 is given by

$$\begin{aligned} \sup_{\eta_1,\eta_2 \in \Theta} \|\dot{l}[y, x, (t\eta_2 + (1-t)\eta_1)]\|_2 &\leq \sup_{\|\theta\|_2 \leq r_n} |-y\dot{u}(x^\top \eta) + \dot{k}(x^\top \eta)| \cdot \|x\|_2 \\ &\leq [|y| g_{r_n}(x) + h_{r_n}(x)] \|x\|_2 := H_{y,x} \end{aligned} \tag{48}$$

under condition (C.2), which implies the excess risk bound in Corollary 10.

Next, we provides two examples of $H_{y,x}$.

Robust logistic regression. We have $u(t) = t,\ k(t) = \log(1 + e^t)$ and $y \in \{0, 1\}$. Note that $H_{y,x}$ in Theorem 4 is given by

$$\sup_{\eta_1,\eta_2 \in \Theta} \|\dot{l}[y, x, (t\eta_2 + (1-t)\eta_1)]\|_2 \leq \sup_s |y + \dot{k}(s)| \cdot \|x\|_2 \leq 2 \|x\|_2 := H_{y,x}$$

from (47) and (48), we have $H_{y,x}^\beta = 2^\beta \|x\|_2^\beta$.

Robust negative binomial regression. The connection of $u(\cdot)$ and $k(\cdot)$ of NBR is $u(t) = t - \log(\eta + e^t)$ and $k(t) = \eta \log(\eta + e^t)$. From (47) and (48), we have via Theorem 4

$$\begin{aligned} \sup_{\eta_1,\eta_2 \in \Theta} \|\dot{l}[y, x, (t\eta_2 + (1-t)\eta_1)]\|_2 &\leq \sup_s |y + \dot{k}(s)| \cdot \|x\|_2 \\ &= (y + \eta) \|x\|_2 := H_{y,x},\ y \geq 0. \end{aligned} \tag{49}$$

Hence, we obtain $H_{y,x}^\beta = |(y + \eta) \|x\|_2|^\beta \leq 2^{\beta-1}[\|yx\|_2^\beta + (\eta \|x\|_2)^\beta]$.

A.7 The proof of Theorem 12

For a fixed L, Lipschitz property of DNN function (Proposition 6 in Taheri et al. (2021)) implies the following excess risk guarantee for elastic net regularization DNN regression estimators. Since the neural network $\mathcal{NN}(N, L)$ in (18) has ReLU activation functions, and it has the approximation error promise (Schmidt-Hieber, 2020) grounded on the smooth assumption of f^*.

Proof Let $\hat{R}_{\psi_\lambda,l,\alpha}(f) := \frac{1}{n\alpha}\sum_{i=1}^n \psi_\lambda[\alpha l(Y_i, f_\theta(X_i))]$. From the definition of $f_{\hat{\theta}_n}$, one has

$$\hat{R}_{\psi_\lambda,l,\alpha}(f_{\hat{\theta}_n}) + \rho\|\hat{\theta}_n\|_2^2 + \gamma\|\hat{\theta}_n\|_1 = \frac{1}{n\alpha}\sum_{i=1}^n \psi_\lambda[\alpha l(Y_i, f_{\hat{\theta}_n}(X_i))] + \rho\|\hat{\theta}_n\|_2^2 + \gamma\|\hat{\theta}_n\|_1$$

$$\le \frac{1}{n\alpha}\sum_{i=1}^n \psi_\lambda[\alpha l(Y_i, f_{\theta^*_{\mathcal{N}}}(X_i))] + \rho\|\theta^*_{\mathcal{N}}\|_2^2 + \gamma\|\theta^*_{\mathcal{N}}\|_1 =: \hat{R}_{\psi_\lambda,l,\alpha}(\theta^*_{\mathcal{N}}) + \rho\|\theta^*_{\mathcal{N}}\|_2^2 + \gamma\|\theta^*_{\mathcal{N}}\|_1$$

which yields

$$\begin{aligned}\hat{R}_{\psi_\lambda,l,\alpha}(f_{\hat{\theta}_n}) - \hat{R}_{\psi_\lambda,l,\alpha}(\theta^*_{\mathcal{N}}) &\le \gamma(\|\theta^*_{\mathcal{N}}\|_1 - \|\hat{\theta}_n\|_1) + \rho(\|\theta^*_{\mathcal{N}}\|_2^2 - \|\hat{\theta}_n\|_2^2)\\ &\le \rho\|\theta^*_{\mathcal{N}}\|_2^2 + \gamma\|\theta^*_{\mathcal{N}}\|_1. \end{aligned} \tag{50}$$

Let $\mathcal{F} = \mathcal{NN}(N, L)$. The excess risk $R_l(f_{\hat{\theta}_n}) - R_l(f^*)$ can be decomposed and bounded by,

$$\begin{aligned} R_l(f_{\hat{\theta}_n}) - R_l(f^*) &= \underbrace{R_l(f_{\hat{\theta}_n}) - \hat{R}_{\psi_\lambda,l,\alpha}(f_{\hat{\theta}_n})}_{Genaralization}\\ &+ \underbrace{\hat{R}_{\psi_\lambda,l,\alpha}(f_{\hat{\theta}_n}) - \hat{R}_{\psi_\lambda,l,\alpha}(f_{\theta^*_{\mathcal{N}}})}_{Optimaztion} + \underbrace{\hat{R}_{\psi_\lambda,l,\alpha}(f_{\theta^*_{\mathcal{N}}}) - R_l(f_{\theta^*_{\mathcal{N}}})}_{Concentration} + \underbrace{R_l(f_{\theta^*_{\mathcal{N}}}) - R_l(f^*)}_{Approximation}\\ &\le [R_l(f_{\hat{\theta}_n}) - \hat{R}_{\psi_\lambda,l,\alpha}(f_{\hat{\theta}_n})] + [\hat{R}_{\psi_\lambda,l,\alpha}(f_{\theta^*_{\mathcal{N}}}) - R_l(f_{\theta^*_{\mathcal{N}}})]\\ &+ \inf_{f\in\mathcal{F}} |R_l(f) - R_l(f^*)| + \rho\|\theta^*_{\mathcal{N}}\|_2^2 + \gamma\|\theta^*_{\mathcal{N}}\|_1, \end{aligned} \tag{51}$$

where the last inequality is from (50) and $R_l(f_{\theta^*_{\mathcal{N}}}) - R_l(f^*) \le \inf_{f\in\mathcal{F}} |R_l(f) - R_l(f^*)|$.

