On the Variance of the Adaptive Learning Rate and Beyond

arXiv, 8 August 2019

Liyuan Liu et al.

Kyung-Jae Cho VUNO Inc.

RAdam found to be useful by some users

> "...I tested it on ImageNette and quickly got new high accuracy scores for the 5 and 20 epoch 128px leaderboard scores, so I know it works... https://forums.fast.ai/t/meet-radam-imo-the-new-state-of-the-art-ai-optimizer/52656
> — Less Wright August 15, 2019

> Thought "sounds interesting, I'll give it a try" - top 5 are vanilla Adam, bottom 4 (I only have access to 4 GPUs) are RAdam... so far looking pretty promising! pic.twitter.com/irvJSeoVfx
> — Hamish Dickson (@_mishy) August 16, 2019

> RAdam works great for me! It’s good to several % accuracy for free, but the biggest thing I like is the training stability. RAdam is way more stable! https://medium.com/@mgrankin/radam-works-great-for-me-344d37183943
> — Grankin Mikhail August 17, 2019

> "... Also, I achieved higher accuracy results using the newly proposed RAdam optimization function....
> https://towardsdatascience.com/optimism-is-on-the-menu-a-recession-is-not-d87cce265b10
> — Sameer Ahuja August 24, 2019

> "... Out-of-box RAdam implementation performs better than Adam and finetuned SGD...
> https://twitter.com/ukrdailo/status/1166265186920980480
> — Alex Dailo August 27, 2019

Motivation

- Many new methods (adaptive optimizers) have been proposed to accelerate optimization
- However, in many cases these optimization methods converge to bad/suspicious local optima

- **Need to use warmup heuristic**
  - Using small learning rate in the first few epochs
  - Removing warmup increases training perplexity

