Learning with Non-Convex Truncated Losses by SGD

Yi Xu[1]*, Shenghuo Zhu[2], Sen Yang[2], Chi Zhang[2], Rong Jin[2], Tianbao Yang[1]*
1. Department of Computer Science, The University of Iowa, Iowa City, IA 52246
2. Machine Intelligence Technology, Alibaba Group, Bellevue, WA 98004
*Correspondence to: yi-xu@uiowa.edu, tianbao-yang@uiowa.edu

Abstract

Learning with a *convex loss* function has been a dominating paradigm for many years. It remains an interesting question how non-convex loss functions help improve the generalization of learning with broad applicability. In this paper, we study a family of objective functions formed by truncating traditional loss functions, which is applicable to both shallow learning and deep learning. Truncating loss functions has potential to be less vulnerable and more robust to large noise in observations that could be adversarial. More importantly, it is a generic technique without assuming the knowledge of noise distribution. To justify non-convex learning with truncated losses, we establish excess risk bounds of empirical risk minimization based on truncated losses for heavy-tailed output, and statistical error of an approximate stationary point found by stochastic gradient descent (SGD) method. Our experiments for shallow and deep learning for regression with outliers, corrupted data and heavy-tailed noise further justify the proposed method.

1 INTRODUCTION

A fundamental problem in machine learning can be described as follows. Let $Z = (X, Y) \sim \mathbb{D}$ denote a random data following an unknown distribution of $\mathbb{D}$, where $X \in \mathcal{X} \subseteq \mathbb{R}^d$ denotes a random input and $Y \in \mathcal{Y} \subseteq \mathbb{R}$ denotes its corresponding output. Let $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$ denote a hypothesis class and $\ell(\cdot, Y)$ denote a loss function. Given a set of training data $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$, the problem is to find a hypothesis $h_n \in \mathcal{H}$ close to a hypothesis that minimizes the expected risk $P(h) := \mathrm{E}_Z[\ell(h(X), Y)]$. A classical approach is empirical risk minimization (ERM):

$$h_n = \arg\min_{h \in \mathcal{H}} \left\{ P_n(h) := \frac{1}{n} \sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i)) \right\}. \quad (1)$$

For large-scale problems with large n, an iterative stochastic algorithm (e.g., the stochastic gradient descent (SGD) method) can be easily employed to find an approximate solution to the above problem [8]. A central question in statistical learning theory is to characterize how close is the empirical risk minimizer h_n or its approximate solution to the optimal hypothesis $h_* \in \mathcal{H}$ that minimizes $P(h)$. In machine learning community, one is usually concerned with the *excess risk* $P(h_n) - P(h_*)$. In statistics community, one usually assumes a statistical model between Y and X, e.g., $Y = h_*(X) + \varepsilon$, where $h_* \in \mathcal{H}$, and ε is a zero-mean random noise, and studies the *statistical error* $\|h_n - h_*\|$ measured in some norm. There are extensive results of excess risk bounds for ERM with general loss functions and statistical error bounds of ERM with a square loss function [3, 9, 35, 46, 25, 27, 37, 49]. However, most of them need to assume the data (X, Y) and noise ε have normal behavior or formally sub-Gaussian tails. When distribution of data or noise deviates from sub-Gaussian, minimizing the standard convex loss functions [1] might yield poor performance [10].

Previous works for handling this issue either suffer from requiring strong knowledge of deviation or has high computational costs (see Related Work). In practice, it is rarely the case that the knowledge of data abnormality is given a-prior. Thus, the methods assuming this knowledge are not applicable. In this paper, we consider a generic method by minimizing non-convex truncated losses. The intuition is that if a particular data point $(X_i, Y_i, \varepsilon_i)$ deviates from normal behavior, the loss $\ell(h(X_i), Y_i)$ could be very large and therefore can be truncated to mitigate its effect on misleading the learn-

[1] the convexity of ℓ with respect to the prediction $h(\mathbf{x}_i)$.

ing process. In particular, we consider a family of non-convex truncated function $\phi(\ell)$ with a varied truncation level, and minimize the following ERM problem with truncated loss:

$$\hat{h}_n = \arg\min_{h\in\mathcal{H}} \left\{ \hat{P}_n(h) := \frac{1}{n}\sum_{i=1}^{n} \phi(\ell(h(\mathbf{x}_i), y_i))) \right\}. \tag{2}$$

There are several noticeable merits of this method: (i) the truncation can be used with any standard convex loss functions (e.g., square loss, absolute loss); (ii) the problem is still of a finite-sum form which enables one to employ simple SGD to find an approximate solution; (iii) it does not depend on knowledge of abnormality. Although minimizing truncated losses has been considered and adopted by practitioners [4], several fundamental questions have not been well addressed: (i) what is the excess risk round of $\hat{h}_n$ under abnormality of data; (ii) how to quantitively understand the benefit of truncation; (iv) if an iterative stochastic algorithm (e.g., SGD) is used to find an approximate solution to (2), what optimization and statistical guarantees can be provided on the obtained approximate solution. In this work, we provide partial answers to these questions. In particular, our contributions are summarized below:

- We establish an excess risk bound in the order of $O(1/\sqrt{n})$ for $\hat{h}_n$ - *the empirical minimizer* of the non-convex learning problem (2), when the output Y has a heavy-tailed distribution with bounded second-order moments. This result is applicable to learning with Lipschitz loss functions $\ell(z, y)$ (e.g., absolute loss) and the square loss function, linear models and non-linear deep models.

- We establish a statistical error bound of an *approximate stationary point* found by SGD that depends on the noise distribution, when the square loss ℓ is used in (2) for learning a linear model. We quantitatively analyze the benefit of truncation. In particular, our analysis shows that within a certain range of truncation levels, larger truncation could yield smaller statistical error. More importantly, truncation can tolerate much higher noise for enjoying consistency than without truncation.

- We consider the convergence of SGD for minimizing truncated Lipschitz losses without smoothness assumption. We show that SGD can still converge to points that are close to ϵ-stationary points with an iteration complexity of $O(1/\epsilon^4)$, which is the same as SGD for minimizing smooth functions.

2 RELATED WORK

Recent advances have sparked increasing interests in **non-convex learning (NCL)** (i.e., learning with non-convex objective functions and/or constraints). Below we will focus on review of non-convex learning for tackling data abnormality, in particular corruptions in Y and X, heavy-tailed noise ε.

Numerous studies have considered corruptions in the output Y [42, 5, 6, 43, 15]. A well-studied corruption model is to *assume that* $\mathbf{y} = \mathbf{X}\mathbf{w}_* + \varepsilon + \mathbf{b} \in \mathbb{R}^n$, where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ are sub-Gaussian noises, and $\mathbf{b} = (b_1, \ldots, b_n)^\top$ is a sparse vector with non-zero components corresponding to corrupted outputs. Recently, [5, 6] have studied minimizing a non-convex problem for recovering $\mathbf{w}_*$ for sub-Gaussian inputs $\mathbf{x}_i$. For example, the method proposed in [5] based on iterative hard-thresholding is motivated by solving a non-convex problem $\min_{\mathbf{w},\|\mathbf{b}\|_0\leq k_*} \|\mathbf{X}^\top\mathbf{w} - \mathbf{y} - \mathbf{b}\|_2^2 = \min_{\|\mathbf{b}\|_0\leq k_*} \|(I - P_X)(\mathbf{y} - \mathbf{b})\|_2^2$, where $P_X = \mathbf{X}(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}^\top$, where k_* is a assumed sparsity level of $\mathbf{b}$. Consistency of the learned model was proved in [5].

