DOI 10.1007/s10958-021-05256-y
Journal of Mathematical Sciences, Vol. 253, No. 4, March, 2021

PRINCIPLE OF MINIMIZING EMPIRICAL RISK AND AVERAGING AGGREGATE FUNCTIONS

Z. M. Shibzukhov

UDC 519.7

Abstract. In this paper, we propose an extended version of the principle of minimizing empirical risk (ER) based on the use of averaging aggregating functions (AAF) for calculating the ER instead of the arithmetic mean. This is expedient if the distribution of losses has outliers and hence risk assessments are biased. Therefore, a robust estimate of the average risk should be used for optimization of the parameters. Such estimates can be constructed by using AAF that are solutions of the problem of minimizing the penalty function for deviating from the mean value. We also propose an iterative reweighting scheme for the numerical solution of the ER minimization problem. We give examples of constructing a robust procedure for estimating parameters in a linear regression problem and a linear separation problem for two classes based on the use of an averaging aggregating function that replaces the α-quantile.

***Keywords and phrases*:** empirical risk, averaging function, aggregation function, loss function, iterative reweighing algorithm.

***AMS Subject Classification*:** 68T05

1. Introduction. The solution of many problems of machine learning is based on the *empirical risk minimization principle* (see [15]). It consists of minimizing the magnitude of the average losses from erroneous functioning trained system on a given final set of precedents. The magnitude of the *empirical risk* is estimated as the *arithmetic average of losses*:

$$\mathsf{ER}(\boldsymbol{w}) = \frac{1}{N}\sum_{k=1}^{N} \ell_k(\boldsymbol{w}), \tag{1}$$

where $\ell_k(\boldsymbol{w})$ is the loss function associated with the kth precedent. The loss functions must be unimodal. The values $\boldsymbol{w}^*$ of the parameters in search minimize the empirical risk:

$$\mathsf{ER}(\boldsymbol{w}^*) = \min_{\boldsymbol{w}} \mathsf{ER}(\boldsymbol{w}).$$

Usually $\ell_k(\boldsymbol{w}) = L(r_k(\boldsymbol{w}))$, where $L(r)$ is the loss function, $r_k(\boldsymbol{w})$ the "*closing error*" function for the kth precedent. As an example we consider the problem of regression and classification.

The problem of regression. It is required to restore an unknown dependence $y = y(\boldsymbol{x})$, where $\boldsymbol{x} \in \mathbb{R}^n$. A finite set of entrances is given $\tilde{\boldsymbol{X}} = \{\tilde{\boldsymbol{x}}_1, \dots, \tilde{\boldsymbol{x}}_N\} \subset \mathbb{R}^n$, for which the values in search are known $\tilde{\boldsymbol{Y}} = \{\tilde{y}_1, \dots, \tilde{y}_N\} \subset \mathbb{R}$. Among parametric dependences $f(\boldsymbol{x}, \boldsymbol{w})$ we seek the one that approximates $y(\boldsymbol{x})$. And we need to find the set of parameters $\boldsymbol{w}^*$ which minimizes the mean quadratic error:

$$\mathsf{ER}(\boldsymbol{w}) = \frac{1}{N}\sum_{k=1}^{N} \Big(f(\tilde{\boldsymbol{x}}_k, \boldsymbol{w}) - \tilde{y}_k\Big)^2. \tag{2}$$

This corresponds to the classical *ordinary least squares method* (OLS).

Translated from Itogi Nauki i Tekhniki, Seriya Sovremennaya Matematika i Ee Prilozheniya. Tematicheskie Obzory, Vol. 154, Proceedings of the International Conference "Actual Problems of Applied Mathematics and Physics," Kabardino-Balkaria, Nalchik, May 17–21, 2017, 2018.

Classification problem (two classes). It is required to divide the finite set of points in $\mathbb{R}^n$ ito two classes. The set of points $\tilde{\boldsymbol{X}} = \{\tilde{\boldsymbol{x}}_1, \dots, \tilde{\boldsymbol{x}}_N\} \subset \mathbb{R}^n$ for which the marks of the classes $\tilde{\boldsymbol{Y}} = \{\tilde{y}_1, \dots, \tilde{y}_N\} \subset \{-1, 0, +1\}^N$ are known is given. For the division into two classes, the parametric dependence $f(\boldsymbol{x}, \boldsymbol{w})$ and the following rule are used:

$$y(\boldsymbol{x}) = \begin{cases} +1, & f(\boldsymbol{x}, \boldsymbol{w}) > 1, \\ 0, & |f(\boldsymbol{x}, \boldsymbol{w})| \le 1, \\ -1 & f(\boldsymbol{x}, \boldsymbol{w}) < -1. \end{cases}$$

We search for a set of parameters $\boldsymbol{w}^*$ that minimizes the following function:

$$\mathsf{ER}(\boldsymbol{w}) = \frac{1}{N} \sum_{k=1}^{N} \Big(1 - \tilde{y}_k f(\tilde{\boldsymbol{x}}_k, \boldsymbol{w})\Big)_+, \tag{3}$$

where $(S)_+ = \max\{S, 0, \}$. This corresponds to the classification method on the basis of the *support vector machine* (SVM).

Thus, functions (2) and (3) represent the examples (1).

2. Problem of outliers. Under the conditions of *outliers* in the empirical distribution of losses, the estimate (1) often turns out to be biased. At the same time, outliers can be associated both with distortions in the source data and with the inadequacy of the model used (for example, when a linear model is used instead of an unknown a priori nonlinear model). The problem of outliers could be solved by eliminating them, if we were able to identify the outliers, or using a weighted version of the empirical risk:

$$\mathsf{ER}(\boldsymbol{w}) = \sum_{k=1}^{N} v_k \ell_k(\boldsymbol{w}), \tag{4}$$

where $v_1, \dots, v_N \ge 0$, $v_1 + \dots + v_N = 1$. However, the difficulty of finding adequate values of the weights $v_1, \dots, v_N$ that would compensate for the contribution of outliers is comparable to the complexity of solving the problem of identifying outliers.

The presence of outliers in the data leads to the fact that the empirical risk estimate is distorted, since the arithmetic average is unstable with respect to the outliers:

$$\Big|\mathsf{M}\big\{z_1, \dots, z_N + \Delta\big\} - \mathsf{M}\big\{z_1, \dots, z_N\big\}\Big| = \frac{1}{N}|\Delta|, \tag{5}$$

i.e. the arithmetic average is distorted by an amount proportional to the magnitude of the distortion itself. With a large number of distortions of arguments with a relative value comparable to the value of the mean value $\bar{z} = \mathsf{M}\big\{z_1, \dots, z_N\big\}$ a significant distortion $\bar{z}$ can occur. This, in turn, can lead to the "displacement" of the desired parameters when solving the problem of minimizing the empirical risk.

Let us illustrate this on a simple example of restoring the linear regression. The direct line is restored using OLS or the more robust *least absolute deviation method* (LAD)

$$\mathsf{ER}(\boldsymbol{w}) = \frac{1}{N} \sum_{k=1}^{N} \Big| f\big(\tilde{\boldsymbol{x}}_k, \boldsymbol{w}\big) - \tilde{y}_k \Big|.$$

The first graph in Fig. (1) illustrates the use of OLS in the presence of slight noise. The second graph illustrates the use of OLS with 20% of outliers: here the OLS method strongly shifts the straight line. A more robust LAD can overcome the influence of 20% of outliers. The third graph at 50% outliers illustrates the displacement of a straight line obtained with the help of OLS. In the last graph, at 80% of the outliers, the LAD stops working as well.

