STOCHASTIC GRADIENT ALGORITHMS FROM ODE SPLITTING PERSPECTIVE – REPRODUCIBILITY CHALLENGE

Yasin Erdogan, Oguz Erol, Arda Berk Kilickaya
ye2n19, oe1n19, abk1u19

ABSTRACT

In this report, we have reproduced the original findings of the paper published by Daniil M. and Ivan o. in ICLR 2020 DeepDiffEq. The main idea behind the original paper is to employ ODE splitting scheme in gradient descent algorithms and to compare this improved algorithm with the classical one. We have analyzed and reproduced experiments conducted in the original paper in this work.

1 INTRODUCTION

The original paper first analyses the similarities between stochastic gradient descent method and gradient flow form of the Euler discretization of the ODE. Since vanilla stochastic gradient descent method consists sequential steps in the direction of the gradient.

$$\theta_{k+1} = \theta_k - h_k \nabla f_i \tag{1}$$

$$\frac{d\theta}{dt} = -\nabla f(\theta) \tag{2}$$

Paper mentions that when dealing with continuous time, usually noise is added to the right-hand side of (2). However, if the full gradient is replaced by its minibatch variant the distribution of noise is not known. Instead the paper proposes a first-order splitting scheme for (2). With this new approach researches believe that the approximation of the full gradient flow could be calculated more efficiently.

In order to demonstrate how first-order splitting scheme corresponds to the SGD the researchers firstly consider simple ODE which consists two summands on its right-hand-side.

$$\frac{d\theta}{dt} = -\frac{1}{2}(g_1(\theta) + g_2(\theta)) \tag{3}$$

For the same purpose the researchers consider and illustrative example of Gradient Flow equation (4), where the right-hand side of ODE is the sum of operators acting of θ.

$$\frac{d\theta}{dt} = -\frac{1}{2}\sum_{i=1}^{2} \nabla f_i(\theta) = -\frac{1}{2}\nabla f_1(\theta) - \frac{1}{2}\nabla f_2(\theta) \tag{4}$$

Table 1: The table describes the correspondence between splitting scheme for discretized Gradient Flow ODE and epoch of SGD

SGD Epoch	**First-order splitting**
$\theta_{SGD} = \theta_0 - h\nabla f_1(\theta_0)$	$\theta_I = \theta_0 - h/2\nabla f_1(\theta_0)$
$\theta_{SGD} = \theta_{SGD} - h\nabla f_2(\theta_{SGD})$	$\theta_I = \theta_I - h/2\nabla f_1(\theta_I)$

2 METHODOLOGY

2.1 OPTIMIZATION STEP WITH ODE SOLVER

The researchers propose to integrate local problem more precisely. The proposed method solves the local ODE problem by replacing gradient in the right-hand side of (3) with batch gradient version. In the paper this new approach has been tested with three different problems.

Problem	Loss Function	Batch Gradient	Initial Local ODE
Linear Least Squares	$f(\theta)=\frac{1}{n}\sum_{i=1}^{m}\| X_i\theta-y_i \|_2^2$	$1/bX_i^T(X_i\theta-y_i)$	$\frac{d\theta}{dt}=-\frac{1}{n}X_i^T(X_i\theta-y_i)$
Binary Logistic Regression	$f(\theta)=-\frac{1}{n}\sum_{i=1}^{n}(y_i\ln\sigma(\theta^T x_i)+(1-y_i)\ln(1-\sigma(\theta^T x_i)))$	$1/bX_i^T(\sigma(X_i\theta)-y_i)$	$\frac{d\theta}{dt}=-\frac{1}{n}X_i^T(\sigma(X_i\theta)-y_i)$
One FC Layer + Softmax	$f(\Theta)=-\frac{1}{n}\sum_{i=1}^{n}\log(\frac{y_i^T e^{\Theta^T x_i}}{1^T e^{\Theta^T x_i}})$	$1/bX_i^T(s(\Theta^T X_i^T)-Y_i)^T$	$\frac{d\theta}{dt}=-\frac{1}{n}X_i^T(s(\Theta^T X_i^T)-Y_i)^T$

Table 2: The table presents ODE, which we need to solve at each step of the algorithm.

