View Reviews

Paper ID

3001

Paper Title

Convergent Adaptive Gradient Methods in Decentralized Optimization

Reviewer #2

Questions

1. Please summarize the main claim(s) of this paper in two or three sentences.

The paper proposes a method for converting a centralized adaptive gradient algorithm into a convergent incarnation of its consensus-based decentralized version. It performs the convergence analysis under the generic setting, and then specifically for AMSgrad.

2. Merits of the Paper. What would be the main benefits to the machine learning community if this paper were presented at the conference? Please list at least one.

The community would learn about a new method to ensure convergence under the decentralized consensus optimization setting for adaptive gradient methods, and be aware of the convergence problem when the adaptive learning rates in different nodes are far from each other.

3. Please provide an overall evaluation for this submission.

Borderline paper, but has merits that outweigh flaws.

4. Score Justification Beyond what you've written above as "merits", what were the major considerations that led you to your overall score for this paper?

Positive:
* The paper proposes the sound idea of applying consensus to the adaptive learning rates in addition to the optimization variable, and analytically demonstrates convergence.
* The empirical improvement in convergence under heterogeneous data distribution is interesting and potentially practically relevant.

Negative:
* Convergence requires certain conditions on the centralized version of the adaptive gradient algorithm, and it seems that the method may not converge for all possible adaptive gradient methods, including Adam and RMSprop (see detailed comments). However, the paper gives the impression that the method would work in decentralizing any adaptive gradient algorithm.
* Numerical results are limited, which casts doubt into the empirical success of the proposed method. Experiments over larger networks and comparison with respect to more baselines are needed (see detailed comments).

5. Detailed Comments for Authors Please comment on the following, as relevant: - The significance and novelty of the paper's contributions. - The paper's potential impact on the field of machine learning. - The degree to which the paper substantiates its main claims. - Constructive

criticism and feedback that could help improve the work or its presentation. - The degree to which the results in the paper are reproducible. - Missing references, presentation suggestions, and typos or grammar improvements.

POST-AUTHOR-RESPONSE:

I thank the authors for their response, which addresses my concerns. I still favor accepting the paper, however I will not be able to increase my score because of the limited scope of the experiments (single experiment, small neural network, small dataset).

=====

The paper explores a reasonable and promising idea, and after specific improvements and clarifications can become a compelling paper. Specific comments:

* In Theorem 2, convergence requires that the last term in (3) grows sub-linearly with respect to iterations. It is expected that this is satisfied for AMSgrad due to decreasing effective learning rate, but it is not clear that this will be satisfied for Adam or RMSprop (most likely it won't). This means that the authors' claim that this is a general framework for converting any adaptive gradient method to its decentralized counterpart seems unsubstantiated, and should be toned down.

* Experiments should definitely be expanded. First, the 5-node ring (the current experiment setup) might be too small to truly see the effect of consensus of adaptive learning rates, since the effect of the update on one node will propagate to all the others in at most two steps, which makes things easier. On a larger network, learning rate consensus might be slower and hurt convergence. This effect should be explored better. Second, since plain AMSgrad ensures monotonic decrease of the adaptive learning rate, it implicitly brings the adaptive learning rates closer together. It is not clear how much of the gain that the authors are observing under heterogeneous data is due to the switch from Adam to AMSgrad, and how much of it is due to their decentralized method. The plain decentralized AMSgrad (without learning rate consensus) as well as authors' method + Adam should be explored empirically, for the paper to be convincing.

* The main idea in the paper (consensus of adaptive learning rates) should be emphasized from the beginning, and the DADAM divergence example should clearly be placed in this context, along with a demonstration of how learning rate consensus would have fixed the convergence problem.

* The language throughout the paper is very poor, and hampers readability. The grammatical errors should be fixed.

6. Please rate your expertise on the topic of this submission, picking the closest match.

I have published one or more papers in the narrow area of this submission.

7. Please rate your confidence in your evaluation of this paper, picking the closest match.

I tried to check the important points carefully. It is unlikely, though possible, that I missed something that could affect my ratings.

8. Datasets If this paper introduces a new dataset, which of the following norms are addressed? (For ICML 2020, lack of adherence is not grounds for rejection and should not affect your score; however, we have encouraged authors to follow these suggestions.)

This paper does not introduce a new dataset (skip the remainder of this question).

12. I agree to keep the paper and supplementary materials (including code submissions and Latex source) confidential, and delete any submitted code at the end of the review cycle to comply with the confidentiality requirements.

Agreement accepted

13. I acknowledge that my review accords with the ICML code of conduct (see https://icml.cc/public/CodeOfConduct).
Agreement accepted

Reviewer #3

Questions

1. Please summarize the main claim(s) of this paper in two or three sentences.
The authors derive the decentralized adaptive stochastic method by incorporating the adaptive stochastic gradient with a decentralized training technique. The convergence of the proposed algorithm is also provided.

2. Merits of the Paper. What would be the main benefits to the machine learning community if this paper were presented at the conference? Please list at least one.
The authors provide the first decentralized adaptive methods.

