This paper talks about the improvements in already existing optimization algorithms such as ADAM or RMSProp using a method known as “**warmup**” wherein the training happens with a small learning rate for the first few epochs. The author claims that the lack of samples in the early stage of training learning rate has an undesirably large variance which leads to suspicious/ bad local optima. The paper introduces a new method known as Rectified Adam with the implementation of these warmups and states the motivation to do so.



This is the previously available structure of an adaptive learning rate optimization algorithm. By specifying different phi’s and theta’s different algorithms are obtained.

The paper then experiments with two different models that are slightly modified versions of ADAM .

The first method is adam2k in which the learning rate is updated in the first 2000 iterations while momentum and other parameters are fixed, and the remaining iterations take place according to conventional ADAM. The second slightly modified version is where the value of epsilon is changed from 10-8  to 10-4. Adam-2k: Additional 2k samples helped avoid the convergence problem of vanilla-Adam. Also, the additional samples prevent the gradient distribution from being distorted which shows that if there is sufficient data in the early stages the convergence problem can be avoided.

Adam-eps: Prevents the gradient distribution from being distorted. The seriousness of the

convergence problem is much lesser compared to vanilla-Adam which proves that reducing the variance of the adaptive learning rate can solve the convergence problem.



𝑟t has a similar form to the heuristic linear warmup. i.e. setting 𝑟! as min 𝑡,𝑇" /𝑇"

This confirms the observation from earlier which shows warmup reduces variance

RAdam deactivates 𝜓(. ) when variance is divergent, thus avoiding instability, RAdam is independent of model architectures and can be combined with other stabilization techniques. On resnet18, resnet20, and cifar10 datasets it is observed that RAdam outperforms Adam in all three datasets. 𝑟t does make it slower than Adam in the first few epochs but converges faster after that. RAdam improves the robustness of model training by making the model less sensitive to the learning rate parameter.

**References:** *Liu, Liyuan, et al. On the variance of the adaptive learning rate and beyond. ICLR 2020.*