Reviewer 1:

On the theoretical front, it is not clear to me that the optimism is shown to bring any advantage in the non-convex setting. Since we are interested in applying the algorithm on non-convex problems, I would have preferred to see such an advantage here when the hints are good estimates of the real gradients.

Also, for the convex regret bound, I am worried there is something missing here. As far as I can see, no part of the algorithm actually enforces that the iterates stay in the bounded domain Θ and as a result I am not sure it is valid to bound distances between iterates by D_∞ as done in the regret bound. Is this easy to fix, perhaps by some appropriate projections?

On the experimental front, obviously it would be better to try the algorithm on larger datasets like Imagenet. However, even subject to limited computational budget I think a valuable baseline that is missing is what happens if you just run an optimistic algorithm that does not have any of the extra features of AMSGrad. That is, is the good performance of OPT-AMSGrad due to incremental advantage over AMSGrad, or is all of the work being done by the optimistic predictions?

Reviewer 2:

- The main algorithm seems to be rather incremental, combining two well-known algorithms, AMSGrad and optimistic mirror descent. It is unclear where the challenge lies in this combination, it seems to me to be a rather straightforward exercise.

- While the experiments show promise, the method of generating predictions via linear regressions seems to be rather computationally heavy, and the authors don't provide wall clock times for the algorithms, so it is unclear if AMSGrad would be competitive with OPT-AMSGrad if they're both given the same amount of time to run.

- The writing of the paper needs improvement, especially in proofreading, there are numerous typos in the paper.

Reviewer 3:

- I find the convergence analysis on non-convex functions to not really add much to the paper. In its current form, it is hard to parse how the accuracy of the gradient prediction affects the result and whether there is a situation in which Opt-AMSGrad can be provably shown to be faster AMSGrad (at least in the convergence of the 1/T term).

- The above could be due to assumption H3, where the authors assume that the gradient prediction is reasonably accurate. I think there needs to be more discussion on how reasonable this assumption is, especially for the gradient prediction algorithm used (algorithm 3). Does algorithm 3 provide any guarantees on how accurate the gradient prediction is?

- My other main concern is that the experimental evaluation of the method is quite poor. The experiments are not trained to convergence. For example in figure 3, both the train and test accuracies can be seen to be still rising. The ResNet models considered are typically trained to 200 epochs for CIFAR datasets, instead of the 100 training epochs used here. Furthermore, no learning rate decay was employed as far as I can tell, so the accuracy scores achieved by all methods are much lower than those reported in previous papers. This makes it hard to evaluate whether the baselines have been tuned sufficiently and thus whether this is a fair comparison.

Reviewer 4:

The test error improvement is not convincing. Since all the experiments are standard machine learning benchmark datasets and models, the competing methods also perform quite well in terms of test error. Ideally, one would identify a case (at least one) in which this method clearly outperforms the others in terms of test performance.

Since no confidence intervals are provided for the 5 runs of each experiment (c.f. Line 245) it is difficult to assess the test performance.

While the training improvement is clearer, the NeurIPS community is focused on the test performance of algorithms.