Robust training of neural network via minimizing robust estimates of the average of loss functions

Z M Shibzukhov\textsuperscript{1,2}, M A Kazakov\textsuperscript{1} and D P Dimitrichenko\textsuperscript{1}

\textsuperscript{1} Institute of applied mathematics and automation KBSC RAS, Shortanova 89a, Nalchik, Russia
\textsuperscript{2} Institute of mathematics and informatics MPGU, Krasnoprudnaya 4, Moscow, Russia

E-mail: szport@gmail.com

Abstract. The paper suggests an extended version of principle of empirical risk minimization and principle of smoothly winsorized averages minimization for robust neural networks learning. It's based on using of M-averaging and WM-averaging functions instead of the arithmetic mean for empirical risk estimation. These approaches generalize robust algorithms based on using median and quantiles for estimation of mean losses. An iteratively reweighted schema for minimization of differentiable robust estimations of the averages of loss functions is proposed. This schema allows to use weighted version of traditional back-propagation algorithms for neural networks learning in presence of outliers.

1. Introduction

We are prepared to look into the problem on the approximation of functions of several variables using multilayer neural networks (NN), which are universal approximants \cite{1} capable of approximating any continuous functions of many variables, that is also confirmed when solving applied problems.

Applied problems data, as a rule, is distorted or erroneous. The first type of characteristics distortion is noise effectively removed by modern of approximation of continuous functions (including NN), based on minimizing the sum of squared errors between observed and predicted values. The second type of distortion is outliers - significant deviations, related both to large errors in the data, and to the fact that the data reflect a mixture of different processes. Usually such distortions can cover from 1\% to 10\% of the data. To correctly recover data functional dependencies, it is desirable to identify and remove the outliers out from the data used for approximation.

This approach is not always possible without first trying to approximate the desired functional dependency, which would provide outliers detection according to error or loss distribution.

Approximation methods for minimizing the arithmetic means or the sum of squared errors are not stable towards the outliers. The outliers could shift the desired parameters so badly that their detection by the error distribution becomes impossible. Cases are known in linear regression where a very small number of outliers can significantly distort coefficients \cite{2,3}.

Regularization methods like Ridge Regression or Lasso Regression potentially can help, if to move the outliers for shifting the desired parameters towards an increase in the norm. However, if, the desired parameters due to outliers are shifted towards a decrease in the norm, then the regularization is of no use.

To overcome this challenge, a number of robust methods for linear regression recovery have been developed that are resistant to a relatively large number of errors. Among them, the most popular are
LMedS (Least Median of Squares) and LTS (Least Trimmed Squares) [4], they provide resistant (stable) results in the presence of up to 50% of outliers. Such results are virtually impossible with M-estimator methods. [5]. This turned out to be possible because we can consider the median and the truncated sum as robust estimation for the mean and sum, respectively.

Nonlinear regression techniques have also been studied in the presence of outliers with the M-estimation approach [6]. NN training based on the M-estimator approach is considered in [7, 8, 9]. NN training based on LTS is considered in [10, 11]. LMedS based NN training is considered in [12, 13]. An interesting PCLTS method is proposed in [14].

LMedS and LMedA methods are individual cases for solving a regression problem via minimizing the functional

$$Q(\mathbf{w}) = \text{med}\{g(r_1(\mathbf{w})), ..., g(r_N(\mathbf{w}))\},$$

where $g(r)$ is the nonnegative quasi-convex function, $r_k(\mathbf{w}) = f(x_k, \mathbf{w}) - y_k$, $\{x_1, ..., x_N\} \subset \mathbb{R}$ and $\{y_1, ..., y_N\} \subset \mathbb{R}$ are given input vectors and expected output, $f(\mathbf{x}, \mathbf{w})$ is the conversion function of the parameterized model for the dependency to be restored.

LTS and LTA estimators are individual cases for solving a regression problem based on minimizing the functional

$$Q(\mathbf{w}) = \sum_{k=1}^{N-p} g(r_{(k)}(\mathbf{w}))$$

where $r_{(1)}$, ..., $r_{(N)}$ is the sequence $r_1$, ..., $r_N$ in ascending order, $p > N/2$. For LMedS and LTS estimators $g(r) = r^2$, and for LMedA and LTA $g(r) = |r|$

Robust linear and non-linear regression recovery with a very large number of outliers in the data (up to 40-50%) is an important task in cases where the outlier detection is very complicated. However, if we could manage to remove outliers in the source data, the amount of data remained would be enough to restore the dependency. In such cases, robust regression techniques might prove effective. They allow us at least to have an approximation, by errors distribution that contribute to outliers detection, and at most, to restore the desired dependencies, despite the outliers.