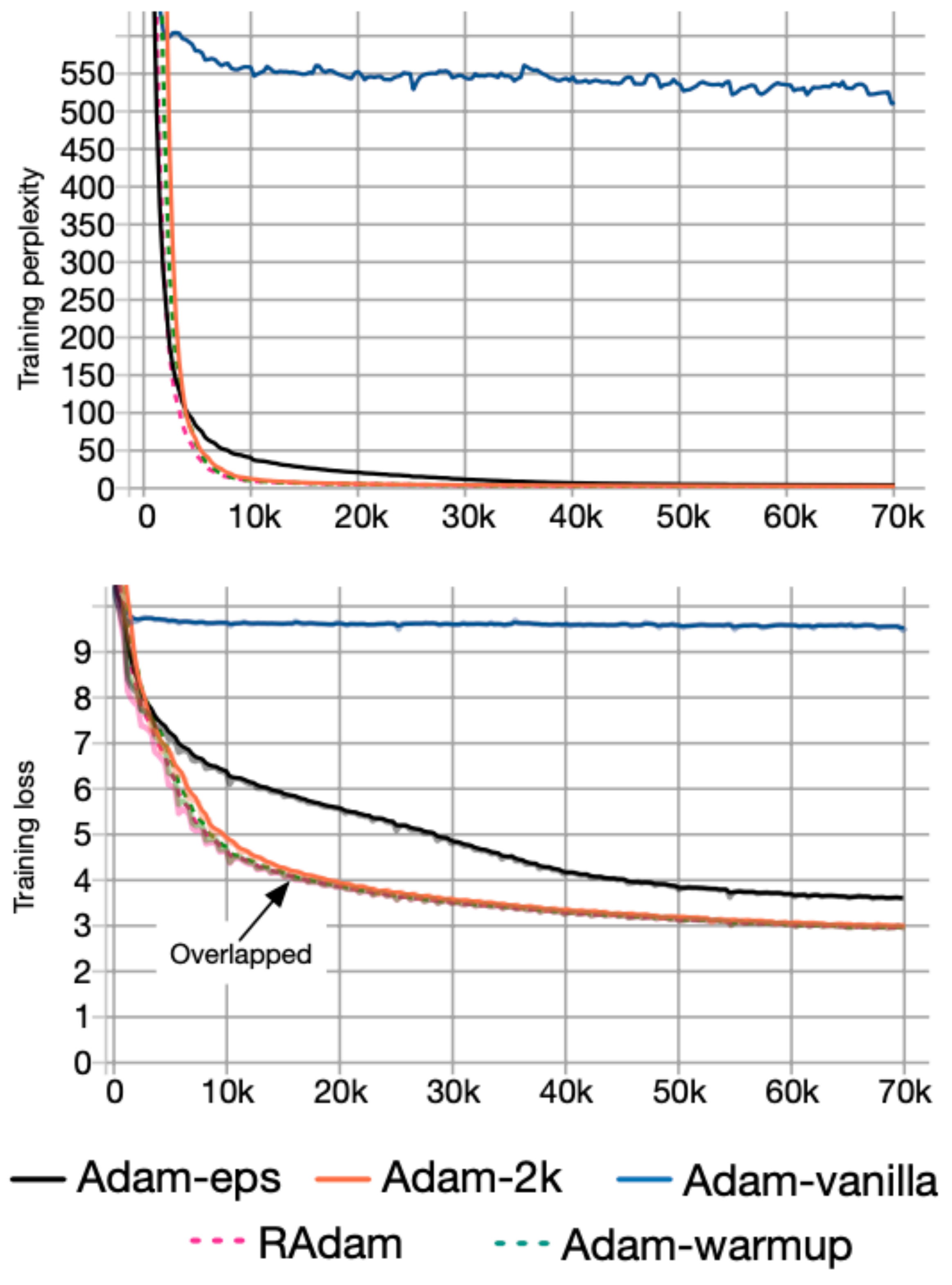


Figure 1: Training of Transformers on the De-En IWSLT'14 dataset. Up: Training perplexity w.r.t. gradient update iterations; Bottom: Training loss w.r.t. gradient update iterations.

Motivation (cont.)

- Limitations
  - Theoretical underpinnings of the warmup heuristic are lacking
  - There is neither guarantee that it always work in various ML settings nor guidance
  - Researchers use different settings in different applications **(trial-and-error approach)**
- Contribution
  - Conducts both theoretical and empirical analysis of the convergence issue
    - Root cause: **adapting learning rate has undesirably large variance in the early stage of training**
  - Propose a new variant of Adam (i.e., RAdam), which rectifies the variance and compares favorably with heuristic warmup

Adam algorithm and Learning rate warmup

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**

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

return θ_T

Momentum

Adaptive learning rate

$$\phi(g_1, \cdots, g_t) = \frac{(1-\beta_1)\sum_{i=1}^{t} \beta_1^{t-i} g_t}{1-\beta_1^t} \quad \text{and} \quad \psi(g_1, \cdots, g_t) = \sqrt{\frac{1-\beta_2^t}{(1-\beta_2)\sum_{i=1}^{t} \beta_2^{t-i} g_i^2}}. \quad (1)$$

Accelerates training by increasing the dimension whose gradients point in the same direction

- Perform smaller updates for parameters associated with frequently occurring features
- Perform larger updates for parameters associated with infrequent features

Learning rate warmup

$$\alpha_t = t\,\alpha_0 \text{ when } t < T_w$$

- sets α_t as some small values in the first few epochs

Learning rate warmup

- **Without applying warmup, the gradient distribution is distorted** to have a mass center in relatively small values within 10 updates
  - Trapped in bad/suspicious local optima
- Warmup reduces the impact of these problematic updates to avoid the convergence problem

