Gradient Computations in the Mini-Batch Gradient Descent with Regularization, DropRule, and AdaBound (MBGD-RDA) Algorithm

Dongrui Wu, *Senior Member, IEEE*
Ministry of Education Key Laboratory of Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China.
Email: drwu@hust.edu.cn.

The detailed MBGD-RDA algorithm for regression problems is given in [1]. This document computes the gradients for different MF shapes. The key notations are summarized in Table I.

TABLE I
KEY NOTATIONS USED IN THIS PAPER.

Notation	Definition
N	The number of labeled training examples
$\mathbf{x}_n = (x_{n,1}, ..., x_{n,M})^T$	M-dimension feature vector of the nth training example. $n \in [1, N]$
y_n	The groundtruth output corresponding to $\mathbf{x}_n$
R	The number of rules in the TSK fuzzy system
$X_{r,m}$	The MF for the mth feature in the rth rule. $r \in [1, R]$, $m \in [1, M]$
$w_{r,0}, ..., w_{r,M}$	Consequent parameters of the rth rule. $r \in [1, R]$
$y_r(\mathbf{x}_n)$	The output of the rth rule for $\mathbf{x}_n$. $r \in [1, R]$, $n \in [1, N]$
$\mu_{X_{r,m}}(x_{n,m})$	The membership grade of $x_{n,m}$ on $X_{r,m}$. $r \in [1, R]$, $m \in [1, M]$, $n \in [1, N]$
$f_r(\mathbf{x}_n)$	The firing level of $\mathbf{x}_n$ on the rth rule. $r \in [1, R]$, $n \in [1, N]$
$y(\mathbf{x}_n)$	The output of the TSK fuzzy system for $\mathbf{x}_n$
L	ℓ_2 regularized loss function for training the TSK fuzzy system
λ	The ℓ_2 regularization coefficient in ridge regression, MBGD-R, MBGD-RA, and MBGD-RDA
M_m	Number of Gaussian MFs in each input domain
N_{bs}	Mini-batch size in MBGD-based algorithms
K	Number of iterations in MBGD training
α	The initial learning rate in MBGD-based algorithms
P	The DropRule rate in MBGD-D, MBGD-RD and MBGD-RDA
β_1, β_2	The exponential decay rates for moment estimates in AdaBound
ϵ	A small positive number in AdaBound to avoid dividing by zero

A. The TSK Fuzzy System for Regression Problems

Assume the input $\mathbf{x} = (x_1, ..., x_M)^T \in \mathbb{R}^{M\times 1}$, and the TSK fuzzy system has R rules:

$$\text{Rule}_r: \text{IF } x_1 \text{ is } X_{r,1} \text{ and } \cdots \text{ and } x_M \text{ is } X_{r,M}, \text{ THEN } y_r(\mathbf{x}) = w_{r,0} + \sum_{m=1}^{M} w_{r,m} x_m, \tag{1}$$

where $X_{r,m}$ ($r = 1, ..., R$; $m = 1, ..., M$) are fuzzy sets, and $w_{r,0}$ and $w_{r,m}$ are consequent parameters.

Let $\mu_{X_{r,m}}(x_m)$ be the membership grade of x_m on $X_{r,m}$. The firing level of Rule_r is:

$$f_r(\mathbf{x}) = \prod_{m=1}^{M} \mu_{X_{r,m}}(x_m), \tag{2}$$

and the output of the TSK fuzzy system is:

$$y(\mathbf{x}) = \frac{\sum_{r=1}^{R} f_r(\mathbf{x}) y_r(\mathbf{x})}{\sum_{r=1}^{R} f_r(\mathbf{x})}. \tag{3}$$

Or, if we define the normalized firing levels as:

$$\bar{f}_r(\mathbf{x}) = \frac{f_r(\mathbf{x})}{\sum_{k=1}^{R} f_k(\mathbf{x})}, \quad r = 1, ..., R \tag{4}$$

then, (3) can be rewritten as:

$$y(\mathbf{x}) = \sum_{r=1}^{R} \bar{f}_r(\mathbf{x}) \cdot y_r(\mathbf{x}). \tag{5}$$

B. Regularization

Assume there are N training examples $\{\mathbf{x}_n, y_n\}_{n=1}^{N}$, where $\mathbf{x}_n = (x_{n,1}, ..., x_{n,M})^T \in \mathbb{R}^{M\times 1}$.