This paper proposes a generalized approach that covers LMedS and LMedA, LTS and LTA estimations. It is based on minimizing robust and differentiable M-means and W-sums. Thus, we can employ gradient minimization and methods similar to IRLS (Iteratively Reweighted Least Squares) [15] to search for optimal values for dependence parameters in linear regression model.

The main idea is to employ robust, but smooth estimating functions instead of robust and non-smooth versions of the latter. Minimizing robust mean or sum estimates is more preferable, since, it is the instability of mean (as a rule, the arithmetic average) or the sum estimation methods in relation with outliers is the main reason why the required parameters shift in the problem of minimizing the approximate functional dependencies employing empirical risk minimization principle.

2. Materials and methods

2.1. Classical empirical risk minimization principle

Regression and classification problems are often formulated as empirical risk minimization problem:

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} Q(\mathbf{w})$$

$$Q(\mathbf{w}) = \frac{1}{N} \sum_{k=1}^{N} g(r_k(\mathbf{w})).$$
The problem of outliers arises when the empirical distribution of losses \( z_1 = \varrho(r_1(w)), \ldots, z_N = \varrho(r_N(w)) \) contains outliers. They cause distortion of the values of the desired parameters. First robust approach to suppress outliers is based on choice of function \( \varrho(r) \), which grows slower than \( r^2 \), i.e. \( |\varrho'(r)| \leq 1 \). However, this approach does not always lead to success in order to explain let’s consider a sum:

\[
S\{r_1, \ldots, r_N\} = \varrho(r_1) + \cdots + \varrho(r_N)
\]

The following equality

\[
|S\{r_1, \ldots, r_N + \Delta\} - S\{r_1, \ldots, r_N\}| = \varrho'(\tilde{r})\Delta,
\]

where \( \tilde{r} \in [z_N, z_N + \Delta] \), can explain: when \( \Delta \) is large or there are large number of outliers.

The alternative is using functions of summing or averaging, which are resistant to outliers. Well-known examples are median and quantile. But they are not continuously differentiable. Therefore, gradient based approaches to training are not possible. However differentiable averaging functions, which are resistant to outliers can be defined.

2.2. Differentiable M-averages minimizing

Thus, for estimating the average we propose to use the M-averages, which are defined as follows.

\[
M_{\varrho}\{z_1, \ldots, z_N\} = \arg \min_u \sum_{k=1}^{N} \rho(z_k - u)
\]

where \( \rho(r) \) is the strictly convex and twice continuously differentiable function.

Here are some examples of well-known M-averages:

- Arithmetic mean: \( \rho(r) = r^2 \);
- Median: \( \rho(r) = |r| \);
- Quantile:

\[
\rho(r) = \begin{cases} 
\alpha r, & \text{if } r \geq 0 \\
((\alpha - 1)r, & \text{if } r < 0,
\end{cases}
\]

where \( 0 < \alpha < 1 \).

In the case when \( \rho'(r) \) and \( \rho''(r) \) are exist the M-average \( M_{\varrho} \) is differentiable:

\[
\frac{\partial M_{\varrho}}{\partial z_k} = \frac{\rho''(z_k - \bar{z}_\varrho)}{\rho''(z_1 - \bar{z}_\varrho) + \rho''(z_N - \bar{z}_\varrho)}
\]

where \( \bar{z}_\varrho = M_{\varrho}\{z_1, \ldots, z_N\} \). Wherein \( \sum_{k=1}^{N} \frac{\partial M_{\varrho}}{\partial z_k} = 1 \).

