Minimisation of robust estimates of the sums of parametrised functions

Z M Shibzukhov[1,2], M A Kazakov[1] and D P Dimitrichenko[1]

[1]Institute of Applied Mathematics and Automation KBSC RAS, Shortanova str. 89A, Nalchik, Russia, 360000
[2]Moscow Pedagogical State University, Malaya Pirogovskaya str. 1, Moscow, Russia, 119991

Abstract. A robust approach to the design of machine learning algorithms, based on minimising finite sums of the parametrised functions is considered. This method implies using robust finite-sum differentiable aggregating functions that are resistant to outliers.

1. Introduction

The majority of machine learning problems can be reduced to problem of minimising finite sums of parametrised functions:

$$Q(\mathrm{w}) = \sum_{k=1}^{N} v_k \, \ell_k(\mathrm{w}),$$

where $\ell_1(\mathbf{w}), \ldots, \ell_N(\mathbf{w})$ – are non-negative basis functions, $\mathbf{w}$ – is the vector of unknown parameters, $\mathbf{w} \in \mathbf{W} \subseteq \mathbb{R}^m$, $v_1, \ldots, v_N \geq 0$ – is nonnegative weights. Most often $v_k = \mathrm{const}$, for example, 1 (arithmetic sum) or $1/N$ (arithmetic mean). Q target function is minimised by the optimal set of parameters $\mathbf{w}^*$:

$$Q(\mathbf{w}^*) = \min_{\mathbf{w} \in \mathbf{W}} Q(\mathbf{w}).$$

Most of algorithms for neural networks (NN) learning are based on this principle. In particular back propagation (BP) algorithm is based on minimisation of arithmetic mean squared errors.

However, if the distribution of the basic functions values contains outliers, the minimization of $Q(\mathbf{w})$, as a rule, leads to a distortion of $\mathbf{w}^*$. This is due to the fact that the arithmetic sum and the arithmetic mean are not resistant to the outliers.

Of course, the problem of outliers could be solved by choosing the values of the weights $v_1, \ldots, v_N$, which, on the one hand, would suppress the values of outliers, and on the other hand, leave the rest left unchanged. However, the selection of such weights is difficult task and is essentially equivalent by the complexity to identifying the outliers in the empirical distribution $\{\ell_1(\mathbf{w}^*), \ldots, \ell_N(\mathbf{w}^*)\}$.

One of the effective way of dealing with this problem is to use robust aggregation functions to calculate the sum or average. Thus, we get definitions for the function Q:

$$Q(\mathbf{w}) = \mathrm{med}_{k=1,\ldots N} \ell_k(\mathbf{w})$$

for robust estimation of a mean

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

and for robust estimation of a sum. Here $z_{(1)}, \ldots, z_{(N)}$ is the sequence of numbers obtained by arranging the initial sequence $z_1, \ldots, z_N$ in ascending order. For example, to build a robust regression with $\ell_k(\mathbf{w}) = (f(\mathbf{x}_k, \mathbf{w}) - y_k)^2$ there have been proposed LMedS and LTS (Least Trimmed Squares) [1,2].

Last sum could be rewritten as trimmed arithmetical mean:

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

Minimizing the above estimates on data with outliers (up to 50%) allows finding adequate estimates for $\mathbf{w}^*$. However, minimization algorithms for LTS and LMS include a combinatorial component form of $\mathbf{w}^*$ search through subsets, since their gradients are singular. It makes application of gradient based algorithms almost impossible. This also reduces the scalability of such algorithms and their application in training neural networks and in problems with big data.

Another way of robust estimation of $\mathbf{w}^*$ is using a winsorized sum

$$Q(\mathbf{w}) = \sum_{k=1}^{N} \max\{\ell_k(\mathbf{w}), \overline{\ell}(\mathbf{w})\}$$

or an average

$$Q(\mathbf{w}) = \frac{1}{N} \sum_{k=1}^{N} \max\{\ell_k(\mathbf{w}), \overline{\ell}(\mathbf{w})\},$$

where $\overline{\ell}(\mathbf{w})$ is threshold value for the empirical distribution $\{\ell_1(\mathbf{w}), \ldots, \ell_N(\mathbf{w})\}$.