We also illustrate this problem with a simple example of finding a straight line separating two classes (see Fig. 2). The first graph of Fig. 2 shows an example of division into two classes using the

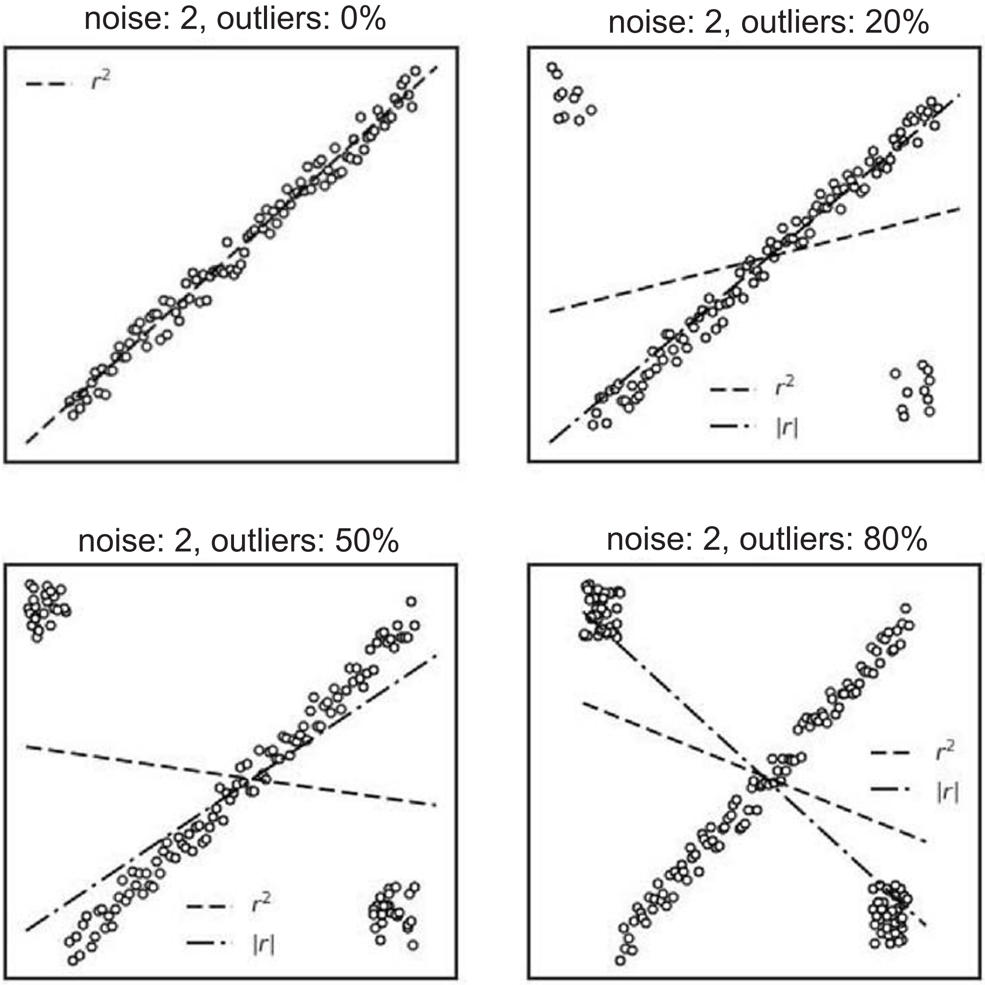

Fig. 1. Linear regression in the situation of outliers.

SVM method with no outliers. The second graph shows that at 100% outliers, the separation line constructed with the help of the SVM, even with the optimal values of the training parameters, shifts significantly.

The example with the restoration of linear regression clearly shows what problems will inevitably be encountered when learning NS in conditions of significant outliers.

3. Robust M-method. The more robust M-method (see [4]) tries to solve the outlier problem by using a scalar functional transformation, which in some cases can "suppress" outliers. So,

$$\mathsf{ER}(\boldsymbol{w}) = \frac{1}{N}\sum_{k=1}^{N} \varrho(\ell_k(\boldsymbol{w})), \tag{6}$$

where ϱ is a nonnegative quasi-differentiable function with a unique minimum equal to zero. The loss functions $\ell_k(\boldsymbol{w}) = L(r_k(\boldsymbol{w}))$ are such that, in the absence of outliers, the problem of minimizing the empirical risk (1) can be solved quite effectively. To suppress the effects of outliers, it is important that $\varrho(z)$ grow "slower" than the linear function.

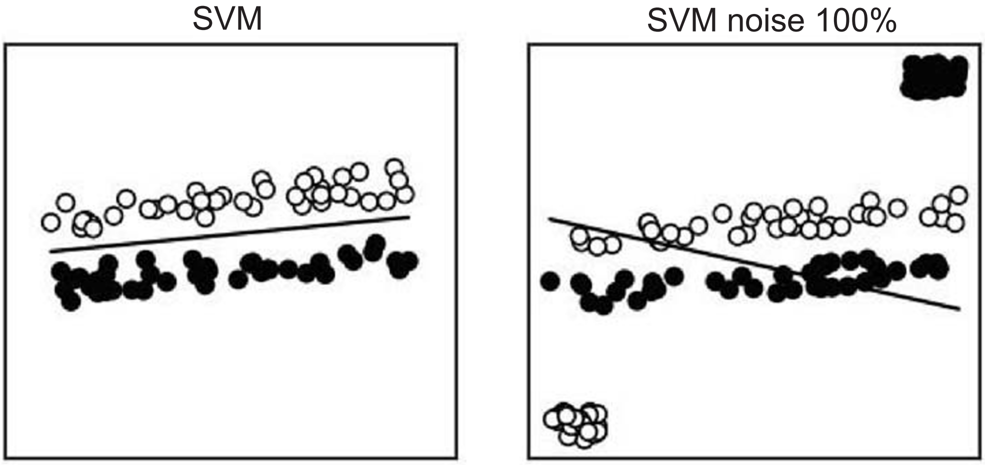

Fig. 2. Linear division into two classes in situation of outliers.

The minimization allows us to reduce (6) to the solution of the equation

$$\sum_{k=1}^{N} \varrho'\big(\ell_k(\boldsymbol{w})\big)\,\mathrm{grad}\,\ell_k(\boldsymbol{w}) = 0.$$

For its solution, the iterative reweighing scheme, IRS, is often used:

```
procedure IRS($\boldsymbol{w}_0$)
    $t \leftarrow 0$
    $\boldsymbol{v} = (1/N, \dots, 1/N)$
    repeat
        $\boldsymbol{w}_{t+1} \leftarrow \arg\min_{\boldsymbol{w}} \sum_{k=1}^{N} v_k \ell_k(\boldsymbol{w})$
        $\boldsymbol{v} \leftarrow \Big(\varrho'\big(\ell_1(\boldsymbol{w}_t)\big)/S, \dots, \varrho'\big(\ell_1(\boldsymbol{w}_t)\big)/S\Big)$, where $S \leftarrow v_1 + \dots + v_N$
        $t \leftarrow t + 1$
    until $\{\boldsymbol{w}_t\}$ does not converge
return $\boldsymbol{w}_t$
end procedure
```

In IRS on each tth step we solve the task of minimization of weighted empirical risk of the form

$$\boldsymbol{w}_t = \arg\min_{\boldsymbol{w}} \sum_{k=1}^{N} v_k \ell_k(\boldsymbol{w}),$$

where $v_1 + \dots + v_N = 1$. Then the weights are recalculated using the formula

$$v_k = \frac{\varrho'(\ell_k(\boldsymbol{w}_t))}{\varrho'\big(\ell_1(\boldsymbol{w}_t)\big) + \dots + \varrho'\big(\ell_N(\boldsymbol{w}_t)\big)}.$$

To solve this problem, there are proven effective algorithms. For example, for linear models with respect to parameters, the weighted least squares method is usually used to solve the regression problem, and for the classification problem, the linear programming method or the quadratic programming method is used.