Several corruption models of input X have been considered [31, 36]. For example, Loh & Wainwright [31] considered three different corruption models, i.e., additive noise, multiplicative noise, and missing values. They proposed to minimize a non-convex quadratic objective based on estimates of $\mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{d\times d}$ and $\mathbf{X}\mathbf{X}^\top\mathbf{w}_*$ *using the knowledge of noise distribution*. The statistical error of the global optimum to the non-convex problem was established and it was also shown that projected gradient descent will converge in polynomial time to a small neighborhood of global minimizers.

The methods mentioned above could achieve superior performance when the corruption of data indeed follows the assumed model. However, in practice it is usually not clear how data is corrupted. A weaker assumption is to consider that the distribution of X or Y or ε is heavy-tailed with bounded moments. Several approaches with excess risk guarantee have been developed based on two popular mean estimators for heavy-tailed data, namely Catoni's mean estimator [12, 2] and median-of-means estimator [41, 1]. Brownless et. al. [10] learn a hypothesis based on minimizing Catoni's mean estimator $\hat{\mu}(h)$, i.e.,

$$\begin{aligned} &\min_{h\in\mathcal{H}} \hat{\mu}(h), \\ &\text{s.t. } \frac{\alpha}{n}\sum_{i=1}^{n} \phi((\ell(h(\mathbf{x}_i), y_i) - \hat{\mu}(h))/\alpha) = 0, \end{aligned} \tag{3}$$

where $\alpha > 0$ is a parameter and $\phi(\cdot) = sign(x)\log(1 + |x| + x^2/2)$. They established $O(1/\sqrt{n})$ excess risk bound. However, *their method is computationally expen-*

sive. In particular, it needs to find a scalar $\hat{\mu}(h)$ that satisfies the equality in (3) given a $h \in \mathcal{H}$, then to search for h that minimizes $\hat{\mu}(h)$. Although SGD could be used for the root finding problem (3), the minimization of $\hat{\mu}(h)$ does not have a nice structure to allow for an efficient solver. Some studies have provided efficient algorithms based on different estimators for learning with heavy-tailed data [21, 22]. But their results are only applicable in restrictive settings (e.g., for smooth and strongly convex losses), which preclude learning with non-convex objectives (e.g., deep learning).

Audibert and Catoni [2] proposed a method for learning a linear model based on solving a non-convex min-max problem and proved an excess risk bound of $O(1/n)$ for heavy-tailed data with a bounded fourth-order moment for noise and a bounded fourth-order moment for input. They proposed *a polynomial time* algorithm based heuristics to solve the non-convex min-max problem. However, it is unclear whether the approximate solution found by the heuristics-based approach satisfy the claimed excess risk bound. There also exists a bulk of studies focusing on understanding the excess risk bound of (regularized) ERM under certain conditions for unbounded loss (e.g., small ball condition, Bernstein condition, v-central condition, etc.) or in restricted settings (e.g., linear least-squares regression) [2, 14, 39, 30, 26, 40, 29, 28, 20, 17].

Different from these aforementioned studies, this paper focus on understanding the model learned by minimizing truncated losses without prescribing strong assumptions on data corruption. We note that this work is not the first one considering truncating the loss functions. In robust statistics, M-estimators based on non-convex truncated losses have been studied (e.g., Tukey's biweight [34], Cauchy loss [7]). However, conventional analysis of these estimators is usually restricted to asymptotic consistency of global minimizers of learning linear models [13]. In contrast, we provide excess risk bounds for learning general non-linear models as well, which is applicable to deep learning. The truncation function was also exploited in recent studies through different ways from ERM [10, 2]. However, their formulations are difficult to be solved, which even preclude simple SGD solvers. In contrast, SGD can be still used for solving our NCL formulation with truncated losses. To justify this approach, we also analyze the statistical error of a model learned by SGD. It is notable a recent work [50] studied ℓ_1-regression with heavy-tailed distribution and established an excess risk bound of ERM with a truncated loss. They focused on the statistical property of ERM and didn't design an efficient optimization algorithm such as SGD for solving their formulation. By contrast, our excess risk bound is applicable to general models, and we employ SGD to solve our formulation with theoretical guarantees.

Finally, we note that statistical error was also analyzed for high-dimensional robust M-estimator in [32]. Their analysis focus on understanding the sufficient conditions for robust linear regression such that the statistical error can be established for local stationary points. However, it is still unclear how truncation helps improve performance of learning without truncation, given that [2] has established the statistical error of linear least-squares regression without truncation. In contrast, our results are complementary, which not only establish the excess risk bounds for learning non-linear models, but also exhibit that truncation can tolerate much larger noise than without truncation (e.g., it allows noise increase as the number of samples but still maintains consistency)

3 NCL WITH TRUNCATED LOSSES

In this section, we first present some preliminaries in subsection 3.1. We present the excess risk bounds of NCL with truncated losses in subsection 3.2. In subsection 3.3, we consider a restricted setting for analyzing the statistical error of an approximate solution found by SGD. Finally, we consider the optimization properties of SGD for solving (2).

3.1 Preliminaries and Notations

For simplicity of presentation, we define $\mathcal{F} = \{f : Z \to \ell(h(X), Y), h \in \mathcal{H}\}$ and $\min_{h\in\mathcal{H}} P(h)$ is equivalent to the following problem:

$$f_* = \arg\min_{f\in\mathcal{F}} \{P(f) := \mathrm{E}_{Z\sim\mathbb{D}}[f(Z)]\}\,. \tag{4}$$

Let T be a (pseudo) metric space and D be a distance metric. An increasing sequence of $(\mathcal{A}_n)$ of partitions of T is said to be admissible if for all $n = 0, \ldots, \#\mathcal{A}_n \leq 2^{2^n}$. For any $t \in T$, let $A_n(t)$ be the unique element of $\mathcal{A}_n$ that contains t. Denote by $\Delta(A, D)$ the diameter of the set $A \subset \mathcal{T}$ under the metric D. Define

$$\gamma_\beta(\mathcal{T}, D) = \inf_{\mathcal{A}_n} \sup_{t\in\mathcal{T}} \sum_{n\geq 0} 2^{n/\beta}\Delta(A_n(t), D),$$

where the infimum is taken over all admissible sequences.

