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:

We thank the reviewers for the remarks on our paper. The idea of optimistic learning, i.e. the fact of integrating a sequence of gradient prediction to improve the convergence of the method, has been well studied in the literature of online learning. As presented in page 2 of our contribution, several studies have used this idea in the context of regret learning. This terminology is proper to the online learning domain where data comes in the streaming fashion. For instance, the benchmark denotes the best value of the parameter w that leads to the optimal loss function. The main novelty of our contribution is to close the gap between the online setting and the stochastic and finite setting, as usually found in modern deep learning tasks. Hence, we introduce OPT-AMS, a stochastic optimization method for finite-sum objective function, used for training deep neural networks. The online learning introduction serves as a mention to prior work where optimistic updates have been used, while our contribution starts Section 3 (which no longer tackles online learning problems but finite-sum optimization as pointed by the reviewer).

- 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. We want 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:

We thank the reviewers for the analysis of our contribution and respond to his/her concerns.

* Assumption H1 is rather usual in current stochastic optimization 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 results.

* All experiments do start from the same initial points for fair comparison. This will be added in the revision.

- 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 majorant 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 baseline for completeness.

2 Message to AC

Dear Area Chair,

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

Best Regards, Authors of 766.