ON THE VARIANCE OF THE ADAPTIVE LEARNING RATE AND BEYOND

Siddharth Sankaran

University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences

Contents

Contents

Introduction

- Goal of deep learning researchers is to create ‘Fast and stable optimization algorithms’
- Many optimization algorithms like SGD, Momentum, RMSProp, Adam, etc have been created
- Adaptive learning rate algorithms like Adam, Adadelta, Nadam have fast convergence rates
- However, above methods may converge to bad/suspicious local minima
- Above methods use **Warmup** – train with small learning rate for the first few epochs
- But there is no theoretical guarantee that warmup provide consistent improvements nor a guide on when and how to conduct warmup
- Warmup is generally used on a trial-and-error basis making it time consuming

Introduction

- There is a need to perform extensive analysis on the convergence issue
- To understand the impact of adaptive learning rate on the variance of gradients during model training
- To justify the use of warmup as a method reduce the variance
- Taking the above learnings, we introduce a new optimization technique called **Rectified Adam** (RAdam)

Contents

Previous work and motivation

Algorithm 1: Generic adaptive optimization method setup. All operations are element-wise.

Input: $\{\alpha_t\}_{t=1}^{T}$: step size, $\{\phi_t, \psi_t\}_{t=1}^{T}$: function to calculate momentum and adaptive rate, θ_0: initial parameter, $f(\theta)$: stochastic objective function.

Output: θ_T: resulting parameters

while $t = 1$ *to* T **do**

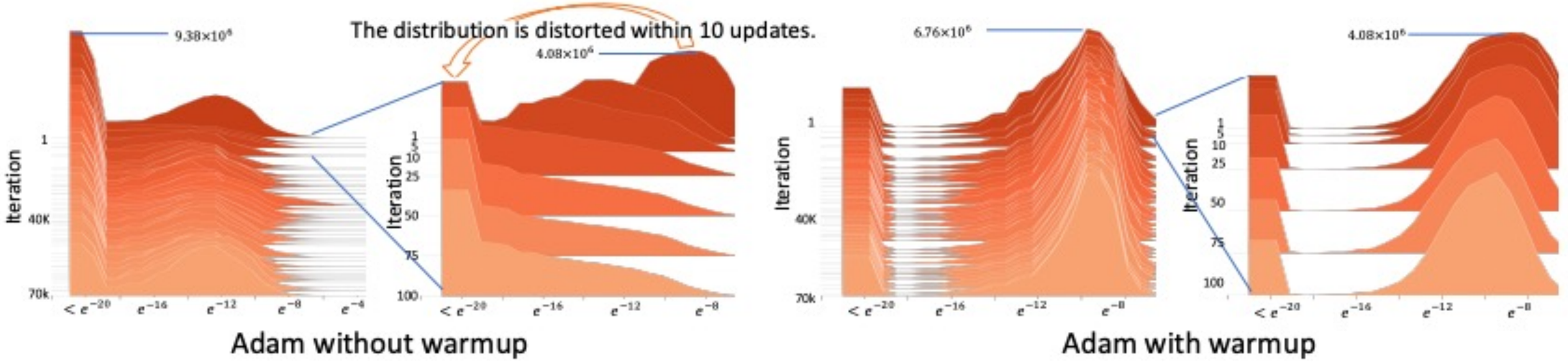
- $g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (Calculate gradients w.r.t. stochastic objective at timestep t)
- $m_t \leftarrow \phi_t(g_1, \cdots, g_t)$ (Calculate momentum)
- $l_t \leftarrow \psi_t(g_1, \cdots, g_t)$ (Calculate adaptive learning rate)
- $\theta_t \leftarrow \theta_{t-1} - \alpha_t m_t l_t$ (Update parameters)

return θ_T

- This is the general structure of an adaptive learning rate optimization algorithm
- By specifying different choices of $\varphi(.)$ and $\phi(.)$ different optimization algorithms are obtained. $\varphi(.)$ is adaptive learning rate and $\phi(.)$ is the momentum at time t

Previous work and motivation

Effect of warmup



- The warmup strategy sets learning rate α_t to small values initially and slowly increases them until $t < T_w$. For example, linear warmup might follow $\alpha_t = t\alpha_0$
- Without warmup, the distribution of absolute value of gradients becomes distorted with time. This is the reason for bad/suspicious local minima
- Warmup reduces the impact of this to avoid any convergence problems

