**Revision Report of Submission NEUCOM-D-16-03093**

We thank the comments with cares and insights made by the reviewers, which are helpful for improving the quality and readability of our paper. In our revised paper, we have made detailed explanations and changes in response to all the reviewers’ comments. Here, we explain our revisions based on every comments and suggestions.

**Response to Reviewer 1’s Comments**

1. **Comments:** *The proposed method is described in Algorithms 1-3. As I understand,* *Algorithm 2 is run on each server, and Algorithm 3 is run on each worker. However, it is not clear on which node(s) Algorithm 1 is run: maybe, a central node collecting information from all the servers?*

**Response:** Thanks for your careful reading. Yes, Algorithm 2 is run on each server, and Algorithm 3 is run on each worker. Algorithm 1 presents the details of our proposed variance reduced SGD, but it does not contain the specific details in the distributed settings, e.g. the distributed communication and aggregation etc. In other words, it is mainly used to present the logical design of the proposed distributed variance reduced SGD. We have reorganized Section 4 and polished the presentation of Algorithm 1 in Section 4.1 carefully.

1. **Comments:** *Moreover, how many servers are used by the method? Does each server correspond with a cluster?*

**Response:** Thanks for your careful reading. There is no limitation on the number of the servers. It is OK for our methods when one or multiple servers are used. As illustrated in Figure 1, we use multiple servers in our implementation, that is, more than one severs are used in a cluster. One of the reasons is to reduce the overloads of the communication links between the workers and the servers. If there is only one server but many workers, the communication overloads will be extremely large for the links which are close to the server. Therefore, the communication time will be large due to the heavy overload, which increases the wait time for the workers. By contrast, if multiple servers are used, the time consumption due to the communication will be decreased.

Inspired by your suggestion, we present more details about the architecture of our method in Section 3.2. There are a server group and a worker group in our design. Either the server group or the worker group consists of one or multiple machines. All the servers in the server group maintain a globally shared parameter table. The workers pull the global parameters from one of the servers in the group, and push its corresponding update to one of the servers. When the servers receive those updates, they aggregate them into the global parameter.

1. **Comments:** *Finally, some formulas in the three algorithms should be checked carefully, and possibly corrected: for instance, what is* *omega\_i in Algorithm 1? I.e., to which iteration does it refer?*

**Response:** Thanks for your careful reading. We have revised Algorithm 1 in Section 4.1 carefully. In the previous submission,  should be corrected to be . It is the initial parameter in an epoch, which is used to obtain a stale full gradient. Additionally, all the symbols used in the new revision are checked carefully. To make it clear, we present all the symbols and their meanings in Section 3.1.

1. **Comments:** *Formula (1) refers to a machine learning problem without regularization, which is justified for huge training sets. However, it would be worth discussing if the proposed method can be still applied to the case of machine learning problems with regularization, such as the ones described in:*

*Cucker, Smale, “On the mathematical foundations of learning”, Bullettin of the American mathematical society, vol. 39, no. 1, pp. 1-49, 2001.*

*Gnecco, Gori, Melacci, Sanguineti, “Learning with mixed hard/soft pointwise constraints”, IEEE Transactions on Neural Networks and Learning Systems, vol. 26, no. 9, pp. 2019-2032, 2015.*

**Response:** Thanks for your careful reading. We have revised Formula 1 and added the regularization item in the optimization objective. Besides, we have reorganized Section 1 and presented some examples of the regularization items in the second paragraph. The L1 and L2 regularization items are presented as the illustrative examples. Some other regularizations can be added in our optimization objective, but it is out of the scope of the paper, and we leave it as the future work.

1. **Comments:** *The paper should explain more clearly why the proposed algorithm computes a variance-reduced gradient: does this depend on an averaging effect in step 4 of Algorithm 1? There is such a sort of explanation at the end of Section 4, but it should be anticipated.*

**Response:** Thanks for your careful reading. The step 4 of Algorithm 1 in the previous submission is used to compute the full gradient for reducing the variance when the parameter is updated at every iteration. Since the stochastic gradient usually leads to much stochastic noise, it is vitally important to use the full gradient to decrease the noise, and accelerate the convergence of SGD.