The second term in (51) is the concentration error bound, which can be bounded from the same proof in Lemma 17 to get for all $f_{\theta^*_{\mathcal{N}}}$ with $\theta^* \in \Theta^*_{\mathcal{N}}$

$$\mathbb{P}\left\{\hat{R}_{\psi_\lambda,l,\alpha}(f_{\theta^*_{\mathcal{N}}}) - R_l(f_{\theta^*_{\mathcal{N}}}) \le \frac{f(\alpha)}{\alpha} R_{\lambda\circ l}(f_{\theta^*_{\mathcal{N}}}) + \frac{1}{n\alpha}\log\frac{1}{\delta}\right\} \ge 1 - \delta. \tag{52}$$

The first term in (51) is the generalization error bound with a upper bound from the same proof in Lemma 18 from Lipschitz property of loss function. For every $x \in \mathbb{R}^p$ and parameters $\theta_1 = (W_0, \ldots, W_L), \theta_2 = (V_0, \ldots, V_L)$ in (18). One can check Lipschitz property of DNNs

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \le c_{\text{Lip}}(x)\,\|\theta_2 - \theta_1\|_{\text{F}} \tag{53}$$

for a given function $c_{\text{Lip}}(x) := 2\sqrt{L}\|x\|_2 \max_{l\in\{0,\ldots,L\}} \prod_{j\in\{0,\ldots,L\}, j\ne l} \sigma_{\max}(W^j) \vee \sigma_{\max}(V^j)$ by Proposition 6 in Taheri et al. (2021).

Next, we consider the s_n-sparse parameter space Θ given in (22). Suppose that $l(\cdot,\cdot)$ satisfies Lipschitz condition with a function $D_{x,y}$

$$|l(y, f_{\theta_2}(x)) - l(y, f_{\theta_1}(x))| \le D_{x,y}|f_{\theta_2}(x) - f_{\theta_1}(x)|,\ \theta_1,\ \theta_2 \in \Theta.$$

From (53), the loss function has Lipschitz property

$$|l(y, f_{\theta_2}(x)) - l(y, f_{\theta_1}(x))| \le D_{x,y}c_{\mathrm{Lip}}(x)\|\theta_2 - \theta_1\|_{\mathrm{F}} \le 2W^L\sqrt{L}D_{x,y}\|x\|_2\|\theta_2 - \theta_1\|_{\mathrm{F}},$$

which shows that

$$H_{y,x} := 2W^L\sqrt{L}\|x\|_2 D_{x,y}. \tag{54}$$

Examples: For robust DNN LAD regression, we have $D_{x,y} = 1$ and thus $H_{y,x} = 2W^L\sqrt{L}\|x\|_2$. For robust DNN logistic regression, it gives $D_{x,y} = y + 1 \le 2$ and $H_{y,x} = 4W^L\sqrt{L}\|x\|_2$. For robust DNN NBR, we get $D_{x,y} = y + \eta$ and $H_{y,x} = 2W^L\sqrt{L}\|x\|_2(y + \eta)$.

By using Lipschitz constant (54), under (C.1)-(C.4) and $f(t) = |t|^\beta$ and $c_2 = 2^{\beta-1}$, one has by Lemma 18

$$\begin{aligned}
&R_l(f_{\hat{\theta}_n}) - \hat{R}_{\psi_\lambda,l,\alpha}(f_{\hat{\theta}_n}) \le 2\kappa\mathbb{E}H_{Y,X} + \frac{2^{\beta-1}f(\alpha)}{\alpha}R_{\lambda\circ l}(\Theta) + \frac{2^{\beta-1}f(\alpha\kappa)}{\alpha}\mathbb{E}[\lambda(H_{Y,X})] + \frac{1}{n\alpha}\log\frac{N(\Theta,\kappa)}{\delta}\\
&= 4\kappa W^L\sqrt{L}\mathbb{E}[\|X\|_2D_{X,Y}] + (2\alpha)^{\beta-1}R_{\lambda\circ l}(\Theta) + (2\alpha)^{\beta-1}(2\kappa W^L\sqrt{L})^\beta\frac{\mathbb{E}|\|X\|_2D_{X,Y}|^\beta}{\beta} + \frac{1}{n\alpha}\log\frac{N(\Theta,\kappa)}{\delta}
\end{aligned}$$

with probability at least $1 - \delta$ for any $\kappa > 0$. For the last term with covering number, we apply Lemma 16 to conclude that

$$\log N(\Theta, \kappa) \le \log N(B_2^p(r_n) \cap B_0^p(s_n), \kappa) \le \log\left(\frac{1}{\sqrt{2es_n}}\right) + s_n\log\left[\frac{e(\kappa + 2r_n)p}{\kappa s_n}\right],$$

which shows that

$$\begin{aligned}
&R_l(f_{\hat{\theta}_n}) - \hat{R}_{\psi_\lambda,l,\alpha}(f_{\hat{\theta}_n}) \le 4\kappa W^L\sqrt{L}\mathbb{E}[\|X\|_2D_{X,Y}] + (2\alpha)^{\beta-1}R_{\lambda\circ l}(\Theta)\\
&+ (2\alpha)^{\beta-1}(2\kappa W^L\sqrt{L})^\beta\frac{\mathbb{E}|\|X\|_2D_{X,Y}|^\beta}{\beta} + \frac{1}{n\alpha}\left\{s_n\log\left[e(\kappa + 2r_n)\frac{p}{\kappa s_n}\right] + \log\left(\frac{\delta^{-1}}{\sqrt{2es_n}}\right)\right\}
\end{aligned} \tag{55}$$

with probability at least $1 - \delta$ for any $\kappa > 0$.