Contents

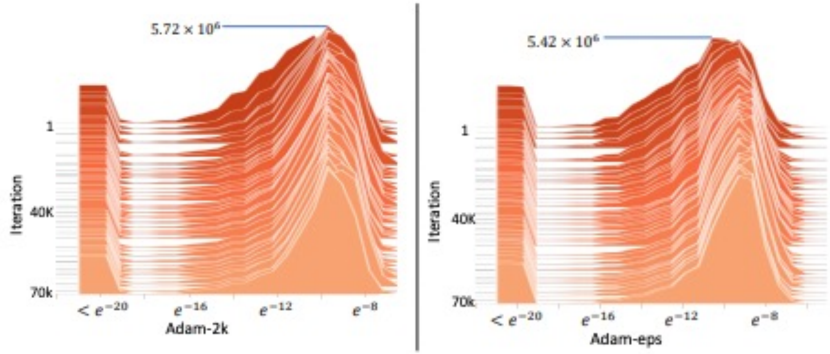
Variance of the adaptive learning rate

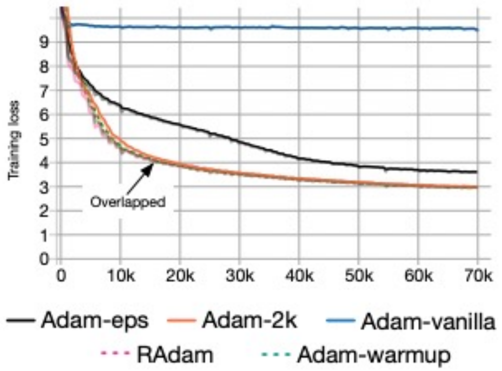
- We will try to prove our hypothesis – 'Due to the lack of samples in the early stage, the adaptive learning rate has an undesirably large variance, which leads to suspicious/bad local optima'
- Consider a case where adaptive learning rate $\varphi(g_1) = \sqrt{1/{g_1}^2}$ at $t = 1$
- Set of gradients $\{g_{1,} g_{2,} \dots g_n\}$ are i.i.d gaussian random variables distributed normally $N(0, \sigma^2)$
- So φ is subject to the scaled inverse chi-squared distribution and its variance is divergent
- This means that the variance is undesirably large initially
- Thus, it is the unbounded variance of φ that causes the above problem

Warmup as variance reduction

- We will experiment with two new model that are slightly modified from Adam
- Adam-2k
  - Only updates the learning rate $\varphi(.)$ in the first 2000 iterations, while momentum and other parameters are fixed
  - These are 2000 additional iterations, the remaining iterations take place similar to Adam
- Adam-eps
  - Increase the value of epsilon from 10^{-8} to a non-negligible 10^{-4}

Warmup as variance reduction





- Adam-2k: Additional 2k samples helped avoid convergence problem of vanilla-Adam. Also, the additional samples prevents the gradient distribution from being distorted
- Shows that if there is sufficient data in 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.
- This proves that reducing the variance of adaptive learning rate can solve convergence problem

Warmup as variance reduction

- It is also seen that the performance of Adam-eps is worser than Adam-2k and Adam-warmup
- This is because the large epsilon induces a large bias and slows down optimization process
- Thus, there is a need to create a more principled way to control the variance of adaptive learning rate

Analysis of Adaptive Learning Rate Variance

- We know Adam uses the exponential moving average to calculate the adaptive learning rate
- In the initial stages the difference of the exponential weights is relatively small. So, we can approximate the distribution of the exponential moving average as the distribution of the simple average

$$p(\psi(.)) = p\left(\sqrt{\frac{1-\beta_2^t}{(1-\beta_2)\sum_{i=1}^{t}\beta_2^{t-i}g_i^2}}\right) \approx p\left(\sqrt{\frac{t}{\sum_{i=1}^{t}g_i^2}}\right).$$

where β is the hyperparameter used in calculating exponential moving average

- Since g_i is a normal distribution $t/\sum_{i=1}^{t} {g_i}^2$ is subject to scaled inverse chi-square distribution. So, $\frac{1-\beta_2^t}{(1-\beta_2)\sum_{i=1}^{t}\beta_2^{t-i}g_i^2}$ is also subject to scaled inverse chi-square distribution with ρ degrees of freedom

Analysis of Adaptive Learning Rate Variance

Theorem

- With the previous assumption we can calculate Var[$\varphi^2(.)$] and the PDF of $\varphi^2(.)$
- Let us analyze the square root variance Var[$\varphi(.)$] and show that it reduces with increase in degrees of freedom ρ
- For all ρ > 4, we have the square root variance as:

$$\mathrm{Var}[\psi(.)] = \mathbb{E}[\psi^2(.)] - \mathbb{E}[\psi(.)]^2 = \tau^2\left(\frac{\rho}{\rho-2} - \frac{\rho\, 2^{2\rho-5}}{\pi}\mathcal{B}\left(\frac{\rho-1}{2}, \frac{\rho-1}{2}\right)^2\right),$$