In this paper, we use the following ℓ_2 regularized loss function:

$$L = \frac{1}{2}\sum_{n=1}^{N_{bs}} [y_n - y(\mathbf{x}_n)]^2 + \frac{\lambda}{2}\sum_{r=1}^{R}\sum_{m=1}^{M} w_{r,m}^2, \tag{6}$$

where $N_{bs} \in [1, N]$, and $\lambda \geq 0$ is a regularization parameter. Note that $w_{r,0}$ $(r = 1, ..., R)$ are not regularized in (6).

C. Mini-Batch Gradient Descent (MBGD)

In MBGD, each time we randomly sample $N_{bs} \in [1, N]$ training examples, compute the gradients from them, and then update the antecedent and consequent parameters of the TSK fuzzy system. Let $\boldsymbol{\theta}_k$ be the model parameter vector in the kth iteration, and $\partial L/\partial \boldsymbol{\theta}_k$ be the first-order gradients. Then, the update rule is:

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} - \alpha \frac{\partial L}{\partial \boldsymbol{\theta}_{k-1}}, \tag{7}$$

where $\alpha > 0$ is the learning rate (step size).

When $N_{bs} = 1$, MBGD degrades to the stochastic GD. When $N_{bs} = N$, it becomes the batch GD.

D. MBGD-RDA Using Gaussian MFs

The membership grade of x_m on a Gaussian MF $X_{r,m}$ is:

$$\mu_{X_{r,m}}(x_m) = \exp\left(-\frac{(x_m - c_{r,m})^2}{2\sigma_{r,m}^2}\right), \tag{8}$$

where $c_{r,m}$ is the center of the Gaussian MF, and $\sigma_{r,m}$ the standard deviation.

The gradients of the loss function (6) are given in (9)-(11), where $\Phi(r, m)$ is the index set of the rules that contain $X_{r,m}$, $x_{n,0} \equiv 1$, and $I(m)$ is an indicator function:

$$I(m) = \begin{cases} 0, & m = 0 \\ 1, & m > 0 \end{cases}$$

$I(m)$ ensures that $w_{r,0}$ $(r = 1, ..., R)$ are not regularized.

$$\begin{aligned} \frac{\partial L}{\partial c_{r,m}} &= \frac{1}{2}\sum_{n=1}^{N_{bs}}\sum_{k=1}^{R} \frac{\partial L}{\partial y(\mathbf{x}_n)} \frac{\partial y(\mathbf{x}_n)}{\partial f_k(\mathbf{x}_n)} \frac{\partial f_k(\mathbf{x}_n)}{\partial \mu_{X_{k,m}}(x_{n,m})} \frac{\partial \mu_{X_{k,m}}(x_{n,m})}{\partial c_{r,m}} \\ &= \sum_{n=1}^{N_{bs}} \sum_{k\in\Phi(r,m)} \left[(y(\mathbf{x}_n) - y_n) \frac{y_k(\mathbf{x}_n)\sum_{i=1}^{R} f_i(\mathbf{x}_n) - \sum_{i=1}^{R} f_i(\mathbf{x}_n) y_i(\mathbf{x}_n)}{\left[\sum_{i=1}^{R} f_i(\mathbf{x}_n)\right]^2} f_k(\mathbf{x}_n) \frac{x_{n,m} - c_{r,m}}{\sigma_{r,m}^2} \right] \end{aligned} \tag{9}$$

$$\frac{\partial L}{\partial \sigma_{r,m}} = \frac{1}{2}\sum_{n=1}^{N_{bs}}\sum_{k=1}^{R} \frac{\partial L}{\partial y(\mathbf{x}_n)} \frac{\partial y(\mathbf{x}_n)}{\partial f_k(\mathbf{x}_n)} \frac{\partial f_k(\mathbf{x}_n)}{\partial \mu_{X_{k,m}}(x_{n,m})} \frac{\partial \mu_{X_{k,m}}(x_{n,m})}{\partial \sigma_{r,m}}$$

$$= \sum_{n=1}^{N_{bs}} \sum_{k\in\Phi(r,m)} \left[(y(\mathbf{x}_n) - y_n) \frac{y_k(\mathbf{x}_n) \sum_{i=1}^{R} f_i(\mathbf{x}_n) - \sum_{i=1}^{R} f_i(\mathbf{x}_n) y_i(\mathbf{x}_n)}{\left[\sum_{i=1}^{R} f_i(\mathbf{x}_n)\right]^2} f_k(\mathbf{x}_n) \frac{(x_{n,m} - c_{r,m})^2}{\sigma_{r,m}^3} \right] \tag{10}$$