We have revised Section 3, and added Section 3.2 to present more details of the variance reduced SGD.

1. **Comments:** *Finally, the following ar**e corrections to some typos/errors:*

- p. 1: "iterative convergent" -> "iteratively convergent";

- p. 2: "can be described the optimization problem" -> "can be described by the optimization problem"; moreover, the comma after formula (1) should be moved above, likewise the dot after formula (2);

- p. 3: "In this paper. We design" -> "In this paper, we design"; "It is noting" -> "It is worth noting";

- p. 4: "i.e.,PetuumSGD" -> "i.e., PetuumSGD";

- p. 5: "all the servers maintains" - "all the servers maintain"; "which is called" -> "which are called";

- p. 6: "send a copy of the global parameters": to which node? to a central node, or to the workers?

- p. 8: "at a constant" -> "at a constant rate"; "converges faster" -> "converge faster";

- p. 10: "is appropriate" -> "are appropriate";

- p. 11: "have designed" -> "have been designed"; "to solve this problem. Those variants" -> "to solve this problem, those variants"; "to proposing" -> "to propose";

- p. 17: "We concludes" -> "We conclude".

**Response:** Thanks for your careful reading. Those errors have been revised in the new submission. We have polished the paper carefully, and those errors and typos have been corrected in the new version of the paper.

**Responses to Reviewer 2's Comments**

1. **Comments:** *I think the paper could contain some publishable results, especially, considered the experimental section where the proposed method seems to show good performance. However, sections 2, 3, 4, 5 (the core of the work) are really too short. The authors must explain in more detail the analysis and solution proposed, because, in the present form, the article resembles more a technical report rather than a scientific journal paper.*

**Response:** Thanks for your careful reading. We have reorganized and enriched the content of the paper in the new submission.

We have added more related work in the Section 2. Some new related researches such as FlexRR have been included. The discussion between our method and the most related method named SSGD has been enriched. Besides, the new variants of the variance reduced SGD such as SAGA, S2GD, SVRG++ have been presented in the new version.

We have reorganized the paper and added the preliminaries in Section 3 in the new submission. First, all the symbols and their meanings used in the paper are presented in Section 3.1 in order to make the paper more clearly and understandable. Second, the parameter server architecture in the Section 3.2 is presented. We give many explanations about the parameter server system and the architecture of our method. Besides, we present the variance reduced SGD in Section 3.3. The advantage of the variance reduced gradient has been shown. In specific, the variance reduced SGD uses a variance reduced gradient to reduce the stochastic noise during the update of the parameter. It reduces the stochastic noise significantly, but leads to less computational cost. Extensive empirical studies show that the variance reduced gradient leads to the comparable computational cost of SGD, but obtains the equivalent convergence performance of gradient descent.

We have reorganized Section 4, and enriched the presentation of our method. Our method named DisSVRG is a distributed algorithm which runs on multiple machines in a cluster. The machines in the clusters are partitioned into two groups, i.e. server and worker group. In the server group, the servers maintain a globally shared parameter table, and update the global parameters. In the worker group, the workers pull the parameters from the server group and save a copy of the parameters. During the update of the parameters, every worker uses a variance reduced gradient to reduce the stochastic noise. After that, every worker pushes their update to the server group, and waits for pulling the newest parameters from the server.

1. **Comments:** *Concerning minor points, I noticed that often the authors insert commas "," and full stops "." after a display math formula in the following line, but this is incorrect.*

**Response:** Thanks for your careful reading. We have revised the paper carefully, and revised the errors in all formulas in the new submission. The presentation of the math formulas in the new version have been polished carefully.

1. **Comments:** *F**urthermore, I suggest to replace the mathematical signs indicating "for every" and "ther**e exists" with the corresp**onding few words.*

**Response:** Thanks for your careful reading. All the mathematical signs indicating “for every” and “there exists” have been replaced by the corresponding words.