We will consider several distance metrics for the class $\mathcal{F}$. For $f, g \in \mathcal{F}$, let $d_m(f, g)$, $d_e(f, g)$, and $d_s(f, g)$ be

defined as follows:

$$d_m(f,g) = \max_{Z\in\mathcal{Z}} |f(Z) - g(Z)|,$$
$$d_e(f,g) = (\mathrm{E}[f(Z) - g(Z)]^2)^{1/2},$$
$$d_s(f,g) = \left[\frac{1}{n}\sum_{i=1}^{n}(f(Z_i) - g(Z_i))^2\right]^{1/2}.$$

Let $N(\mathcal{F}, \epsilon, d)$ be ϵ-covering number of the class $\mathcal{F}$ under the distance metric d, i.e., the minimal cardinality N of any set $\{f_1, \ldots, f_N\} \subset \mathcal{F}$ such that for all $f \in \mathcal{F}$ there exists $f_i \in \{f_1, \ldots, f_N\}$ with $d(f, f_i) \leq \epsilon$. Let $\Delta(\mathcal{F}, d_e)$ be diameter of the class F under the distance metric d_e. It is notable that $\gamma_\beta(\mathcal{F}, D) \leq \int_0^1 \log N(\mathcal{F}, \epsilon, D)^{1/\beta} d\epsilon$ [45].

Throughout the paper, we will focus on regression tasks and use the following statistical model between Y and X to demonstrate our results:

$$Y = h_*(X) + \varepsilon \tag{5}$$

where $\varepsilon \in \mathbb{R}$ is random noise independent of X, whose distribution is not necessarily sub-Gaussian. It is notable that the above model also capture some corruption models in X. For example, if $h_*(\mathbf{x}) = \mathbf{w}_*^\top \mathbf{x}$, then with an additive corruption model $\widehat{\mathbf{x}} = \mathbf{x} + \mathbf{x}_n$ we have $Y = \mathbf{w}_*^\top X + \mathbf{w}_*^\top X_n + \varepsilon = \mathbf{w}_*^\top X + \hat{\varepsilon}$. In fact, the theoretical results in Subsection 3.2 does not hinge on the statistical model in (5). The results can be applied to many other settings such as clustering problem in [10].

We consider the following definition of a truncation function.

Definition 1. *A function $\phi_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ is a* **truncation function** *parameterized by $\alpha > 0$ if (i) $\phi_\alpha(\cdot)$ is smooth, i.e., there exists a constant $L > 0$ and for any $x_1, x_2 \in \mathbb{R}_+$, $|\phi_\alpha(x_1) - \phi_\alpha(x_2) - \phi'_\alpha(x_2)(x_1 - x_2)| \leq \frac{L}{2}(x_1 - x_2)^2$; (ii) $\phi'_\alpha(x) = 1$ if $x = 0$ and $\phi'_\alpha(x) = 0$ if $x = \infty$; (iii) $\phi'_\alpha(x)$ is a monotonically decreasing function, i.e., $\phi'_\alpha(x_1) \geq \phi'_\alpha(x_2)$ if $x_1 \leq x_2$; (iv) there exists a universal constant $M > 0$ such that $|\phi_\alpha(x) - x| \leq \frac{Mx^2}{\alpha}$, and for any $\alpha_1 \leq \alpha_2$, we have $\phi'_{\alpha_1}(x) \leq \phi'_{\alpha_2}(x)$.*

According to the definition, we can see that $\phi''_\alpha(x) \leq 0$, which implies the non-convexity of ϕ_α. The parameter α determines the truncation level, i.e., the larger the α the smaller the truncation. From (iv) of the definition, we can see that when $\alpha = \infty$, we have $\phi_\alpha(x) = x$ meaning no truncation. Below, we will give some examples of truncation function.

Example 1. $\phi_\alpha^{(1)}(x) = \alpha \log(1 + \frac{x}{\alpha})$. Applying this truncation to a square loss yields Cauchy loss for regression [7]. We can verify that it is a truncation function and $|\phi_\alpha(x) - x| \leq \frac{x^2}{2\alpha}$ (see Appendix).

Example 2. $\phi_\alpha^{(2)}(x) = \alpha \log(1 + \frac{x}{\alpha} + \frac{x^2}{2\alpha^2})$. This truncation has been considered by [10] for computing a mean estimator under heavy-tailed distribution of data. One could consider a more general function $\phi_\alpha^{(m)}(x) = \alpha \log(1 + \sum_{k=1}^m \frac{x^k}{\alpha^k k!})$. See Appendix for verification of this function.

Example 3. The following function can be shown to be a truncation function (see Appendix):

$$\phi_\alpha^h(x) = \begin{cases} \frac{\alpha}{3}\left[1 - (1 - \frac{x}{\alpha})^3\right] & \text{if } 0 \leq x < \alpha, \\ \frac{\alpha}{3} & \text{otherwise.} \end{cases}$$

We plot the curves of the three truncation functions with varying α in Figure 1.

3.2 Excess Risk Bounds of NCL with Truncated Losses

This section concerns the excess risk bounds of NCL with truncated losses. Define:

$$\widehat{f} = \arg\min_{f\in\mathcal{F}} \left\{ P_n(\phi_\alpha(f)) := \frac{1}{n}\sum_{i=1}^{n} \phi_\alpha(f(Z_i)) \right\}. \tag{6}$$

Our analysis and results in this section are based on the following assumption.

Assumption 1. *There exists a constant $\sigma > 0$ such that $\mathrm{E}[f(Z)^2] \leq \sigma^2$ for any $f \in \mathcal{F}$.*

Remark. Please notice that the random function $f(Z)$ is not necessarily bounded, but it is reasonable to have a bounded mean and variance so that its second order moment is bounded. This assumption also made in many previous works [10, 22, 11, 14]. Next section will use a relaxed assumption for learning a linear model. Below, we will use the statistical model (5) to demonstrate the above assumption could hold under heavy-tailed distribution of Y. It is worth mentioning that there is **no convexity assumption** on function f in this subsection.

Theorem 2. *Under Assumption 1 and $\phi_\alpha(\cdot)$ is a truncation function, for any $\alpha > 0$ and for all $\delta \in (0, 1)$, with a probability at least $1 - \delta$ we have $P(\widehat{f}) - P(f^*) \leq \frac{2M\sigma^2}{\alpha} + C\beta(\mathcal{F}, \alpha)\log\left(\frac{1}{\delta}\right)\left(\frac{\gamma_2(\mathcal{F}, d_e)}{\sqrt{n}} + \frac{\gamma_1(\mathcal{F}, d_m)}{n}\right)$, where C is a universal constant, M is a constant appearing in Definition 1, $\beta(\mathcal{F}, \alpha) \in (0, 1]$ is a non-decreasing function of α.*

To understand the above result, we first present a corollary and an example below.

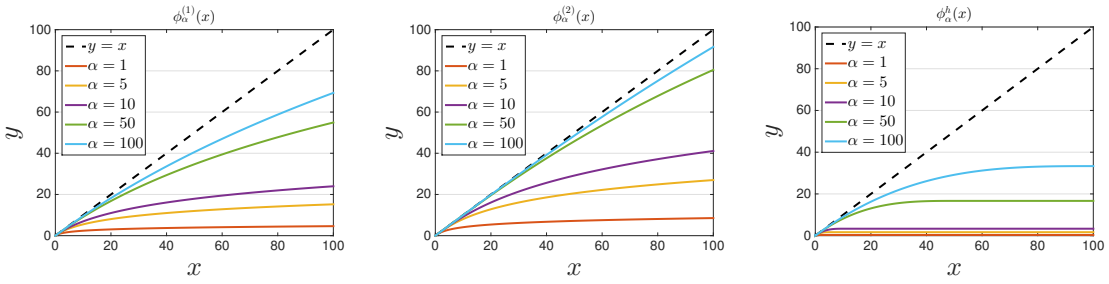

Figure 1: Visualization of different truncation losses with different α.