$$\frac{\partial L}{\partial w_{r,m}} = \frac{1}{2} \sum_{n=1}^{N_{bs}} \frac{\partial L}{\partial y(\mathbf{x}_n)} \frac{\partial y(\mathbf{x}_n)}{\partial y_r(\mathbf{x}_n)} \frac{\partial y_r(\mathbf{x}_n)}{\partial w_{r,m}} + \frac{\lambda}{2} \frac{\partial L}{\partial w_{r,m}} = \sum_{n=1}^{N_{bs}} \left[(y(\mathbf{x}_n) - y_n) \frac{f_r(\mathbf{x}_n)}{\sum_{i=1}^{R} f_i(\mathbf{x}_n)} \cdot x_{n,m} \right] + \lambda I(m) w_{r,m} \tag{11}$$

E. MBGD-RDA Using Trapezoidal MFs

The membership grade of x_m on a trapezoidal MF $X_{r,m}$ is:

$$\mu_{X_{r,m}}(x_m) = \begin{cases} \frac{x_m - a_{r,m}}{b_{r,m} - a_{r,m}}, & x_m \in (a_{r,m}, b_{r,m}) \\ 1, & x \in [b_{r,m}, c_{r,m}] \\ \frac{d_{r,m} - x_m}{d_{r,m} - c_{r,m}}, & x_m \in (c_{r,m}, d_{r,m}) \\ 0, & otherwise \end{cases}, \tag{12}$$

where $a_{r,m} < b_{r,m} \leq c_{r,m} < d_{r,m}$ determine a trapezoidal MF, as shown in Fig. 1.

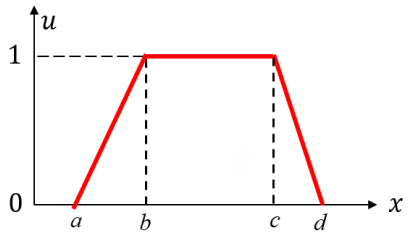

Fig. 1. A trapezoidal MF.

The gradients of the loss function (6) are given in (13)-(17). After updating, the relationship $a_{r,m} < b_{r,m} \leq c_{r,m} < d_{r,m}$ may be violated, so we need to sort them to make sure $a_{r,m} < b_{r,m} \leq c_{r,m} < d_{r,m}$.

$$\frac{\partial L}{\partial a_{r,m}} = \frac{1}{2} \sum_{x_n \in (a_{r,m}, b_{r,m})} \sum_{k=1}^{R} \frac{\partial L}{\partial y(\mathbf{x}_n)} \frac{\partial y(\mathbf{x}_n)}{\partial f_k(\mathbf{x}_n)} \frac{\partial f_k(\mathbf{x}_n)}{\partial \mu_{X_{k,m}}(x_{n,m})} \frac{\partial \mu_{X_{k,m}}(x_{n,m})}{\partial a_{r,m}}$$

$$= \sum_{x_n \in (a_{r,m}, b_{r,m})} \sum_{k\in\Phi(r,m)} \left[(y(\mathbf{x}_n) - y_n) \frac{y_k(\mathbf{x}_n) \sum_{i=1}^{R} f_i(\mathbf{x}_n) - \sum_{i=1}^{R} f_i(\mathbf{x}_n) y_i(\mathbf{x}_n)}{\left[\sum_{i=1}^{R} f_i(\mathbf{x}_n)\right]^2} \frac{f_k(\mathbf{x}_n)}{\mu_{X_{r,m}}(x_{n,m})} \frac{x_{n,m} - b_{r,m}}{(b_{r,m} - a_{r,m})^2} \right] \tag{13}$$

$$\frac{\partial L}{\partial b_{r,m}} = \frac{1}{2} \sum_{x_n \in (a_{r,m}, b_{r,m})} \sum_{k=1}^{R} \frac{\partial L}{\partial y(\mathbf{x}_n)} \frac{\partial y(\mathbf{x}_n)}{\partial f_k(\mathbf{x}_n)} \frac{\partial f_k(\mathbf{x}_n)}{\partial \mu_{X_{k,m}}(x_{n,m})} \frac{\partial \mu_{X_{k,m}}(x_{n,m})}{\partial b_{r,m}}$$

