An Optimistic Acceleration of AMSGrad for Nonconvex Optimization

Jun-Kun Wang Xiaoyun Li Belhal Karimi Ping Li

1 Rebuttal

We thank the four reviewers for their valuable feedback. Our-point-by-point responses goes as follows:

- Reviewer 1:

Thanks for reading through our paper and your remarks. We totally understand that our paper might not fit best with your background. We hope our response can answer your questions. Online learning setting has been considered in many works, e.g. the Adam and AMSGrad paper. See e.g. page 4 of [Kingma and Ba, 2015] and page 2 of [Reddi et al., 2018] for more introduction on the background. Basically, in online learning, we need to analyze the regret since data points come sequentially, and the nature of the sequence is unknown a priori. While our motivation stems from the online learning setting, we extend it to stochastic optimization problems (finite-sum objective loss function), as usually found in modern deep learning tasks. We accelerate stochastic optimization by optimistic learning techniques, which was mainly used for online games in prior literature. We believe this could lead to future research in this direction. Starting from a classical online convex regret analysis, we develop convergence bound for nonconvex stochastic optimization. Lastly, experiments show remarkably better performance of the proposed OPT-AMS method.

- Reviewer 2:

We thank the reviewers for the valuable comments. We will add several baselines of interest, including SGD with momentum, in the revised paper. In our experiments, our primary goal is to stress on the comparison with the AMS method without adding optimistic information, in order to show its benefit, and with the only known work in adaptive optimization using optimistic updates, namely OPT-Adam.

- Reviewer 3:

Thanks for nicely summarizing our contributions and valuable feedback.

* Assumption H1 is rather usual in literature. As we agree with the reviewers that ensuring H1 for every task and model is challenging, we stress on our result in Lemma 2 of Section 4.3 that verifies H1 for a class of deep neural networks, giving a sense of how feasible verifying H1 can be in practice too. To the best of our knowledge, no other results in the related literature bypass this assumption H1 and neither verifies it like we attempt to do in Lemma 2.

* For the convex regret analysis, the bound in Corollary 1 can indeed be arbitrarily

large. The term $||g_t - m_t||^2_{\psi^*_{t-1}}$ implies that if the prediction m_t of the next gradient is very bad (far from the true g_t), then the rate will be slow. This term corresponds to the theoretical benefit of integrating such optimistic update.

* Assumption H3 is a constraint on the quality of the prediction m_t of the next gradient. We believe that the case where this prediction is arbitrarily bad is not worth studying. Hence, for the theoretical analysis, we consider that this prediction vector is in general reasonable, in the sense that m_t has acute angle with g_t. Boundedness of m_t is classical; in the stochastic optimization literature where bounding the gradient vectors is necessary to establish any result.

* All experiments do start from the same initial points for fair comparison.

- Reviewer 4:

* We thank the reviewer for the reference [ref1]. As for OPTIMISTIC-ADAM, [ref1] is specifically designed to optimize two-player games such as GAN. As we detail in the introduction of our paper, several papers have shown that if both players use an Optimistic type of update, then accelerating the convergence to the equilibrium of the game is possible. We emphasize that in this paper we propose a novel method, namely OPT-AMS, in order to accelerate stochastic nonconvex finite-sum optimization, which is different from the minmax problem in GAN.

*In Lemma 2, the constant T serves as an upperbound for the norm of the gradient of the multilayer model. It simply states that the gradient needs to be bounded, giving the existence of a single upper bound T is thus enough to satisfy that assumption. T does not correspond to the iteration index here, we will modify that in the revision. The boundedness of the weights is established uniformly on the parameter λ which is stronger. No matter the value of the regularization parameter, the weights are guaranteed to be bounded via Lemma 2. We present a regularized loss for generality, λ can be set to zero as an instance of this setting, and the result will still hold.

* We will include the Adam optimizer as a baseline for completeness in our numerical experiments.

2 Message to AC

Dear Meta-Reviewer and Program Chairs,

We are writing about several reviews we received on our submission 766. In particular, the quality of reviewers R1 and R4, while showing good or fairly good confidence in their evaluation, are rather poor in our opinion.

Regarding Reviewer 1, we note that the review does not contain any concrete and informative questions or remarks. Rather, sentences such as "I do not understand the online setting and the online objectives are not sufficiently explained to me. For example, the regret is explained as the composition of some action, some loss and some benchmark. I don't know what benchmark and what even the idea of the regret is." or "It could be that this paper is written for another audience, but for me, who has worked in nonconvex optimization and deep learning, it is not clear what the main focus/contribution of the paper is. I would assume that the online setting is a setting in which data is coming in like a stream, where constant updates are necessary. Yet, I'm not so sure about this, reading the paper. First of all, as I already said, the regret is not clear to me. I didn't understand what the benchmark w* is." give the strong impression that the reviewer is not in full capacity to review our optimization paper. Indeed, notions such as regret, online optimization or benchmark parameter are more that standard concepts to know in order to take on any optimization articles. The several assertions beginning with "I am not sure..." in the review, shows a real discrepancy between the knowledge of the reviewer and the content of our paper, and is in contradiction with the score given for the confidence.

Regarding Reviewer 4, while the confidence here is a bit less strong, which we acknowledge and we understand, the score being "Reject" is likely to be stronger than the superficial remarks the reviewer has provided. In particular, the reviewer suggested to compare with a reference that we did not include ([ref1]) because [ref1] deals with the training of GANs, as two-player game, which we explicitly stated as out of the scope of our contribution, in our introduction while introducing the baseline OPT-Adam. The second remark is also weak in the sense that the understanding of the statement of Lemma 2 is completely missed, both on the level of notation of the upper bound T, which we will of course modify to avoid such issue, and on the consideration of a regularized loss function, which is broader than unregularized loss, since it includes the case when the regularization parameter is equal to 0.

For all the reasons above, we would like to be able to obtain additional reviews as a replacement of Review 1 and 4, and/or have the AC participate in a thorough review of our contribution.

We appreciate your attention and thank you again for handling the review of our paper.

Best Regards, Authors of 766.