Under the two high provability events in (52) and (55), inequality (51) shows that

$$\begin{aligned}
&R_l(f_{\hat{\theta}_n}) - R_l(f^*)\\
&\le \inf_{f\in\mathcal{F}}|R_l(f) - R_l(f^*)| + \rho\|\theta^*_{\mathcal{N}}\|_2^2 + \gamma\|\theta^*_{\mathcal{N}}\|_1 + \left(\alpha^{\beta-1}R_{\lambda\circ l}(f_{\theta^*_{\mathcal{N}}}) + \frac{1}{n\alpha}\log\frac{1}{\delta}\right) + \frac{1}{n\alpha}\log\frac{N(\Theta,\kappa)}{\delta}\\
&+ 4\kappa W^L\sqrt{L}\mathbb{E}[\|X\|_2D_{X,Y}] + (2\alpha)^{\beta-1}R_{\lambda\circ l}(\Theta) + (2\alpha)^{\beta-1}\kappa^\beta(2W^L\sqrt{L})^\beta\frac{\mathbb{E}|\|X\|_2D_{X,Y}|^\beta}{\beta}\\
&\le \rho\|\theta^*_{\mathcal{N}}\|_2^2 + \gamma\|\theta^*_{\mathcal{N}}\|_1 + \frac{1}{n\alpha}\left\{s_n\log\left[e(\kappa + 2r_n)\frac{p}{\kappa s_n}\right] + \log\left(\frac{\delta^{-2}}{\sqrt{2es_n}}\right)\right\} + (1 + 2^{\beta-1})\alpha^{\beta-1}R_{\lambda\circ l}(\Theta)\\
&+ 4\kappa W^L\sqrt{L}\mathbb{E}[\|X\|_2D_{X,Y}] + (2\alpha)^{\beta-1}(2\kappa W^L\sqrt{L})^\beta\frac{\mathbb{E}|\|X\|_2D_{X,Y}|^\beta}{\beta} + \inf_{f\in\mathcal{F}}|R_l(f) - R_l(f^*)|
\end{aligned} \tag{56}$$

with probability at least $1 - 2\delta$.

Let $\kappa = 1/n$ and put $\frac{1}{n\alpha}\left\{\log(\delta^{-2}/\sqrt{2es_n}) + s_n \log\left[e(\kappa+2r_n)p/(\kappa s_n)\right]\right\} = (1+2^{\beta-1})\alpha^{\beta-1}R_{\lambda\circ l}(\Theta)$ in (56), and it gives $\alpha = \frac{1}{n^{1/\beta}}\left(\frac{\log(\delta^{-2}/\sqrt{2es_n})+s_n\log[e(1+2nr_n)p/s_n]}{(2^{\beta-1}+1)R_{\lambda\circ l}(\Theta)}\right)^{1/\beta}$. Hence, (56) implies by taking $\inf_{\theta^*_{\mathcal{N}}\in\Theta^*_{\mathcal{N}}}$ on the upper bound (56)

$$\begin{aligned}
R_l(f_{\hat{\theta}_n}) - R_l(f^*) \le & \inf_{\theta^*_{\mathcal{N}}\in\Theta^*_{\mathcal{N}}}(\rho\|\theta^*_{\mathcal{N}}\|_2^2 + \gamma\|\theta^*_{\mathcal{N}}\|_1) + \frac{2}{n\alpha}\left\{\log(\delta^{-2}/\sqrt{2es_n}) + s_n\log\left[(1+2nr_n)ep/s_n\right]\right\} \\
& + \frac{4W^L\sqrt{L}}{n}\mathbb{E}[\|X\|_2 D_{X,Y}] + \frac{2^{\beta-1}\alpha^{\beta-1}}{n^\beta}\frac{\mathbb{E}|\|X\|_2 D_{X,Y}|^\beta}{\beta} + \inf_{f\in\mathcal{F}}|R_l(f) - R_l(f^*)| \\
\le & \inf_{\theta^*_{\mathcal{N}}\in\Theta^*_{\mathcal{N}}}(\rho\|\theta^*_{\mathcal{N}}\|_2^2 + \gamma\|\theta^*_{\mathcal{N}}\|_1) + \frac{2\left\{\log(\delta^{-2}/\sqrt{2es_n}) + s_n\log\left[e(1+2nr_n)p/s_n\right]\right\}^{1-\beta^{-1}}}{n^{1-\beta^{-1}}(2^{\beta-1}+1)^{1-\beta^{-1}}R_{\lambda\circ l}^{1-\beta^{-1}}(\Theta)}(2^{\beta-1}+1)R_{\lambda\circ l}(\Theta) \\
& + \frac{2^{\beta-1}(2W^L\sqrt{L})^\beta}{n^{1-\beta^{-1}+\beta}}\left(\frac{\log(\delta^{-2}/\sqrt{2es_n}) + s_n\log\left[e(1+2nr_n)p/s_n\right]}{(2^{\beta-1}+1)R_{\lambda\circ l}(\Theta)}\right)^{1-\beta^{-1}}\frac{\mathbb{E}|\|X\|_2 D_{X,Y}|^\beta}{\beta} \\
& + \frac{4W^L\sqrt{L}}{n}\mathbb{E}[\|X\|_2 D_{X,Y}] + \inf_{f\in\mathcal{F}}|R_l(f) - R_l(f^*)| \\
\le & \frac{4W^L\sqrt{L}}{n}\mathbb{E}[\|X\|_2 D_{X,Y}] + \frac{F_{\beta,L,W}(R_{\lambda\circ l})}{n^{\frac{1-\beta}{\beta}}}\left[\log(\delta^{-2}/\sqrt{2es_n}) + s_n\log\left[e(1+2nr_n)p/s_n\right]\right]^{\frac{1-\beta}{\beta}} \\
& + \inf_{\theta^*_{\mathcal{N}}\in\Theta^*_{\mathcal{N}}}(\rho\|\theta^*_{\mathcal{N}}\|_2^2 + \gamma\|\theta^*_{\mathcal{N}}\|_1) + \inf_{f\in\mathcal{F}}|R_l(f) - R_l(f^*)|
\end{aligned}$$

with probability at least $1-2\delta$, where

$$F_{\beta,L,W}(R_{\lambda\circ l}) := \left[2(2^{\beta-1}+1)R_{\lambda\circ l}(\Theta) + \tfrac{(4W^L\sqrt{L})^\beta}{2\beta}\mathbb{E}|\|X\|_2 D_{X,Y}|^\beta\right]/[(2^{\beta-1}+1)R_{\lambda\circ l}(\Theta)]^{\frac{\beta-1}{\beta}}.$$

