← Back to **Author Console** (/group?id=NeurIPS.cc/2021/Conference/Authors#your-submissions)

Convergence and Alignment of Gradient Descent with Random Backpropagation Weights

PDF **(/pdf?id=1QhRTsqYPB)**

Anonymous

21 May 2021 (modified: 28 May 2021) NeurIPS 2021 Conference Blind Submission Readers: Conference, Paper10585 Senior Area Chairs, Paper10585 Area Chairs, Paper10585 Reviewers, Paper10585 Authors

Abstract: Stochastic gradient descent with backpropagation is the workhorse of artificial neural networks. It has long been recognized that backpropagation fails to be a biologically plausible algorithm. Fundamentally, it is a non-local procedure---updating one neuron's synaptic weights requires knowledge of synaptic weights or receptive fields of downstream neurons. This limits the use of artificial neural networks as a tool for understanding the biological principles of information processing in the brain. Lillicrap et al. (2016) propose a more biologically plausible "feedback alignment" algorithm that uses random and fixed backpropagation weights, and show promising simulations. In this paper we study the mathematical properties of the feedback alignment procedure by analyzing convergence and alignment for two-layer networks under squared error loss. In the overparameterized setting, we prove that the error converges to zero exponentially fast, and also that regularization is necessary in order for the parameters to become aligned with the random backpropagation weights. Simulations are given that are consistent with this analysis and suggest further generalizations. These results contribute to our understanding of how biologically plausible algorithms might carry out weight learning in a manner different from Hebbian learning, with performance that is comparable with the full non-local backpropagation algorithm.

Supplementary Material: ⤓ zip (/attachment?id=1QhRTsqYPB&name=supplementary_material)

Submission History: No

Revealed to Ganlin Song, Ruitu Xu, John Lafferty

21 May 2021 (modified: 28 May 2021) NeurIPS 2021 Conference Submission

Authors: *Ganlin Song (/profile?id=~Ganlin_Song1), Ruitu Xu (/profile?id=~Ruitu_Xu1), John Lafferty (/profile?id=~John_Lafferty2)*

Checklist: Yes, we completed the NeurIPS 2021 paper checklist, and have included it in our PDF.

Code Of Conduct: I certify that all co-authors of this work have read and commit to adhering to the NeurIPS Statement on Ethics, Fairness, Inclusivity, and Code of Conduct.

Add **Official Comment** **Withdraw**

Author Discussion **Committee Discussion** **All**

Reply Type: all Author: everybody Visible To: all readers **5 Replies**

Hidden From: nobody

[−] Official Review of Paper10585 by Reviewer giv6

NeurIPS 2021 Conference Paper10585 Reviewer giv6

Official Review 19 Jul 2021 👁 Program Chairs, Paper10585 Senior Area Chairs, Paper10585 Area Chairs, Paper10585 Reviewers Submitted, Paper10585 Authors

Summary:

This paper studies the convergence of Feedback Alignment, and proves that in the over-parameterized setting the error converges to zero exponentially fast. This is made possible by taking inspirations from recent work on the NTK, albeit with specific challenges as some quantities are not obviously positive semi-definite. Furthermore, it makes the somewhat surprising finding that regularization helps alignment, proving this in linear networks, and confirming this finding in simulations on Gaussian data and MNIST.

Main Review:

For ease of answering, I have annotated my various points with O.1/Q.1/etc...

This is a solid and very well written paper, with a clear contribution to the field of alternative training methods.

The paper could be better positioned in the existing literature, in particular by mentioning all the works that has been done around Direct Feedback Alignment, such as Refinetti et al., 2020. Moreover, I think the surprising finding around regularization & alignment should be discussed more in depth.

Accordingly, **this paper currently stands marginally above the acceptance threshold (6)** . If the two points above were to be addressed, I would be willing to increase my score to a clear accept (8).

Originality

The paper is an interesting application of NTK methods to an open-problem in alternative training methods, the convergence of Feedback Alignment. It is well placed in the immediate literature, but misses a number of important papers on the theory of such methods.