The average value can be found as a solution of nonlinear equation:

\[
\sum_{k=1}^{N} \rho'(z_k - u) = 0.
\]

In cases when

\[
\sum_{k=1}^{N} \varphi'(z_k - \bar{z})(z_k - \bar{z}) < \sum_{k=1}^{N} \varphi(z_k - \bar{z}),
\]

where \( \varphi(r) = \varrho'(r)/r \), the averaging value can be numerically calculated with the help of the iterative procedure:
\[ z_{t+1} = \frac{\sum_{k=1}^{N} \phi(z_k - \bar{z}_t)z_k}{\sum_{k=1}^{N} \phi(z_k - \bar{z}_t)}. \]

All considered below functions \( \rho \) for definition of resistant M-averages have satisfied this condition. Therefore, such iterative procedure is more preferable.

So now we able to formulate the problem of searching the optimal parameter values of \( \mathbf{w}^* \) as minimization of the following function

\[ Q(\mathbf{w}) = M_{\theta}\{\varrho(r_1(\mathbf{w})), ..., \varrho(r_N(\mathbf{w}))\}, \]

which can be solved numerically using gradient descent and variants. Gradient of \( Q(\mathbf{w}) \) has the following form:

\[ \nabla Q(\mathbf{w}) = \sum_{k=1}^{N} v_k(\mathbf{w}) \nabla \ell_k(\mathbf{w}), \]

where \( \ell_k(\mathbf{w}) = \varrho(r_k(\mathbf{w})) \) and

\[ v_k(\mathbf{w}) = \frac{\partial M_{\theta}\{\ell_1(\mathbf{w}), ..., \ell_N(\mathbf{w})\}}{\partial z_k} \bigg|_{z_k = \ell_k(\mathbf{w})}. \]

It’s follows from the properties of \( \frac{\partial M_{\theta}}{\partial z_k} \) that

\[ \sum_{k=1}^{N} v_k(\mathbf{w}) = 1. \]

Solution \( \mathbf{w}^* \) and \( \bar{z}^* \) of the minimization problem are solutions of system:

\[ \begin{cases} \sum_{k=1}^{N} \rho'(\ell_k(\mathbf{w}) - \bar{z}) = 0 \\ \sum_{k=1}^{N} v_k(\mathbf{w}) \nabla \ell_k(\mathbf{w}) = 0. \end{cases} \]

In order to minimize the overhead of calculation of averaging value an iteratively re-weighted schema is applied. The averaging value is calculated iteratively:

\[ \mathbf{w}_{t+1} \leftarrow \arg \min_{\mathbf{w}} \sum_{k=1}^{N} v_k \ell_k(\mathbf{w}), \]

where

\[ v_k = \frac{\rho''(z_k - \bar{z}_t)}{\rho''(z_1 - \bar{z}_t) + \cdots + \rho''(z_N - \bar{z}_t)}. \]

So, to minimize the functional, we can apply the IR-ERM learning algorithm (Iteratively Reweighted Empirical Risk Minimization) \cite{16} which in pseudocode language can be expressed as:

```plaintext
procedure IR-ERM(\mathbf{w}_0)
    \text{t} \leftarrow 0
    \text{repeat}
        \text{z}_1 = \ell_1(\mathbf{w}_t), ..., \text{z}_N = \ell_N(\mathbf{w}_t)
        \text{z}_t \leftarrow \text{M}[\text{z}_1, ..., \text{z}_N]
        \text{for} k = 1, ..., N \text{ do}
            \text{v}_k = \frac{\rho''(z_k - \text{z}_t)}{\rho''(z_1 - \text{z}_t) + \cdots + \rho''(z_N - \text{z}_t)}
            \text{w}_{t+1} \leftarrow \arg \min_{\mathbf{w}} \sum_{k=1}^{N} v_k \ell_k(\mathbf{w}),
        \text{end for}
    \text{end repeat}
```
Consider a smooth summing Winsorized Absolutes resist where

\[ v_k = \frac{\rho''(z_k - \bar{z})}{\rho''(z_1 - \bar{z}) + \cdots + \rho''(z_N - \bar{z})} \]

end

\[ w_{t+1} \leftarrow \arg \min_w \sum_{k=1}^{N} v_k \ell_k(w) \]

\[ t \leftarrow t + 1 \]

until \{\bar{z}_t\} and \{w_t\} stabilize

end

2.3. Robust differentiable M-means

We give examples of differentiable averages, to be used as an adequate replacement for the median:

