**My paper is 157: On the convergence speed of AMSGRAD and beyond, Tao Tan, Shiqun Yin, Kunling Liu, Man Wan**

[METHODS - TOOLS-1] – {Nov. 4 @ 15.15pm- 17.15pm}

1 Hello everyone:

My name is Kunling Liu, It’s my great to have the opportunity to take part in the meeting. I’m from Southwest University, Chongqing, China. Our paper’s name is On the convergence speed of AMSGRAD and beyond.

2 Now I will introduce this paper from 5 sections as show. Introduction, Related works, ACADG algorithm, Experiments and Conclusion.

3 Firstly, introduction.

4 Nowadays, most learning algorithms in deep learning are based on iterative ideas, and the purpose is to find a set of network parameters, optimize the model weights, and minimize the objective function. The existing optimization algorithms are mainly based on the idea of SGD algorithm, and can be divided into two categories. One is the momentum method, such as Momentum and NAG algorithms. The other is the adaptive method, which mainly includes RMSPROP, ADAGRAD, ADADELTA, ADAM and AMSGRAD algorithms.

5 In this paper, we find two problems in AMSGRAD algorithm. Firstly, the AMSGRAD algorithm is easy to oscillate. Secondly, the AMSGRAD algorithm converges slowly. In order to solve the above two problems, we proposes the ACADG algorithm, which is a new adaptive learning rate optimization algorithm.

6 Secondly, Related work.

7 In physics, people use momentum to simulate the inertia of an object as it moves. Through derivation, it is found that the momentum idea can be used in the deep learning optimization algorithm to improve the convergence speed. Then the Momentum algorithm is proposed, and the update rules of model weights are designed as show, where *w* is the model weights, *t* is the number of iteration steps, *ρ* is the momentum factor, *η* is the learning rate, *g* is the parameter gradient.

8 AMSGRAD algorithm is a new exponential moving average gradient optimization algorithm, and its purpose is to solve the convergence problem of ADAM algorithm. The specific implementation steps of AMSGRAD algorithm are as follows:

1. Calculate the biased first-order and second-order moment estimators of parameter gradient, respectively.
2. Use the biased first-order and second-order moment estimators to calculate the unbiased first-order and second-order moment estimators of parameter gradient, respectively.
3. Record the historical maximum of unbiased second-order moment estimator of parameter gradient.
4. Update the model weights with the following formula.

9 The third section is our algorithm: ACADG.

10 Through the derivation of mathematical formulas, the AMSGRAD algorithm proves that it can overcome the non-convergence problem in ADAM algorithm, and guarantees that the optimization algorithm converges with the increase of the number of iterations. However, this paper finds the following two problems from the convergence process of the AMSGRAD algorithm:

11 Firstly, In order to ensure the convergence of the AMSGRAD algorithm, the AMSGRAD algorithm must record the historical maximum of the unbiased second-order moment estimate of the parameter gradient, and use it to calculate the adaptive learning rate. However, the adaptive learning rate of the AMSGRAD algorithm is lower than that of the ADAM algorithm during the iterative process, which will increase convergence time, reduce convergence speed.

12 Secondly, When *gt−1gt ≤ 0*, the angle between the parameter gradient vector in the previous step and the parameter gradient vector in the current step is greater than 90 degrees. Using the AMSGRAD algorithm to update the model weights will cause the unbiased first-order moment estimate of the parameter gradient in the previous step to have a negative effect on the current parameter gradient vector. Sometimes, this negative effect will last for a long time, and even cause the objective function to experience severe oscillations during convergence.

13 In order to solve the above two problems of AMSGRAD algorithm, this paper designs ACADG algorithm as shown. The ACADG algorithm divides the training process into two cases by computing *gt*

*gt-1*

14 When *gt-1gt > 0*, the momentum term has a positive effect on the first-order moment estimation of the parameter gradient in the current steps. Through the momentum term, the ACADG algorithm can speed up the decline of the loss function, and make the loss function quickly reach a satisfactory local optimal solution. Therefore, the ACADG algorithm accelerates the convergence speed of the model.

15 When *gt-1gt ≤ 0*, whether the AMSGRAD algorithm or the ADAM algorithm, the first-order moment estimation of the parameter gradient in the previous steps will have a negative effect on the parameter gradient vector in the current steps. Sometimes, this negative effect will last for a long time, and even cause the objective function to experience severe oscillations during convergence. In this case, the ACADG algorithm uses the SGD algorithm to update the model weights. Regardless of the negative effect, the ACADG algorithm will not be affected. Therefore, the ACADG algorithm reduces the oscillation amplitude of the loss function.

16 The fourth section is experiments.

17 In order to verify the convergence, convergence speed and oscillation amplitude of ACADG algorithm in synthetic experiment. This paper designs two different objective functions as shown.

18 Figure 1 is the performance of the three algorithms on the synthetic data.

19 As can be seen from Figure1, the three algorithms have the following three phenomena in the first objective function and the second objective function:

1. As the number of iterations increases, the ADAM algorithm converges to 1, but both the AMSGRAM algorithm and ACADG algorithm converge to -1.
2. The convergence speed of the objective function in the ACADG algorithm is significantly faster than that of the AMSGRAD algorithm.
3. During the convergence process, the oscillation amplitude of the AMSGRAD algorithm is much larger than that of the ACADG algorithm.

20 In order to verify the performance of ACADG algorithm on convex optimization problem and non-convex optimization problem. This paper uses the Mnist data set to perform experiments on logistic regression model and DNN model, respectively. The model and experiments results are as shown.

21 Figure2 is the training performance of the three algorithms on logistic regression model

22 Figure3 is the test performance of the three algorithms on logistic regression model

23 Figure4 is the training performance of the three algorithms on DNN model

24 Figure5 is the test performance of the three algorithms on DNN model.

25 From the above figure, whether it is a convex optimization problem or a non-convex optimization problem, the ACADG algorithm is superior to AMSGRAD and ADAM algorithms in terms of the convergence speed, the oscillation amplitude, the accuracy of training and test data sets.

26 In order to verify the performance of ACADG algorithm on complex neural networks. This paper performs experiment on CNN model using the Mnist and Cifar-10 data sets, respectively.The model and experiments results are as shown.

27 Figure6 is the Mnist training performance of the three algorithms on CNN model

28 Figure7 is the Mnsit test performance of the three algorithms on CNN model

29 Figure8 is the Cifar-10 training performance of the three algorithms on CNN model

30 Figure9 is the Cifar-10 test performance of the three algorithms on CNN model.

31 From the above figure, whether it is Mnist or Cifar-10, the ACADG algorithm is superior to AMSGRAD and ADAM algorithms in terms of the convergence speed, the oscillation amplitude, the accuracy of training and test data sets.

32 The last is conclusion.

33 The main contributions of this paper are as follows:

1. This paper finds that there are two problems in the convergence process of AMSGRAD algorithm, one is oscillation amplitude, the other is convergence speed.
2. This paper proposes ACADG algorithm, which is an adaptive gradient optimization algorithm. The ACADG algorithm not only can improve the convergence speed, suppress the oscillation amplitude of objective function, but also can improve the accuracy of training and test data sets.
3. Through some comparative experiments, it is found that ACADG algorithm is superior to AMSGRAD and ADAM algorithms in the convergence speed, the oscillation amplitude, the accuracy of training and test data sets, whether it is convex optimization problems or non-convex optimization problems.

34 That’s all. Thank you.