In this paper, we consider general approach where for estimation of average empirical losses it will be used M-averaging aggregation functions (M-averages). This approach generalises M-regression method [5] and provides universal technique for solving the problem of the empirical risk minimisation in presence of outliers. It allows to use differentiable M-averages that could be treated as a sort of approximations of median and quantiles. In such cases a general gradient based procedure could be constructed for NN robust training.

2. Minimisation of M-averages from parametrised functions

For the median case, the problem can be solved using M-averages [5,7-10], which are differentiable and, in a sense, are approximate median:

$$\mathsf{M}_\rho\{z_1, \ldots, z_N\} = \operatorname*{argmin}_u \sum_{k=1}^{N} \rho(z_k - u),$$

where ρ – is the nonnegative strictly convex function, $\rho(0) = 0$.

Here are some examples of M-averages:

- Collection of symmetrical averages:

$$\mathsf{M}^\gamma\{z_1, \ldots, z_N\} = \operatorname*{argmin}_u \sum_{k=1}^{N} |z_k - u|^{1+\gamma},$$

- where $0 \le \gamma \le 1$ (M^0 is median, M^1 is arithmetical mean).
- Collection of non-symmetrical averages:

$$\mathsf{M}^\gamma_\alpha\{z_1, \ldots, z_N\} = \operatorname*{argmin}_u \sum_{k=1}^{N} |z_k - u|^{1+\gamma}_\alpha,$$

- where $|u|^{1+\gamma}_\alpha = (\alpha - [u > 0])u|u|^\gamma$, $0 \le \gamma \le 1$ (M^0_α is α-quantile, M^1_α is α-expectile).

Here is the sufficient condition: if ρ – is twice differentiable, then $\mathsf{M}_\rho\{z_1, \ldots, z_N\}$ has all partial derivatives:

$$\frac{\partial \mathsf{M}_\rho}{\partial z_k} = \frac{\rho''(z_k - \overline{z})}{\rho''(z_1 - \overline{z}) + \cdots + \rho''(z_N - \overline{z})}.$$

Besides, $\frac{\partial \mathsf{M}_\rho}{\partial z_k} \geq 0$ and

$$\sum_{k=1}^{N} \frac{\partial \mathsf{M}_\rho}{\partial z_k} = 1.$$

In order to find out in which cases the function M-average M_ρ can be stable with respect to outliers we consider the following inequality:

$$|\mathsf{M}_\rho\{z_1,\ldots,z_N+\Delta\} - \mathsf{M}_\rho\{z_1,\ldots,z_N\}| = \frac{\rho''(\tilde{z}-u_{\tilde{z}})\Delta}{\sum_{k=1}^{N-1}\rho''(z_k-u_{\tilde{z}})+\rho''(\tilde{z}-u_{\tilde{z}})} < \rho''(\tilde{z}-u_{\tilde{z}})\Delta,$$

where $\rho(r)$ is convex, $\rho''(r)$ is continouous function, $\Delta > 0$ is value of distortion, $\tilde{z} \in [z_N, z_N + \Delta], u_{\tilde{z}} = \mathsf{M}_\rho\{z_1,\ldots,z_{N-1},\tilde{z}\}$. Let M_ρ be some M-averaging function. We define empirical risk based on M-averaging function M_ρ, as follows:

$$\mathcal{Q}_\rho(\mathbf{w}) = \mathsf{M}_\rho\{\ell_1(\mathbf{w}),\ldots,\ell_N(\mathbf{w})\}.$$

The classical empirical risk is a special case when M_ρ is arithmetical mean. The best set of the parameters for $\mathbf{w}^*$ have to minimize the function with respect to the minimization principle:

$$\mathcal{Q}_\rho(\mathbf{w}^*) = \min_{\mathbf{w}} \mathsf{M}_\rho\{\ell_1(\mathbf{w}),\ldots,\ell_N(\mathbf{w})\}.$$

Since the median and quantile are not continuously differentiable, the gradient procedures for minimisation of the risk functional are not practical. However, instead of median we can use continuously differentiable parametric family of M-average functions based on the dissimilarity function $\rho_\varepsilon(z-u)$ that satisfy the following requirements:

1. $\lim_{\varepsilon\to 0}\rho_\varepsilon(z-u) = |z-u|$;
2. $\lim_{\varepsilon\to 0}\rho'_\varepsilon(z-u) = \operatorname{sign}(z-u)$;
3. $\lim_{\varepsilon\to 0}\rho''_\varepsilon(z-u) = \delta(z-u)$ (Dirac's δ-function).

