On the Robustness and Generalization of Cauchy Regression*

Tongliang Liu and Dacheng Tao
Centre for Quantum Computation and Intelligent Systems
Faculty of Engineering and Information Technology
University of Technology, Sydney
Ultimo, NSW 2007, Australia
tongliang.liu@student.uts.edu.au, dacheng.tao@uts.edu.au

***Abstract*— It was recently highlighted in a special issue of Nature [1] that the value of big data has yet to be effectively exploited for innovation, competition and productivity. To realize the full potential of big data, big learning algorithms need to be developed to keep pace with the continuous creation, storage and sharing of data. Least squares (LS) and least absolute deviation (LAD) have been successful regression tools used in business, government and society over the past few decades. However, these existing technologies are severely limited by noisy data because their breakdown points are both zero, i.e., they do not tolerate outliers. By appropriately setting the turning constant of Cauchy regression (CR), the maximum possible value (50%) of the breakdown point can be attained. CR therefore has the capability to learn a robust model from noisy big data. Although the theoretical analysis of the breakdown point for CR has been comprehensively investigated, we propose a new approach by interpreting the optimization of an objective function as a sample-weighted procedure. We therefore clearly show the differences of the robustness between LS, LAD and CR. We also study the statistical performance of CR. This study derives the generalization error bounds for CR by analyzing the covering number and Rademacher complexity of the hypothesis class, as well as showing how the scale parameter affects its performance.**

***Index Terms*— Cauchy regression, robustness, generalization error bound, covering number, Rademacher complexity.**

I. Introduction

The explosive growth in sensors and the rapid development of internet technology have prompted the collection, sharing, combination and use of massive amounts of data. The era of big data has arrived, and it represents a new frontier for innovation and productivity [2]. As a correlate of big data, big learning - the use, exploration, exploitation and integration of large amount of data using database systems, data mining, pattern recognition, statistics, and distributed and parallel systems in academia and industry (see e.g., [3], [4], [5], [6], [7]) - has achieved significant successes for business, government and society.

The sheer volume, velocity, variety and veracity of big data challenge the more popular machine learning techniques; the explosion of data obtained through social networks or special data-collecting channels within organizations and sensor networks are good examples. Big data can easily be corrupted by noise during collection, convection and combination due to its highly distributed, dynamic and unstructured nature. Robust analytic tools are therefore essential for big learning.

Data regression is a fundamental problem in modern data analysis, and regression methods need to be carefully re-investigated for big data purposes. One well-known measure of robustness for regression is the finite sample breakdown point, defined as the minimum proportion of incorrect observations (e.g., outliers) an estimator can handle before giving an arbitrarily large result [8]. Although least squares (LS) and least absolute deviation (LAD) perform well for regression, their breakdown points are both zero, which means that they do not tolerate outliers. Here, we introduce Cauchy regression (CR) [9] to big learning. CR assumes the observations are perturbed by independent but identical Cauchy distribution noise and learns the parameter based on maximum likelihood estimation. Since the Cauchy distribution is fat-tailed, CR is much more robust to outliers and has breakdown points reaching the maximum possible value (50%) with an appropriate tuning constant [10].

Consider the linear generative model $y = \langle w, x\rangle + \epsilon$, where ϵ is the additive noise. CR assumes that ϵ is distributed independently from a Cauchy distribution centering at zero with spread σ (the scale parameter). Let ρ be a probability distribution on $\mathbb{R}^d \times \mathbb{R}$ and $Z = (x_i, y_i)_{i=1}^n \in (\mathbb{R}^d \times \mathbb{R})^n$ an independent and identically distributed (i.i.d.) sample with size n. Using maximum likelihood estimation rule, we can optimize w and σ to solve the following minimization problem:

$$\min_{w,\sigma} F(w,\sigma,Z) = \sum_{i=1}^{n} \ln\left(1 + \left(\frac{y_i - w^T x_i}{\sigma}\right)^2\right) + k\ln\sigma,$$

where k is the tuning constant. The loss function of CR is $\ell(z,w,\sigma) = \ln\left(1 + \left(\frac{y_i - w^T x_i}{\sigma}\right)^2\right)$ [1].