Corollary 3. *Under the same condition in Theorem 2, and $\ell(z, y)$ is a Lipschitz continuous function w.r.t the first argument and $\max_{X\in\mathcal{X},h,h'\in\mathcal{H}} |h(X) - h'(X)|$ is bounded. By setting $\alpha \geq \Omega(\sqrt{n})$, for all $\delta \in (0, 1)$, with a probability at least $1 - \delta$ we have $P(\widehat{f}) - P(f^*) \leq O\left(\frac{\log(1/\delta)[\gamma_2(\mathcal{F},d_e)+\gamma_1(\mathcal{F},d_m)/\sqrt{n}]}{\sqrt{n}}\right)$.*

Remark. Please note that α grows with the increasing of n, meaning that the truncation becomes "smaller" as n increases. The truncation is indeed "small" since intuitively we only need to truncate a small part of samples having a heavy-tailed noise but without changing the original function too much.

Let us consider the statistical model (5) as a special example. To learn a predictive function, we can use an absolute loss function $\ell(z, y) = |z - y|$. By assuming that $\sup_{h\in\mathcal{H},X\in\mathcal{X}} h(X) < \infty$ and $\mathrm{E}[Y^2] \leq \sigma^2$ (please note that the distribution of Y or ε could be heavy-tailed), then we have $\mathrm{E}[f(Z)^2] \leq 2\mathrm{E}[h(X)^2] + 2\sigma^2$ and the conditions in Corollary 3 hold. As a result, the empirical minimizer $\widehat{f}$ of (6) with $\alpha \geq \Omega(\sqrt{n})$ has an excess risk bound of $\widetilde{O}(1/\sqrt{n})$. Other loss functions that are Lipchitz continuous for a regression problem include ϵ-insensitive loss [44], piecewise-linear loss [24], and huber loss [23]. In comparison, Brownless et al. [10] have derived a similar order of excess risk bound for Lipschitz continuous losses. However, their solution is based on solving a difficult problem (3), while our solution is empirical minimizer of the truncated losses.

It is notable that the result in Theorem 2 is restricted to Lipschitz continuous loss functions, which precludes some non-Lipschitz continuous loss functions for heavy-tailed data. One example is the square loss for regression $\ell(z, y) = (z - y)^2$. The reason for this restriction is that the analysis for Theorem 2 hinges on the covering number of $\mathcal{F}$ under the metric d_m. Next, we present a result that relies on metrics d_e and d_s, which could imply an $\widetilde{O}(1/\sqrt{n})$ excess risk bound of $\widehat{f}$ for square loss.

Theorem 4. *Under the same condition in Theorem 2, for any $\delta \in (0, 1)$, let Γ_δ satisfy $\Pr(\gamma_2(\mathcal{F}, d_s) > \Gamma_\delta) \leq \delta/8$. With a probability at least $1 - 3\delta$, we have $P(\widehat{f}) - P(f^*) \leq C\beta(\mathcal{F}, \alpha)\max(\Gamma_\delta, \Delta(\mathcal{F}, d_e))\sqrt{\frac{\log(8/\delta)}{n}} + \frac{2M\sigma^2}{\alpha}$.*

Remark: It is not difficult to see that the above result only uses distance metrics d_e and d_s of $\mathcal{F}$, which makes it possible to derive an $\widetilde{O}(1/\sqrt{n})$ excess risk bound of $\widehat{f}$ for least-squares regression without the Lipschitz continuous assumption.

In particular, let us consider the regression model (5) and assume that $\mathrm{E}[Y^4] \leq \sigma^2$ (heavy-tailed) and $\sup_{h\in\mathcal{H},X\in\mathcal{X}} h(X) \leq \infty$. Let $\ell(h(X), Y) = (h(X) - Y)^2$. Then $\mathrm{E}[f(Z)^2] \leq 8\sigma^2 + 8\sup_{h\in\mathcal{H},X\in\mathcal{X}} h(X)^4 \triangleq \sigma_f^2$. By setting

$$\Gamma_\delta = 2\sqrt{2}\sqrt{\Delta^2(\mathcal{H}, d_m) + \mathrm{E}[Y^2] + \sqrt{\frac{8\sigma^2}{n\delta}}}\gamma_2(\mathcal{H}, d_m),$$

it was shown [10] that $\Pr(\gamma_2(\mathcal{F}, d_s) > \Gamma_\delta) \leq \delta/8$ and $\Gamma_\delta \geq \Delta(\mathcal{F}, d_e)$. By assuming $\sup_{h\in\mathcal{H},X\in\mathcal{X}} h(X) \leq \infty$, then $\Delta^2(\mathcal{H}, d_m)$ $\gamma_2(\mathcal{H}, d_m)$ are bounded. As a result, Theorem 4 implies an excess risk bound of $\widetilde{O}(1/\sqrt{n})$ for truncating the square loss to learn $\widehat{f}$ with $\alpha > \sqrt{n}$.

For comparison, we compare this result with that by Audibert and Catoni [2], which focuses on learning a linear model with a square loss function. They obtained an $\widetilde{O}(d/n)$ bound of regular ERM based on square losses for sufficiently large n, and also obtained $\widetilde{O}(1/n)$ bound for a non-convex min-max estimator. In contrast, our excess risk bound is worse but our bound is not pertained to linear model and square loss and is applicable to nonlinear models and non-square losses, and therefore has broader applicability (e.g., deep learning). In addition, our formulation could enjoy faster solver, e.g., SGD. For linear models using a square loss function, in next section we will establish a stronger result than that by Audibert and Catoni [2].

Finally, we mention how the truncation level parameter α enters into the excess risk bounds in Theorems 2, 4.

In particular, let us compare learning with truncation and without truncation. Indeed, $\beta(\mathcal{F}, \alpha)$ is related to Lipchitz constant of $\phi_\alpha(f)$ in terms of f. Without truncation $\alpha = \infty$, the first term in both bounds dominates and $\beta(\mathcal{F}, \alpha) = 1$. With truncation (e.g., $\alpha \leq \infty$), the first term could be scaled down by $\beta(\mathcal{F}, \alpha)$, making it possible to lower the overall bound. Notice that a smaller α gives a smaller $\beta(\mathcal{F}, \alpha) \leq 1$, thus reducing the first term. It implies that within a certain truncation level (i.e., not very small α such that the first term in the excess risk bound dominates), a larger truncation (i.e., a smaller α) could yield a smaller excess risk bound. However, it is difficult to quantify $\beta(\mathcal{F}, \alpha)$ due to that the analysis is based on a uniform bound for any $f \in \mathcal{F}$. To address this issue, we will present a different analysis below to demonstrate the benefits of truncation.

