Project Executive Summary
Track 3: Reproducibility Challenge
"ON THE CONVERGENCE OF ADAM AND BEYOND"

Barleen Kaur, Jacob Shnaidman and Hamed Layeghi

In this project, an attempt was made to reproduce the results of the experiment in the recently published ICLR 2018 paper "ON THE CONVERGENCE OF ADAM AND BEYOND [1]" .

In the project, we first studied the theory to understand the basic advantages of AMSGRAD over other stochastic optimization methods. The authors provide an online optimization setting which in Machine Learning literature is very similar to Empirical Risk Minimization. In this context, a synthetic example is provided to show an essential flaw in the performance of ADAM which converges to the maximizing point for the regret function and so the average regret does not go to zero. This contradicts the original convergence theorem for ADAM [2].

I. SYNTHETIC EXAMPLES

The first set of experiments that were reproduced were the synthetic experiments. The obtained results completely matched the one on the paper. The learning rates were not specified. Although the theory suggested that a decaying learning rate should be used, we found the best performance on the stochastic synthetic experiment with a constant $\alpha_t = 5 \times 10^{-4}$.

II. LOGISTIC REGRESSION

For logistic regression, an extensive hyperparameter tuning was done to find the optimal ones for both AMSGRAD and ADAM. The learning rates were decaying with time and $\beta_1 = 0.9$ and $\beta_2 \in [0.99, 0.999]$. A grid search was performed on hyper-parameters α and β_2 to find the minimum final validation loss. It was noticed after initial grid search that especially for ADAM, α is the deciding parameter for this purpose. The resulting hyper-parameters were used to compare the train and test performance of logistic regression classification over the MNIST dataset. The hyper-parameters with the best validation performance found were $\alpha_t = \frac{0.005}{\sqrt{t}}$ and $\beta_2 = 0.999$

III. FEED FORWARD NEURAL NETWORK

For the feed-forward neural network experiment, the architecture was easily reproduced. The relative performance of Amsgrad was better than Adam using the architecture described in the paper; however, it proved unfeasible to attain losses as small as seen in the results in the original paper given our constraints on time and computational resources. A grid search was done over the learning rates and β_2's as was done by the authors to attain similar results.

IV. CIFARNET

Two architectures of CIFARNET were implemented as the exact architectural details used in the paper were not clear. An extensive hyper-parameter tuning was carried out using grid search to find the best hyper-parameter values α and β_2 which gave best generalization over the validation set. After training our CIFARNET models for 70 epochs for both the architectures, our experimental results show that AMSGrad performs better in terms of train loss than Adam. However, the test loss increases with number of iteration due to overfitting of the models. It seems that the authors used a very small learning rate for training the CIFARNET model as the number of iteration are too high, 2e+06 and still the model didn't converge fully. It was not possible for us to run our model for such high number of iterations due to time and computation constraints. Although the graphs are not exact replication of the results of CIFARNET as in the paper, still they strongly indicate the success of AMSGrad optimizer over Adam.

V. CONCLUSIONS

In this project, we were able to reproduce the results of the experiments in [1] to variable degrees. For the synthetic experiment, the results in our work match perfectly the ones in the paper and it clearly illustrates a basic flaw in ADAM with respect to AMSGRAD. For all experiments, the tuning of hyper-parameters for the optimizers played a key role in reproducing the results. We also learned the importance of reporting exact architectural details and range of hyper-parameters in the paper. The results of the neural network and CIFARNET supported the claims the authors made about the relative performance of AMSGrad compared to Adam; however, it proved computationally infeasible to reproduce the absolute performance as seen in the paper.

REFERENCES

[1] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," in *International Conference on Learning Representations*, 2018.

[2] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.