Previous theoretical analyses of CR have focused on breakdown points. An analysis of statistical performance is still lacking. In this paper, we first propose an approach to

*This work was partially supported by ARC DP-140102164 and ARC FT-130101457.

[1]The loss function excludes the term $k\ln\sigma/n$, because this term will not contribute to the generalization performance.

978-1-4799-4808-6 /14/$31.00 ©2014 IEEE

study the robustness of an algorithm from an optimization view and clearly show the superiority of CR to LS and LAD regarding to robustness. Then, we derive the generalization error bounds for CR stemmed from statistical learning theory [11]. We use the covering number [12] and Rademacher complexity [13] to measure the complexity of the hypothesis class of CR and obtain dimensionality-dependent and dimensionality-independent generalization error bounds, respectively. Although the bound obtained by the covering number is comparable to that obtained by the Rademacher complexity due to the existence of the scale parameter, these two bounds are complementary to each other.

The rest of the paper is organized as follows. In Section II, we compare the robustness between LS, LAD and CR from an optimization view. In Section III, we derive the generalization error bounds for CR. We discuss how the scale parameter affects the generalization performance in Section IV. Proofs of the generalization error bounds are presented in Section V. We conclude the paper and give suggestions for future works in Section VI.

II. Compare the robustness between LS, LAD and CR

Besides the breakdown point, many other concepts, such as the influence function [14] and robustness[2] [15], are also extensively studied to measure how an algorithm tolerates to noise (or outliers). However, most of them are very rough and they cannot provide a thorough comparison between LS, LAD and CR. We present a new approach to compare the robustness between LS, LAD and CR regarding to an optimization view.

Let $F(w)$ be the objective function of an regression problem. Let $f(t) = F(tw)$. We can verify that optimizing the objective function $F(w)$ is equal to finding a w such that $f'(1) = 0$, where $f'(t)$ denotes the derivative of $f(t)$.

Let

$$F_{\text{LS}}(w) = \sum_{i=1}^{n} (y_i - w^T x_i)^2$$

and

$$F_{\text{LAD}}(w) = \sum_{i=1}^{n} \left| y_i - w^T x_i \right|$$

and

$$F_{\text{CR}}(w) = \sum_{i=1}^{n} \ln \left(1 + \left(\frac{y_i - w^T x_i}{\sigma} \right)^2 \right) + k \ln \sigma$$

[2]This robustness measures the property that if a testing sample is "similar" to a training sample, then the testing error is close to the training error. In the rest of the paper, the concept of robustness measures how much an algorithm tolerates to noise (and outliers).

be the objective functions of LS, LAD and CR, respectively. We then have

$$f'_{\text{LS}}(1) = \sum_{i=1}^{n} 2(y_i - w^T x_i)(-w^T x_i) \tag{1}$$

and

$$\delta f_{\text{LAD}}(1) = \sum_{i=1}^{n} \frac{1}{|y_i - w^T x_i|} (y_i - w^T x_i)(-w^T x_i), \tag{2}$$

where $\delta f(x)$ is the subgradient of $f(x)$. We also define that $\frac{0}{0}$ can be any real value in $[-1, 1]$. And

$$f'_{\text{CR}}(1) = \sum_{i=1}^{n} \frac{2}{\sigma^2 + (y_i - w^T x_i)^2} (y_i - w^T x_i)(-w^T x_i). \tag{3}$$

Now, we can thoroughly compare the robustness of LS, LAD and CR by comparing equations (1), (2) and (3). Given a training sample set, we want to optimize the objective functions of LS, LAD and CR by finding ws such that $f'_{\text{LS}}(1)$, $\delta f_{\text{LAD}}(1)$ and $f'_{\text{CR}}(1)$ equal to zero. Let

$$c(w, z_i) = (y_i - w^T x_i)(-w^T x_i)$$

be the contribution of the ith sample to the optimization procedure. We see that for LS, all $c(w, z_i), i = 1, \ldots, n$ has the same weights 2. However for LAD, the weights are different, they are

$$\frac{1}{|y_i - w^T x_i|}, i = 1, \ldots, n$$

and for CR, the weights are

$$\frac{2}{\sigma^2 + (y_i - w^T x_i)^2}, i = 1, \ldots, n.$$

Let

$$e(w, z_i) = |y_i - w^T x_i|, i = 1, \ldots, n$$

be an error function. By the generative model of regression and the traditional definition of an outlier, we see that $e(w, z_i)$ represents the noise error added to z_i or the very large error introduced by an outlier z_i. Thus, LS, LAD and CR can be interpreted as sample-weighted procedures and they have different weighting strategies.

