# An Optimistic Acceleration of AMSGrad for Nonconvex Optimization

Jun-Kun Wang Xiaoyun Li Belhal Karimi Ping Li

## 1 Rebuttal

— 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 objectives), as usually found in modern deep learning tasks. We accelerate stochastic optimization by optimistic learning techniques, which were 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: Thanks for your valuable suggestions and support.

We will add several baselines of interest, including SGD with momentum, in the revised paper. In our experiments, our primary goal is to stress 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 for two-player games, namely OPT-Adam.

— Reviewer 3: Thanks for your valuable feedback.

* Assumption H1 is common in literature, e.g. [Duchi et al. 2011], [Kingma and Ba, 2015], [Reddi et al., 2018]. As we agree with the reviewers that ensuring H1 for every task and model is challenging, we stress on our result in Lemma 2 that verifies H1 for a class of deep neural networks. To the best of our knowledge, no other results in the related literature bypass this assumption H1, and neither verifies it like we we establish in Lemma 2. In practice, and for the cases not covered by our theoretical results, techniques like layer normalization or weight decay could ensure that the parameter remains bounded.

* Instead of bounding the term $||g_t - m_t||^2_{\psi^*_{t-1}}$, we place it in the convergence rate which implies that if the prediction $m_t$ of the next gradient is very bad (very large $||g_t - m_t||^2_{\psi^*_{t-1}}$), then the rate will be slow. This term corresponds to the theoretical

benefit of the optimistic step, see Corollary 1 and its discussion.

* Assumption H3: The main purpose of assuming a positive inner product is to illustrate in simple terms the impact that the gradient prediction quality has on the overall convergence. It can be handled too if we lose this assumption (allowing negative cosines). We will clarify this in the paper and proof. The boundedness of $m_t$ is classical, e.g., in the stochastic optimization literature where bounding the gradient vectors is one of the standard assumptions.

* Yes, all our experiments start from the same initial points, and are averaged over 5 runs.

— Reviewer 4:Thanks for your valuable comments.

* [ref1] is specifically designed to optimize two-player games such as GAN, and 1) uses Adagrad as the backbone; and 2) simply uses the previous gradient as prediction. Our method uses AMSGrad as the underlying algorithm, with more advanced gradient prediction mechanism. As we highlighted in the paper, OPT-AMS is a novel method to accelerate stochastic nonconvex finite-sum minimization, as in training deep nets for classification purposes for instance, which is different from the minimax problem addressed in GAN.

*In Lemma 2, the constant $T$ serves as an upperbound for the norm of the gradient of the multi-layer 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 clarify this in the revision. The boundedness of the weights is established uniformly on the parameter $\lambda$, 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, $\lambda$ can be set to zero as an instance of this setting, where the result will still hold.

* Though Adam is very similar to AMSGrad, we will include it as a baseline for completeness in the comparison. Thanks for the suggestion. In our experiments, our primary goal is to stress the comparison with the vanilla AMSGrad to show the benefit of the optimistic step, which constitutes the main algorithmic novelty of our paper.

## 2 Message to AC

We believe R1 lacks the background to properly review our paper, e.g. "this paper is written for another audience, but not for me". While we appreciate his/her effort, the quality of the review is below acceptable. Also, we feel R4 may have missed the key idea of this paper -since we have addressed his/her main concern in Section 1 and 5 about optimistic GAN, a two-player game – and basic formulation in our Lemmas -T is just an upperbound. We hope our rebuttal solves well the concerns. Thank you.

Dear 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. Instead, sentences such as "I do not understand the online setting [...] 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 [...]. I would assume that the online setting is a setting in which data is coming in like a stream [...] as I already said, the regret is not clear to me" 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, show a real discrepancy between the knowledge of the reviewer and the content of our paper, and is in contradiction with the score given for its confidence.

Regarding Reviewer 4, while the confidence here is a 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 any unregularized loss, since it includes the case when the regularization parameter is equal to 0.

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

Best Regards, Authors of 766.