O.1 : Notably, the authors completely ignore the large body of literature around Direct Feedback Alignment (Nøkland, 2016). DFA is an extension of FA, where the random projection of the error is directly sent to each layer, enabling parallelization of weight updates after the loss is calculated. In particular, Refinetti et al., 2020 have done an extensive analysis of the dynamics of convergence & alignment in DFA. This analysis brings a number of important findings around alignment, which it would be interesting to see discussed and compared with the work here done. Furthermore, Frenkel et al. 2019 have also provided a theoretical analysis for a variant of DFA, Direct Random Target Projection. To go a bit further, this could also lead to bridging the gap between some of the theoretical findings made here and the real-world behaviors of FA/DFA: open questions remain, such as FA/DFA failing on convolutions (Bartunov et al., 2018) but working in many other architectures (Launay et al., 2020). A discussion of the impacts of the findings made in this paper in terms of our general understanding of FA/DFA would be very valuable.

Quality

The theoretical contributions are sound and well backed. The experiments to confirm the findings around regularization are interesting, although I think the authors do not go in depth enough in their analysis.

Q.1 : In results both on synthetic data and MNIST (Figure 2 & Figure 3), it is interesting to see that despite lower alignment in non-regularized nets, end-task performance is not necessarily improved by higher alignment values (the lambda = 0.0 of Figure 3 probably overfits). This warrants a more in depth discussion. Could the higher alignment values be an artifact of the weights being pushed closer to zero? (This is noted by the authors l185.) This would motivate disentangling the components of alignment due to L2 regularization, and the "true alignment" of the raw gradients. This would be really interesting to measure and plot, and could provide further elements to understand this "surprise" of regularization helping alignment. The authors also briefly mention experiments with dropout, highlighting that a study of other regularization methods (dropout, etc.) could also prove interesting.

Clarity

The paper is well written and the mathematical notations clear. The structure is also good, making it an enjoyable paper to read.

Significance

The findings of the authors are interesting, and contribute to a broader understanding of feedback alignment methods.

I do feel that the addition of a discussion of the potential impacts of these findings on the current understand of FA/DFA, as well as a more in depth discussion around the "surprising" regularization finding would help make this paper more impactful for the community.

Limitations And Societal Impact:
Yes, the authors accurately describe the limitations and potential impact of their work.

Needs Ethics Review: No

Time Spent Reviewing: 3

Rating: 6: Marginally above the acceptance threshold

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: While performing my duties as a reviewer (including writing reviews and participating in discussions), I have and will continue to abide by the NeurIPS code of conduct.

Add **Official Comment**

[−]
Official Review of Paper10585 by Reviewer 9xDE

NeurIPS 2021 Conference Paper10585 Reviewer 9xDE

Official Review 17 Jul 2021 👁 Program Chairs, Paper10585 Senior Area Chairs, Paper10585 Area Chairs, Paper10585 Reviewers Submitted, Paper10585 Authors

Summary:
The paper uses analysis techniques from Neural Tangent Kernels to prove that the feedback alignment algorithm converges in neural networks with a single non-linear hidden layer. In addition, they prove that the forward weights do not converge to alignment with the random backward weights.

Main Review:
This is a well-written paper that appears to make a significant contribution to the understanding of the feedback aliignment / random backpropagation. Previous work has shown that this algorithm converges for special cases (deep chains of linear neurons, etc.), but this proof adresses a much more general case. However, I was not able to check the proof due to time constraints.

The second result about the weights failing to align is less surprising, but still interesting. A recent ICML paper by Refinetti, et al. "Align, then memorise: the dynamics of learning with feedback alignment" provides an analysis that is complementary to the results presented here.

Limitations And Societal Impact:
The authors adequately address this issue.

Needs Ethics Review: No

Time Spent Reviewing: 1

Rating: 7: Good paper, accept

Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

Code Of Conduct: While performing my duties as a reviewer (including writing reviews and participating in discussions), I have and will continue to abide by the NeurIPS code of conduct.

Add **Official Comment**

[−]
Official Review of Paper10585 by Reviewer 4iby

NeurIPS 2021 Conference Paper10585 Reviewer 4iby

Official Review 16 Jul 2021 👁 Program Chairs, Paper10585 Senior Area Chairs, Paper10585 Area Chairs, Paper10585 Reviewers Submitted, Paper10585 Authors

Summary:

The paper provides a theoretical analysis of the feedback alignment algorithm (FA), a biologically plausible approximation to backpropagation. The authors consider a two-layered network, and derive training error bounds and error alignment for a variety of cases, using the neural tangent kernel method. The results show that FA-trained network can achieve zero training error for infinitely wide networks, and have some (strictly positive) degree of alignment between backprop and FA error vectors when proper regularization is used.