Let us treat the approximation error bound of $\inf_{f\in\mathcal{F}}|R_l(f) - R_l(f^*)|$ under (D.2). By the Lipschitz condition (D.1), we have

$$\begin{aligned}
|R_l(f) - R_l(f^*)| :&= |\mathbb{E}[l(Y, f(X)) - l(Y, f^*(X))]| \le \mathbb{E}[D_{X,Y}\cdot|f(X) - f^*(X)|] \\
&\le (\mathbb{E}D_{X,Y}^2)^{1/2}\cdot(\mathbb{E}|f(X) - f^*(X)|^2)^{1/2} = (\mathbb{E}D_{X,Y}^2)^{1/2}\,\|f(X) - f^*(X)\|_{L^2(\nu)}
\end{aligned}$$

from which,

$$\inf_{f\in\mathcal{F}}|R_l(f) - R_l(f^*)| \le (\mathbb{E}D_{X,Y}^2)^{1/2}\inf_{f\in\mathcal{F}}\|f(X) - f^*(X)\|_{L^2(\nu)}\,. \tag{57}$$

The sparse Relu DNN has approximation power in terms of $\inf_{f\in\mathcal{F}}\|f(X) - f^*(X)\|_{L^2(\nu)}$ by tuning the width and depth. Theorem 5 in Schmidt-Hieber (2020) shown the approximation ability of the sparse ReLU DNN under Hölder functional class $\mathcal{C}^\gamma([0,1]^d, B)$ with smoothness index γ. We extend the space $\mathcal{C}^\gamma([0,1]^d, B)$ to $\mathcal{C}^\gamma([0,a_n]^d, B)$ easily, which asserts that *Given a continuous function $f^*\in\mathcal{C}^\gamma([0,a_n]^d, B)$ with a sequence $a_n\ge 1$, there exists a function f implemented by a ReLU network with width $N = 6(d+\lceil\gamma\rceil)M$, ($M\ge(\gamma+1)^d\vee(B+1)e^d$), and depth*

$$L = 8 + (m+5)\left(1+\lceil\log_2(d\vee\gamma)\rceil\right) \textit{ with an integer } m\ge 1$$

and number of parameters $s \leq 141(d+\gamma+1)^{3+d}M(m+6)$ *such that*

$$\|f-f^*\|_{L^\infty([0,a_n]^d)} \leq \frac{(2B+1)\left(1+d^2+\gamma^2\right)M6^d a_n^\gamma}{2^m} + \frac{K3^\gamma a_n^\gamma}{N^{\frac{\gamma}{d}}}. \tag{58}$$

Under (D.3), we define the event $\mathcal{A}_n := \{\|X\|_\infty \leq a_n\}$ for a given non-decreasing and positive sequence $\{a_n\}$. Then, Markov's inequality shows

$$\mathbb{P}\left(\mathcal{A}_n\right) \geq 1 - \frac{b}{a_n}. \tag{59}$$

Under the restriction $\|f-f^*\|_\infty \leq F < \infty$ with $f \in \mathcal{NN}(N,L)$, it leads to

$$\begin{aligned}
\mathbb{E}|f(X)-f^*(X)|^2 &= \int_{\{\|x\|_\infty \leq a_n\}} |f(x)-f^*(x)|^2\mu(dx) + \int_{\{\|x\|_\infty > a_n\}} |f(x)-f^*(x)|^2\mu(dx) \\
&\leq \|f-f^*\|^2_{L^\infty([0,a_n]^d)} + F^2\mathbb{P}(\mathcal{A}_n^c) \\
&\leq \left[\frac{(2B+1)\left(1+d^2+\gamma^2\right)6^d M}{2^m} + \frac{B3^\gamma}{N^{\frac{\gamma}{d}}}\right]^2 a_n^{2\gamma} + F^2\frac{b}{a_n},
\end{aligned}$$

where the last inequality is from (58) and (59); and $X \sim \mu$. The (57) gives

$$\begin{aligned}
\inf_{f\in\mathcal{F}} |R_l(f)-R_l(f^*)| &\leq (\mathbb{E}D^2_{X,Y})^{\frac{1}{2}} \left(\left[\frac{(2B+1)\left(1+d^2+\beta^2\right)6^d M}{2^m} + \frac{B3^\gamma}{N^{\frac{\gamma}{d}}}\right]^2 a_n^{2\gamma} + F^2\frac{b}{a_n}\right)^{\frac{1}{2}} \\
&\leq (\mathbb{E}D^2_{X,Y})^{\frac{1}{2}} \left(\left[\frac{(2B+1)\left(1+d^2+\gamma^2\right)6^d M}{2^m} + \frac{B3^\gamma}{N^{\frac{\gamma}{d}}}\right] a_n^{\gamma} + (\frac{b}{a_n})^{\frac{1}{2}}F\right)
\end{aligned}$$

Finally, we put $\left[\frac{(2B+1)\left(1+d^2+\beta^2\right)6^d M}{2^m} + \frac{3^\gamma B}{N^{\frac{\gamma}{p}}}\right] a_n^\gamma = (\frac{b}{a_n})^{\frac{1}{2}}F$, which implies

$$a_n = \frac{b^{\frac{1}{2\gamma+1}} F^{\frac{2}{2\gamma+1}}}{\left[\frac{(2B+1)\left(1+d^2+\gamma^2\right)6^d M}{2^m} + \frac{3^\gamma B}{N^{\gamma/d}}\right]^{\frac{2}{2\gamma+1}}}.$$

Then

$$\begin{aligned}
\inf_{f\in\mathcal{F}} |R_l(f)-R_l(f^*)| &\leq 2(\mathbb{E}D^2_{X,Y})^{\frac{1}{2}}(b/a_n)^{\frac{1}{2}}F \\
&= 2(\mathbb{E}D^2_{X,Y})^{\frac{1}{2}} F^{\frac{2\gamma}{2\gamma+1}} b^{\frac{r}{2\gamma+1}} \left[\frac{(2B+1)\left(1+d^2+\gamma^2\right)6^d M}{2^m} + \frac{3^\gamma B}{N^{\gamma/d}}\right]^{\frac{1}{2\gamma+1}}.
\end{aligned}$$

■

A.8 The proof of Corollary 14

Proof Based on (23), if the depth and width of DNNs is increasing with n, i.e. $L = L_n$ and $N = N_n$, it allows $\frac{W^{L_n}\sqrt{L_n}}{n} \le C(s_n \log(npr_n)/n)^{(\beta-1)/\beta}$ for a constant $C > 0$. Then the inequality

$$\log W \cdot L_n + \frac{1}{2}\log L_n \le \log C + \frac{\beta-1}{\beta}[\log(s_n \log(npr_n)) - \log n] + \log n \tag{60}$$

guarantees the rate $O(s_n \log(npr_n)/n)^{(\beta-1)/\beta}$ excess risk if $L_n \lesssim \frac{\log(s_n \log(npr_n))}{\log W} + \frac{\log n}{\log W}$ for $W > 1$. The sparsity level $s_n = o(n)$ performs as an effective dimension in DNN parameter that plays a key role to determine the main term in the excess risk bound. According to restriction $s_n = o(n)$, we put a *depth-sample condition*:

$$L_n \lesssim \frac{\log n + \log(s_n \log(npr_n))}{\log W} \tag{61}$$

for a sufficient large n and constant.