- \( \rho_{\text{sqrt},\varepsilon}(r) = \sqrt{e^2 + r^2} - \varepsilon; \)
- \( \rho_{\text{ln},\varepsilon}(r) = |r| - \varepsilon \ln(|r|) - \varepsilon \ln \varepsilon; \)
- \( \rho_{\text{atan},\varepsilon}(r) = r \cdot \arctan \left( \frac{r}{\varepsilon} \right) - \frac{1}{2} \varepsilon \ln \left( 1 + \left( \frac{r}{\varepsilon} \right)^2 \right). \)

In order to define differentiable replacements for quantiles let’s define

\[ \rho_\alpha(r) = \begin{cases} \alpha \rho(r), & \text{if } r \geq 0 \\ (1 - \alpha) \rho(r), & \text{if } r < 0, \end{cases} \]

where \( \rho(r) \) is a function for definition of replacement for median.

The key to understanding the empirical mean resistance to outliers can be the following inequality:

\[ |M_\rho(z_1, \ldots, z_N + \Delta) - M_\rho(z_1, \ldots, z_N)| < \rho''(\bar{z} - u_2) \Delta, \]

where \( \rho(r) \) is the convex function, \( \rho''(r) \) is the continuous function, \( \Delta > 0 \) is the distortion value, \( \bar{z} = [z_N, z_N + \Delta], u_2 = M_\rho(z_1, \ldots, \bar{z}). \)

We give the following estimates for \( \rho''(r) \Delta: \)

1. \( \rho'_{\text{sqrt},\varepsilon}(r) = \frac{e^2}{(e^2 + r^2)^{3/2}} \cdot \rho'_{\text{sqrt},\varepsilon}(r) \Delta < \frac{e^2 \Delta}{r^3}; \)
2. \( \rho'_{\text{ln},\varepsilon}(r) = \frac{e}{(e + |r|)^{3/2}} \cdot \rho'_{\text{ln},\varepsilon}(r) \Delta < \frac{e^2 \Delta}{r^3}; \)
3. \( \rho'_{\text{atan},\varepsilon}(r) = \frac{e}{e^2 + r^2} \cdot \rho'_{\text{atan},\varepsilon}(r) \Delta < \frac{e \Delta}{r^2}. \)

It is easy to see that for the sufficiently small \( \varepsilon \) the M-averages with \( \rho_{\text{sqrt},\varepsilon}, \rho_{\text{ln},\varepsilon} \) and \( \rho_{\text{atan},\varepsilon} \) can be resistant to relatively high number of outliers. For example, with \( \rho_{\text{ln},\varepsilon} \) if \( \Delta < r^2/e \) then \( \rho_{\text{ln},\varepsilon}(r) \Delta < \varepsilon \).

So any value \( z_k \) such that \( |z_k - \bar{z}| > r \) could be replaced with value from \([z_k, z_k + \Delta]\) so that

\[ |M_\rho(z_1, \ldots, z_k + \Delta, \ldots, z_N) - M_\rho(z_1, \ldots, z_N)| < \varepsilon. \]

2.4. Smooth WM-averages minimizing

Along with the LTS and LTA, we can also employ LWS (Least Winsorized Squares) and LWA (Least Winsorized Absolutes) estimations, which are based on Winsorized Sum (WS) – the robust method of summing:

\[ WS_u \{z_1, \ldots, z_N\} = \sum_{k=1}^{N} \min\{z_k, u\}. \]

Consider a smooth version of winsorized sum. To do this, define the smoothed version:
\[ w_p(z, u) = \frac{1}{2} \left( z + u - \rho(z - u) \right). \]

Set \( w_p(z) = w_p(z, \tilde{z}_p) \), where \( \tilde{z}_p = M_p(z_1, ..., z_N) \) and define the averaging function

\[ WM_p\{z_1, ..., z_N\} = \frac{1}{N} \sum_{k=1}^{N} w_p(z_k). \]

Calculate partial derivatives as follows:

\[ \frac{\partial WM_p}{\partial z_k} = \frac{1}{2N} \left( 1 - \rho'(z_k - \tilde{z}_p) \right) + \frac{1}{2} \frac{\partial M_p}{\partial z_k}. \]

Training is carried out using well known iteratively re-weighing schema that implies solving a problem step-by-step:

\[ w_{t+1} \leftarrow \arg \min_w \sum_{k=1}^{N} v_k \ell_k(w), \]

Where

\[ v_k = \frac{1}{2N} \left( 1 - \rho'\left( \ell_k(w_t) - \tilde{z}_t \right) \right) + \frac{1}{2} \frac{\partial M_p}{\partial z_k} \left( \ell_1(w_t), ..., \ell_N(w_t) \right), \]

and \( \tilde{z}_t = M_p\{\ell_1(w_t), ..., \ell_N(w_t)\} \).