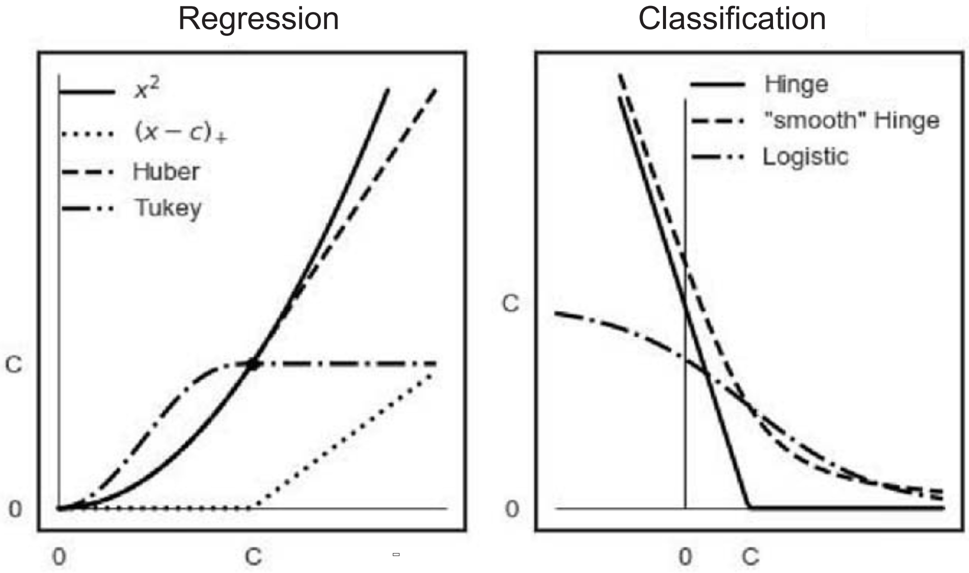

Fig. 3. Examples of functions ϱ.

Often in the regression problem we have $\ell_k(\boldsymbol{w}) = \big(f(\tilde{\boldsymbol{x}}_k, \boldsymbol{w}) - \tilde{y}_k\big)^2$. For such loss functions one can give the following functions as examples of ϱ:

$$(1) \quad \varrho(z) = \max\{\sqrt{z} - \sqrt{c}, 0\};$$

$$(2) \quad \varrho(z) = \begin{cases} 2\sqrt{z/c} - 1, & \text{if } z \geq c, \\ z/c, & \text{if } z < c \end{cases} \qquad \text{(the Huber function)};$$

$$(3) \quad \varrho(z) = \begin{cases} 1, & \text{if } z \geq c, \\ 1 - \big(1 - z/c\big)^3, & \text{if } z < c \end{cases} \qquad \text{(the Tukey function)}.$$

In the classification problem

$$\ell_k(\boldsymbol{w}) = -\tilde{y}_k f\big(\tilde{\boldsymbol{x}}_k, \boldsymbol{w}\big),$$

the following functions can be examples of ϱ:

$$(1) \quad \varrho(z) = (c + z)_+ \qquad \text{(the Hinge function)};$$

$$(2) \quad \varrho(z) = \frac{1}{2}\left(1 + \frac{z}{c} + \sqrt{1 + \left(1 + \frac{z}{c}\right)^2}\right) \qquad \text{(the smooth Hinge function)};$$

$$(3) \quad \varrho(z) = \frac{1}{1 + e^{-z/c}} \qquad \text{(the logistic function)}.$$

The robustness of the M-method or its absence can be explained using the following relation:

$$\Delta\mathsf{M}(\Delta) = \Big|\mathsf{M}\big\{\varrho(z_1), \ldots, \varrho(z_N)\big\} - \mathsf{M}\big\{\varrho(z_1), \ldots, \varrho(z_N + \Delta)\big\}\Big| = \frac{1}{N}\big|\varrho'(\tilde{z})\big|\Delta.$$

For stability with respect to large Δ, the boundedness of $\varrho(z)$ (or $\varrho'(z) \to 0$ for $z \to \infty$) is necessary. However, this leads to a greater dependence on the initial approximation $\boldsymbol{w_0}$ in the search procedures for the minimum of $\mathsf{ER}(\boldsymbol{w})$. If $|\varrho'(z)|$ is bounded below, then the M-method may be unstable with respect to outliers with large values of Δ.

For instance, in the regression problem

(i) if ϱ is a truncated linear function, or the Huber function, then $\varrho'(z)$ is bounded;

(ii) if ϱ is the Tukey function, then $\varrho'(z) \to 0$, as $z \to \infty$.

In the classification problem

(i) if ϱ is the Hinge function (in particular, smooth), then $\varrho'(z)$ is bounded;

(ii) if ϱ is a logistic function, then $\varrho'(z) \to 0$ as $z \to \infty$.

However, there are tasks for which the M-method does not allow us to overcome the outlier problem. In these cases, in a number of problems a different approach is used, based on the empirical average estimates that are resistant to outliers, for example, the medians (see [10, 11]), quantile (see [7]), expectile (see [9]) instead of arithmetic mean.

However, the main thing is that the M-method is equivalent to the Kolmogorov average minimization:

$$\mathsf{ER}_\rho(\boldsymbol{w}) = \mathsf{M}_{[\varrho]}\{\ell_1(\boldsymbol{w}), \dots, \ell_N(\boldsymbol{w})\},$$

where

$$\mathsf{M}_{[\varrho]}\{z_1, \dots, z_N\} = \varrho^{-1}\left(\frac{1}{N}\sum_{k=1}^{N} \varrho(z_k)\right),$$

i.e., the M-method is based on minimizing the arithmetic mean of losses that were previously converted to another "scale" using the function ϱ.

In this regard, in this paper we consider a generalizing approach, when arbitrary averaging aggregation functions can be used to estimate average losses, which we call M-mean. This approach in a certain sense generalizes the M-method and provides a universal method for solving the problem of minimizing the empirical risk in the situation of outliers.

The further presentation is constructed as follows. First, the class of standard averaging aggregating functions is determined, the method of their representation and approximate calculation is described. Next, we introduce the extended concept of empirical risk, the value of which is calculated as the value of the averaging aggregating function of the magnitude of the losses. To solve the problem of minimizing the parameterized extended empirical risk, a method for calculating the gradient of the averaging aggregating function and gradient procedures are determined to solve the problem of minimizing the extended empirical risk. At the end, examples of the construction of a robust regression and a robust version of the support vector machine for solving the classification problem based on the application of a certain parametrically given "approximation" of the median and quantile are considered.

4. M-means. The arithmetic average is not a robust average, but the median is. Therefore, when solving problems with outliers, they sometimes began to use the median to estimate the empirical risk

$$\mathsf{ER}(\boldsymbol{w}) = \operatorname{med}\{\ell_1(\boldsymbol{w}), \dots, \ell_N(\boldsymbol{w})\}$$

instead of the arithmetical.