During the optimization procedures, robust algorithms should give a small weight to a large error training sample and a large weight to a small error training sample. Comparing the weight functions of LS, LAD and CR, we can conclude that LAD is more robust than LS and that CR is more robust than LAD.

III. Generalization error bounds

We first provide the stochastic framework for RC, upon which our results are based.

Definition 1 (Expected risk): The expected risk of Cauchy regression is defined as

$$R(w,\sigma) \triangleq \int_{\mathcal{X}} \ell(z,w,\sigma) d\rho(z).$$

However, the probability density function of ρ may be unknown. We define the empirical risk as follows to approximate the expected risk.

Definition 2 (Empirical risk): The empirical risk of Cauchy regression is defined as

$$R_n(w,\sigma) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(Z_i,w,\sigma),$$

which is a discretization of the expected risk $R(w,\sigma)$.

A probabilistic bound on the defect $R(w,\sigma) - R_n(w,\sigma)$ is called the *generalization error bound*.

The approach, which involves computing the complexity of hypothesis class, is one of the traditional ways to derive generalization error bounds in statistical learning theory [11]. Fat-shattering dimension [16], covering number [12] and Rademacher complexity [13] are the most frequently used complexity measures of the hypothesis class for regression. Fat-shattering dimension, a generalization of VC dimension, is usually bounded using a covering number [16]. We bound the covering number and Rademacher complexity to derive dimensionality-dependent and dimensionality-independent generalization error bounds, respectively.

The definition of the covering number is obtained from [12].

Definition 3 (Covering number): Let B be a metric space with metric d. Given observation $X = (x_1,\ldots,x_n)$, and vectors $f(X) = (f(x_1),\ldots,f(x_n)) \in B^n$, the covering number in p-norm, denoted as $\mathcal{N}_p(F,\xi,X)$, is the minimum number m of a collection of vectors $v_1,\ldots,v_m \in B^n$, such that $\forall f \in F, \exists v_j$:

$$\|d(f(X),v_j)\|_p = \left[\sum_{i=1}^{n} d(f(x_i),v_j^i)^p\right]^{1/p} \leq n^{1/p}\xi,$$

where v_j^i is the i-th component of vector v_j. We also define $\mathcal{N}_p(F,\xi,n) = \sup_X \mathcal{N}_p(F,\xi,X)$.

We bound the covering number of the loss class induced by the loss function for CR.

Lemma 1: Let $|y| \leq A, \|x\| \leq R$ and $W = \{w | \|w\| \leq B\}$. Let F_W be the loss function class induced by the loss function $\ell(Z,w,\sigma)$ with $w \in W$ for Cauchy regression. Then

$$\begin{aligned}\mathcal{N}_1(F_W,\xi,2n) &\leq \left(\frac{4\sqrt{d}BR(A+BR)}{\xi\sigma^2}\right)^d \\ &= \frac{1}{\sigma^{2d}}\left(\frac{4\sqrt{d}BR(A+BR)}{\xi}\right)^d.\end{aligned}$$

Detailed proof is given in Section IV. A dimensionality-dependent generalization error bound can be obtained using Lemma 1.

Theorem 1: Let $Z = (x_i,y_i)_{i=1}^n$ be an i.i.d. sample. If $|y| \leq A, \|x\| \leq R$ and $W = \{w | \|w\| \leq B\}$, for any $n \geq 8$ and $\delta > 0$, with probability at least $1-\delta$ we have for all w in W learned by CR that

$$\begin{aligned}&|R(w,\sigma) - R_n(w,\sigma)| \\ &\leq C\sqrt{\frac{32\left(d\ln\left(\frac{4\sqrt{d}BR(A+BR)}{C\sigma^2}\right) - \ln\frac{\delta}{8}\right)}{n}},\end{aligned}$$

where $C = \ln(1+\frac{2A^2+2B^2R^2}{\sigma^2})$.
Detailed proof is given in Section IV.