We demonstrate, for example, that for the role of "approximate" median the following functions can be used:

- $\rho_\varepsilon(r) = \sqrt{\varepsilon^2 + r^2} - \varepsilon$;
- $\rho_\varepsilon(r) = |r| - \varepsilon\ln(\varepsilon + |r|) - \varepsilon\ln\varepsilon$.

Such M-averages M_{ρ_α} are continuously differentiable and robust with sufficiently small ε. This implies that they are resistant to outliers (in some cases up to 50%).
To approximate the α-quantile, one can use the function

$$\rho_{\varepsilon,\alpha}(r) = \begin{cases} (1-\alpha)\rho_\varepsilon(r), & \text{if } r < 0 \\ \alpha\rho_\varepsilon(+0) + (1-\alpha)\rho_\varepsilon(-0), & \text{if } r = 0 \\ \alpha\rho_\varepsilon(r), & \text{if } r > 0. \end{cases} \quad (1)$$

The $\mathbf{w}^*$ search algorithm is an IR-ERM (Iteratively Re-weighted Empirical Risk Minimization) [3]:

procedure IR-ERM($\mathbf{w}_0$)
  $t \leftarrow 0$
  repeat
    $z_1 = \ell_1(\mathbf{w}_t), \ldots, z_N = \ell_N(\mathbf{w}_t)$
    $\overline{z}_t \leftarrow \mathsf{M}\{z_1,\ldots,z_N\}$
    for $k = 1,\ldots,N$ **do**
      $$v_k = \frac{\rho''(z_k - \overline{z}_t)}{\rho''(z_1-\overline{z}_t)+\cdots+\rho''(z_N-\overline{z}_t)}$$
    end
    $$\mathbf{w}_{t+1} \leftarrow \operatorname{argmin}_{\mathbf{w}} \sum_{k=1}^{N} v_k\,\ell_k(\mathbf{w})$$
  $t \leftarrow t+1$
  until $\{\overline{z}_t\}$ and $\{\mathbf{w}_t\}$ stabilize
end

At the heart of IR-ERM is the process of iterative re-weighting, as well as in the IRLS [4]. The IR-ERM algorithm differs from the IRLS in the way of recalculation of the weights.

To demonstrate the possibility of empirical risk minimisation [6] based on a robust estimate and IR-ERM algorithm, here is an example of linear regression problem with a large number of outlies.
We have a straight line through data points with an evenly distributed small error. For the linear regression recovering we use the least squares method, the absolute-error-minimising method, and the robust differentiable estimate minimising technique by means of the M-average through the function

$$\rho_\alpha(r) = |u - z| - \alpha\ln(\alpha + |u - z|) + \alpha\ln\alpha,$$

where $\alpha = 0.001$. Fig. 1 explain advantage of robust linear regression recovery. In both cases, robust differentiable average estimate minimising technique made it possible to avoid the influence of outliers.

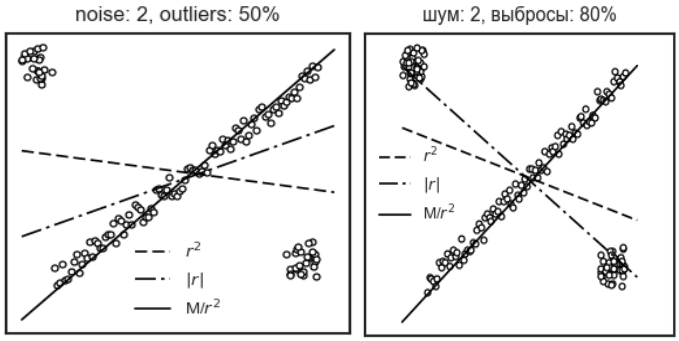

Figure 1. Recovery examples for linear regression with 50% and 80% of outliers from the amount of data without outliers.