The median can be defined as follows. Consider the function

$$P(z_1, \dots, z_m, u) = \sum_{j=1}^{m} |z_j - u|$$

and denote by $\boldsymbol{M}_{z_1,\dots,z_m}$ the set of all u for which $P(z_1, \dots, z_m, u)$ takes its maximum value. The set $\boldsymbol{M}_{z_1,\dots,z_m}$ is a singleton or a connected segment. Then

$$\operatorname{med}\{z_1, \dots, z_m\} = \arg\min_u P(z_1, \dots, z_m, u),$$

if $\boldsymbol{M}_{z_1,\dots,z_m}$ is a singleton, and

$$\operatorname{med}\{z_1, \dots, z_m\} = \frac{a+b}{2},$$

if $\boldsymbol{M}_{z_1,\dots,z_m}$ is a connected segment with the borders a and b.

The median is resistant to outliers in the following sense: if we increase the initial arguments that exceed the median, then its value on the new set will remain unchanged. So you can “distort” the values up to 50% without damage to its value. However, there are cases when it is better to use α-quantile $(0 < \alpha < 1)$ instead of the median. Such cases occur when the part α of the data is undistorted. The quantile with $\alpha = 0.5$ coincides with the median; it can be defined in the same way as the median, if we put

$$P(z_1, \dots, z_m, u) = \sum_{j=1}^{m} |z_j - u|_\alpha,$$

where $|x|_\alpha = \bigl(\alpha - [\![x < 0]\!]\bigr)x$, $[\![s]\!] = \max\{0, s\}$.

Since the median and the quantile are nondifferentiable functions, algorithms with inevitable elements of search are actively used to minimize $\mathsf{ER}(\boldsymbol{w})$. It is difficult to use such algorithms to train the NS.

Arithmetic mean, median, and α-quantile are examples of M-averages. Among them there are robust and differentiable averages, which could be used instead of the median and the α-quantile.

M-average form a subclass of *aggregating functions* (AF; see [3, 8]). Take $\mathbb{I} = [A, B] \subseteq \mathbb{R}$, where $-\infty \le A < B \le \infty$. The aggregating function M assigns to each set $\{z_1, \dots, z_N\} \subset \mathbb{I}$ the value $\mathsf{M}\{z_1, \dots, z_N\} \in \mathbb{I}$. It satisfies the following conditions:

$$\inf_{z_1,\dots,z_N} \mathsf{M}\{z_1, \dots, z_N\} = \inf \mathbb{I}, \qquad \sup_{z_1,\dots,z_N} \mathsf{M}\{z_1, \dots, z_N\} = \sup \mathbb{I};$$

$$\text{if } z_1' \le z_1'',\ \dots,\ z_N' \le z_N'',\ \text{then } \mathsf{M}_N\bigl\{z_1', \dots, z_N'\bigr\} \le \mathsf{M}_N\bigl\{z_1'', \dots, z_N''\bigr\}.$$

The *averaging AF* satisfies the additional requirement

$$\min\{z_1, \dots, z_N\} \le \mathsf{M}\{z_1, \dots, z_N\} \le \max\{z_1, \dots, z_N\}.$$

Any averaging AF under certain conditions can be determined using the corresponding *penalty function* (see [1, 2]).

By definition, the function $P(z_1, \dots, z_N, u)$ is a *penalty function* if it satisfies the following requirements:

(1) $P(z_1, \dots, z_N, u) \ge 0$ for all u and $z_1, \dots, z_N$;
(2) $P(z_1, \dots, z_N, u) = 0$, only if $z_1 = \dots = z_N = u$;
(3) the set

$$\boldsymbol{M}_{z_1 \dots z_N} = \bigl\{u : P(z_1, \dots, z_N, u) = P_{\min}\bigr\},$$

where

$$P_{\min} = \min_u P(z_1, \dots, z_N, u),$$

is a singleton or a connected segment.

the averaging AF M_P, besed on a penalty function, is defined as follows:

$$\mathsf{M}_P\{z_1, \dots, z_N\} = \arg\min_u P(z_1, \dots, z_N, u), \tag{7}$$

if $\boldsymbol{M}_{z_1 \dots z_N}$ is a singleton, and

$$\mathsf{M}_P\{z_1, \dots, z_N\} = \frac{a + b}{2},$$

if $\boldsymbol{M}_{z_1 \dots z_N}$ is a segment with the ends a and b.

Consider penalty functions of the form

$$P(z_1, \dots, z_N, u) = \sum_{k=1}^{N} \rho(z_k, u), \tag{8}$$

where $\rho(z,u) = g\big(h(z) - h(u)\big)$, $\rho(z,u)$ is the dissimilarity function, g is nonnegative and convex ($g(0) = 0$), and h is a monotone reversible function. By definition, the function $\rho(z,u)$ is a dissimilarity function if

(i) $\rho(z,u) = 0 \Leftrightarrow z = u$;
(ii) $\rho(z_1,u) \geq \rho(z_2,u)$, when $z_1 \geq z_2 \geq u$ or $z_1 \leq z_2 \leq u$.

The averaging AF, based on a penalty function P of the form (8), defines the M-*mean.*

The M-means form a sufficiently wide class of functions for calculating the mean. It contains the following families of averaging functions:

(I) *a family of symmetric means*:

$$\mathsf{M}^{\gamma}\{z_1,\dots,z_N\} = \arg\min_u \sum_{k=1}^{N} |z_k - u|^{1+\gamma}, \tag{9}$$

where $0 \leq \gamma \leq 1$; here M^0 is the median and M^1 is the arithmetic mean;

(II) *a family of nonsymmetric means*:

$$\mathsf{M}^{\gamma}_{\alpha}\{z_1,\dots,z_N\} = \arg\min_u \sum_{k=1}^{N} |z_k - u|^{1+\gamma}_{\alpha}, \tag{10}$$

where

$$|u|^{1+\gamma}_{\alpha} = (\alpha - [u > 0])u|u|^{\gamma}, \quad 0 \leq \gamma \leq 1;$$

here M^0_{α} is the α-quantile and M^1_{α} is the α-expectile;

(III) *a family of symmetric means of Kolmogorov type*:

$$\mathsf{M}^{\gamma}_{g}\{z_1,\dots,z_N\} = \arg\min_u \sum_{k=1}^{N} |g(z_k) - g(u)|^{1+\gamma},$$

where g is a reversible function, $0 \leq \gamma \leq 1$; here M^0_g is the scalable median

$$\underset{g}{\mathrm{med}}\{z_1,\dots,z_N\} = g^{-1}\big(\mathrm{med}\{g(z_1),\dots,g(z_N)\}\big)$$

and M^1_g is the Kolmogorov mean:

$$\mathsf{M}_g\{z_1,\dots,z_N\} = g^{-1}\left(\frac{g(z_1) + \dots + g(z_N)}{N}\right).$$

If there exist $\rho''_{uz}(z,u)$ and $\rho''_{uu}(z,u)$, then there exist partial derivatives M_{ρ}:

$$\frac{\partial \mathsf{M}_{\rho}\{z_1,\dots,z_N\}}{\partial z_k} = \frac{-\rho''_{uz}(z_k,\bar{u})}{\rho''_{uu}(z_1,\bar{u}) + \dots + \rho''_{uu}(z_N,\bar{u})}, \tag{11}$$

where $\bar{z} = \mathsf{M}_{\rho}\{z_1,\dots,z_N\}$.

