A Closer Look At The Convergence of Adam and AMSGrad: A Reproduction Study

Tamir Bennatan
tamir.bennatan@mail.mcgill.ca
260614526

Lea Collin
lea.collin@mail.mcgill.ca
260618407

Emmanuel Ng Cheng Hin
emmanuel.ngchenghin@mail.mcgill.ca
260615964

I. INTRODUCTION

The authors of the paper present issues with the popular stochastic gradient descent optimizers: RMSProp and ADAM, focusing mainly on ADAM. ADAM uses exponential moving averages of squared past gradients, which limits the reliance of parameter updates to only the last few gradients. Though ADAM has been proven to be very useful in many settings, it has also been shown to fail to converge to optimal solutions in other settings. The usual problem in these other settings is that large, informative gradients during updates occur infrequently. Because ADAM limits the reliance of parameter updates to only the past few gradients, the influence of these informative gradients quickly die out due to the use of exponential moving averages, leading to poor convergence.

An adversarial example is presented where there is a clear optimal solution yet ADAM fails to find it and actually converges to the worst solution. The example is as follows:

$$f_t(x) = \begin{cases} Cx, & \text{for } t \text{ mod } 3 = 1 \\ -x, & \text{otherwise} \end{cases}$$

where $C > 2$. It is easy to see that the value of x that leads to the minimum regret is -1, however, the authors show that ADAM converges to the highly suboptimal solution of $x = +1$. This elucidates the intuition that the influence of the large gradient C disappears too quickly to counteract the gradient of -1, which moves the algorithm in the wrong direction.

II. METHODOLOGY

III. RESULTS

IV. DISCUSSION