Contents

Rectified Adaptive Learning Rate

Estimating ρ

- The previous section gives the analytical form of Var[$\varphi(.)$] with ρ degrees of freedom
- As mentioned earlier, the exponential moving average (EMA) can be approximated as the simple moving average (SMA)

$$p\left(\frac{(1-\beta_2)\sum_{i=1}^{t}\beta_2^{t-i}g_i^2}{1-\beta_2^t}\right) \approx p\left(\frac{\sum_{i=1}^{f(t,\beta_2)} g_{t+1-i}^2}{f(t,\beta_2)}\right)$$

- Here $f(t, \beta_2)$ is the length of the SMA which makes it have the same centre of mass as EMA

$$\frac{(1-\beta_2)\sum_{i=1}^{t}\beta_2^{t-i}\cdot i}{1-\beta_2^t} = \frac{\sum_{i=1}^{f(t,\beta_2)}(t+1-i)}{f(t,\beta_2)}$$

Rectified Adaptive Learning Rate

Estimating ρ

- We assumed that EMA is subject to the scaled-inverse chi-square distribution. Since EMA is approximated to SMA and gradient distribution is normal, SMA is also subject to scale-inv-χ^2
- Thus Scale-Inv-$\chi^2(\mathrm{f}(\mathrm{t}, \beta_2), \frac{1}{\sigma^2})$ is an approximation of Scale-Inv-$\chi^2(\rho, \frac{1}{\sigma^2})$
- Therefore, $\mathrm{f}(\mathrm{t}, \beta_2)$ is an estimate of ρ. $\mathrm{f}(\mathrm{t}, \beta_2)$ is marked as ρ_t for convenience

Rectified Adaptive Learning Rate

Variance estimation and rectification

- It is seen that the value of variance is significantly larger in initial stage than in later stages
- Let the minimal variance be represented as C_{var}
- In order to ensure consistent variance, we modify the variance at t^{th} timestamp as:

$$Var[r_t \psi(g_1, g_2, \ldots g_t)] = C_{var}, \text{ with } r_t = \sqrt{C_{var}/Var[\psi(g_1, g_2, \ldots g_t)]}$$

- By using first order approximation, we get the rectification term as:

$$r_t = \sqrt{(\rho_t - 4)(\rho_t - 2)\rho_\infty \,/\, (\rho_\infty - 4)(\rho_\infty - 2)\rho_t}$$

Rectified Adaptive Learning Rate

Modified Algorithm

Algorithm 2: Rectified Adam. All operations are element-wise.

Input: $\{\alpha_t\}_{t=1}^{T}$: step size, $\{\beta_1, \beta_2\}$: decay rate to calculate moving average and moving 2nd moment, θ_0: initial parameter, $f_t(\theta)$: stochastic objective function.

Output: θ_t: resulting parameters

$m_0, v_0 \leftarrow 0, 0$ (Initialize moving 1st and 2nd moment)

$\rho_\infty \leftarrow 2/(1-\beta_2) - 1$ (Compute the maximum length of the approximated SMA)

while $t = \{1, \cdots, T\}$ **do**

- $g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (Calculate gradients w.r.t. stochastic objective at timestep t)
- $v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) g_t^2$ (Update exponential moving 2nd moment)
- $m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$ (Update exponential moving 1st moment)
- $\widehat{m_t} \leftarrow m_t/(1-\beta_1^t)$ (Compute bias-corrected moving average)
- $\rho_t \leftarrow \rho_\infty - 2t\beta_2^t/(1-\beta_2^t)$(Compute the length of the approximated SMA)
- **if** *the variance is tractable, i.e.,* $\rho_t > 4$ **then**
  - $l_t \leftarrow \sqrt{(1-\beta_2^t)/v_t}$ (Compute adaptive learning rate)
  - $r_t \leftarrow \sqrt{\frac{(\rho_t-4)(\rho_t-2)\rho_\infty}{(\rho_\infty-4)(\rho_\infty-2)\rho_t}}$ (Compute the variance rectification term)
  - $\theta_t \leftarrow \theta_{t-1} - \alpha_t r_t \widehat{m_t} l_t$ (Update parameters with adaptive momentum)
- **else**
  - $\theta_t \leftarrow \theta_{t-1} - \alpha_t \widehat{m_t}$ (Update parameters with un-adapted momentum)

return θ_T

Rectified Adaptive Learning Rate

Comparison with Warmup and other stabilization techniques

- r_t has a similar form to the heuristic linear warmup. i.e. setting r_t as $\min(t, T_w)\ /T_w$
- This confirms observation from earlier which shows warmup reduces variance
- RAdam deactivates $\psi(.)$ when variance is divergent, thus avoiding instability
- RAdam is independent of model architectures and can be combined with other stabilization techniques