Theorem 1 looks a little complicated. However, we can simplify it by using the fact that $\ln(1+\frac{2A^2+2B^2R^2}{\sigma^2}) \leq \frac{2A^2+2B^2R^2}{\sigma^2}$. We therefore have

$$\begin{aligned}&|R(w,\sigma) - R_n(w,\sigma)| \\ &\leq C\sqrt{\frac{32\left(d\ln\left(\frac{4\sqrt{d}BR(A+BR)}{2A^2+2B^2R^2}\right) - \ln\frac{\delta}{8}\right)}{n}}.\end{aligned}$$

For a given scale parameter σ, its upper bound is of order $\mathcal{O}(\sqrt{d\ln d/n})$.

We will analyze the upper bound by comparing it with a dimensionality-independent upper bound (shown by Theorem 2) in Section III.

Theorem 1 provides a dimensionality-dependent generalization bound. We can derive a dimensionality-independent bound by bounding the Rademacher complexity of the loss class. The definition of Rademacher complexity is taken from [13].

Definition 4 (Rademacher complexity): Given observation $X = (x_1,\ldots,x_n)$, the empirical Rademacher complexity of a function class F is

$$\mathfrak{R}_n(F) = E_g \sup_{f\in F} \frac{2}{n}\sum_{i=1}^{n} g_i f(x_i),$$

where $g_i,\ldots,g_n$ are independent Rademacher variables, which are uniformly distributed on $\{-1,1\}$. The Rademacher complexity of the function class is

$$\mathfrak{R}(F) = E_X \mathfrak{R}_n(F).$$

Lemma 2: Let $|y| \leq A, \|x\| \leq R$ and $W = \{w | \|w\| \leq B\}$. Let F_W be the loss function class induced by the loss function $\ell(Z, w, \sigma)$ with $w \in W$ for Cauchy regression. Then

$$\mathfrak{R}(F_W) \leq \frac{(4A + 4BR)BR}{\sqrt{n}\sigma^2}.$$

Detailed proof is given in Section IV. The following dimensionality-independent bound is derived through Lemma 2.

Theorem 2: Let $Z = (x_i, y_i)_{i=1}^n$ be an i.i.d. sample. If $|y| \leq A, \|x\| \leq R$ and $\|w\| \leq B$, for any $\delta > 0$, with probability at least $1 - \delta$ we have for all w in W learned by CR that

$$|R(w, \sigma) - R_n(w, \sigma)| \leq \frac{(4A + 4BR)BR}{\sigma^2\sqrt{n}} + \sqrt{\frac{C \ln 1/\delta}{2n}},$$

where C is the same as in Theorem 1.

Detailed proof is given in Section IV.

Although the upper bound of Theorem 2 is of order $\mathcal{O}(\sqrt{1/n})$, it depends much on the value of A, B and R.

IV. The scale parameter

The scale parameter σ is an important parameter for CR. Figure 1 shows that the upper bounds of Theorems 1 and 2 are comparable because of different orders of the scale parameter σ. Although the bound obtained by the covering number is comparable to that obtained by the Rademacher complexity due to the existence of the scale parameter, these two bounds are complementary to each other.

For a small σ, Theorem 2 needs a large number of training samples to guarantee a satisfactory generalization error bound. We know that the function $f(\sigma^2) = \ln(1 + 1/\sigma^2)$ decreases faster than the function $f(\sigma^2) = 1/\sigma^2$ as σ^2 decreases, which implies when σ^2 decreases, the upper bounds of $R(w, \sigma)$ and $R_n(w, \sigma)$ decrease faster than that of the upper bound of Theorem 2. Therefore, to achieve satisfactory upper bound (when σ is small) for Theorem 2, the sample size n should be large. Fortunately, big learning has plenty of samples, which enable the successful use of CR for its applications.

V. Proof

In this section, we give the detailed proofs for the results in Section II.

A. Auxiliary Results

The following two concentration inequalities, well-known as the Hoeffding's inequality [17] and MicDiarmid's inequality [18], are widely used for deriving generalization error bounds.