3.3 A Statistical Error of SGD for NCL with Truncated Losses

One shortcoming of excess risk bound analysis in last section is that it is restricted to the empirical minimizers $\widehat{f}$, which might not be obtained in practice due to the problem (6) is non-convex. It is well-known that non-convex problems could have bad local minimum or stationary points, and commonly used solvers (e.g., SGD) may stuck at local minimum and even stationary points [18, 51, 48]. In this section, we provide a direct analysis of SGD for solving (6) to show that truncation has a clear advantage for reducing statistical error. It should be noted that it will be difficult to analyze SGD for a general problem (6). Instead, we consider a statistical model $y_i = \mathbf{w}_*^\top \mathbf{x}_i + \varepsilon_i, \quad (i = 1, \ldots, n)$, and minimizing truncated square losses:

$$\min_{\mathbf{w}\in\mathbb{R}^d} F_\alpha(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \phi_\alpha\left(\frac{1}{2}(\mathbf{w}^\top \mathbf{x}_i - y_i)^2\right), \quad (7)$$

The update of SGD for minimizing (7) is

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla\phi_\alpha\left(\frac{1}{2}(\mathbf{w}_t^\top \mathbf{x}_{i_t} - y_{i_t})^2\right), \quad (8)$$

where i_t is a random sampled index. Considering ε_i is independent of $\mathbf{x}_i$, then $\mathbf{w}_*$ is the global minimizer of $\min_{\mathbf{w}\in\mathbb{R}^d} \mathrm{E}[\frac{1}{2}(\mathbf{w}^\top \mathbf{x}_i - y_i)^2)]$. We first show that SGD can find an approximate stationary point of $F_\alpha(\mathbf{w})$ with $O(1/\epsilon^4)$ iteration complexity.

Proposition 1. *Assume ϕ_α is a truncation function satisfying that there exists a constant $\kappa > 0$ such that $|x^2\phi''_\alpha(x^2/2)| \leq \kappa$ for any x, $\|\mathbf{x}_i\|_2 \leq R$ and $\mathrm{E}[\|\nabla\phi_\alpha(\frac{1}{2}(\mathbf{w}_t^\top \mathbf{x}_{i_t} - y_{i_t})^2) - \nabla F_\alpha(\mathbf{w}_t)\|^2] \leq \sigma_\alpha$ for all $t = 1, \ldots$. Then SGD finds an approximate stationary point $\mathbf{w}_R$ satisfying $\mathrm{E}_R[\|\nabla F_\alpha(\mathbf{w}_R)\|^2] \leq \epsilon^2$ with a complexity of $T = O(\sigma_\alpha^2/\epsilon^4)$, where R is a uniform random variable supported on $\{1, \ldots, T\}$.*

Remark: The condition $|x^2\phi''_\alpha(x^2/2)| \leq \kappa$ can be satisfied by the three examples presented before. The variance condition can be verified. Indeed, we can prove $\mathbf{w}_t$ reside in a bounded ball meaning this condition holds. In order to focus on our theme, we omit detailed discussion here.

Next, we present a result showing the statistical error of an approximate stationary solution found by SGD that depends on the distribution of ε for $\alpha < \infty$. For ease of understanding, we present a result for a particular truncation function. The result can be generalized to other truncation functions such that $|x^2\phi''_\alpha(x^2/2)| \leq \kappa$ as done in [32].

Theorem 5. *Suppose SGD returns an approximate stationary point $\mathbf{w}_\alpha$ such that $\|\mathbf{w}_\alpha - \mathbf{w}_*\|_2 \leq r$ and $\|\nabla F_\alpha(\mathbf{w}_\alpha)\| \leq \epsilon$. c and c' are universal constants. Assume $\mathbf{x}_i$ follows a sub-Gaussian distribution with parameter σ_x^2 and covariance matrix Σ_x, whose minimal eigen-value $\lambda_{min}(\Sigma_x) > 0$, $\phi_\alpha(x) = \alpha\log(1 + x/\alpha)$, $n \geq \Omega(d\log d)$ and the noise ε_i follows a distribution such that*

$$\begin{aligned} &c\sigma_x^2(\Pr(\varepsilon_i^2 \geq T^2/4)^{1/2} + \exp(-c'T^2/(2\sigma_x^2 r^2))) \\ \leq &\frac{\lambda_{min}(\Sigma_x)}{20} \end{aligned} \quad (9)$$

for $T \leq \sqrt{2\alpha}/2$, then with high probability $1 - c\exp(-c'\log d)$ we have

$$\|\mathbf{w}_\alpha - \mathbf{w}_*\|_2 \leq O\left(\sqrt{\frac{\alpha d\log d}{n}} + \frac{d\log d}{n}\frac{T^2}{r} + \epsilon\right) \quad (10)$$

Remark: The proof of above theorem builds on some results established in [32]. Please note that the above result can be applied to stationary points found by other algorithms than SGD. As SGD is emplyoed in this paper, we only state the results for SGD in above theorem. Below, we discuss new insights brought by the above results to justify the truncation. In particular, we focus on how truncation helps reduce the statistical error, which is missing in most robust statistics literature such as [32, 38].

First, it is notable that the noise ε_i could be heavy-tailed. The condition (9) imposes a lower bound for α due to the constraint $T \leq \sqrt{2\alpha}/2$ (i.e., truncation could not be arbitrarily large). An appropriate value should depend on the distribution of noise. Within a certain truncation level, the statistical error bound in (10) implies that smaller α may yield a smaller error.

Second, we show that the above result of an approximate stationary point to minimizing truncated square

losses can achieve a similar order of statistical error as linear least-squares regression without truncation established by Audibert and Catoni [2] under similar assumptions. In particular, under the assumptions that $\mathrm{E}[\varepsilon_i^4] \leq \sigma$ and a boundedness assumption of inputs, they achieved $F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) \leq O(d/n)$, where F is expected square loss, $\widehat{\mathbf{w}}$ is the optimal empirical solution to minimizing square losses. Under an eigen-value condition $\lambda_{min}(\Sigma_x) > 0$ as in above theorem it implies that $\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2 \leq O(\sqrt{d/n})$. In contrast, assuming $\mathrm{E}[\varepsilon_i^4] \leq \sigma$, we have $\Pr(\varepsilon_i^2 \geq T^2/4) \leq \frac{4\mathrm{E}[\varepsilon_i^2]}{T^2} \leq 4\sigma^2/T^2$ by Markov inequality. Therefore by choosing a large enough α (e.g., $\alpha = \Theta(\max(1/\lambda_{\min}(\Sigma_x), \log(1/\lambda_{\min}(\Sigma_x))))$, we can make (9) holds by setting $T = \sqrt{2\alpha}/2$. Then the statistical error bound of $\mathbf{w}_\alpha$ becomes $O(\sqrt{d \log d/n})$, which is comparable to $\widehat{\mathbf{w}}$. We note that mismatch of the $\log d$ factor is caused by different assumptions on the inputs. Nevertheless, $O(\sqrt{d \log d/n})$ is the minimax optimal rate when ε_i follows a Gaussian distribution [5].