In order to generalize this method let’s introduce a differentiable function \( \sigma(z, u) \), which satisfies the following conditions:

- \( \lim_{z \to \infty} \sigma(z, u) = u; \)
- \( \lim_{z \to -\infty} \sigma(z, u) / z = 1. \)

This is monotone function, which has two asymptotes: horizontal (for \( z \to \infty \)) and with slope 45° (for \( z \to -\infty \)).

So we can formulate the following generalized form of winsorized mean:

\[ WM_p\{z_1, ..., z_N\} = \frac{1}{N} \sum_{k=1}^{N} \sigma(z_k, \tilde{z}_p). \]

Partial derivatives of \( WM_p \) can be calculated as follows:

\[ \frac{\partial WM_p}{\partial z_k} = \frac{1}{N} \sigma'_k(z_k, \tilde{z}) + \frac{1}{N} \frac{\partial M_p}{\partial z_k} \sum_{m=1}^{N} \sigma'_m(z_m, \tilde{z}_p). \]

In particular, for \( \sigma(z, u) = \min(z, u) \) it can be formulated the following form of the partial derivatives:

\[ \frac{\partial WM_p}{\partial z_k} = \begin{cases} \frac{1}{N} \#\{z_k > \tilde{z}_p\} \frac{\partial M_p}{\partial z_k} \quad & \text{if } z_k \leq \tilde{z}_p, \\ \frac{1}{N} \#\{z_k > \tilde{z}_p\} \frac{\partial M_p}{\partial z_k} \quad & \text{if } z_k > \tilde{z}_p, \end{cases} \]

where \( \#\{z_k > \tilde{z}_p\} \) is the number of \( z_k \), which are greater than \( \tilde{z}_p \). It can be easily shown that \( \sum_{k=1}^{N} \frac{\partial WM_p}{\partial z_k} = 1 \). So that this method of averaging has same property as the \( M_p \).
Learning algorithm IR-WERM (Iteratively Reweighted Winsorized Empirical Risk Minimization) in pseudocode interpretation transforms into:

```
procedure IR-WERM(\( \mathbf{w}_0 \))
  \( t \to 0 \)
  repeat
    \( z_1 = \ell_1(\mathbf{w}_t), ..., z_N = \ell_N(\mathbf{w}_t) \)
    \( \bar{z}_t = M_\rho[\ell_1(\mathbf{w}_t), ..., \ell_N(\mathbf{w}_t)] \)
    for \( k = 1, ..., N \) do
      \( v_k = \frac{1}{N} \sigma'_z(z_k, \bar{z}_t) + \frac{1}{N} \frac{\partial M_\rho}{\partial z_k} \sum_{m=1}^{N} \sigma'_u(z_m, \bar{z}_t) \)
    end
    \( \mathbf{w}_{t+1} \leftarrow \arg \min_{\mathbf{w}} \sum_{k=1}^{N} v_k \ell_k(\mathbf{w}) \),
    \( t \leftarrow t + 1 \)
  until \( \{\bar{z}_t\} \) and \( \{\mathbf{w}_t\} \) stabilize
end
```

We can further generalize this approach by the way of using of arbitrary M-average instead of the arithmetical average.

Let’s \( M_\rho \) is M-averaging function for replacement of the arithmetical averaging. Then let’s define averaging function:

\[
WM_{\rho M_\rho}(z_1, ..., z_N) = M_\rho[\sigma(z_1, \bar{z}_p), ..., \sigma(z_N, \bar{z}_p)].
\]

Partial derivatives can be calculated according to following formula:

\[
\frac{\partial WM_{\rho M_\rho}}{\partial z_k} = \frac{\partial M_{\rho \sigma}}{\partial v_k} \sigma'_z(z_k, \bar{z}_p) + \frac{\partial M_\rho}{\partial z_k} S,
\]

where \( v_k = \sigma(z_k, \bar{z}_p) \) and

\[
S = \sum_{m=1}^{N} \frac{\partial M_{\rho \sigma}}{\partial v_m} \sigma'_u(z_m, \bar{z}_p).
\]

3. Results and discussion

Let’s consider results of numerical examples of training different parametrized models on the base of training data that contains outliers. We use presented above algorithms IR-ERM and IR-WERM.

