We sincerely thank the four reviewers for their valuable feedback.

Reviewer 1: We thank the reviewer for valuable comments. Our point-to-point response is as follows:
Comparison with OPT-FTRL: As stated in our introduction, we believe that the optimistic estimates presents an advantage over vanilla AMSGrad like OPT-FTRL presents over vanilla FTRL. Keeping in mind that the overall goal of our paper is to introduce a new method to solve a stochastic optimization problem where the objective function is a (large) finite-sum, OPT-FTRL is not an option, hence our motivation for developing a counterpart of OPT-FTRL in our settings.

Reviewer 2: We thank the reviewer for valuable comments. A proofreading is being done we clarify that:
Novelty of the contribution: Although combining gradient prediction to AMSGrad update seems natural, as pointed out in the first paragraph of Section 3, we would like to stress on how the embedding of the prediction process (represented Figure 1) led to the two-stage algorithm OPT-AMSGrad (unlike the sequential structure of the original AMSGrad) where, first an auxiliary variable $\tilde{w}$ is updated and then the global model w. Also, as discussed in the paper, optimistic learning is typically used in two-player-games, which is an online learning problem, and to the best of our knowledge, this is the first proposal to apply optimistic acceleration to stochastic optimization problems (e.g. training deep neural networks). Introducing the online optimization framework is a natural way to introduce our method, providing related literature on optimistic methods.
Comparison with other gradient prediction methods: Comparing the way we predict the gradients is indeed important in our study. We have devoted in Section F of the appendix an illustrative example of how this process can impact the performances of our method.

Reviewer 3: We thank the reviewer for the thorough analysis. Our remarks are listed below:
Gradient prediction algorithm: We will add more explanations on why the average of the last gradients can be a good approximation of the next one. While this may be counterintuitive, we invite the reviewer to read the citation we make regarding our extrapolation method: "Regularized nonlinear acceleration" by Scieur, d'Aspremont and Bach, NIPS 2016. We chose the latter reference mainly due to its success in training deep networks as observed in some prior works. Of course, there is room for improvement regarding this prediction process and can be the object of further research papers

Numerical Experiments: The learning rates have been tuned over a grid search for all methods and the best performances over the 5 repetitions have been reported. The main motivation behind those plots is to show that adding an optimistic update to the vanilla AMSGrad actually speed up the convergence in terms of both losses and accuracies. Given the well-known advantages of Adam-type methods as ADAM or AMSGrad, we did not compare to slower methods.

Reviewer 5: We thank the reviewer for valuable comments and typos. Our response is as follows:
Discussion on the bounds:

Global convergence analysis: The term *Global* is employed in the sense that it does not restrict the initialization of the algorithm and our bound is true for any iteration (finite-time). In other words the result is global since it is true for any initial point. Of course this is not related to the stationary point, as the objective function is nonconvex, no guarantees are given regarding the nature of the obtained stationary point.

Numerical experiments: There are several works considering applying Alg. 3 in deep learning, e.g. [Nonlinear Acceleration of Deep Neural Networks, Scieur et al., 2018], with positive results. As noted in their paper, in practice extrapolation on CPU is faster than a forward pass on mini-batch and can be further accelerated on GPU. Moreover, note that at each iteration, we only change one past gradient, so we do not need to compute the whole linear system every time leading to practical efficiency. Secondly, the main focus of our paper is essentially the framework of integrating optimistic learning with AMSGrad. We chose Algorithm 3 mainly because of the empirical success reported in prior works. The choice of gradient prediction method is actually flexible. So, OPT-AMSGrad will definitely benefit from an algorithm with faster running time and good prediction quality.

Reviewer 6: We thank the reviewer for valuable comments and typos. Our response is as follows:
OPT-ADAM: OPT-Adam has been developed in the particular case of an online problem, where the observations are being presented in a streaming fashion. Here, our goal is to develop a method for the finite-sum stochastic optimization problem. To the best of our knowledge, OPT-Adam has not been studied and no guarantees are given to claim that it also performs well under this latter settings. Empirical evaluations actually show that OPT-AMSGrad is better.

Boundedness of the iterates: Lemma 2 is needed to ensure that H1 is verified. We will change the notations to avoid any confusion. Lemma 2 indeed bounds the iterates of Algorithm 2 (OPT-AMSGrad) and should read $\|w_t^{(\ell)}\| \leq A_{(\ell)}$ where w_t are the weight estimates at iteration t. Then Lemma 2 holds for all iteration $t > 0$.