Lastly, we show that the result in Theorem 5 is stronger than previous results on heavy-tailed noise (including Audibert and Catoni's results), especially with large noise. In particular, we could let $\mathrm{E}[\varepsilon_i^k]$ (where $k \in 2\mathbb{N}^+$) grows as n. For example, assume that $\mathrm{E}[\varepsilon_i^k] = n^c$. Let us set $\alpha = n^\beta$ with $\beta < 1$ and $T = n^{\beta/2}/2$. By Markov inequality, we have $\Pr(\varepsilon_i^2 \geq T^2/4) \leq O(\frac{\mathrm{E}[\varepsilon_i^k]}{T^k}) \leq O(n^{c-\beta k/2})$ Assuming that $c < \beta k/2$ and n is large enough, the inequality in (9) could hold. As a result, the statistical error becomes $O(\sqrt{d \log d/n^{1-\beta}})$, which still implies consistency of a stationary solution to minimizing the truncated losses. In contrast, most previous results on heavy-tailed noise assume $\mathrm{E}[\varepsilon_i^k]$ is bounded by a constant [10, 22, 11, 14, 2].

3.4 Optimization Issues For Minimizing Truncated Losses

Finally, we discuss the complexity of SGD for finding stationary points of averaged truncated losses beyond the setting of square loss and linear model as in last section. We assume the hypothesis is characterized by $\mathbf{w}$ and denote the loss function by $\ell(\mathbf{w}; \mathbf{x}, y)$ and the objective function in (2) becomes $F_\alpha(\mathbf{w}) = 1/n \sum_{i=1}^n \phi_\alpha(\ell(\mathbf{w}; \mathbf{x}_i, y_i))$. Note that $F_\alpha(\mathbf{w})$ is non-convex due to that ϕ_α is non-convex. We consider two cases depending on whether $\ell(\mathbf{w}; \mathbf{x}, y)$ is a convex function or not.

The complexity of SGD has been extensively studied in literature, especially when $F_\alpha(\mathbf{w})$ is smooth. When $\ell(\mathbf{w}; \mathbf{x}, y)$ is a non-convex function of $\mathbf{w}$ (e.g., for learning deep neural networks), if it is a smooth and Lipschitz continuous function of $\mathbf{w}$, then by the smoothness of $\phi_\alpha(\cdot)$ we can show that $F_\alpha(\mathbf{w})$ is a smooth function with Lipschitz continuous gradient. Hence it can find a ϵ-stationary point satisfying $\mathrm{E}[\|\nabla F_\alpha(\mathbf{w})\|^2] \leq \epsilon^2$ with a complexity of $O(1/\epsilon^4)$ [19].

If $\ell(\mathbf{w}; \mathbf{x}, y)$ is non-smooth and non-convex, characterizing the complexity of SGD becomes difficult, though it was shown that SGD can still converge to stationary points for a broad family of non-smooth non-convex functions [16]. Nevertheless, if $\ell(\mathbf{w}; \mathbf{x}, y)$ is a non-smooth convex function, e.g., for learning a linear model with absolute loss, ϵ-insensitive loss, piecewise linear loss, we can still characterize the complexity of SGD even without smoothness of the loss function. It is notable that gradient of non-smooth non-convex function may not be defined at some points. However, we can define sub-differentiable of a non-smooth non-convex function $g(\mathbf{w})$. Let $\partial g(\mathbf{w})$ denote the sub-differentiable of $g(\mathbf{w})$, which consists of a set of points $\mathbf{v}$ satisfying:

$$g(\mathbf{u}) \geq g(\mathbf{w}) + \mathbf{v}^\top(\mathbf{u} - \mathbf{w}) + o(\|\mathbf{u} - \mathbf{w}\|), \text{as } \mathbf{u} \to \mathbf{w}.$$

For a convex function $\ell(\mathbf{w})$ and a smooth truncation function $\phi_\alpha(\ell)$, we have $\partial\phi_\alpha(\ell(\mathbf{w})) = \phi'_\alpha(\ell(\mathbf{w}))\partial\ell(\mathbf{w})$. With this, we can define $\partial F_\alpha(\mathbf{w}) = 1/n \sum_i \partial\phi_\alpha(\ell(\mathbf{w}^\top\mathbf{x}_i - y_i))$. A point $\mathbf{w}$ is said to be stationary point of $F_\alpha(\mathbf{w})$ if $\mathrm{dist}(0, \partial F_\alpha(\mathbf{w})) = 0$, where dist denotes the distance from a point to a set. For our problem, we can establish the following convergence result of SGD for minimizing $F_\alpha(\mathbf{w})$ with Lipschitz continuous convex losses.

Proposition 2. *Assume $\ell(\mathbf{w}; \mathbf{x}, y)$ is convex and satisfies $\|\partial\ell(\mathbf{w}^\top\mathbf{x}_i - y_i)\| \leq G$, then SGD for minimizing $F_\alpha(\mathbf{w})$ can find a point $\mathbf{w}_\alpha$ that is close to a point $\widetilde{\mathbf{w}}_\alpha$ such that $\mathrm{E}[\|\mathbf{w}_\alpha - \widetilde{\mathbf{w}}_\alpha\|_2^2] \leq \epsilon^2$, and $\mathrm{E}[dist(0, \partial F_\alpha(\widetilde{\mathbf{w}}_\alpha))^2] \leq \epsilon^2$ with a complexity $O(1/\epsilon^4)$.*

Remark: The result implies even for non-smooth loss functions, SGD for learning with truncated losses can converge to a point that is close to an approximate stationary point. The idea for proving this result is that we prove $F_\alpha(\mathbf{w})$ is a weakly convex function and then the result of SGD for minimizing weakly convex function is applicable [16].

4 EXPERIMENTS

We provide some empirical results to demonstrate the effectiveness of the proposed approach for learning both linear and deep models. We use SGD to find approximate solutions of ERM with truncation and without truncation. A standard regularization term $\lambda\|\mathbf{w}\|^2$ is also added to ERM. The values of λ and α are selected by cross-validation. Two loss functions will be considered, namely absolute loss and square loss. The truncation function is $\phi_\alpha^{(1)}$. Other truncation functions offer simi-

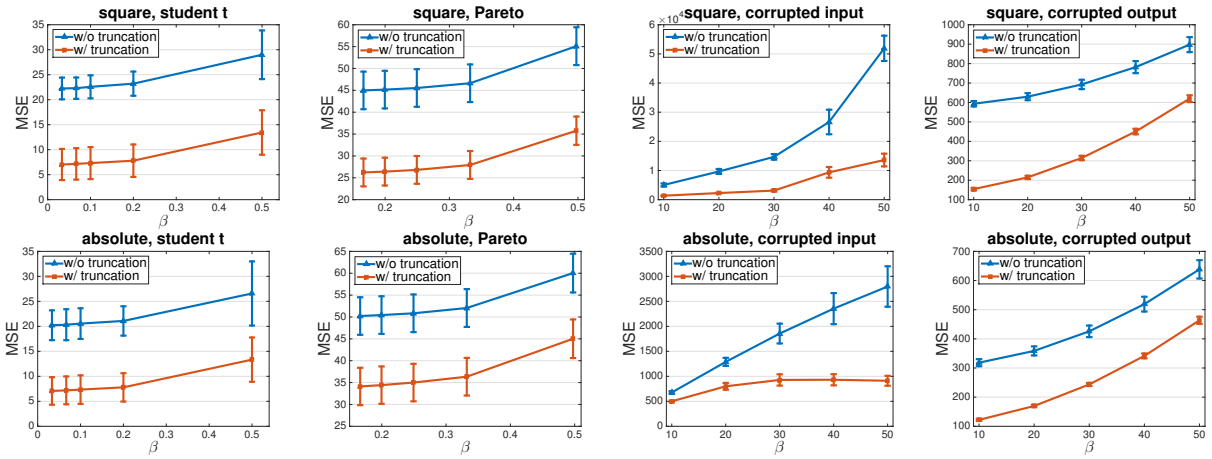

Figure 2: Comparisons of Testing Error for w/ and w/o truncation with varying noise level.