First, let’s consider two simple examples of linear regression and linear separation of 2 classes with application of IR-ERM. They demonstrate ability of the proposed algorithm IR-ERM for training on the base of data, which contain large number of big outliers. These examples also demonstrate the inability of well-known classic robust training algorithms to overcome influence of the large number of big outliers.

3.1. Linear regression example

There is artificially generated example that demonstrate the robust capabilities of the proposed approach. The exact linear relationship is \( y = 3x \). The distortions are caused by 80% of big outliers and 100% of regular data. To design a robust technique for solving linear regression problem it used \( M_{\rho_{\alpha,\varepsilon}} \) with \( \rho_\varepsilon(x) = \sqrt{x^2 + \varepsilon^2} - \varepsilon, \varepsilon \approx 10^{-2} \div 10^{-3} \), where \( \rho_{\alpha,\varepsilon} \) is dissimilarity function introduced above.
3.2. Linear separation of 2 classes

Let’s consider another artificially generated example of samples of two classes of points on 2d plain. It contains 80% of outliers. SVC (Support Vector Classification) algorithm fail to linearly separate these 2 classes. IR-ERM algorithm successfully separate classes. Figure 2 demonstrates the resistant property of the proposed method.

Below we present here several examples on the use of the IR-WERM algorithm for NN training with one hidden layer using training datasets with outliers:

\[ y = w_0 + w_1 u_1 + \cdots + w_m u_m \]
\[ u_j = \text{softplus}(w_0 + w_{j1} x_1 + \cdots + w_{jn} x_n), \]

where softplus(s) = ln(1 + e^s), j = 1, ..., m.
3.3. Boston dataset example

Consider the boston housing dataset. The neural network contains the hidden layer with 7 neurons inside. This example is used to demonstrate the ability of robust training to reduce the absolute value of a large number of errors.

First, the neural network learns to reduce the sum of squared errors. Next, NN is trained by minimizing the sum of the absolute values of the errors. And finally, NN trained using IR-WERM with \( \rho_{\alpha,\varepsilon} \), where \( \rho_{\varepsilon}(r) = \sqrt{r^2 + \varepsilon^2} - \varepsilon \), \( \alpha = 0.9 \), \( \varepsilon = 0.0001 \). Fig. 3 shows the distribution of target pairs: the predicted value (on the Y axis) versus the original one (on the X axis).

Figure 3 demonstrates the ability of IR-WERM training to reduce absolute errors on the training data by 80% compared to minimizing techniques for the sums of squares and absolute values of the error.

Two artificial datasets have further been employed, according to [15]. They represent some function \( f(\mathbf{x}) \), where \( \mathbf{x} \in \mathbb{R}^3 \). Training and test sets are generated evenly and randomly in the cube \([-2,2]^3\], the number of points in the training and testing sets is 1000 and 1000, respectively. In addition, uniformly distributed noise with \( \varepsilon \in [-0.1,0.1] \) is added. Outliers are supplemented to the training set according to the following model. Outliers points chosen uniformly and randomly make up 30% and 50% of the training data, respectively.

All these examples clearly illustrate that minimizing smoothed winsorized sum of absolute errors on data provides error distribution with number of absolute values at points that do not exceed a given value, larger than in the case where the minimizing of the sum of squared errors is applied. The above also shows that minimizing the sum of absolute errors prevents a successful result being achieved.

![Figure 3](image-url)

**Figure 3.** Error distribution for the trained NN on the Boston housing dataset with three approaches: Least Squares, Least Absolute Errors and Least Winsorized Absolute Errors.

3.4. Example 1

In this example, the function \( f(\mathbf{x}) = \|\mathbf{x}\|^{2/3} \), where \( \mathbf{x} \in \mathbb{R}^3 \).

For all outlier data points in the training sets, each target value is replaced by 20 (note that \( \max f(\mathbf{x}) < 3 \) in \([-2,2]^3\)). Results are presented in figure 4.