Theorem 3: Let $X_1, \ldots, X_n$ be independent random variables with the range $[a_i, b_i]$ for $i = 1, \ldots, n$. Let $S_n = \sum_{i=1}^n X_i$. Then for any $\epsilon > 0$, the following inequalities hold:

$$\Pr\{S_n - ES_n \geq \epsilon\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$

$$\Pr\{ES_n - S_n \geq \epsilon\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

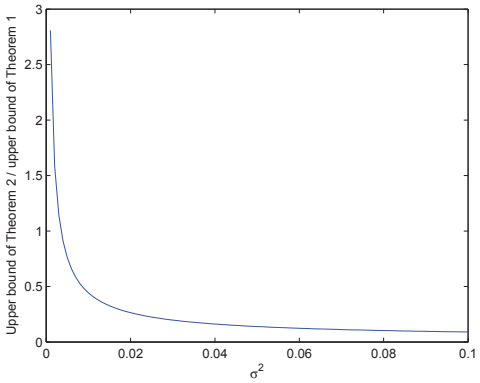

Fig. 1
Comparison of the upper bounds of Theorems 1 and 2. Setting $A = B = R = 1$, $d = 100$ and $\delta = 0.01$, the upper bound of Theorem 1 is equal to that of Theorem 2 around $\sigma^2 = 0.005$.

Theorem 4: Let $X = (x_1, \ldots, x_n)$ be a sample set of independent random variables and X^i a new sample set with the i-th sample in X being replaced by an independent random variable x_i'. If there exists $c_1, \ldots, c_n > 0$ such that $f : \mathcal{X}^n \to \mathbb{R}$ satisfies the following conditions:

$$|f(X) - f(X^i)| \leq c_i, \forall i \in \{1, \ldots, n\}.$$

Then for any $X \in \mathcal{X}^n$ and $\epsilon > 0$, the following inequalities hold:

$$\Pr\{f(X) - Ef(X) \geq \epsilon\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n (c_i)^2}\right),$$

$$\Pr\{Ef(X) - f(X) \geq \epsilon\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n (c_i)^2}\right).$$

The Glivenko-Cantelli theorem [19] is often used to analyze the non-asymptotic uniform convergence of the empirical risk to the expected risk. A relatively small complexity of the loss class is essential to prove a Glivenko-Cantelli class. Rademacher complexity and covering numbers are the most frequently used complexity measures when using concentration inequalities for deriving generalization error bounds.

The following lemma will be used throughout the proofs.

Lemma 3: Let $f(x) = \ln(1+x)$, we can easily verify that f is Lipschitz continuous with Lipschitz constant $L = 1$ on $[0, \infty)$.

B. Proof of Lemma 1 and Theorem 1

The following well-known theorem [19] obtained using MicDiarmid's inequality and the covering number plays a central role in proving Theorem 1.

Theorem 5: Let $X_1^{2n} = \{x_1, \ldots, x_{2n}\}$ be an i.i.d. sample. For a function class F with the range $[a, b]$, let $Ef = \int f(x)d\rho(x)$ and $E_m f = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$. For any $\xi > 0$ and any $n \geq \frac{8(b-a)^2}{\xi^2}$, we have

$$P\left\{\sup_{f\in F}|Ef - E_n f| \geq \xi\right\} \leq 8E\mathcal{N}_1(F, \xi/8, X_1^{2n}) \exp -\frac{n\xi^2}{32(b-a)^2}.$$

We are going to bound the covering number of the loss class induced by the loss function.

Proof of Lemma 1. We will bound the covering number of the loss function class F_W by bounding the covering number of the parameter class W. Cutting the subspace $[-B, B]^d \subset \mathbb{R}^d$ into small d-dimensional regular solids with width ξ, there are a total of

$$\left\lceil\frac{2B}{\xi}\right\rceil^d \leq \left(\frac{2B}{\xi} + 1\right)^d \leq \left(\frac{4B}{\xi}\right)^d$$

such regular solids. If we pick out the centers of these regular solids and use them to make up w, there are

$$\left\lceil\frac{2B}{\xi}\right\rceil^d \leq \left(\frac{4B}{\xi}\right)^d$$

choices, denoted by $\mathcal{S}$. $|\mathcal{S}|$ is the upper bound of the ξ-cover of the parameter class W.