3. Please provide an overall evaluation for this submission.
Borderline paper, but the flaws may outweigh the merits.

4. Score Justification Beyond what you've written above as "merits", what were the major considerations that led you to your overall score for this paper?
1. This paper directly incorporates the adaptive stochastic gradient with a decentralized training technique, which limits the novelty of this work. In particular, In Algorithm 2, the adaptive learning rate (in line 9) is estimated with a delayed $\hat{v}_{t-1}$. Hence, the adaptive learning rate $1/\sqrt{ \hat{u}_{t-1}}$ is independent with m_{t} with stochastic variable ξ_{t}, which makes the convergence of Algorithm 2. However, how did the authors guarantee the delayed $\hat{v}_{t-1}$ can contribute to faster convergence results?
2. In Algorithm 3 the adaptive learning rate $1/\sqrt{ \hat{u}_{t-1}}$ is estimated with delayed $\hat{v}_{t-1}$. When the number of nodes reduces to 1 (N=1), the proposed algorithms cannot reduce to AMSGrad. Hence, it is not appropriate to call Algorithm 3 as decentralized AMSGrad.
3. To establish the convergence of Algorithm 2 and Algorithm 3, hyperparameters ϵ are introduced (Algorithm 2 line 8, Algorithm 9). Actually, the adaptive learning rate is highly affected by ϵ. For example, if ϵ is sufficiently large, both Algorithm 2 and Algorithm 3 reduce to standard decentralized SGD. If ϵ is sufficiently small, the convergence rates of Algorithm 2 and Algorithm 3 are poor. We recommend the authors apply the proposed algorithm to solve the counter-examples in [1] to show the sensitivity of ϵ.
4. The experiments are too simple to show the effectiveness of the proposed algorithms.

[1] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. arXiv preprint arXiv:1904.09237, 2019.

5. Detailed Comments for Authors Please comment on the following, as relevant: - The significance and novelty of the paper's contributions. - The paper's potential impact on the field of machine learning. - The degree to which the paper substantiates its main claims. - Constructive criticism and feedback that could help improve the work or its presentation. - The degree to which the results in the paper are reproducible. - Missing references, presentation suggestions, and typos or grammar improvements.
see the above comments.

6. Please rate your expertise on the topic of this submission, picking the closest match.

I have published one or more papers in the narrow area of this submission.

7. Please rate your confidence in your evaluation of this paper, picking the closest match.
I am very confident in my evaluation of the paper. I read the paper very carefully and I am very familiar with related work.

8. Datasets If this paper introduces a new dataset, which of the following norms are addressed? (For ICML 2020, lack of adherence is not grounds for rejection and should not affect your score; however, we have encouraged authors to follow these suggestions.)
This paper does not introduce a new dataset (skip the remainder of this question).

12. I agree to keep the paper and supplementary materials (including code submissions and Latex source) confidential, and delete any submitted code at the end of the review cycle to comply with the confidentiality requirements.
Agreement accepted

13. I acknowledge that my review accords with the ICML code of conduct (see https://icml.cc/public/CodeOfConduct).
Agreement accepted

Reviewer #4

Questions

1. Please summarize the main claim(s) of this paper in two or three sentences.
The paper considers decentralized optimization with adaptive gradient methods. They propose an algorithm that is convergent to a stationary point in the non-convex case. They also provide numerical experiments.

2. Merits of the Paper. What would be the main benefits to the machine learning community if this paper were presented at the conference? Please list at least one.
Decentralized optimization is an important area in distributed learning since it is a communication-efficient method for training learning models without the use of a parameter server.

3. Please provide an overall evaluation for this submission.
Below the acceptance threshold, I would rather not see it at the conference.

4. Score Justification Beyond what you've written above as "merits", what were the major considerations that led you to your overall score for this paper?
Here are the main issues:
- I don't see any advantage in using adaptive methods in decentralized optimization. The achieved rate (O(1/sqrt(T))) is still the same as DGD.
- There are no results for the convex/strongly convex case.
- Do you have theoretical result on the consensus term itself and its rate of convergence? Without proving consensus the result is meaningless in non-convex setting. By that I mean can we show that \|x(i)_T - \bar{x}_T\| -> 0 as T gets large, and with what rate? (\bar{x} is the average of x(i)s)
- The gradient bounded assumption is too strong for an unconstrained optimization problem. Most results in the literature try to avoid this assumption and that is indeed a major technical challenge that recent papers in decentralized optimization try to resolve.

- In the experimental results as well, there seems to be no real gain compared to DGD. What is the motivation then?
- MNIST data set is too simple. For a conference like ICML, it is better to use CIFAR or ImageNet ...

5. Detailed Comments for Authors Please comment on the following, as relevant: - The significance and novelty of the paper's contributions. - The paper's potential impact on the field of machine learning. - The degree to which the paper substantiates its main claims. - Constructive criticism and feedback that could help improve the work or its presentation. - The degree to which the results in the paper are reproducible. - Missing references, presentation suggestions, and typos or grammar improvements.

Please see above.

6. Please rate your expertise on the topic of this submission, picking the closest match.

I have published one or more papers in the narrow area of this submission.

7. Please rate your confidence in your evaluation of this paper, picking the closest match.

I tried to check the important points carefully. It is unlikely, though possible, that I missed something that could affect my ratings.

12. I agree to keep the paper and supplementary materials (including code submissions and Latex source) confidential, and delete any submitted code at the end of the review cycle to comply with the confidentiality requirements.

Agreement accepted

13. I acknowledge that my review accords with the ICML code of conduct (see https://icml.cc/public/CodeOfConduct).

Agreement accepted