For $W < 1$, (60) holds for some large L_n.

Under the Hölder functional class, we assume that the approximation error has a same or smaller order as the statistical error $O((s_n \log(npr_n)/n)^{(\beta-1)/\beta})$. Suppose that

$$\tfrac{(2B+1)\left(1+d^2+\gamma^2\right)6^d M}{2^m} \lesssim \tfrac{3^\gamma B}{N_n^{\gamma/d}},$$

which implies $(\gamma/d)\log_2 N_n \lesssim m$.

The condition of L_n in Theorem 3 in Schmidt-Hieber (2020) requires that $L_n \lesssim 8 + (\log n + \log(s_n \log(npr_n)) + 5)\,(1 + \lceil \log_2(d \vee \gamma)\rceil)$, which coincides (61). To obtain the rate $O(s_n \log(npr_n)/n)^{(\beta-1)/\beta}$ approximation error, one must have

$$\inf_{f\in\mathcal{F}} |R_l(f) - R_l(f^*)| \lesssim b^{\frac{r}{2\gamma+1}} \left[\frac{3^\gamma B}{N_n^{\gamma/d}}\right]^{\frac{1}{2\gamma+1}} \lesssim (s_n \log(npr_n)/n)^{(\beta-1)/\beta}.$$

It leads to the *width-sample condition*

$$N_n \gtrsim b^d \left[\frac{n}{s_n \log(npr_n)}\right]^{\frac{d(2\gamma+1)}{\gamma}\cdot\frac{\beta-1}{\beta}}.$$

So, $(\gamma/d)\log_2 N_n \lesssim m \lesssim \log n + \log(s_n \log(npr_n))$ and $L_n \lesssim 8 + (\log n + \log(s_n \log(npr_n)) + 5)\,(1 + \lceil \log_2(d \vee \gamma)\rceil)$ implies the rate $(s_n \log(npr_n)/n)^{(\beta-1)/\beta}$ approximation error, under the order of tuning parameters $\rho \vee \gamma \lesssim (s_n \log(npr_n)/n)^{(\beta-1)/\beta}$. ■

A.9 Robust two-component mixed linear regression

For convenience, we define a combined parameter $\theta := (\pi, \eta_0, \eta_1)$, which is restricted in following space $\Theta \subset \mathbb{R}^{1+2d}$ with $r > 1$.

$$\theta \in \Theta := \left\{(a, b_0, b_1) \in \mathbb{R}^{1+2d} : 0 < \rho \le a \le 1-\rho, \max(\|b_1\|_2, \|b_2\|_2) \le u,\ u := \sqrt{(r^2-1)/2}\right\}.$$

Define $\hat{\theta}_n$ by (6) with $\lambda(x) = \frac{1}{\beta}|x|^{\beta}$, $\beta \in (1,2)$, and θ^* is given by (2) with loss $l(y,x,\cdot)$ given by (33).

We now identify the $H_{y,x}$ in Theorem 4. Denote $p_k := p\left(y, x^\top \eta_k\right)$ for $k = 0, 1$. One has

$$\left|\frac{\partial l(y,x,\theta)}{\partial \pi}\right| = \left|-\frac{p(y,x^\top\eta_0) - (y,x^\top\eta_1)}{\pi p(y,x^\top\eta_0) + (1-\pi)p(y,x^\top\eta_1)}\right| = \left|\frac{p_0}{\pi p_0 + (1-\pi)\rho_1} + \frac{p_1}{\pi p_0 + (1-\pi)p_1}\right|$$

$$\leq \frac{p_0}{\pi p_0} + \frac{p_1}{(1-\pi)p_1} = \frac{1}{\pi(1-\pi)} \leq \frac{1}{\rho(1-\rho)}, \quad (\rho \leq \pi \leq 1-\rho).$$

Let $\nabla p(y,x^\top\eta_k) := \frac{\partial}{\partial t} p(y,t)|_{t=x^\top\eta_k}$ and $\nabla \log p(y,x^\top\eta_k) := \frac{\partial}{\partial t} \log p(y,t)|_{t=x^\top\eta_k}$. For $j = 1, \cdots, d$, we have

$$\left|\frac{\partial l(y,x,\theta)}{\partial \eta_{j0}}\right| = \left|\frac{-x_j \dot{p}(y,x^\top\eta_0)}{\pi p(y,x^\top\eta_0) + (1-\pi)p(y,x^\top\eta_1)}\right| \leq \left|\frac{-x_j \dot{p}(y,x^\top\eta_0)}{\pi p(y,x^\top\eta_0)}\right| \leq \frac{|x_j| \cdot \left|\nabla \log p(y,x^\top\eta_0)\right|}{\rho}$$

and

$$\left|\frac{\partial l(y,x,\theta)}{\partial \eta_{j1}}\right| = \left|\frac{-x_j \dot{p}(y,x^\top\eta_1)}{\pi p(y,x^\top\eta_0) + (1-\pi)p(y,x^\top\eta_1)}\right| \leq \left|\frac{-x_j \dot{p}(y,x^\top\eta_0)}{(1-\pi) p(y,x^\top\eta_0)}\right| \leq \frac{|x_j| \cdot \left|\nabla \log p(y,x^\top\eta_1)\right|}{1-\rho}.$$