3. Minimizing robust sums of functions

Consider a number of summation methods resistant to outliers. All M-averages including the arithmetic mean feature:

$$\frac{\partial \mathsf{M}}{\partial z_1} + \cdots + \frac{\partial \mathsf{M}}{\partial z_N} = 1.$$

But the arithmetic summation features the following important property:

$$\frac{\partial \mathsf{S}}{\partial z_1} + \cdots + \frac{\partial \mathsf{S}}{\partial z_N} = N.$$

It is therefore natural that the proposed summation methods can also maintain this property. Consider the following summation method.

3.1. Least Winsorized Sum and Mean

In the Least Winsorized Sum (LWS) method before summing, all values that are greater than the specified threshold value u are replaced by u, i.e.

$$\mathsf{WS}_u\{z_1, \ldots, z_N\} = \sum_{k=1}^{N} \frac{1}{2}(z_k + u - |z_k - u|).$$

Let's call it WS (Winsorized Sum). It has the following property: if u is the arithmetic mean of $z_1, \ldots, z_N$, then $\mathsf{WS}_u\{z_1, \ldots, z_N\} = z_1 + \cdots + z_N$.

The WM (Winsorized Mean) averaging method is defined as

$$\mathsf{WM}_u\{z_1, \ldots, z_N\} = \frac{1}{N}\mathsf{WS}_u\{z_1, \ldots, z_N\}.$$

We generalize the WS summing method as follows. Let M_ρ be M-average on the basis of a twice differentiable strictly convex function ρ. Denote $\overline{z} = \mathsf{M}_\rho\{z_1, \ldots, z_N\}$. Define

$$\mathsf{WS}_\rho\{z_1, \ldots, z_N\} = \sum_{k=1}^{N} \frac{1}{2}(z_k + \overline{z} - \rho(z_k - \overline{z})).$$

Calculate the partial derivatives:

$$\frac{\partial \mathrm{WS}_\rho}{\partial z_k} = \frac{1}{2}(1 - \rho'(z_k - \overline{z})) + \frac{1}{2}\frac{\partial \mathrm{M}_\rho}{\partial z_k}\left(N + \sum_{l=1}^{N} \rho'(z_l - \overline{z})\right).$$

Since, by definition,

$$\sum_{k=1}^{N} \rho'(z_k - \overline{z}) = 0,$$

then

$$\frac{\partial \mathrm{WS}_\rho}{\partial z_k} = \frac{1}{2}(1 - \rho'(z_k - \overline{z})) + \frac{N}{2}\frac{\partial \mathrm{M}_\rho}{\partial z_k}.$$

Therefore

$$\sum_{k=1}^{N} \frac{\partial \mathrm{WS}_\rho}{\partial z_k} = N.$$

If

$$\lim_{|r|\to\infty} \rho(r)/|r| = 1,$$

then the summation method defined here can be considered as a smooth version of WS.

Now we consider the following problem of the objective function minimizing winsorized mean:

$$Q(\mathbf{w}) = \frac{1}{N}\mathrm{WS}_\rho\{\ell_1(\mathbf{w}), \ldots, \ell_N(\mathbf{w})\}$$

to find the optimal set of parameters $\mathbf{w}^*$. Now write down the gradient:

$$\mathrm{grad}Q(\mathbf{w}) = \sum_{k=1}^{N} v_k(\mathbf{w})\mathrm{grad}\ell_k(\mathbf{w}),$$

where

$$v_k(\mathbf{w}) = \frac{1}{2N}\left(1 - \rho'(\ell_k(\mathbf{w}) - \overline{z}(\mathbf{w}))\right) + \frac{1}{2}\frac{\partial \mathrm{M}_\rho}{\partial z_k},$$