The possibility of calculating the gradient M_{ρ} gives a basis for using M_{ρ} for estimating the mean losses.

Estimation of robustness of the M-*average.* To estimate the robustness of the M-average, let us give an inequality that shows how the mean changes if each element changes by some Δ:

$$\Delta\mathsf{M}(\Delta) = \Big|\mathsf{M}_{\rho}\big\{z_1,\dots,z_N + \Delta\big\} - \mathsf{M}_{\rho}\big\{z_1,\dots,z_N\big\}\Big| = \frac{\partial \mathsf{M}\big\{z_1,\dots,z_{N-1},\tilde{z}\big\}}{\partial z_N}\Delta,$$

where $\tilde{z} \in (z_N, z_N + \Delta)$. Thus, if $0 < a < \rho''_{uu}(z,u)$, then

$$\Delta\mathsf{M}(\Delta) < \frac{1}{Na}\rho''_{uz}(z_N,\tilde{z})\Delta.$$

If $\left|\rho''_{uz}(z,u)\right| \le \dfrac{b}{|z-u|}$, then $\Delta\mathsf{M}(\Delta) \le \dfrac{b}{Na}$; this implies the boundedness and independence of Δ in such cases.

Examples of robust and differentiable M*-means.* For illustration, let us consider parametric families of functions (by the parameter $\varepsilon \ge 0$), which are asymptotically equivalent to the median. Such families must meet the following requirements:

(i) $\lim\limits_{\varepsilon\to 0} \rho_\varepsilon(z-u) = |z-u|$;
(ii) $\lim\limits_{\varepsilon\to 0} \rho'_\varepsilon(z-u) = \operatorname{sign}(z-u)$;
(iii) $\lim\limits_{\varepsilon\to 0} \rho''_\varepsilon(z-u) = \delta(z-u)$.

For example,

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

These families can be generalized to the families which are asymptotically equivalent to the α-quantile. They must satisfy the following requirements:

(i) $\lim\limits_{\varepsilon\to 0} \rho_{\alpha,\varepsilon}(z-u) = |z-u|_\alpha$;
(ii) $\lim\limits_{\varepsilon\to 0} \rho'_{\alpha,\varepsilon}(z-u) = \begin{cases} \alpha, & \text{if } z-u>0,\\ \alpha-1/2, & \text{if } z-u=0,\\ \alpha-1, & \text{if } z-u<0;\end{cases}$
(iii) $\lim\limits_{\varepsilon\to 0} \rho''_{\alpha,\varepsilon}(z-u) = \delta(z-u)$.

For example, if $\mathsf{M}_{\rho_\varepsilon}$ with $\rho_\varepsilon(x)$ is asymptotically equivalent to the mediane, then we put by definition

$$\rho_{\alpha,\varepsilon}(x) = \begin{cases} \alpha\rho_\varepsilon(x), & \text{if } x>0,\\ 0, & \text{if } x=0,\\ (1-\alpha)\rho_\varepsilon(x), & \text{if } x<0; \end{cases} \qquad \rho_{\alpha,\varepsilon}(x) = \begin{cases} \alpha\rho'_\varepsilon(x), & \text{if } x>0,\\ \frac{1}{2}\Big(\alpha\rho'_\varepsilon(0^+) + (1-\alpha)\rho'_\varepsilon(0^-)\Big), & \text{if } x=0,\\ (1-\alpha)\rho'_\varepsilon(x), & \text{if } x<0. \end{cases}$$

Gradient procedures for calculating M*-averages.* For calculating the M-average one can apply arbitrary method for minimizing the function

$$P(u) = P(z_1,\dots,z_N,u) = \sum_{k=1}^{N} \rho(z_k,u).$$

If the derivative $\rho'_u(z,u)$ exists, then to calculate the approximate value of $\mathsf{M}_\rho\{z_1,\dots,z_N\}$ one can use any gradient minimization method (8), for example, the full gradient method:

$$u_{t+1} \leftarrow u_t - \tau Q_t,$$

where

$$Q_t \leftarrow \frac{1}{N}\sum_{k=1}^{N} \rho'_u(z_k,u_t).$$

To construct a stochastic analogue with the same order of convergence, you can use the SAG algorithm construction scheme (see [12]). In this scheme, the value Q_t is updated according to the rules

$$Q_t = \frac{1}{N}\sum_{k=1}^{N} Q_{t,k},$$

and the set $\{Q_{t+1,k} : k = 1, \dots, N\}$ is updated by the rule

$$Q_{t+1,k} = \begin{cases} \rho'_u(z_k, u_t), & \text{if } k = k(t), \\ Q_{t,k} & \text{otherwise;} \end{cases}$$

here $k = k(t)$ is the index of value from $\{z_1, \dots, z_N\}$, chosen randomly at the step t. To shorten the computation, it is better to use the following rule:

$$Q_{t+1} = Q_t + \frac{1}{N}\big(\rho'_u(z_k, u_t) - Q_{t,k}\big).$$

In all algorithms, the learning rate parameter τ does not depend on the step number.

To improve convergence, one can apply the AdaM algorithm scheme (see [5]). It uses the following update method u:

$$m_{t+1} = \beta_1 m_t + (1 - \beta_1)Q_t, \quad v_{t+1} = \beta_2 v_t + (1 - \beta_2)Q_t^2,$$

m_t and v_t are the moments of first and second order, respectively,

$$\tilde{m}_{t+1} = \frac{m_{t+1}}{1 - \beta_1^{t+1}}, \quad \tilde{v}_{t+1} = \frac{v_{t+1}}{1 - \beta_2^{t+1}},$$

$\tilde{m}_t$ and $\tilde{v}_t$ are the corrected values of moments,

$$u_{t+1} = u_t - \tau \frac{\tilde{m}_{t+1}}{\sqrt{\tilde{v}_{t+1} + \varepsilon}},$$

$m_0 = v_0 = 0$; $0.5 < \beta_1 \leq \beta_2 < 1$. This scheme allows one to increase the stableness of the search procedure of $\boldsymbol{w}$.

In cases where there is a nondegenerate second derivative $\rho''_{uu}(z, u)$, one can also use the Newton–Raphson method.

Iterative procedures for calculating the values of a standard averaging function. Under certain conditions, the following iterative method can be used to calculate the average value. It follows from the definition that the value $u = \mathsf{M}_\rho\{z_1, \dots, z_N\}$ satisfies the equation

$$\sum_{k=1}^{N} \rho'(z_k - u) = 0. \tag{12}$$

We set $\varphi(x) = \rho'(x)/x$ and assume that $\rho(x)$ is twice differentiable, and $\varphi(x)$ has sense on the whole domain of ρ. Then

$$\sum_{k=1}^{N} \varphi(z_k - u) \cdot (z_k - u) = 0,$$

and hence

$$u = \left(\sum_{k=1}^{N} \varphi(z_k - u) z_k \right) \Big/ \left(\sum_{k=1}^{N} \varphi(z_k - u) \right).$$

This allows us to give the iterational scheme for calculating $\bar{z}$:

$$u_{t+1} = \left(\sum_{k=1}^{N} \varphi(z_k - u_t) z_k \right) \Big/ \left(\sum_{k=1}^{N} \varphi(z_k - u_t) \right). \tag{13}$$

This scheme converges if the following condition holds:

$$\left| \sum_{k=1}^{N} \varphi'(z_k - \bar{z})(z_k - \bar{z}) \right| < \left| \sum_{k=1}^{N} \varphi(z_k - \bar{z}) \right|. \tag{14}$$

For example for the averaging aggregating function given above, M_α and M_α^γ, and $0 < \gamma < 1$, this condition holds.