3.5. Example 2

In this example, the function \( f(\mathbf{x}) = \sin\|\mathbf{x}\|/(\|\mathbf{x}\| + 10^{-8}) \), where \( \mathbf{x} \in \mathbb{R}^3 \).

For all outlier data points in the training set, each target value is also replaced by 20 (note that \( \max f(\mathbf{x}) < 1 \) in \([-2,2]^3\)). Results are presented in figure 5.
Figure 4. Distribution of errors of the trained NN in the data set, Example 1, based on three approaches: Least Squares, Least Absolute Errors and Least Winsorized Absolute Errors.

Figure 5. Distribution of errors of the trained NN in the data set, Example 2, with approaches: Least Squares, Least Absolute Errors and Least Winsorized Absolute Errors.

Both algorithms IR-ERM and IR-WERM are implemented with the help of the open sourced library mlgrad\(^1\). The entire numerical calculation has been performed using mlgrad.

Conclusion

The paper considers a new approach to improving the robustness of neural networks training in the cases when the training data contains a significant number of outliers. It’s based on minimization of differentiable analogues of median, quantiles and winsorized means of loss functions. The presented above techniques are more desirable in cases where gradient-based minimization algorithms are preferable. For example, these methods allow the use of weighted versions of back-propagation

\(^1\) http://bitbucket.org/intellimath/mlgrad
algorithms for robust training of NN. In particular, an iterative re-weighted least squares procedures can be used. In these procedures, a weighted version of the back-propagation algorithm is used for each step. The examples above have clearly shown the proposed approaches and algorithms can be resistant to a large number of outliers.

Acknowledgments
This work have been supported by the RFBR grant №18-01-00050.

References
[1] Hornik K, Stinchcombe M and White H 1989 Multilayer feedforward networks are universal approximators. Neural Networks 2 pp 359-366.
[2] Maronna R, Martin R and Yohai V 2006 Robust Statistics: Theory and Methods (Wiley, New York)
[3] Rousseeuw P J and Leroy A M 2003 Robust Regression and Outlier Detection (Wiley, New York)
[4] Rousseeuw P J 1984 Least median of squares regression. Journal American Statistics Association 79 pp 871-880
[5] Huber P J 1981 Robust Statistics (John Wiley and Sons, New York)
[6] Stromberg A J and Ruppert D 1992 Breakdown in nonlinear regression. Journal of American Statistical Association 87 pp 991-997
[7] Chen D S and Jain R C 1994 A robust back-propagation learning algorithm for function approximation IEEE Trans. Neural Networks 5 pp 467-479
[8] Liano K 1996 Robust error measure for supervised neural network learning with outliers IEEE Trans. on Neural Networks 7 pp 246-250
[9] El-Melegy M T, Essai M H and Ali A A 2009 Robust Training of Artificial Feedforward Neural Networks. Foundations of Computational, Intelligence 1. Studies in Computational Intelligence 201. ed Hassanien A E, Abraham A, Vasilakos A V and Pedrycz W (Springer, Berlin, Heidelberg)
[10] Rusiecki A 2007 Robust LTS backpropagation learning algorithm. In F Sandoval, A Prieto J Cabestany and M Graa, editors. Computational and Ambient Intelligence, Lecture Notes in Computer Science 4507 pp 102-109 (Springer Berlin / Heidelberg)
[11] Jeng J-T, Chuang C-T. and Chuang C-C 2011 Least trimmed squares based CPBUM neural networks. International Conference on System Science and Engineering pp 187-192 (IEEE)
[12] Rusiecki A 2012 Robust Learning Algorithm Based on Iterative Least Median of Squares Article in Neural Processing Letters 36 pp 145-160
[13] Lina Y-L, Hsieh J-G, Jeng J-H and Cheng W-C 2015 On least trimmed squares neural networks Neurocomputing 61 pp 107-112
[14] Beliakov G and Kelarev A and Yearwood J 2012 Robust artificial neural networks and outlier detection https://arxiv.org/abs/1110.0169
[15] Holland P W and Welsch R E 1977 Robust regression using iteratively reweighted least-squares Communications in Statistics Theory and Methods 6 pp 813-827
[16] Shibzukhov Z M 2017 On the principle of empirical risk minimization based on averaging aggregation functions Doklady Mathematics 96(2) pp 494-497