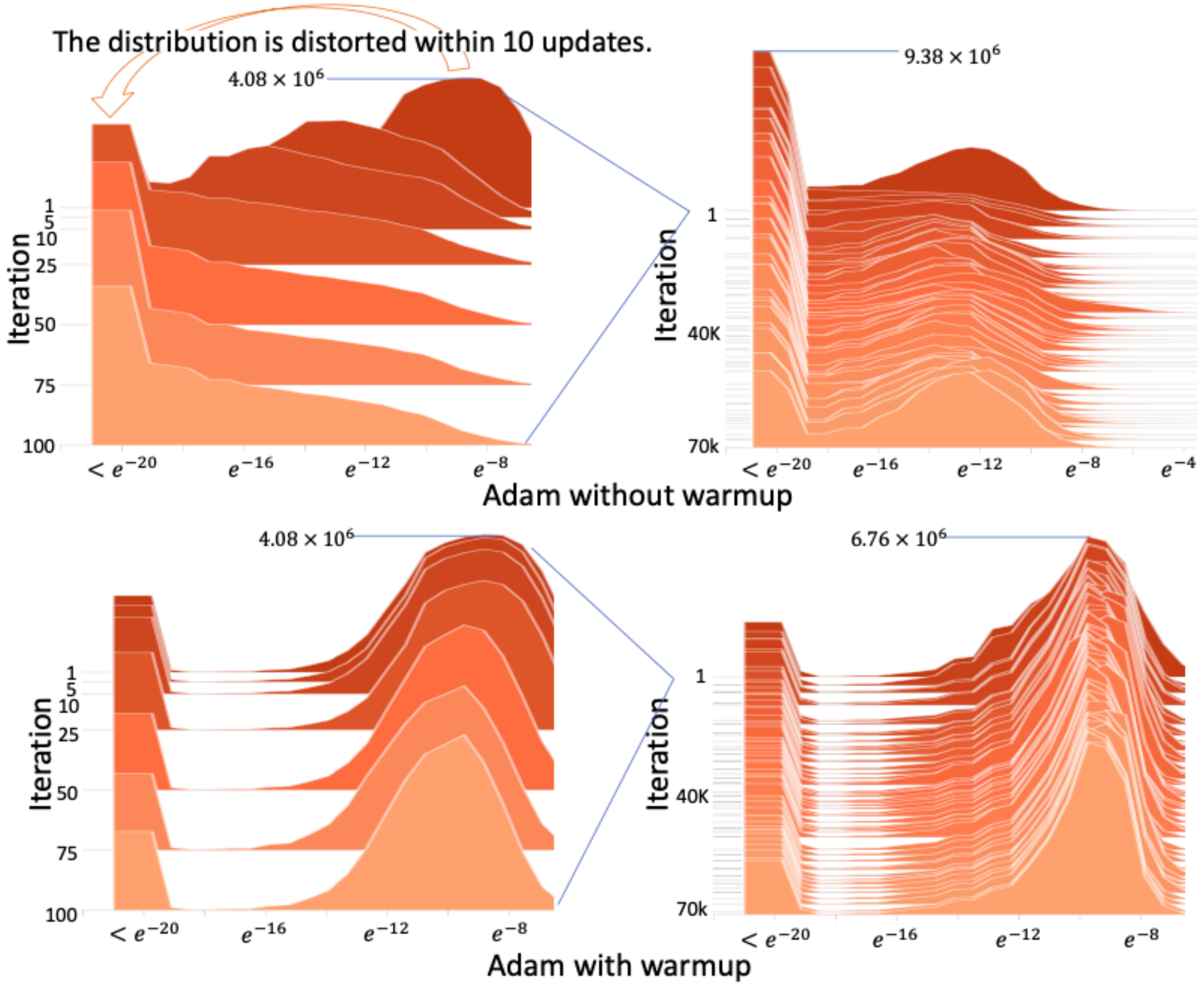


Figure 2: The absolute gradient histogram of the Transformers on the De-En IWSLT' 14 dataset. X-axis is absolute value in the log scale and the height is the frequency. Without warmup, the gradient distribution is distorted in the first 10 steps.