Main Review:
Originality: The paper leverages the theory used for neural tangent kernels, but adapts it in a non-trivial way for feedback alignment. The results presented in the paper are new, and the related work is adequately cited.

Quality: The paper is technically sound.

Clarity: Overall, the paper is well written. I would suggest explaining spectral properties of $G(0)$ and $H(0)$ (around assumption 3.1) in more detail. One option is to mention the result of lemma A.3, and explain intuitively why $G(0)$ has large eigenvalues, and $H(0)$ -- small, as it is important for the whole method. I also disagree with the last line of the abstract, which says “with performance that is comparable with the full non-local backpropagation algorithm” -- we know that FA doesn’t perform well on hard tasks (see Akrout et.al. 2019 paper cited in this work).

Significance: The theoretical results of this paper are important for our understanding of feedback alignment. The results are somewhat limited as we know that FA fails on hard tasks (see Akrout et.al. 2019 mentioned above), but the presented method (and the NTK framework in general) might be useful for studying other alternatives to backprop.

Limitations And Societal Impact:
The authors have adequately addressed the limitations and potential negative societal impact of their work.

Needs Ethics Review: No

Time Spent Reviewing: 3

Rating: 8: Top 50% of accepted NeurIPS papers, clear accept

Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

Code Of Conduct: While performing my duties as a reviewer (including writing reviews and participating in discussions), I have and will continue to abide by the NeurIPS code of conduct.

Add **Official Comment**

[−]

Official Review of Paper10585 by Reviewer pXMq

NeurIPS 2021 Conference Paper10585 Reviewer pXMq

Official Review 12 Jul 2021 Program Chairs, Paper10585 Senior Area Chairs, Paper10585 Area Chairs, Paper10585 Reviewers Submitted, Paper10585 Authors

Summary:
This paper builds upon the work of Lillicrap et al 2016 of attempting to find biologically plausible methods for implementing backpropagation. They thoroughly describe the issues with non-local information and why there are limits to using backprop-based learning to understand more about biological neural systems. They describe the algorithm of feedback alignment using random back-propagated weights. The authors go on to confirm some of the results of the Lillicrap paper, and find a surprising and novel result that regularization is required to produce convergence in over-parameterized settings. The experimental section of the paper shows some work on a more standard ML benchmark task of MNIST classification. They show the results that regularization is required to get good performance on the task.

Main Review:
Overall, I believe this area of research, finding biologically plausible mechanisms, is an important direction for the field. This will not only help us understand biological systems better, but also aid in the development of technologies that are more scalable and efficient. The Lillicrap work was the introduction of a new concept, and this paper attempts to explain why and when the concept of random BP weights applies. However, I don’t feel the authors really did enough in this paper to cross the threshold of a minimum publishable unit of work. The main result of showing that regularization is required for convergence is an interesting result, but there is no theoretical basis as to why this is happening. I’m ok

with empirical results, but they haven't demonstrated that this technique can really work on larger problems either. Most of the paper is devoted to explaining the Lillicrap work and providing a notation. In my view, either more empirical results or more well-motivated concepts are required to publish this work.

The quality of the writing is excellent, and the background provides a good platform to build from. The explanations are clear and concise. The experimental work is a bit sparse. For instance, in figure 3 there is no discussion on why the classification performance on MNIST is best with lambda of 0.1 rather than 0.3. How does the alignment relate to classification performance? Why might the performance drop when alignment clearly increases?

The significance of the work is the main issue I have. If the significance of the concepts presented were of much great magnitude, I would be ok with the level of rigor in the paper. However, with an extension/clarification of a concept, there must an explanation of what new questions the work opens up.

Limitations And Societal Impact:
The authors have not addressed this issue directly. They highlighted that a better understanding of biologically plausible algorithms will lead to a better understanding of the brain. I don't believe there are immediate negative consequences of this work, but the authors could choose to include a statement making this explicit.

Ethical Concerns:
None

Needs Ethics Review: No

Time Spent Reviewing: 4 hours

Rating: 4: Ok but not good enough - rejection

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: While performing my duties as a reviewer (including writing reviews and participating in discussions), I have and will continue to abide by the NeurIPS code of conduct.