We will prove that for every w, there exists a $w' \in W$ such that $|f_w - f_{w'}| \leq \xi'$, where $\xi' = \frac{\sqrt{d}\xi R(A+BR)}{\sigma^2}$.

$$\begin{aligned}
&|f_w - f_{w'}| \\
&= |\ell(x, w, \sigma) - \ell(x, w', \sigma))| \quad \text{(Using Lemma 3)} \\
&\leq \frac{1}{\sigma^2}|(y - w^T x)^2 - (y - w'^T x)^2| \\
&\leq \frac{1}{\sigma^2}|2y\langle w - w', x\rangle| + \frac{1}{\sigma^2}|\langle w + w', x\rangle\langle w - w', x\rangle| \\
&\leq \frac{1}{\sigma^2}\sqrt{d}\xi AR + \frac{1}{\sigma^2}\sqrt{d}B\xi R^2 \\
&= \frac{\sqrt{d}\xi R(A+BR)}{\sigma^2} = \xi'.
\end{aligned}$$

Let the metric d be the absolute difference metric. According to Definition 3, for $\forall f_w \in F_W$, there is a $w' \in W$ such that

$$\begin{aligned}
&\|d(f_w(Z, w, \sigma), f_{w'}(Z, w', \sigma))\|_1 \\
&= \left[\sum_{i=1}^{2n} d(f_w(Z_i), f_{w'}(Z_i))\right] \\
&\leq 2n\xi'.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathcal{N}_1(F_W, \xi', 2n) &\leq |\mathcal{S}| \leq \left(\frac{4B}{\xi}\right)^d \\
&= \left(\frac{4\sqrt{d}BR(1+BR)}{\xi'\sigma^2}\right)^d.
\end{aligned}$$

■

Now, we are ready to proof Theorem 1.

Proof of Theorem 1. According to Theorem 5, let

$$8\left(\frac{4\sqrt{d}BR(A+BR)}{\xi\sigma^2}\right)^d \exp -\frac{n\xi^2}{32C^2} = \delta,$$

where $C = \ln(1 + \frac{2A^2+2B^2R^2}{\sigma^2})$ (the upper bound of CR loss function).

We have

$$\xi = C\sqrt{\frac{32\left(d\ln\left(\frac{4\sqrt{d}BR(A+BR)}{\xi\sigma^2}\right) - \ln\frac{\delta}{8}\right)}{n}}.$$

We know that the upper bound of the generalization error makes sense only when $\xi < C$. Setting ξ on the left hand side equal to C, when $\xi \leq C$ we have

$$\xi \leq C = C\sqrt{\frac{32\left(d\ln\left(\frac{4\sqrt{d}BR(A+BR)}{C\sigma^2}\right) - \ln\frac{\delta}{8}\right)}{n}}.$$

Thus, with probability at least $1 - \delta$, we have

$$\begin{aligned}
&|R(w, \sigma) - R_n(w, \sigma)| \\
&\leq C\sqrt{\frac{32\left(d\ln\left(\frac{4\sqrt{d}BR(A+BR)}{C\sigma^2}\right) - \ln\frac{\delta}{8}\right)}{n}},
\end{aligned}$$

which concludes the proof of Theorem 1. ■

C. Proof of Lemma 2 and Theorem 2

The Rademacher complexity is useful to prove dimensionality-independent generalization error bounds. Thus, it can be used to derive generalization error bounds for kernel methods and has been widely used in the literature. The following theorem is proved using Hoeffding's inequality and Rademacehr complexity plays a central role in proving Theorem 2.

Theorem 6: Let F be a family of functions with the range $[a, b]$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, each of the following holds for all $f \in F$:

$$|Ef - E_n f| \leq \mathfrak{R}(F) + (b-a)\sqrt{\frac{\log 1/\delta}{2n}}$$

and

$$|Ef - E_n f| \leq \mathfrak{R}_n(F) + 3(b-a)\sqrt{\frac{\log 2/\delta}{2n}}.$$

The following property of Rademacher complexity [13] is very useful to upper bound Rademacher complexity.

Lemma 4: If $\phi : \mathbb{R} \to \mathbb{R}$ is Lipschitz with constant L and satisfies $\phi(0) = 0$, then

$$\mathfrak{R}_n(\phi \circ F) \leq 2L\mathfrak{R}_n(F).$$

We first give the proof of Lemma 2.