Variance of adaptive rate

- **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**

Adaptive learning rate

$$\psi(g_1, \cdots, g_t) = \sqrt{\frac{1-\beta_2^t}{(1-\beta_2)\sum_{i=1}^{t}\beta_2^{t-i} g_i^2}}. \quad (1)$$

To begin with, we first analyze a special case. When $t = 1$, we have $\psi(g_1) = \sqrt{\frac{1}{g_1^2}}$. We view $\{g_1, \cdots, g_t\}$ as i.i.d. random variables drawn from a Normal distribution $\mathcal{N}(0, \sigma^2)^{\boxed{2}}$. Therefore, $\frac{1}{g_1^2}$ is subject to the scaled inverse chi-squared distribution, Scale-inv-$\mathcal{X}^2(1, \frac{1}{\sigma^2})$. Noted $\mathrm{Var}[\sqrt{\frac{1}{g_1^2}}] \propto \int_0^\infty x^{-1} e^{-x} dx$ and it is divergent. It means that the adaptive ratio can be undesirably large in the first stage of learning. Meanwhile, setting a small learning rate at the early stage can reduce the variance

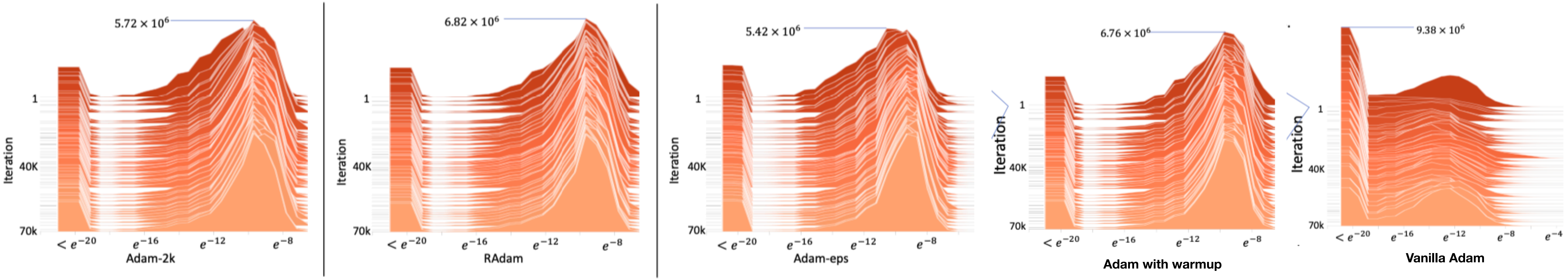
Mean zero normal assumption is valid at the beginning of training, since weights are sampled from normal distributions with mean zero

Warmup as variance reduction

- Convergence can be avoided by reducing the variance of adaptive learning rate on NMT dataset
  - Adam-2k: the first 2k iterations, only the adaptive learning rate is updated
  - Adam-eps: increase epsilon to reduce the variance to a non-negligible value

$$\widehat{\psi}(g_1,\cdots,g_t) = \frac{\sqrt{1-\beta_2^t}}{\epsilon+\sqrt{(1-\beta_2)\sum_{i=1}^t \beta_2^{t-i} g_i^2}}.$$

- **We need a more principled way to control the variance**



Analysis of adaptive learning rate variance

Theorem 1. If $\psi^2(.) \sim \text{Scale-inv-}\mathcal{X}^2(\rho, \frac{1}{\sigma^2})$, $\text{Var}[\psi(.)]$ monotonically decreases as ρ increases.