According to (47),

$$\sup_{\eta_1,\eta_2\in\Theta} \|\dot{l}[y,x,(t\eta_2 + (1-t)\eta_1)]\|_2 \leq \sup_{\theta\in\Theta} \|\dot{l}(y,x^\top\theta)\|_2$$

$$\leq \sup_{\theta\in\Theta} \left[\|x\|_2^2 \left\{\rho^{-2}|\nabla \log p(y,x^\top\eta_1)|^2\right\} + (1-\rho)^{-2}|\nabla \log p(y,x^\top\eta_1)|^2 + \rho^{-2}(1-\rho)^{-2}\right]^{1/2}.$$

$$\leq \frac{1}{\rho(1-\rho)} + \frac{\|x\|_2}{\rho} \sup_{\eta_0\in B_2^d(u)} |\nabla \log p(y,x^\top\eta_0)| + \frac{\|x\|_2}{1-\rho} \sup_{\eta_1\in B_2^d(u)} |\nabla \log p(y,x^\top\eta_1)| := H(y,x).$$

From the proof Theorem 4, one has

$$R_l(\hat{\theta}_n) - \inf_{\theta\in\Theta} R_l(\theta) \leq \frac{2}{n}\left[\frac{\mathrm{E}[\|X\|_2 \sup\limits_{\eta_0\in B_2^d(u)} |\nabla \log p(Y,X^\top\eta_0)| + 1]}{\rho} + \frac{\mathrm{E}[\|X\|_2 \sup\limits_{\eta_1\in B_2^d(u)} |\nabla \log p(Y,X^\top\eta_1)| + 1]}{1-\rho}\right]$$

$$+ \left(\frac{\log(\delta^{-2}) + (2d+1)\log(1+2rn)}{n(2^{\beta-1}+1)}\right)^{\frac{\beta-1}{\beta}} \times \left[\frac{2(2^{\beta-1}+1)}{[R_{\lambda\circ l}(\Theta)]^{-\beta^{-1}}} + \frac{2^{2(\beta-1)}}{[R_{\lambda\circ l}(\Theta)]^{(\beta-1)/\beta}}\right.$$

$$\left.\times \left(\frac{\mathrm{E}[\|X\|_2 \sup\limits_{\eta_0\in B_2^d(u)} |\nabla \log p(Y,X^\top\eta_0)| + 1]^{\beta}}{\rho^{\beta}} + \frac{\mathrm{E}[\|X\|_2 \sup\limits_{\eta_1\in B_2^d(u)} |\nabla \log p(Y,X^\top\eta_1)| + 1]^{\beta}}{(1-\rho)^{\beta}}\right)\right] + \rho\|\Theta^*\|_2^2.$$

with probability at least $1 - 2\delta$.

For example, in Gaussian mixture regressions of two component, we have

$$p(y,x^\top\eta_k) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{\frac{(y-x^\top\eta_k)^2}{2\sigma_k^2}\right\} \text{ and } \nabla \log p(y,x^\top\eta_k) = -\frac{y-x^\top\eta_k}{\sigma_k^2},\ k = 0, 1.$$

Without loss of generality, we assume that variance parameter $\sigma_k^2 \equiv \sigma^2$ is known. To obtain rate $O(((2d+1)\log(nr_n)/n)^{(\beta-1)/\beta})$ excess risk, we require moment conditions

$$\mathrm{E}(\|X\|_2|Y|)^\beta < \infty \text{ and } \mathrm{E}\,\|X\|_2^{2\beta} < \infty$$

by noticing $\mathrm{E}[\|X\|_2 \sup_{\eta_k \in B_2^d(u)} |Y - X^\top \eta_k|/\sigma^2 + 1]^\beta < C_1 \mathrm{E}(\|X\|_2|Y|)^\beta + C_2 \mathrm{E}\|X\|_2^{2\beta} < \infty$, where C_1 and C_2 are some positive constants.

A.10 Simulation results of negative binomial regression models

We use the same noise settings for the NBR models as the simulations of logistic regression. And the dispersion parameter η is set to be 20. For the network configuration, we use the ReLU activated 5-layers DNN model with width $(d, 20d, 15d, 10d, 5d, 1)$ to train the DNN NBR when the real function f^* is the complex function. For the case of DNN-based f^*, we adopt the ReLU activated 2-layers DNN model with width $(d, 0.6d, 0.4d, 1)$ to train the model. In Tables 7 and 8, we compute the average ℓ_2-estimation errors for the predicted coefficients of each normal NBR models with 100 replications. Table 9 presents the absolute average errors (MAEs) of the response predictors $\{\hat{Y}_i\}_{i=1}^n$ with 100 replications, that is defined as

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^n |\hat{Y}_i - Y_i|.$$

Table 7: Comparison of average ℓ_2-estimation error for NBR on Pareto noise model.

				ℓ_2-estimation error for NBR			
		$\varsigma = 0$			$\varsigma = 0.5$		
				$d = 100, n = 200$			
Pareto	β	High-order	Cauchy	Non-truncation	High-order	Cauchy	Non-truncation
1.60	1.5	2.7324(0.0287)	3.1469(0.0401)	3.6866(0.0742)	2.9015(0.1450)	3.1510(0.1566)	3.8726(0.0081)
1.80	1.5	2.6656(0.0228)	2.9910(0.0189)	3.7492(0.0621)	3.0704(0.1125)	3.1675(0.1286)	3.8732(0.0163)
2.01	2.0	2.2859(0.0228)	2.6282(0.0194)	3.6218(0.0637)	3.0658(0.1080)	3.0603(0.1523)	3.8086(0.0470)
4.01	2.0	2.4727(0.0167)	2.7543(0.0159)	3.6441(0.0181)	3.0460(0.1554)	3.0847(0.1161)	3.8646(0.0354)
6.01	2.0	2.4430(0.0124)	2.5573(0.0173)	3.6831(0.0118)	3.0083(0.1802)	3.0642(0.1178)	3.8940(0.0148)
				$d = 200, n = 500$			
Pareto	β	High-order	Cauchy	Non-truncation	High-order	Cauchy	Non-truncation
1.60	1.5	4.1902(0.9830)	4.5767(0.5005)	4.6152(0.4303)	4.3648(0.0845)	4.4987(0.1371)	4.7120(0.0656)
1.80	1.5	4.6652(0.5420)	4.8640(0.5153)	4.9682(0.3303)	4.4122(0.0846)	4.5386(0.1631)	4.7052(0.0807)
2.01	2.0	4.4903(0.1020)	4.8618(0.1221)	4.8653(0.4294)	4.3573(0.0693)	4.5685(0.1403)	4.6675(0.1002)
4.01	2.0	4.3130(0.1572)	4.5597(0.1459)	3.6799(0.2815)	4.4048(0.1123)	4.5182(0.0429)	4.5985(0.0328)
6.01	2.0	4.3355(0.1309)	4.5255(0.1164)	3.6624(0.0214)	4.3008(0.1350)	4.4457(0.0967)	4.6910(0.0552)
				$d = 1000, n = 1000$			
Pareto	β	High-order	Cauchy	Non-truncation	High-order	Cauchy	Non-truncation
1.60	1.5	10.9820(0.1000)	11.0961(0.1007)	11.6957(0.0492)	10.7608(0.1608)	10.9988(0.1585)	11.0076(0.1068)
1.80	1.5	10.9745(0.0654)	11.0259(0.1234)	11.6771(0.0725)	10.7184(0.0757)	10.9624(0.1096)	11.0044(0.0688)
2.01	2.0	10.8991(0.0686)	10.9826(0.0453)	11.6676(0.0548)	10.7373(0.1280)	10.8823(0.1073)	10.9500(0.0702)
4.01	2.0	10.8049(0.0121)	10.9907(0.1753)	11.6273(0.0534)	10.7024(0.0916)	10.7892(0.0985)	10.9563(0.0904)
6.01	2.0	10.9433(0.0876)	10.9645(0.1369)	11.6186(0.0469)	10.6301(0.0708)	10.7471(0.1166)	10.9544(0.0734)