$\overline{z}(\mathbf{w}) = \mathrm{M}_\rho\{\ell_1(\mathbf{w}), \ldots, \ell_N(\mathbf{w})\}$. At that, we note

$$v_1(\mathbf{w}) + \cdots + v_N(\mathbf{w}) = 1.$$

For numerical calculation, we can apply the algorithm IR-SWSM (Iteratively Re-weighted Smoothly Winsorized Sum Minimization) the next version of the IR-SWSM algorithm:

procedure IR-SWSM($\mathbf{w}_0$)
    $t \leftarrow 0$
    repeat
        $z_1 = \ell_1(\mathbf{w}_t), \ldots, z_N = \ell_N(\mathbf{w}_t)$
        $\overline{z}_t \leftarrow \mathrm{M}\{z_1, \ldots, z_N\}$
        for $k = 1, \ldots, N$ **do**
            $v_k = \frac{1}{2N}(1 - \rho'(z_k - \overline{z})) + \frac{1}{2}\frac{\partial \mathrm{M}_\rho}{\partial z_k}$
        end
        $\mathbf{w}_{t+1} \leftarrow \mathrm{argmin}_{\mathbf{w}} \sum_{k=1}^{N} v_k \ell_k(\mathbf{w})$
        $t \leftarrow t + 1$
    until $\{\overline{z}_t\}$ and $\{\mathbf{w}_t\}$ stabilize
end

To illustrate the IR-SWSM algorithm capacity, we consider a neural network with single hidden layer:

$$y = w_0 + w_1 u_1 + \cdots + w_m u_m$$
$$u_j = \mathrm{softplus}(w_{j0} + w_{j1}x_1 + \cdots + w_{jn}x_n),$$

where $\mathrm{softplus}(s) = \ln(1 + e^s)$. NN was trained with the Boston data set.

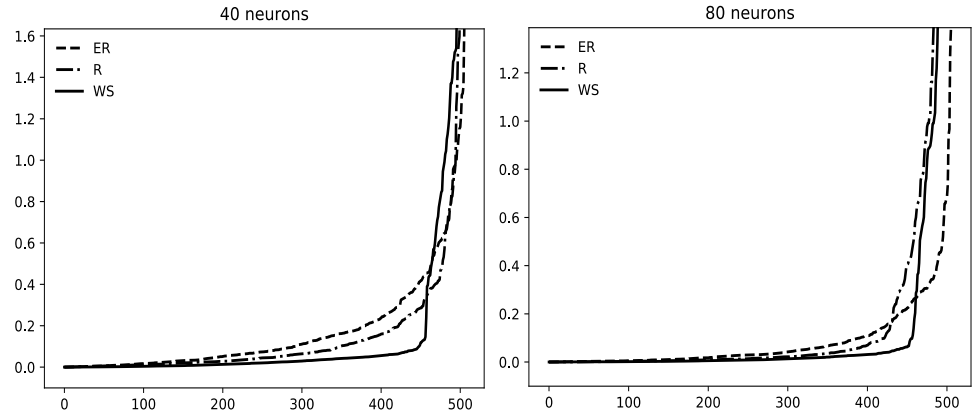

Figure 2. The errors distribution by trained NN with one hidden layer containing 40 and 80 neurons, with respect to the Boston dataset.

For training, error back propagation has been applied, where the mean square error (ER) and mean value of the Huber function with the small parameter (0.001) are minimised, for it to be a continuously differentiable approximation of the module error. IR-SWSM learning algorithm has also been used, where the robust estimate of the error sum WS of squares (WS) are minimised using the function $\rho_{\varepsilon,\alpha}$ type (1), where $\rho_\varepsilon(r) = \sqrt{\varepsilon^2 + r^2} - \varepsilon$, $\alpha = 0.90$.

The Fig. 2 shows the of mean absolute error distribution across the entire dataset. This clearly demonstrate that training neural networks (40 and 80 neurons in the hidden layer) with IR- SWSM algorithm reduce error values to more than 80% of the data.

There is also another experiment with only 7 neurons in the hidden layer. The Fig. 3 shows the of mean absolute error distribution across the entire dataset.