학습 과정이 뒤로갈수록, variance가 점진적으로 감소함

Proof. For ease of notation, we refer $\psi^2(.)$ as x and $\frac{1}{\sigma^2}$ as τ^2. Thus, $x \sim \text{Scale-inv-}\mathcal{X}^2(\rho, \tau^2)$ and:

$$p(x) = \frac{(\tau^2\rho/2)^{\rho/2}}{\Gamma(\rho/2)} \frac{\exp[\frac{-\rho\tau^2}{2x}]}{x^{1+\rho/2}} \quad \text{and} \quad \mathbb{E}[x] = \frac{\rho}{(\rho-2)\sigma^2} \ (\forall\, \rho > 2) \tag{2}$$

where $\Gamma(.)$ is the gamma function. Therefore, we have:

$$\mathbb{E}[\sqrt{x}] = \int_0^\infty \sqrt{x}\, p(x)\, dx = \frac{\tau\sqrt{\rho}\,\Gamma(\rho/2-1)}{\sqrt{2}\,\Gamma(\rho/2)} \ (\forall\, \rho > 4). \tag{3}$$

Based on Equation 2 and 3, for $\forall\, \rho > 4$, we have:

$$\text{Var}[\psi(.)] = \text{Var}[\sqrt{x}] = \mathbb{E}[x] - \mathbb{E}[\sqrt{x}]^2 = \tau^2\left(\frac{\rho}{\rho-2} - \frac{\rho\, 2^{2\rho-5}}{\pi}\mathcal{B}\left(\frac{\rho-1}{2}, \frac{\rho-1}{2}\right)^2\right) \tag{4}$$

To prove the monotonic, we need to show

Lemma 1. for $t \geq 4$, $\frac{\partial}{\partial t}\left(\frac{t}{t-2} - \frac{t\, 2^{2t-5}}{\pi}\mathcal{B}\left(\frac{t-1}{2}, \frac{t-1}{2}\right)^2\right) < 0$

- For ease of analysis
  - we approximate exponential moving average as the distribution of simple average

Exponential moving average — Simple moving average

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

  - Gradients are drawn from zero mean normal distribution

Rectified Adaptive learning rate

- In order to **ensure that the adaptive learning rate has consistent variance, we rectify the variance at the t-th timestamp**

$$\mathrm{Var}[\boxed{r_t}\,\psi(g_1, \cdots, g_t)] = \boxed{C_{\text{var}}} \quad \text{where} \quad r_t = \sqrt{\frac{C_{\text{var}}}{\mathrm{Var}[\psi(g_1, \cdots, g_t)]}}.$$

Since $\psi^2(.) \sim \text{Scale-inv-}\mathcal{X}^2(\rho_t, \frac{1}{\sigma^2})$, we have:

$$\mathrm{Var}[\psi(.)] \approx \frac{\rho_t}{2(\rho_t - 2)(\rho_t - 4)\sigma^2}.$$

$$\boxed{r_t = \sqrt{\frac{(\rho_t - 4)(\rho_t - 2)\rho_\infty}{(\rho_\infty - 4)(\rho_\infty - 2)\rho_t}}.}$$

Although we have the analytic form of $\mathrm{Var}[\psi(.)]$ (*i.e.*, Equation 4), it is not numerically stable. Therefore, we use the first-order approximation to calculate the rectification term. Specifically, by approximating $\sqrt{\psi^2(.)}$ to the first order (Wolter, 2007),

$$\sqrt{\psi^2(.)} \approx \sqrt{\mathbb{E}[\psi^2(.)]} + \frac{1}{2\sqrt{\mathbb{E}[\psi^2(.)]}}(\psi^2(.) - \mathbb{E}[\psi^2(.)]) \quad \text{and} \quad \mathrm{Var}[\psi(.)] \approx \frac{\mathrm{Var}[\psi^2(.)]}{4\,\mathbb{E}[\psi^2(.)]}.$$