Generally, machine learning problems the size of the mini batch (b) is often less than the number of trainable parameters (p), this allows us to implement QR decomposition to reduce dimensionality of the system. QR decomposition only needs to be performed once before the training session.

2.2 UPPER BOUND ON THE GLOBAL SPLITTING ERROR

For calculation of the upper bound on the global splitting error the original paper assumes that we have only two batches and the problem (5) has an exact solution θ_* such as $X\theta_*=y$. The form of the gradient descent can be expressed as (6)

$$f(\theta)=\frac{1}{n}\| X\theta-y \|_2^2=\frac{1}{n}\sum_{i=1}^{s}\| X_i\theta-y_i \|_2^2\to\min_{\theta\in\mathbb{R}^p} \tag{5}$$

$$\frac{d\theta}{dt}=-X^T(X\theta-y)=-X^TX(\theta-\theta_*)=-(X_1^TX_1+X_2^TX_2)(\theta-\theta_*) \tag{6}$$

Here the splitting scheme corresponds to a linear operator splitting

$$A=A_1+A_2,\ A=-X^TX,\ A_i=-X_i^TX_i,\ i=1,2$$

In the above equations A_1 and A_2 are symmetric non-negative definite metrices. If we assume that $\theta_*=0$,

$$A_i=Q_iB_iQ_i^*,$$

where Q_i is an N x r_i matrix with orthonormal columns.

2.3 EXPERIMENTS

Linear Least Squares

For this problem type both random and the real linear systems were tested in the original paper. For random linear systems in order to reduce the computational burden we have reduced the dimensions of the originally generated 10000x500 matrix to 1000x50 with Gaussian noise which has 0.01 magnitude. The real linear system is the standard tomography data from AIRTools II Hansen & Jørgensen (2018). Which constructs 50x50 images from 12780x2500 linear system. With batch sizes 20 and 60 respectively. As for the stopping criterion for the experiments relative error value of 10e-3 was chosen.

Binary Logistic Regression

For this problem type two classes from MNIST LeCun et al. (1998) dataset has been used in the original paper, which corresponds to the 0 and 1 digits. For this experiment the size of the batch is 50 and the test error value as the stopping criterion for the experiment is 10e-3.

Softmax Logistic Regression

For this problem type Fashion MNIST Xiao et al. (2017) dataset with 60000 grayscale pictures from 10 classes has been used. Each example is 28x28 image. For this experiment the size of the batch is 64 and the test error value as the stopping criterion for the experiment is 0.25.

3 RESULTS

3.1 UPPER BOUND ON THE GLOBAL SPLITTING ERROR

For the upper bound on the global splitting error two experiments have been carried out. In the first one there are only two summands in the right-hand side. In the second experiment there 40 summands in the right-hand side. The results of the global error calculation have been given below.

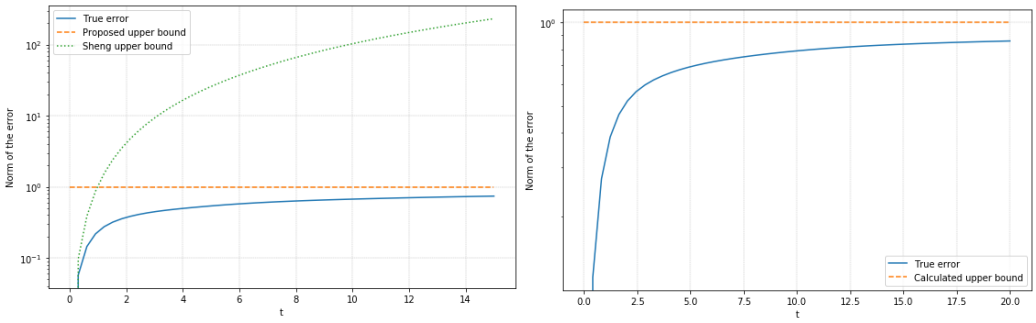

Fig 1. Global error of the splitting scheme for 2 and 40 summands.