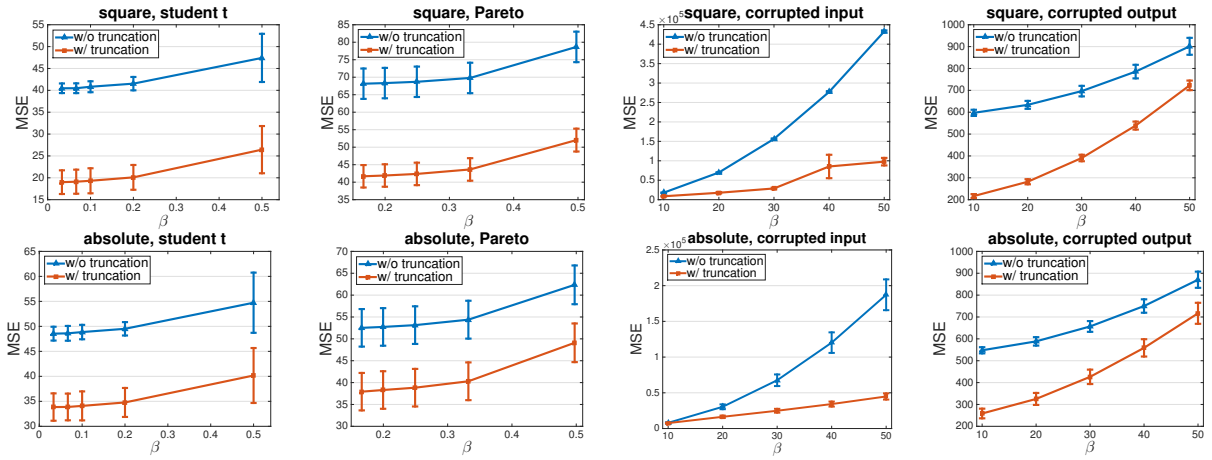

Figure 3: Comparisons of Testing Error for w/ and w/o truncation with varying noise level (without regularizer).

lar trend as reported results. In the presentation of experimetal results, the results of our algorithm are marked as "w/ truncation" and the results of the baseline are marked as "w/o truncation".

Table 1: Statistics of Datasets

data	Pred.Date	Pred. Period	n_{train}	$n^{\text{s}}_{\text{test}}$	$n^{\text{i}}_{\text{test}}$
P1	24, Apr	02-08, May	588956	410	2689
P2	01, May	09-15, May	586761	405	2523
P3	08, May	16-22, May	576386	397	2561
P4	15, May	23-29, May	564145	398	2775

Synthetic data. We conduct experiments on synthetic data first because it allows us to add different corruptions with varying noise level. We consider a linear regression model $y_i = \mathbf{w}_*^\top \mathbf{x}_i + \varepsilon_i$, and two loss functions, i.e., square loss and absolute loss. We generate a random data matrix $X \in \mathbb{R}^{n_{\text{train}} \times d}$ with $n_{\text{train}} = 1000$ and $d = 1000$. The entries of X and $\mathbf{w}_*$ are generated independently with a standard Gaussian and a uniform distribution $U[0, 1]$, respectively. Then we add several types of noise into the statistical model for generating outputs. **(a) student-t noise** where the noise ε follows a Student's t-distribution with degrees of freedom $1/\beta \in \{2, 5, 10, 15, 30\}$. **(b) Pareto noise** whre the noise ε follows a Pareto distribution with tail parameter of $1/\beta \in \{2.01, 3.01, 4.01, 5.01, 6.01\}$, and then following by [10], it is appropriately recentered in order to have zero mean. **(c) Corrupted output:** following by [6], a randomly generated sparse vector $\mathbf{b}$ is added to Gaussian noise ε for generating $\mathbf{y}$. The nonzero entries of $\mathbf{b}$ follow a uniform distribution $U[-\beta, \beta]$ with $\beta \in \{10, 20, 30, 40, 50\}$. The sparsity is set to be 80%. **(d) Corrupted input:** following by [33], $\mathbf{x}$ is corrupted by $\mathbf{z} = \mathbf{x} + \mathbf{x}_\xi$ where $\mathbf{x}_\xi$ is independent of $\mathbf{x}$ and follows a uniform distribution $U[-\beta, \beta]$ with

Table 2: Comparison of Testing Error On Real Datasets

Model	Data	Absolute loss (MAE)		Square loss (MSE)	
		w/o truncation	w/ truncation	w/o truncation	w/ truncation
linear model	house	6.8931	5.1561	66.4300	23.8871
deep model (item-SKU level data)	P1	15.196	14.113	1482	1167
	P2	17.766	16.797	2210	1806
	P3	22.104	18.642	2375	2049
	P4	20.648	14.176	2323	1032
deep model (supplier level data)	P1	76.459	74.190	11726	9515
	P2	87.276	81.292	38618	15247
	P3	121.95	99.161	28396	17571
	P4	137.06	82.542	33106	11913

$\beta \in \{10, 20, 30, 40, 50\}$. Note that these corruptions have been considered in previous works and β controls noise level in the corruption. A testing dataset with the sample sizes of $n_{\text{test}} = 1000$ is generated following the true model $y = \mathbf{w}_*^\top \mathbf{x}$ for evaluation. We report the standard testing mean-square-error (MSE) for both our algorithm (the lines marked as "w/ truncation") and the baseline (the lines marked as "w/o truncation") in different noise levels averaged over 5 random trials in Figure 2. We also perform the experiments without regularizer ($\lambda = 0$) and include the results in Figure 3. The results clearly show that the performance of learning with truncation by SGD are better than learning without truncation.

Real data. We use a real dataset housing from libsvm website[2] with sample size $n = 506$ to train a linear model. We randomly select $n_{\text{train}} = 253$ as for training and cross-validation and the remaining as testing. We also investigate a real-world application of learning deep neural networks in e-commerce, and demonstrate that our theoretical results can be effectively applied to learning a deep non-linear model. The task is to forecast the weekly sales (e.g., future two weeks) of certain products. In online retailing, accurate forecasting is crucial since it helps the platform to design the promotion activities as well as online sellers to optimize the inventory strategies. A dataset of four continuous weeks in May 2017 is used for the experimental demonstration (denoted by P1$\sim$ P4). A total of 324 features including previous sales, consumer preference, and other useful information are collected. The statistics of each weekly data are included in Table 1 for reference. The DNN model has 5 layers, and ReLu is used as the activation function. In each hidden layer, the number of units is 80, and both input and output layers contain 50 units. For models learned with absolute losses, mean-absolute-error (MAE) is used to measure the performance, while for model learned with square losses, MSE is used. The results are shown in Table 2, which again demonstrate that the performance of learning with truncation has a significant improvement over that without truncation transformation for both linear and non-linear models. We also provide the Q-Q plots of the prediction error to show that the considered data do exhibit heavy-tailed nature. Q-Q plot is a graphical method for comparing two probability distributions by plotting their quantiles against each other [47]. If the two distributions are similar, the points in the Q-Q plot will approximately lie on the red line. These results are presented in Figure 4, showing the heavy-tailed nature of the house data.