Table 8: Comparison of average ℓ_2-estimation error for NBR on Uniform noise model.

ℓ_2-estimation error for NBR							
		$\varsigma = 0$			$\varsigma = 0.5$		
$d = 100, n = 200$							
Uniform	β	High-order	Cauchy	Non-truncation	High-order	Cauchy	Non-truncation
0.3	2.0	1.0241(0.0432)	1.6514(0.0471)	3.1721(0.1542)	3.0282(0.1629)	3.0570(0.3486)	3.2361(0.1837)
0.5	2.0	1.2061(0.0514)	1.7425(0.0651)	3.2321(0.1345)	3.0229(0.1415)	3.2562(0.1371)	3.2746(0.0520)
0.8	2.0	1.3454(0.0453)	1.9241(0.0645)	3.5432(0.1124)	3.0570(0.3486)	3.4104(0.1218)	3.5124(0.1273)
$d = 200, n = 500$							
Uniform	β	High-order	Cauchy	Non-truncation	High-order	Cauchy	Non-truncation
0.3	2.0	3.6431(0.0515)	3.9321(0.0541)	5.2354(0.1345)	4.6455(0.1201)	4.9514(0.1011)	5.0917(0.1902)
0.5	2.0	2.9241(0.0762)	3.0235(0.0785)	5.4252(0.1645)	4.8460(0.1309)	5.0286(0.1208)	5.3802(0.1298)
0.8	2.0	4.0212(0.0815)	4.1023(0.0845)	5.7254(0.1432)	5.0538(0.0816)	5.1568(0.1119)	5.4333(0.0692)
$d = 1000, n = 1000$							
Uniform	β	High-order	Cauchy	Non-truncation	High-order	Cauchy	Non-truncation
0.3	2.0	7.2532(0.1547)	8.8754(0.1471)	9.9584(0.1241)	11.0984(0.0595)	11.2696(0.0846)	11.7264(0.0967)
0.5	2.0	7.7652(0.1457)	9.2413(0.1453)	10.125(0.1892)	11.1030(0.0870)	11.3124(0.0460)	12.1964(0.2704)
0.8	2.0	7.9254(0.1745)	9.5432(0.1793)	11.235(0.1346)	11.2125(0.0982)	11.4113(0.0442)	12.4671(0.0754)

Table 9: Comparison of average MAEs for DNN NBR under two noise settings.

MAEs for DNN NBR					
		$d = 6$, $n = 200$ (Complex function)		$d = 100$, $n = 1000$ (DNN)	
β	Pareto (τ)	High-order	Non-truncation	High-order	Non-truncation
1.5	1.60	0.5809(0.0063)	1.5829(0.4448)	0.4702(0.0274)	0.6115(0.0392)
1.5	1.80	0.5586(0.0074)	1.4882(0.3538)	0.4483(0.0379)	0.5232(0.0402)
2.0	2.01	0.4845(0.0056)	1.2000(0.2203)	0.5620(0.0399)	0.6436(0.0294)
2.0	4.01	0.5466(0.0049)	1.2958(0.9950)	0.5251(0.0191)	0.7019(0.0152)
2.0	6.01	0.6001(0.0056)	1.1369(0.2027)	0.5253(0.0131)	0.6718(0.0178)
β	Uniform (π)	High-order	Non-truncation	High-order	Non-truncation
2.0	0.3	0.9842(0.2706)	0.9672(0.3283)	0.5260(0.0309)	0.5978(0.0271)
2.0	0.5	1.6386(0.3332)	1.7000(0.1610)	0.5784(0.0187)	0.6029(0.0088)
2.0	0.8	1.4314(0.4509)	2.7475(0.3379)	0.5824(0.0195)	0.6736(0.0057)

References

Yannick Baraud, Lucien Birgé, and Mathieu Sart. A new method for estimation and model selection: *rho*-estimation. *Inventiones mathematicae*, 207(2):425–517, 2017.

Peter L Bartlett and Shahar Mendelson. Empirical minimization. *Probability theory and related fields*, 135(3):311–334, 2006.

Christian Brownlees, Emilien Joly, Gábor Lugosi, et al. Empirical risk minimization for heavy-tailed losses. *Annals of Statistics*, 43(6):2507–2536, 2015.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

Peng Chen, Xinghu Jin, Xiang Li, and Lihu Xu. A generalized catoni's m-estimator under finite α-th moment assumption with $\alpha \in (1, 2)$. *Electronic Journal of Statistics*, 15(2): 5523–5544, 2021a.

Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Bridging convex and nonconvex optimization in robust pca: Noise, outliers and missing data. *The Annals of Statistics*, 49(5):2948–2971, 2021b.

Zhiyi Chi. A local stochastic lipschitz condition with application to lasso for high dimensional generalized linear models. *arXiv preprint arXiv:1009.1052*, 2010.

Geoffrey Chinot, Guillaume Lecué, and Matthieu Lerasle. Robust statistical learning with lipschitz and convex loss functions. *Probability Theory and related fields*, pages 1–44, 2019.

Jianqing Fan, Quefeng Li, and Yuyan Wang. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 79(1):247, 2017.