As can be seen from the graph above we can see the difference between global upper bound for Sheng (1994) and the derived one.

3.2 EXPERIMENT RESULTS

The experimental results for the problems explained in the section above is given in the figures below

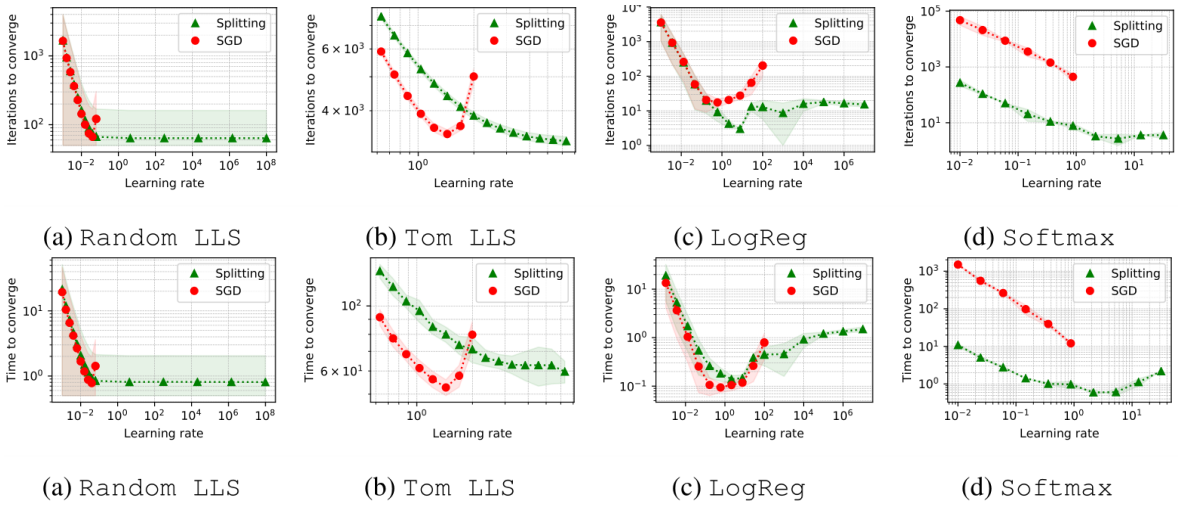

Fig 2. Convergence speed of SGD and the proposed splitting algorithm for different learning rates in different problem types.

As can be seen from the figure above proposed splitting method generally performs better than SGD starting from a specific learning rate value which is different for each problem type.

3.3 DISCUSSION

The original paper is clear in implementation of the proposed method and steps that must be taken to implement them. The proposed method has been coded according to the formulas and the algorithms given in the paper. To reproduce the experiments conducted in the original paper we had to reduce the dimensionality of the generated datasets in order to reduce the computational burden because of hardware limitations. However, since the datasets used in this work is distributed uniformly as in the original paper, we have managed to obtain similar results. The code needed for drawings and the training phase of the model has been taken from the original work. The equations obtained in the original work has been used in this work as it is in the source material.

4 CONCLUSION

In this work we have reproduced the ICLR 2020 paper STOCHASTIC GRADIENT DESCENT ALGORTIHM FROM ODE SPLITTING PERSPECTIVE successfully. The novel idea in the paper is to obtain a better local solver by using the similarities between the gradient flow form of the Euler discretization of the ODE and the stochastic gradient descent method. The performance of the new method proposed in the original paper has been tested in 3 different problem type. In these experiments it has been observed that the proposed splitting scheme can outperform vanilla SGD algorithm when the learning rate reaches a critical point which we were able to recreate successfully. Overall, we can say that the reproduction of the original work was successful.

REFERENCES

Kaczmarz., S. Angen¨aherte aufl¨osung von systemen linearer gleichungen. Bull. Internat. Acad. Polon.Sci.

Kaczmarz., S. Angen¨aherte aufl¨osung von systemen linearer gleichungen. Bull. Internat. Acad. Polon.Sci.