5. Principle of minimizing average losses. Let M_ρ be some M-average. Let us define the empirical risk based on M_ρ as follows:

$$\mathsf{ER}_\rho(\boldsymbol{w}) = \mathsf{M}_\rho\bigl\{\ell_1(\boldsymbol{w}), \dots, \ell_N(\boldsymbol{w})\bigr\}. \tag{15}$$

The classical empirical risk (1) is a special case of (15), when M_ρ is the arithmetic mean. In accordance with the principle of risk minimization the optimal set of parameters $\boldsymbol{w}^*$ minimizes the function

$$\mathsf{ER}_\rho(\boldsymbol{w}^*) = \min_{\boldsymbol{w}} \mathsf{M}_\rho\bigl\{\ell_1(\boldsymbol{w}), \dots, \ell_N(\boldsymbol{w})\bigr\}. \tag{16}$$

This approach has already been used to define aggregating functionals for evaluating the quality of algorithms in [13, 14] when determining aggregate correct operations on algorithms. In [10, 11], when solving the regression problem of estimating the mean square error, the median was used instead of the arithmetic mean, since it is a robust estimate of the mean value. In [7], in constructing one robust version of the SVC method, the median and the α-quantile were used to average losses.

Let $\rho(z,u)$ have derivatives $\rho''_{uz}(z,u)$ and $\rho''_{uu}(z,u)$. We introduce the notation $\bar z = \mathsf{M}_\rho\{z_1, \dots, z_N\}$. Then the kth partial derivative $(1 \le k \le N)$ is calculated as follows:

$$\frac{\partial \mathsf{M}_\rho}{\partial z_k} = \frac{-\rho''_{uz}(z_k, \bar z)}{\sum\limits_{l=1}^{N} \rho''_{uu}(z_l, \bar z)},$$

except when the denominator is zero. If $\rho(z,u)$ is convex with respect to u, then the denominator is zero only if $\{z_1, \dots, z_N\} \subseteq \boldsymbol{M}_{z_1 \dots z_N}$, and in this case, $\mathsf{M}_\rho\{z_1, \dots, z_N\}$ is found trivially. If the function $\rho(z,u)$ is strictly convex in u, then the denominator is always nonzero except for the trivial case $z_1 = \dots = z_N$.

If the gradients $\ell_1(\boldsymbol{w}), \dots, \ell_N(\boldsymbol{w})$ exist, then

$$\operatorname{grad} \mathsf{ER}_\rho(\boldsymbol{w}, \bar z) = \left(\sum_{k=1}^{N} -\rho''_{uz}\bigl(\ell_k(\boldsymbol{w}), \bar z\bigr) \operatorname{grad} \ell_k(\boldsymbol{w})\right) \Big/ \left(\sum_{k=1}^{N} \rho''_{uu}\bigl(\ell_k(\boldsymbol{w}), \bar z\bigr)\right), \tag{17}$$

where $\bar z = \mathsf{M}_\rho\bigl\{\ell_1(\boldsymbol{w}), \dots, \ell_N(\boldsymbol{w})\bigr\}$. For convenience let us write (17) in another form:

$$\operatorname{grad} \mathsf{ER}_\rho(\boldsymbol{w}, \bar z) = \sum_{k=1}^{N} \alpha_k(\boldsymbol{w}, \bar z) \operatorname{grad} \ell_k(\boldsymbol{w}),$$

where

$$\alpha_k(\boldsymbol{w}, \bar z) = \frac{-\rho''_{uz}(\ell_k(\boldsymbol{w}), \bar z)}{\rho''_{uu}(\ell_1(\boldsymbol{w}), \bar z) + \dots + \rho''_{uu}(\ell_N(\boldsymbol{w}), \bar z)}.$$

Note that if $\rho(z,u) = g(zu)$, where g is a strictly convex function, then $\alpha_k(\boldsymbol{w}, \bar z) \ge 0$, $\sum \alpha_k(\boldsymbol{w}, \bar z) = 1$. In this case, the gradient $\mathsf{ER}_\rho(\boldsymbol{w}, \bar z)$ is the weighted arithmetic mean of loss gradients $\ell_1(\boldsymbol{w}), \dots, \ell_N(\boldsymbol{w})$ with variable weights:

$$\alpha_k(\boldsymbol{w}, \bar z) = \frac{g''(\ell_k(\boldsymbol{w}) - \bar z)}{g''\bigl(\ell_1(\boldsymbol{w}) - \bar z\bigr) + \dots + g''\bigl(\ell_N(\boldsymbol{w}) - \bar z\bigr)}.$$

Connection with the M*-method.* The approach proposed here can be considered in a certain sense as a generalization of the M-method (see [4, 16]). Within this method, the minimization problem for functional (6) is solved. If ϱ is a strictly monotone continuous function, then its solution coincides with the solution of the problem of minimizing the Kolmogorov mean:

$$\mathcal{Q}(\boldsymbol{w}) = \mathsf{M}_\rho\bigl\{r_1(\boldsymbol{w}), \dots, r_N(\boldsymbol{w})\bigr\} = \varrho^{-1}\left(\frac{1}{N}\sum_{k=1}^{N} \varrho(r_k(\boldsymbol{w}))\right), \tag{18}$$

where $\rho(z,u) = (\varrho(z) - \varrho(u))^2$. Thus, the solution to the problem of minimizing functional (6) in a certain sense coincides with the solution of the problem of minimizing empirical risk calculated as a Kolmogorov average (18) from losses with the scaling function ϱ.

Gradient schemes for minimizing averaging functional. Solving the problem (16) of finding the optimal set of parameters $\boldsymbol{w}^*$ and the minimum risk u^* can be done numerically using gradient descent methods:

```
procedure PBFG($\boldsymbol{w}_0$)
    $t \leftarrow 0$
    repeat
        $u_t \leftarrow \mathsf{M}_\rho\{\ell_1(\boldsymbol{w}_t), \ldots, \ell_N(\boldsymbol{w}_t)\}$
        $\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - h_t \sum_{k=1}^{N} \alpha_k(\boldsymbol{w}_t, u_t) \operatorname{grad} \ell_k(\boldsymbol{w}_t)$
        $t \leftarrow t+1$
    until $\{u_t\}$ and $\{\boldsymbol{w}_t\}$ is not stabilized
end procedure
```

Here in the calculation of the value $u_t = \mathsf{M}_\rho\big\{\ell_1(\boldsymbol{w}_t), \ldots, \ell_N(\boldsymbol{w}_t)\big\}$ at each step one can use any gradient or interactive procedure, if the sufficient condition of its convergence (14) is satisfied. As an initial approximation for u_t one can use u_{t-1}.