Proof of Lemma 2. We have

$$\begin{aligned}
\mathfrak{R}(F_W) &= E \sup_{w \in W} \frac{2}{n} \sum_{i=1}^{n} g_i \ell(Z_i, w, \sigma) \\
&\quad \text{(Using Lemmas 3 and 4)} \\
&\leq E \sup_{w \in W} \frac{2}{n} \sum_{i=1}^{n} g_i \left(\frac{y_i - w^T x_i}{\sigma} \right)^2 \\
&\quad \text{(Using Lemma 4 again)} \\
&\leq \frac{4A + 4BR}{n\sigma^2} E \sup_{w \in W} \left\langle \sum_{i=1}^{n} g_i x_i, w \right\rangle \\
&\leq \frac{4A + 4BR}{n\sigma^2} E \sup_{w \in W} \left\| \sum_{i=1}^{n} g_i x_i \right\| \|w\| \\
&\leq \frac{(4A + 4BR)B}{n\sigma^2} \sqrt{\sum_{i=1}^{n} E(g_i x_i)^2} \\
&\leq \frac{(4A + 4BR)BR}{\sqrt{n}\sigma^2}.
\end{aligned}$$

■

Theorem 2 can be easily proved by combining Theorem 6 and Lemma 2.

VI. Conclusion and future work

First, we proposed a new approach by interpreting the optimization of an objective function as a sample-weighted procedure, through which we thoroughly compared the robustness of different algorithms and proved that LAD is more robust than LS and that CR is more robust than LAD. Then, we explored the generalization error bounds for CR. By bounding the covering number and Rademacher complexity we derived comparable but complementary dimensionality-dependent and dimensionality-independent generalization error bounds, respectively. CR is suitable for robust analysis since it is tolerant to outliers. Although CR may require a large number of training examples to guarantee a satisfactory generalization error bound, big data with its characteristic of providing a sheer volume of data answers precisely to CR's need and thus provides a broad stage to fully exert CR's potential.

In our future work we will comprehensively investigate CR's performance on real-world big data problems by applying CR's loss to various big learning tasks including label propagation, multi-class classification, multi-label learning, semi-supervised learning, transfer and multi-task learning.

References

[1] *Nature*, ISSN: 0028-0836, EISSN: 1476-4687, 2009.

[2] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela H Byers. Big data: The next frontier for innovation, competition, and productivity. 2011.

[3] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[4] Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. Wikilinks: A large-scale cross-document coreference corpus labeled via links to wikipedia. Technical Report UM-CS-2012-015, University of Massachusetts, 2012.

[5] Evan Sparks, Ameet Talwalkar, Virginia Smith, Xinghao Pan, Joseph Gonzalez, Tim Kraska, Michael I Jordan, and Michael J Franklin. Mli: An api for distributed machine learning. In *International Conference on Data Mining (ICDM)*. IEEE, 2013.

[6] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[7] Naiyang Guan, Dacheng Tao, Zhigang Luo, and John Shawe-Taylor. Mahnmf: Manhattan non-negative matrix factorization. *arXiv preprint arXiv:1207.3438*, 2012.

[8] Peter J Huber. *Robust statistics*. Springer, 2011.

[9] Harry Richard Moore et al. Robust regression using maximum-likelihood weighting and assuming cauchy-distributed random error. Technical Report DTIC Document, Naval Postgraduate School, 1977.

[10] Ivan Mizera and Christine H Müller. Breakdown points of cauchy regression-scale estimators. *Statistics & probability letters*, 57(1):79–89, 2002.

[11] Vladimir Vapnik. *The nature of statistical learning theory*. Springer, 2000.

[12] Tong Zhang. Covering number bounds of certain regularized linear function classes. *The Journal of Machine Learning Research*, 2:527–550, 2002.

[13] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

[14] Andreas Christmann and Arnout Van Messem. Bouligand derivatives and robustness of support vector machines for regression. *The Journal of Machine Learning Research*, 9:915–936, 2008.

[15] Huan Xu and Shie Mannor. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

[16] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge university press, 2009.

[17] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[18] Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[19] Shahar Mendelson. A few notes on statistical learning theory. In *Advanced Lectures on Machine Learning*, pages 1–40. Springer, 2003.