$$= \sum_{x_n \in (a_{r,m}, b_{r,m})} \sum_{k\in\Phi(r,m)} \left[(y(\mathbf{x}_n) - y_n) \frac{y_k(\mathbf{x}_n) \sum_{i=1}^{R} f_i(\mathbf{x}_n) - \sum_{i=1}^{R} f_i(\mathbf{x}_n) y_i(\mathbf{x}_n)}{\left[\sum_{i=1}^{R} f_i(\mathbf{x}_n)\right]^2} f_k(\mathbf{x}_n) \frac{-1}{b_{r,m} - a_{r,m}} \right] \tag{14}$$

$$\frac{\partial L}{\partial c_{r,m}} = \frac{1}{2} \sum_{x_n \in (c_{r,m}, d_{r,m})} \sum_{k=1}^{R} \frac{\partial L}{\partial y(\mathbf{x}_n)} \frac{\partial y(\mathbf{x}_n)}{\partial f_k(\mathbf{x}_n)} \frac{\partial f_k(\mathbf{x}_n)}{\partial \mu_{X_{k,m}}(x_{n,m})} \frac{\partial \mu_{X_{k,m}}(x_{n,m})}{\partial c_{r,m}}$$

$$= \sum_{x_n \in (c_{r,m}, d_{r,m})} \sum_{k\in\Phi(r,m)} \left[(y(\mathbf{x}_n) - y_n) \frac{y_k(\mathbf{x}_n) \sum_{i=1}^{R} f_i(\mathbf{x}_n) - \sum_{i=1}^{R} f_i(\mathbf{x}_n) y_i(\mathbf{x}_n)}{\left[\sum_{i=1}^{R} f_i(\mathbf{x}_n)\right]^2} f_k(\mathbf{x}_n) \frac{1}{d_{r,m} - c_{r,m}} \right] \tag{15}$$

$$\frac{\partial L}{\partial d_{r,m}} = \frac{1}{2} \sum_{x_n \in (c_{r,m}, d_{r,m})} \sum_{k=1}^{R} \frac{\partial L}{\partial y(\mathbf{x}_n)} \frac{\partial y(\mathbf{x}_n)}{\partial f_k(\mathbf{x}_n)} \frac{\partial f_k(\mathbf{x}_n)}{\partial \mu_{X_{k,m}}(x_{n,m})} \frac{\partial \mu_{X_{k,m}}(x_{n,m})}{\partial d_{r,m}}$$

$$= \sum_{x_n \in (c_{r,m}, d_{r,m})} \sum_{k \in \Phi(r,m)} \left[(y(\mathbf{x}_n) - y_n) \frac{y_k(\mathbf{x}_n) \sum_{i=1}^{R} f_i(\mathbf{x}_n) - \sum_{i=1}^{R} f_i(\mathbf{x}_n) y_i(\mathbf{x}_n)}{\left[\sum_{i=1}^{R} f_i(\mathbf{x}_n)\right]^2} \frac{f_k(\mathbf{x}_n)}{\mu_{X_{r,m}}(x_{n,m})} \frac{x_{n,m} - c_{r,m}}{(d_{r,m} - c_{r,m})^2} \right] \tag{16}$$

$$\frac{\partial L}{\partial w_{r,m}} = \frac{1}{2} \sum_{n=1}^{N_{bs}} \frac{\partial L}{\partial y(\mathbf{x}_n)} \frac{\partial y(\mathbf{x}_n)}{\partial y_r(\mathbf{x}_n)} \frac{\partial y_r(\mathbf{x}_n)}{\partial w_{r,m}} + \frac{\lambda}{2} \frac{\partial L}{\partial w_{r,m}} = \sum_{n=1}^{N_{bs}} \left[(y(\mathbf{x}_n) - y_n) \frac{f_r(\mathbf{x}_n)}{\sum_{i=1}^{R} f_i(\mathbf{x}_n)} \cdot x_{n,m} \right] + \lambda I(m) w_{r,m} \tag{17}$$

REFERENCES

[1] D. Wu, Y. Yuan, J. Huang, and Y. Tan, "Optimize TSK fuzzy systems for regression problems: Mini-batch gradient descent with regularization, DropRule and AdaBound (MBGD-RDA)," *IEEE Trans. on Fuzzy Systems*, 2020, in press. [Online]. Available: https://arxiv.org/abs/1903.10951