Note that $\boldsymbol{w}^*$ and u^* are a solution for the system of equations:

$$\sum_{k=1}^{N} \rho'(\ell_k(\boldsymbol{w}), u) = 0, \quad \sum_{k=1}^{N} \alpha_k(\boldsymbol{w}, u) \operatorname{grad} \ell_k(\boldsymbol{w}) = 0.$$

Therefore, to solve this system of equations, one can use an analogue of the Seidel iterative method for solving systems of nonlinear equations:

```
procedure PBFG2($\boldsymbol{w}_0$)
    $t \leftarrow 0$
    repeat
        $u_t \leftarrow \mathsf{M}_\rho\{\ell_1(\boldsymbol{w}_t), \ldots, \ell_N(\boldsymbol{w}_t)\}$
        $\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w} : \sum_{k=1}^{N} \alpha_k(\boldsymbol{w}, u_t) \operatorname{grad} \ell_k(\boldsymbol{w}) = 0.$
        $t \leftarrow t+1$
    until $\{u_t\}$ and $\{\boldsymbol{w}_t\}$ is not stabilized
end procedure
```

Here on each step one calculates first $u_t = \mathsf{M}_\rho\big\{\ell_1(\boldsymbol{w}_t), \ldots, \ell_N(\boldsymbol{w}_t)\big\}$, which is a solution of the first equation with respect to u for a given $\boldsymbol{w}_t$. Then one seeks the solution $\boldsymbol{w}_{t+1}$ of the second equation with respect to $\boldsymbol{w}$ for a given u_t.

To simplify the calculations, it can be reduced to a variant of the iterative re-weighting method, in which the values of the weight functions $v_k = \alpha_k(\boldsymbol{w}, u)$ are calculated before solving the second equation.

```
procedure IRLAL($\boldsymbol{w}_0$)
    $t \leftarrow 0$
    repeat
        $z_1, \ldots, z_N \leftarrow \ell_1(\boldsymbol{w}_t), \ldots, \ell_N(\boldsymbol{w}_t)$
        $u_t \leftarrow \mathsf{M}_p\{z_1, \ldots, z_N\}$
        $(v_1, \ldots, v_N) = \operatorname{grad} \mathsf{M}_p\{z_1, \ldots, z_N\}$
```

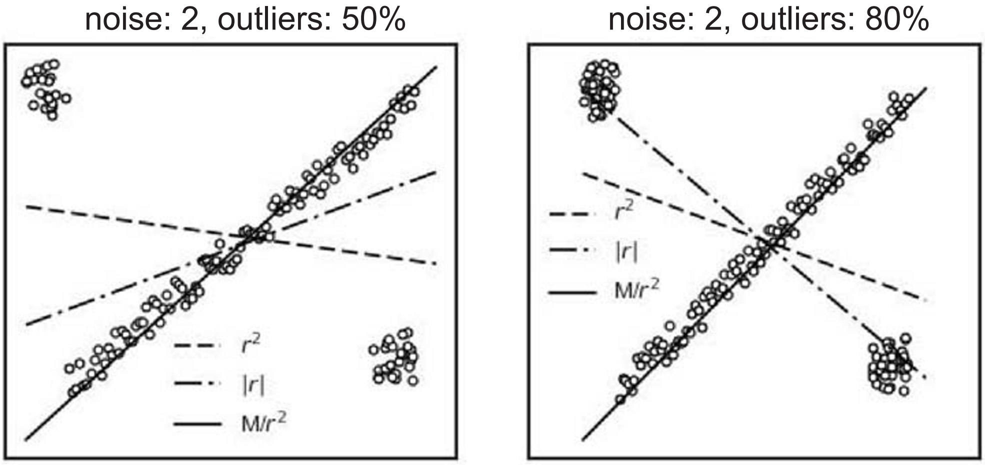

Fig. 4. Outlier data that “unfold” the straight line.

```
        $\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w} : \sum_{k=1}^{N} v_k \operatorname{grad}\ell_k(\boldsymbol{w}) = 0.$
        $t \leftarrow t + 1$
    **until** $\{u_t\}$ and $\{\boldsymbol{w}_t\}$ is not stabilized
**end procedure**
```

Robust estimates of average losses. The arithmetic average, as an empirical estimate of average losses in (1), is statistically adequate on the basis of the maximum likelihood principle, if the values of the losses are distributed according to the normal law. However, even for a normal law, the arithmetic mean is not a robust estimate. Therefore, in order to estimate the mean, in some cases, instead of the arithmetic mean, use *median* (M_0) or even quantile (M_0^γ, $0 < \gamma < 1$). In these cases

$$\mathsf{ER}(\boldsymbol{w}) = \mathsf{M}_0^\gamma\{\ell_1(\boldsymbol{w}), \ldots, \ell_N(\boldsymbol{w})\}. \tag{19}$$

It is shown (see [10, 11]) that in the problem of restoring linear regression by minimizing the median from the square of the error, it is possible to find the desired parameters in conditions when there is up to 50% outliers, which is almost impossible to achieve in the M-method (see [4]). Using the MM-method (see [16]), it is possible to find the desired parameters in principle. However it assumes a successive solution of two problems: first, the task of finding the scale parameter in the distribution of losses, and only then, the problem of minimizing average losses, taking into account the scale parameter found.

Since the median is a nondifferentiable function, the gradient procedures cannot be applied to minimize the risk functional. However, instead of the median, you can use the mean functions from the parametric family of differentiable M-average values given above.

Similarly, one can construct parametrized averaging functions that, for $\alpha \to 0$, converge uniformly to the function $\rho_\alpha(x) = |x|_\alpha$ (see (10)), which defines quantiles.

6. Application in building robust procedures for solving regression and classification problems. Based on the risk estimate (19), we construct methods for estimating stable linear regression parameters and a straight line dividing the two classes on a plane. Examples will be selected to demonstrate the robust capabilities of the proposed approach.

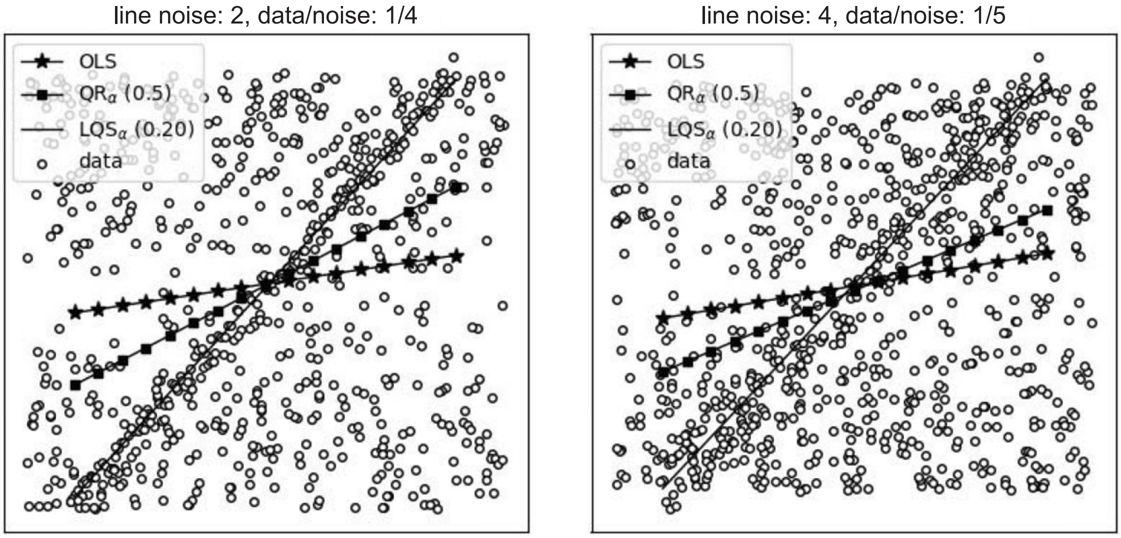

Fig. 5. Strongly noisy data that “unfold” the straight line.