By solving this equation, we have: $f(t, \beta_2) = \frac{2}{1-\beta_2} - 1 - \frac{2t\beta_2^t}{1-\beta_2^t}$. In the previous section, we assume: $\frac{1-\beta_2^t}{(1-\beta_2)\sum_{i=1}^t \beta_2^{t-i} g_i^2} \sim \text{Scale-inv-}\mathcal{X}^2(\rho, \frac{1}{\sigma^2})$. Here, since $g_i \sim \mathcal{N}(0, \sigma^2)$, we have $\frac{\sum_{i=1}^{f(t,\beta_2)} g_{t+1-i}^2}{f(t,\beta_2)} \sim \text{Scale-inv-}\mathcal{X}^2(f(t, \beta_2), \frac{1}{\sigma^2})$. Thus, Equation 5 views $\text{Scale-inv-}\mathcal{X}^2(f(t, \beta_2), \frac{1}{\sigma^2})$ as an approximation to $\text{Scale-inv-}\mathcal{X}^2(\rho, \frac{1}{\sigma^2})$. Therefore, we treat $f(t, \beta_2)$ as an estimation of ρ. For ease of notation, we mark $f(t, \beta_2)$ as ρ_t. Also, we record $\frac{2}{1-\beta_2} - 1$ as ρ_∞ (maximum length of the approximated SMA), due to the inequality $f(t, \beta_2) \le \lim_{t\to\infty} f(t, \beta_2) = \frac{2}{1-\beta_2} - 1$.

Rectified Adam

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**
    $g_t \leftarrow \Delta_\theta f_t(\theta_{t-1})$ (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**
        $\widehat{v_t} \leftarrow \sqrt{v_t/(1-\beta_2^t)}$ (Compute bias-corrected moving 2nd moment)
        $r_t \leftarrow \sqrt{\frac{(\rho_t - 4)(\rho_t - 2)\rho_\infty}{(\rho_\infty - 4)(\rho_\infty - 2)\rho_t}}$ (Compute the variance rectification term)
        $\theta_t \leftarrow \theta_{t-1} - \alpha_t r_t \widehat{m_t}/\widehat{v_t}$ (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

- RAdam is a variance reduction technique, which deactivates the adaptive learning rate when its variance is divergent

Comparing to vanilla adam

- **Does large variance in early stage leading to bad local optima widely exists in other similar tasks and applications?**

Table 1: Perplexity on Language Modeling

Method	One Billion Word
Adam	36.92
RAdam	35.70

Table 2: Accuracy on Image Classification

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

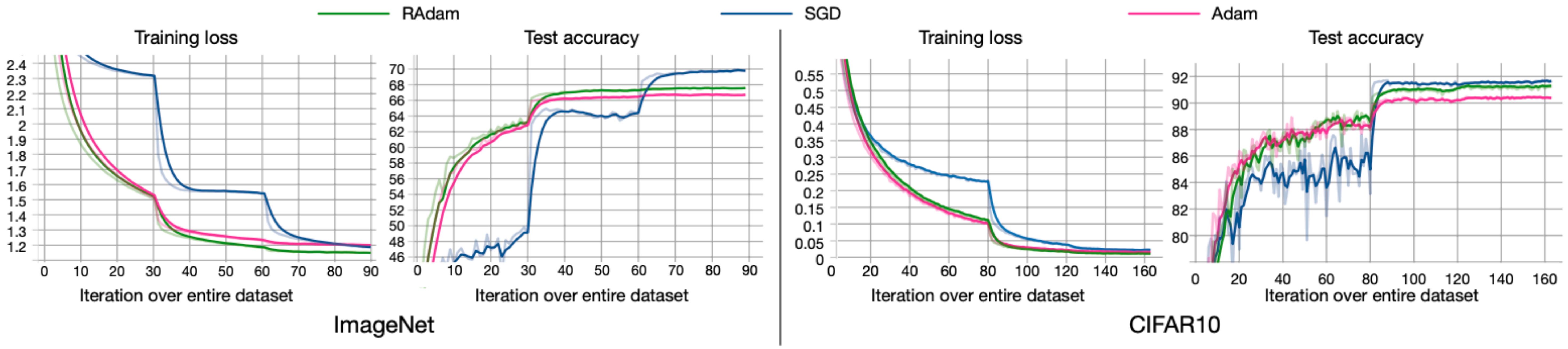


Figure 5: Training of ResNet-18 on the ImageNet and ResNet-20 on the CIFAR10 dataset.