Add **Official Comment**

[−]

Official Review of Paper10585 by Reviewer B3Jb

NeurIPS 2021 Conference Paper10585 Reviewer B3Jb

Official Review 29 Jun 2021 Program Chairs, Paper10585 Senior Area Chairs, Paper10585 Area Chairs, Paper10585 Reviewers Submitted, Paper10585 Authors

Summary:
This paper gives a theoretical analysis of feedback alignment (FA), an algorithm to train neural networks by approximating the gradient of the loss function using random matrices. More precisely, the weights of the network that is being trained are replaced with random matrices in the back-propagation step.

The authors study two-layer neural networks in the over-parametrised regime, where the number of samples n is much larger than the number of neurons in the hidden layer, n, and provide two results:

1. A proof of linear convergence of the loss to 0. This proof follows the proofs of Du et al. (2018), Gao & Lafferty (2020) for standard backprop. The added difficulty of analysing DFA in this setting is that the effective kernel of the network is not a priori positive semi-definite.
2. An analysis of the alignment between the second-layer weights of the network and the feedback matrix used in the feedback alignment algorithm, which highlights the important role of regularisation and finds that (the role of) alignment is more complex than previously thought.

Main Review:

Strengths

- The topic is timely - there has been increasing interest in feedback alignment methods and alternatives to back-prop more broadly, as evidenced by the NeurIPS 2020 workshop on this very topic and a number of recent papers (see below). However, an understanding of the power of these alternatives to backprop remains an open problem, so I see a need for theoretical work in this direction, which the paper addresses.

- The authors give clear and convincing mathematical arguments for two key problems of FA, namely convergence despite training on a surrogate loss, and alignment of the weights to the feedback vectors.
- The authors clarify the mechanics of alignment between weights and feedback matrix, which has been conjectured to be key to the success of the method.
- The paper is well-written: the exposition is clear, and I found that the main arguments were explained clearly.

Weaknesses

While I found the paper insightful and the results interesting, the discussion in its present version ignores several papers that have recently provided analysis of feedback alignment algorithms. A discussion of the following two theoretical papers, and how the results presented here relate to their results, seems particularly appropriate:

- Regarding alignment, Frenkel et al. [Fre19] showed that using only the error sign is sufficient to maintain feedback alignment and to provide learning in the hidden layers of *linear* networks.
- Refinetti et al. [Ref21] analysed the dynamics of FA and DFA for two-layer non-linear networks in the feature learning regime.

Furthermore,

- the definition of alignment that the authors use (Def. 4.1) has been discussed extensively in the literature, cf. [Cra19, Fre19, Ref21]. This should be acknowledged.
- ...it might be worth mentioning that despite the problem with convolutions that the authors mention, DFA has recently been used to train modern neural networks such as Transformers [Lau20]

Let me also note that an important variant of feedback alignment is the direct feedback alignment of Nokland [Nok16], where the error vector is injected into each of the layers directly using a random matrix. For two-layer neural networks, the resulting update equations are the same, so the authors might want to mention that their results cover this relevant case, too.

Some small comments

- What does the \tilde O notation mean? (l. 94)
- In the definition of the NTK after l. 137, the average <.> is taken w.r.t. to which distribution?

References

- [Cra19] Crafton et al, Frontiers in neuroscience (2019). http://arxiv.org/abs/1903.02083 (http://arxiv.org/abs/1903.02083)
- [Fre19] Frenkel et al, (2019) http://arxiv.org/abs/1909.01311 (http://arxiv.org/abs/1909.01311)
- [Lau20] Launay et al. (2020) http://arxiv.org/abs/2006.12878 (http://arxiv.org/abs/2006.12878)
- [Nok16] Nøkland, NeurIPS 2016. http://arxiv.org/abs/1609.01596 (http://arxiv.org/abs/1609.01596)
- [Ref21] Refinetti et al, ICML 2021. https://arxiv.org/abs/2011.12428 (https://arxiv.org/abs/2011.12428)

Limitations And Societal Impact:

- The authors clearly address the range of validity of their theoretical claims.
- Since this is a theoretical work, there are no immediate societal impacts that need to be addressed imho.

Needs Ethics Review: No

Time Spent Reviewing: 4

Rating: 7: Good paper, accept

Confidence: 5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.

Code Of Conduct: While performing my duties as a reviewer (including writing reviews and participating in discussions), I have and will continue to abide by the NeurIPS code of conduct.

Add **Official Comment**