Jianqing Fan, Cong Ma, and Yiqiao Zhong. A selective overview of deep learning. *Statistical Science*, 36(2):264–290, 2021.

Jianqing Fan, Yihong Gu, and Wen-Xin Zhou. How do noise tails impact on deep relu networks? *arXiv preprint arXiv:2203.10418*, 2022.

Igor Fedotenkov. A review of more than one hundred pareto-tail index estimators. *Statistica*, 80(3):245–299, 2020.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2018.

Naiyang Guan, Tongliang Liu, Yangmuzi Zhang, Dacheng Tao, and Larry S Davis. Truncated cauchy non-negative matrix factorization. *IEEE Transactions on pattern analysis and machine intelligence*, 41(1):246–259, 2017.

Peter J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1):492–518, 1964.

Koulik Khamaru and Martin J Wainwright. Convergence guarantees for a class of nonconvex and non-smooth optimization problems. *Journal of Machine Learning Research*, 20(154):1–52, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

Jason M Klusowski, Dana Yang, and WD Brinda. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *IEEE Transactions on Information Theory*, 65(6):3515–3524, 2019.

Keith Knight. Limiting distributions for l_1 regression estimators under general conditions. *Annals of Statistics*, 26(2):755–770, 1998.

Roger Koenker. *Quantile regression.* Cambridge University Press, New York, 2005.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica*, pages 33–50, 1978.

Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

Clifford Lam and Wenyu Cheng. Robust mean and eigenvalues regularized covariance matrix estimation. *London School of Economics and Political Science*, 2021.

Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: upper and minimax bounds. *Topics in Learning Theory-Societe Mathematique de France,(S. Boucheron and N. Vayatis Eds.)*, 2013.

Johannes Lederer. Risk bounds for robust deep learning. *arXiv:2009.06202*, 2020.

Kyungjae Lee, Hongjun Yang, Sungbin Lim, and Songhwai Oh. Optimal algorithms for stochastic multi-armed bandits with heavy tailed rewards. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Matthieu Lerasle. Lecture notes: Selected topics on robust statistical learning theory. *arXiv preprint arXiv:1908.10761*, 2019.

Tongliang Liu and Dacheng Tao. On the robustness and generalization of cauchy regression. In *2014 4th IEEE International Conference on Information Science and Technology*, pages 100–105. IEEE, 2014.

Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(2):866–896, 2017.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Timothée Mathieu and Stanislav Minsker. Excess risk bounds in robust empirical risk minimization. *Information and Inference: A Journal of the IMA*, 2021.

P McCullagh and John A Nelder. *Generalized Linear Models*, volume 37. CRC Press, 1989.

Song Mei, Yu Bai, Andrea Montanari, et al. The landscape of empirical risk for nonconvex losses. *Annals of Statistics*, 46(6A):2747–2774, 2018.

Stanislav Minsker. Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics*, 46(6A):2871–2903, 2018.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Ilsang Ohn and Yongdai Kim. Nonconvex sparse regularization for deep neural networks and its optimality. *Neural Computation*, 34(2):476–517, 2022.

Dmitrii M Ostrovskii and Francis Bach. Finite-sample analysis of m-estimators using self-concordance. *Electronic Journal of Statistics*, 15(1):326–391, 2021.

Oscar Hernan Madrid Padilla, Wesley Tansey, and Yanzhen Chen. Quantile regression with relu networks: Estimators and minimax rates. *Journal of Machine Learning Research*, 23(247):1–42, 2022.

Liang Peng and Yongcheng Qi. *Inference for heavy-tailed data: applications in insurance and finance*. Academic press, 2017.

Sidney I Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Guohao Shen, Yuling Jiao, Yuanyuan Lin, Joel L Horowitz, and Jian Huang. Deep quantile regression: Mitigating the curse of dimensionality through composition. *arXiv preprint arXiv:2107.04907*, 2021a.

Guohao Shen, Yuling Jiao, Yuanyuan Lin, and Jian Huang. Non-asymptotic excess risk bounds for classification with deep convolutional neural networks. *arXiv preprint arXiv:2105.00292*, 2021b.

Guohao Shen, Yuling Jiao, Yuanyuan Lin, and Jian Huang. Robust nonparametric regression with deep neural networks. *arXiv preprint arXiv:2107.10343*, 2021c.

Qiang Sun. Do we need to estimate the variance in robust mean estimation? *arXiv preprint arXiv:2107.00118*, 2021.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

Mahsa Taheri, Fang Xie, and Johannes Lederer. Statistical guarantees for regularized neural networks. *Neural Networks*, 142:148–161, 2021.

John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

Roman Vershynin. On the role of sparsity in compressed sensing and random matrix theory. In *2009 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 189–192. IEEE, 2009.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Lili Wang, Chao Zheng, Wen Zhou, and Wen-Xin Zhou. A new principle for tuning-free huber regression. *Statistica Sinica*, 32:1–25, 2022.

Ming Wen, Yixi Xu, Yunling Zheng, Zhouwang Yang, and Xiao Wang. Sparse deep neural networks using $l^{1,\infty}$-weight normalization. *Statistica Sinica*, 31:1397–1414, 2021.

Yi Xu, Shenghuo Zhu, Sen Yang, Chi Zhang, Rong Jin, and Tianbao Yang. Learning with non-convex truncated losses by sgd. In *Uncertainty in Artificial Intelligence*, pages 701–711. PMLR, 2020.

Xiaowei Yang, Shuang Song, and Huiming Zhang. Law of iterated logarithm and model selection consistency for generalized linear models with independent and dependent responses. *Frontiers of Mathematics in China*, pages 1–32, 2021.

Mingyang Yi, Ruoyu Wang, and Zhi-Ming Ma. Non-asymptotic analysis of excess risk via empirical risk landscape. *arXiv preprint arXiv:2012.02456*, 2020.

Huiming Zhang and Jinzhu Jia. Elastic-net regularized high-dimensional negative binomial regression: Consistency and weak signals detection. *Statistica Sinica*, 32:181–207, 2022.

Lijun Zhang and Zhi-Hua Zhou. ℓ_1-regression with heavy-tailed distributions. In *NeurIPS*, 2018.

Ziwei Zhu and Wenjing Zhou. Taming heavy-tailed features by shrinkage. In *International Conference on Artificial Intelligence and Statistics*, pages 3268–3276. PMLR, 2021.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.