Contents

- Introduction
- Previous work and Motivation
- Variance of adaptive learning rate
- Rectified adaptive learning rate
- **Experiments**
- Conclusion

Experiments

Figure 4: Language modeling (LSTMs) on the One Billion Word.

Table 1: Image Classification

	Method	Acc.
CIFAR10	SGD	91.51
	Adam	90.54
	RAdam	91.38
ImageNet	SGD	69.86
	Adam	66.54
	RAdam	67.62

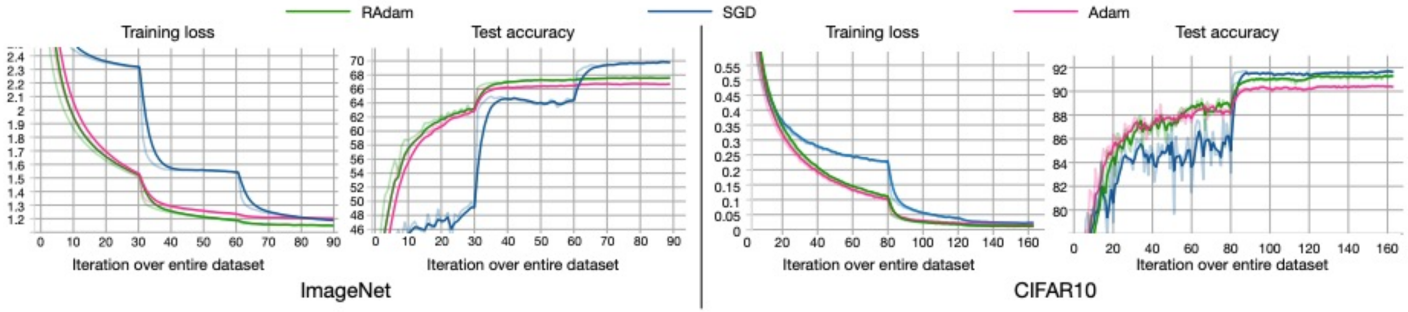


Figure 5: Training of ResNet-18 on the ImageNet and ResNet-20 on the CIFAR10 dataset.

- It is seen RAdam outperforms Adam in all three datasets
- r_t does make the it slower than Adam in first few epochs, but converges faster after that
- In CIFAR10, test accuracy of SGD is slightly higher that RAdam