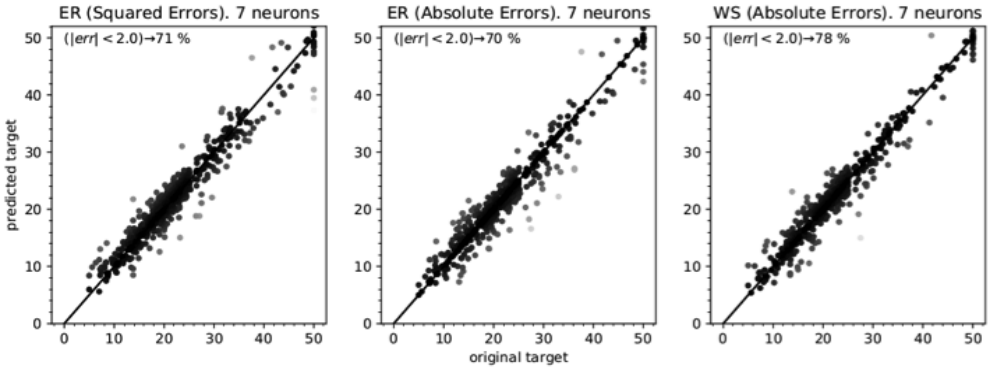

Figure 3. Errors distributions for NN training by boston dataset using three approaches: Least Squares (ER), Least Absolute Errors (ER) and Least Winsorized Squares (WS).

4. Conclusion

In this paper, we propose a method and algorithms for minimizing robust differentiable estimate of means and sums that are potentially resistant to outliers and errors that can lead to a shift in the parameters of the trainees. It is based on minimization of differentiable robust analogs of median, quantiles and winsorized sums of loss functions.

The above approaches are preferable in the cases when application of gradient based minimisation procedures are preferable. For example, these approaches made possible application of weighted variants of back propagation algorithms for NN robust learning. Construction of robust learning algorithms of NN are important in a sense of many applications [12-15]. In particular an iteratively re-weighted procedures are proposed.

In these procedures at each step a weighted variant of back propagation algorithm is used. Examples presented above clearly show that proposed approaches and algorithms can be resistant to a large amount of outliers.

5. References

[1] Rousseeuw P J 1984 Least median of squares regression *American Statistical Association* **79** 871-880
[2] Rousseeuw P J 1987 *Robust regression and outlier detection* (NY: John Wiley and Sons)
[3] Shibzukhov Z M 2017 On the principle of empirical risk minimization based on averaging aggregation functions *Doklady Mathematics* **96(2)** 494-497
[4] Andersen R 2008 *Modern Methods For Robust Regression* (Thousand Oaks: SAGE Publications)
[5] Huber P J 1981 *Robust Statistics* (NY: John Wiley and Sons)
[6] Vapnik V 2000 *The Nature of Statistical Learning Theory. Information Science and Statistics* (Springer-Verlag)
[7] Mesiar R, Komornikova M, Kolesarova A and Calvo T 2008 Aggregation functions: A revision *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models* (Springer, Berlin, Heidelberg)
[8] Grabich M, Marichal J-L and Pap E 2009 Aggregation Functions *Encyclopedia of Mathematics and its Applications* **127**
[9] Beliakov G, Sola H and Calvo T 2016 *A practical guide to averaging functions* (Springer)
[10] Calvo T and Beliakov G 2010 Aggregation functions based on penalties *Fuzzy Sets and Systems* **161(10)** 1420-1436
[11] Yohai V J 1987 High breakdown-point and high effciency robust estimates for regression *The Annals of Statistics* **15** 642-656
[12] Nikonorov A V, Petrov M V, Bibikov S A, Kutikova V V, Morozov A A and Kazanskiy N L 2017 Reconstruction of the images in difractional-optical systems on the base of convolutional neural neworks and deconvolution *Computer Optics* **41(6)** 875-887 DOI: 10.18287/2412-6179-2017-41-6-875-887
[13] Nikitin M Y, Konushin V S and Konushin A S 2017 Neural network model for human recognition by face in video sequences with estimation of frame's utility *Computer Optics* **41(5)** 732-742 DOI: 10.18287/2412-6179-2017-41-5-732-742
[14] Spicyn V G, Bolotova Y A, Fan Hgok Hoang and Bui Thi 2016 Chang Recognition of symbols on the base of wavelet transformations, principal components and neural networks *Computer Optics* **40(2)** 249-257 DOI: 10.18287/2412-6179-2016-40-2-249-257
[15] Poletaev S D and Volotovsky S G 2016 Precision laser recording of microstructures on molybdenum films for generating a diffractive microrelief *Computer Optics* **40(3)** 422-426 DOI: 10.18287/2412-6179-2016-40-3-422-426

Acknowledgments
The work was supported by a grant from the Russian Foundation for Basic Research 18–01–00050.