The regression problem is usually reduced to the problem of minimizing the mean square error:

$$\mathsf{ER}(\boldsymbol{w}) = \frac{1}{N}\sum_{k=1}^{N} r_k(\boldsymbol{w})^2,$$

where $r_k(\boldsymbol{w}) = f(\boldsymbol{x}_k, \boldsymbol{w}) - y_k$ is the closing error, $f(\boldsymbol{x}, \boldsymbol{w})$ a linear function.

The problem of linear separation of two classes can be reduced to the problem of minimizing the functional

$$\mathsf{ER}(\boldsymbol{w}) = \frac{1}{N}\sum_{k=1}^{N} (1 - d_k(\boldsymbol{w}))_+,$$

where $d_k(\boldsymbol{w}) = f(\boldsymbol{x}_k, \boldsymbol{w})y_k$ is the shift, $f(\boldsymbol{x}, \boldsymbol{w})$ a linear function, $(S)_+ = [S \geq 0]S$.

To construct a robust procedure for solving the regression problem and classification, instead of the arithmetic mean, we will use the averaging function M_{ρ_α} with $\rho_\alpha(x)$, $\alpha \approx 10^{-2}$–10^{-3}, where ρ_α are the dissimilarity functions listed above.

The results of the calculations are presented below in the figures with the description of the source data. In the case of linear regression, the LQS$_\alpha$ algorithm is constructed, a variant of the LQS method (least quantile of squares; see [10, 11]), which is based on minimizing the averaging function med$_\alpha$ ($\alpha = 0.01$) from the squares of the error. In the case of the classification problem, the LQHS$_\alpha$ algorithm (least quantile of hinge of squares) is built, which is based on minimizing the averaging function med$_\alpha$ ($\alpha = 0.01$) from the values of the Hinge function on the shift value. For numerical calculations we use the library `mlgrad`[1] and the algorithm `mlgrad.PbFG` to minimize the average risk, based on PBFG3 procedure outlined above. The stochastic version of `mlgrad.PbSAG`, which is similar to the SAG, is also used, but it is more sensitive to the choice of setting the pace of learning, although sometimes it converged faster.

Example 1 (see Fig. 4)**.** In this example, the data is artificially selected so that the use of the least squares method and the M-method leads to the “unfold” of the straight line. The undistorted linear

[1] See `http://bitbucket.org/intellimath/mlgrad`

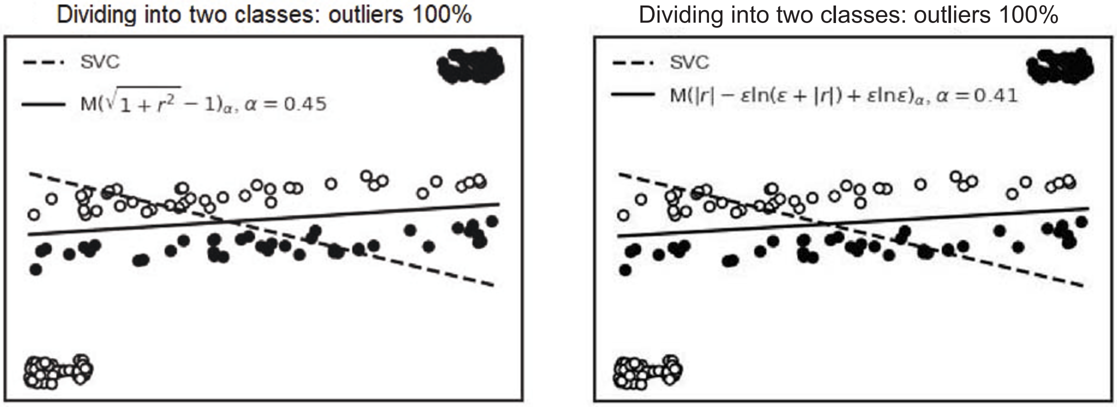

Fig. 6. Strongly noisy data that “unfold” the straight line separating two classes.

dependence is $y = 3x$. Distortions include: uniform noise with amplitude 3 and 6, respectively; outliers from the top and bottom of the original straight line are 100% and 150%, respectively.

Example 2 (see Fig. 5)**.** In this example, the data is artificially selected so that the use of the least squares method and the M-method leads to the “unfold” of a straight line. The undistorted linear dependence is $y = 3x$. The distortions include: uniform noise with amplitudes 2 and 4, respectively; the ratios of the volume of data for restoring by the least squares method to the distortions that represent noise “tending” the upper and lower borders of the graph are 1/4 and 1/5, respectively. These distortions also certainly “unfold” the straight line when recovering with method QR_α (quantile regression) with the best selection of the parameter α.

Example 3 (see Fig. 6)**.** In this example, the data are artificially selected so that the use of the SVM method leads to the “unfold” of the straight line that divides the two classes.

Acknowledgment. This work was partially supported by the Russian Foundation for Basic Research (project No. 15-01-03381) and a Fundamental Scientific Project of the Department for Nanotechnology and Information Technology of the Russian Academy of Sciences.

REFERENCES

1. G. Beliakov, H. Sola, and T. Calvo, *A Practical Guide to Averaging Functions*, Springer-Verlag (2016).
2. T. Calvo and G. Beliakov, “Aggregation functions based on penalties,” *Fuzzy Sets Syst.*, **161**, No. 10, 1420–1436 (2010).
3. M. Grabich, J.-L. Marichal, E. Pap, “Aggregation Functions,” in: *Encycl. Math. Appl.*, **127**, Cambridge Univ. Press (2009).
4. P. J. Huber, *Robust Statistics*, Wiley, New York (1981).
5. D. P. Knigma and Ba J., *Adam: A method for stochastic optimization*, e-print `arXiv1412.6980`.
6. R. Koenker, *Quantile Regression*, Cambridge Univ. Press, New York (2005).
7. Y. Ma, L. Li, X. Huang, and S. Wang, “Robust support vector machine using least median loss penalty,” *IFAC Proc. Vols. 18th IFAC World Congr.*, **44**, No. 1, 11208–11213 (2011).
8. R. Mesiar, M. Komornikova, A. Kolesarova, and T. Calvo, “Aggregation functions: A revision,” in: *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models* (H. Bustince, F. Herrera, and J. Montero, eds.), Springer-Verlag, Berlin–Heidelberg (2008).

9. W. Newey and J. Powell, "Asymmetric least square estimation and testing," *Econometrica*, **55**, No. 4, 819–847 (1987).
10. P. J. Rousseeuw, "Least median of square regression," *J. Am. Stat. Assoc.*, **79**, 871–880 (1984).
11. P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, Wiley, New York (1987).
12. M. Schmidt, N. Le Roux, and F. Bach, *Minimizing finite sums with the stochastic average gradient*, e-print arXiv1309.2388.
13. Z. M. Shibzukhov, "Correct aggregate operations with algorithms," *Pattern Recogn. Image Anal.*, **24**, No. 3, 377–382 (2014).
14. Z. M. Shibzukhov, "Aggregating correct operations on algorithms," *Dokl. Ross. Akad. Nauk*, **462**, No. 6, 649–652 (2015).
15. V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag (2000).
16. V. J. Yohai, "High breakdown-point and high efficiency robust estimates for regression," *Ann. Stat.*, **15**, 642–656 (1987).

Z. M. Shibzukhov
Institute of Applied Mathematics and Automation,
Kabardino-Balkar Research Center of the Russian Academy of Sciences, Nalchik, Russia
E-mail: szport@gmail.com