Experiments

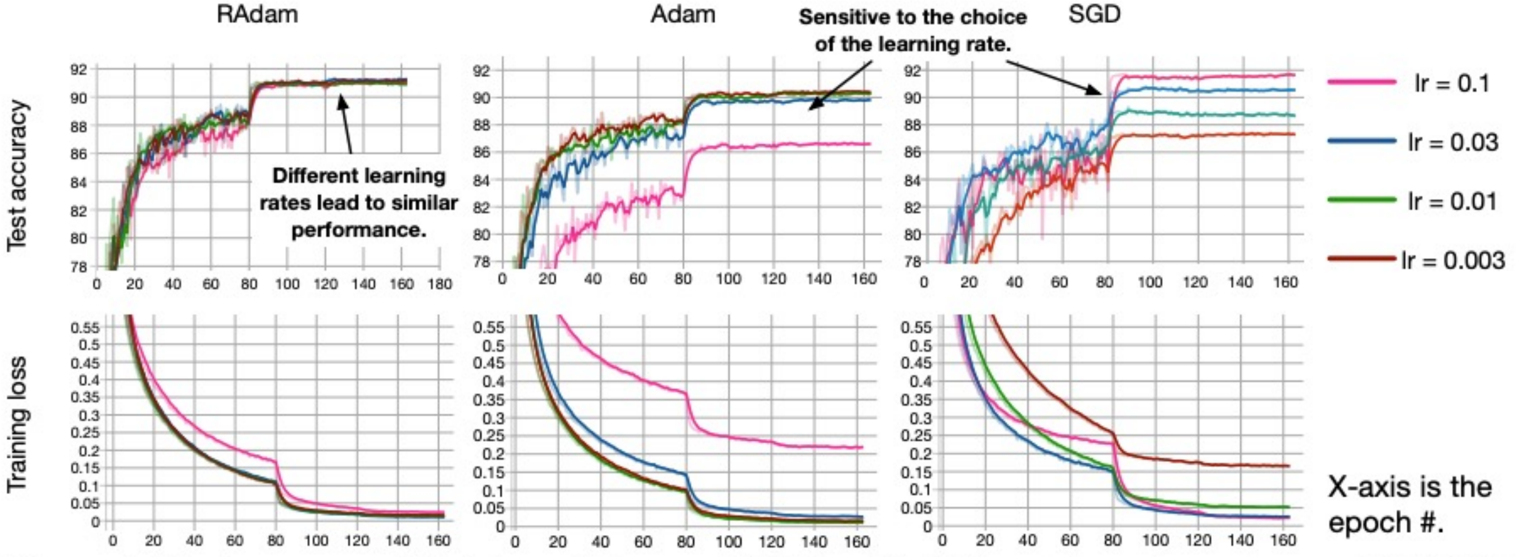


Figure 6: Performance of RAdam, Adam and SGD with different learning rates on CIFAR10.

- RAdam improves robustness of model training by making model less sensitive to learning rate parameter
- Both Adam and SGD can be seen as sensitive to learning rate

Experiments

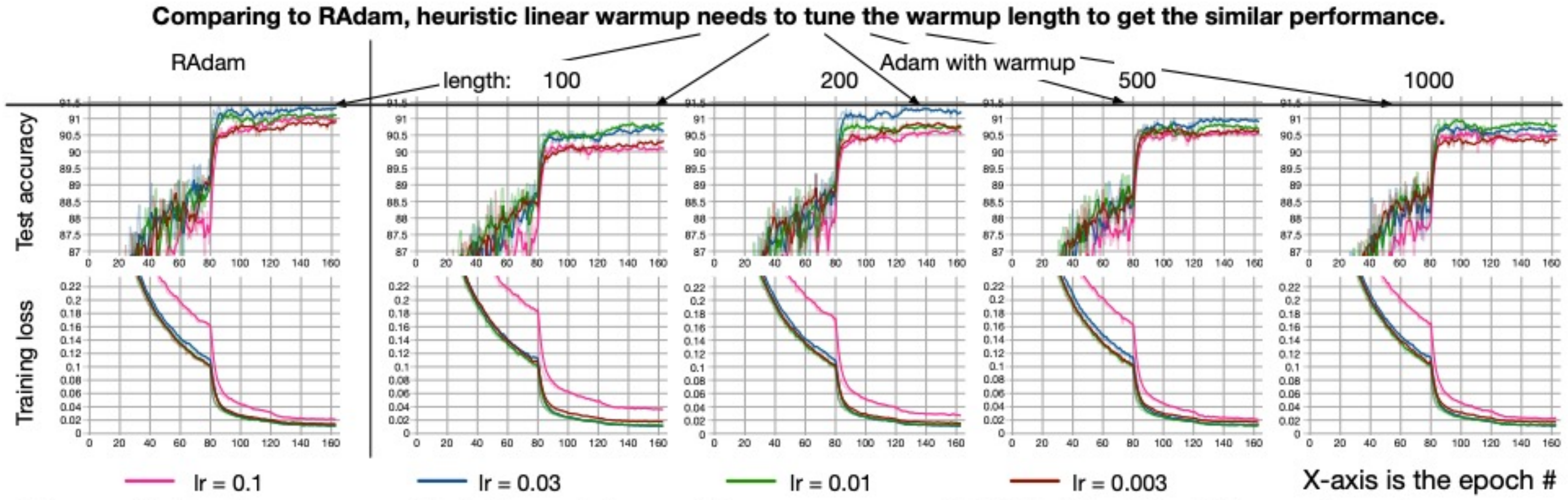


Figure 7: Performance of RAdam, Adam with warmup on CIFAR10 with different learning rates.

- Though test accuracy is similar, RAdam requires less hyperparameter tuning
- Adam is more sensitive to warmup length as well as learning rate

Contents

Conclusion

- Due to limited data during initial model training, the adaptive learning rate has large variance and can cause the model to converge at bad/suspicious local optima
- Supported by empirical and theoretical analysis
- RAdam is a new variant of Adam which rectifies adaptive learning rate to maintain consistent variance
- Experimental results show the effectiveness of RAdam over vanilla Adam