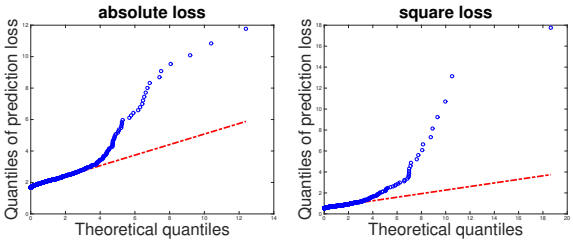

Figure 4: Q-Q plots on housing data

5 CONCLUSIONS

In this paper, we have considered non-convex learning with truncated losses from various perspectives and justified the benefit of truncation in the presence of large noise in data. For future work, we will consider analyze the **statistical error of stationary points for other losses** and develop stochastic algorithms for solving the involved problem with better time complexity.

Acknowledgements

The authors thank the anonymous reviewers for their helpful comments. Part of this work was done when Y. Xu was a research intern at Alibaba. Y. Xu and T. Yang are partially supported by National Science Foundation (IIS-1545995).

[2]https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

References

[1] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58:137–147, 1999.

[2] J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *Annals of Statistics*, 39(5):2766–2794, 2011.

[3] P. L. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006.

[4] V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab. Robust optimization for deep regression. In *International Conference on Computer Vision*, pages 2830–2838, 2015.

[5] K. Bhatia, P. Jain, P. Kamalaruban, and P. Kar. Consistent robust regression. In *Advances in Neural Information Processing Systems*, pages 2107–2116, 2017.

[6] K. Bhatia, P. Jain, and P. Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729, 2015.

[7] M. J. Black and P. Anandan. The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63:75–104, 1996.

[8] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*, pages 177–187, 2010.

[9] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

[10] C. Brownlees, E. Joly, and G. Lugosi. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 2015.

[11] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

[12] O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 1148–1185. Institut Henri Poincaré, 2012.

[13] L. Chang, S. Roberts, and A. Welsh. Robust lasso regression using tukey's biweight criterion. *Technometrics*, 60(1):36–47, 2018.

[14] C. Cortes, S. Greenberg, and M. Mohri. Relative deviation learning bounds and generalization with unbounded loss functions. *arXiv:1310.5796*, 2013.

[15] A. S. Dalalyan and Y. Chen. Fused sparsity and robust estimation for linear models with unknown variance. In *Advances in Neural Information Processing Systems*, pages 1268–1276, 2012.

[16] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[17] V. C. Dinh, L. S. Ho, B. Nguyen, and D. Nguyen. Fast learning rates with heavy-tailed losses. In *Advances in Neural Information Processing Systems*, pages 505–513, 2016.

[18] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points — online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[19] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[20] P. D. Grünwald and N. A. Mehta. Fast rates for general unbounded loss functions: from ERM to generalized Bayes. *arXiv:1605.00252*, 2016.

[21] D. Hsu and S. Sabato. Heavy-tailed regression with a generalized median-of-means. In *International Conference on Machine Learning*, pages 37–45, 2014.

[22] D. Hsu and S. Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40, 2016.

[23] P. J. Huber. Robust estimation of a location parameter. *The annals of mathematical statistics*, pages 73–101, 1964.

[24] R. Koenker. *Quantile regression*. Number 38. Cambridge university press, 2005.

[25] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer, 2011.

[26] G. Lecué and S. Mendelson. General nonexact oracle inequalities for classes with a subexponential envelope. *The Annals of Statistics*, 40(2):832–860, 2012.

[27] G. Lecué and S. Mendelson. Learning subgaussian classes: upper and minimax bounds. *Topics in Learning Theory-Societe Mathematique de France*, 2013.

[28] G. Lecué and S. Mendelson. Regularization and the small-ball method II: complexity dependent error rates. *Journal of Machine Learning Research*, 18(1):5356–5403, 2017.

[29] G. Lecué and S. Mendelson. Regularization and the small-ball method I: sparse recovery. *The Annals of Statistics*, 46(2):611–641, 2018.

[30] T. Liang, A. Rakhlin, and K. Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285, 2015.

[31] P. Loh and M. J. Wainwright. Corrupted and missing predictors: Minimax bounds for high-dimensional linear regression. In *International Symposium on Information Theory*, pages 2601–2605. IEEE, 2012.

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

[33] P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, pages 1637–1664, 2012.

[34] R. A. Maronna, D. R. Martin, and V. J. Yohai. *Robust Statistics: Theory and Methods*. John Wiley and Sons, New York, 2006.

[35] P. Massart. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

[36] B. McWilliams, G. Krummenacher, M. Lucic, and J. M. Buhmann. Fast and robust least squares estimation in corrupted linear models. In *Advances in Neural Information Processing Systems*, pages 415–423, 2014.

[37] N. A. Mehta and R. C. Williamson. From stochastic mixability to fast rates. In *Advances in Neural Information Processing Systems*, pages 1197–1205, 2014.

[38] S. Mei, Y. Bai, A. Montanari, et al. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

[39] S. Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.

[40] S. Mendelson. Learning without concentration for general loss functions. *Probability Theory and Related Fields*, 2017.

[41] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience series in discrete mathematics. Wiley, Chichester, New York, 1983.

[42] N. H. Nguyen and T. D. Tran. Exact recoverability from dense corrupted observations via l1-minimization. *IEEE Trans. Information Theory*, 59(4):2017–2035, 2013.

[43] N. H. Nguyen and T. D. Tran. Robust lasso with missing and grossly corrupted observations. *IEEE Trans. Information Theory*, 59(4):2036–2058, 2013.

[44] L. Rosasco, E. D. Vito, A. Caponnetto, M. Piana, and A. Verri. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004.

[45] M. Talagrand. *The Generic Chaining: Upper and Lower Bounds of Stochastic Processes*. Springer, 2005.

[46] S. A. Van de Geer. *Applications of empirical process theory*, volume 91. Cambridge University Press Cambridge, 2000.

[47] M. B. Wilk and R. Gnanadesikan. Probability plotting methods for the analysis for the analysis of data. *Biometrika*, 55(1):1–17, 1968.

[48] Y. Xu, R. Jin, and T. Yang. First-order stochastic algorithms for escaping from saddle points in almost linear time. In *Advances in Neural Information Processing Systems*, pages 5535–5545, 2018.

[49] L. Zhang, T. Yang, and R. Jin. Empirical risk minimization for stochastic convex optimization: $O(1/n)$- and $O(1/n^2)$-type of risk bounds. In *Conference on Learning Theory*, pages 1954–1979, 2017.

[50] L. Zhang and Z.-H. Zhou. ℓ_1-regression with heavy-tailed distributions. In *Advances in Neural Information Processing Systems*, pages 1084–1094, 2018.

[51] Y. Zhang, P. Liang, and M. Charikar. A hitting time analysis of stochastic gradient langevin dynamics. In *Conference on Learning Theory*, pages 1980–2022, 2017.