- The result shows that **RAdam outperforms Adam in all three datasets**
- Although, RAdam is slower than Adam in the first few epochs, it converge faster after that
- By reducing variance of adaptive learning rate, **it gets both faster convergence and better performance**

Robustness to Learning rate change

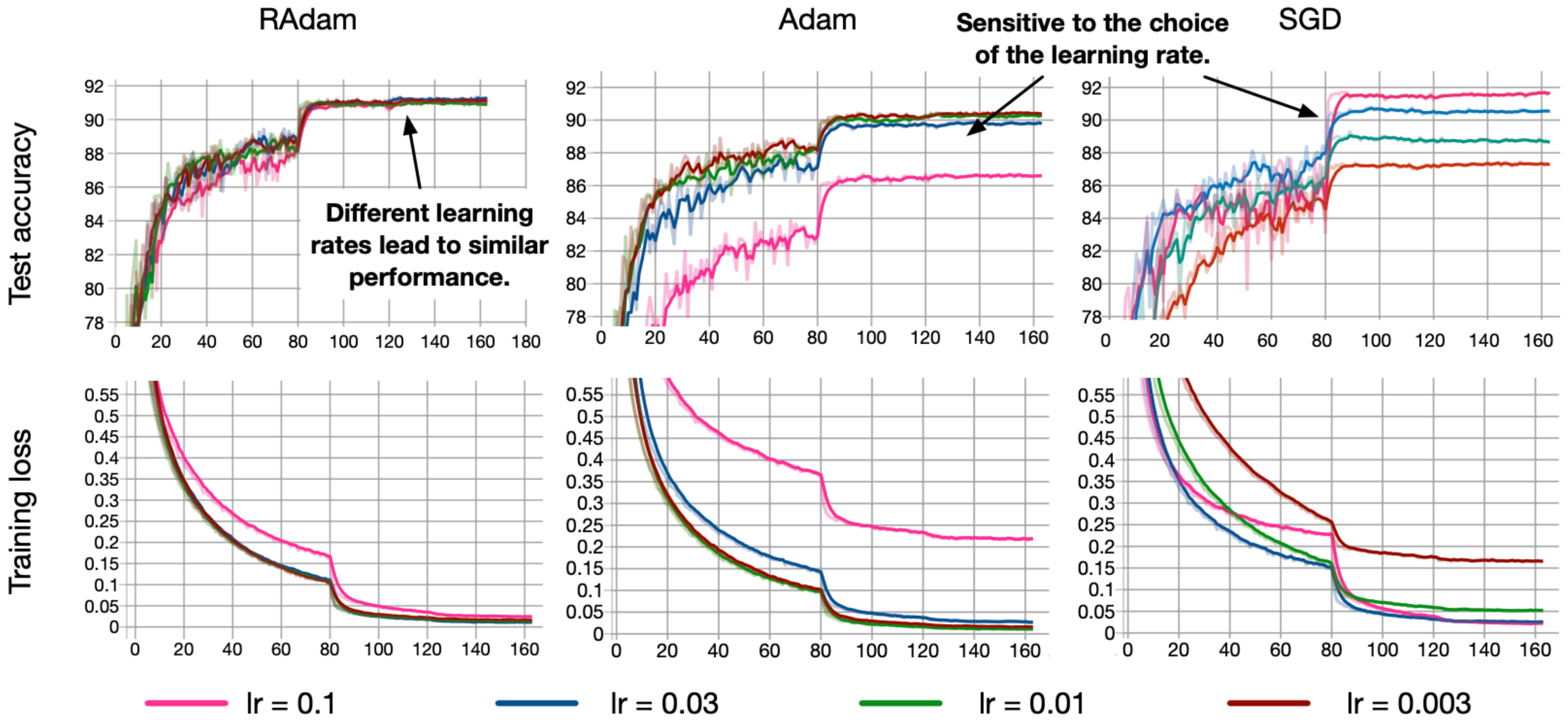


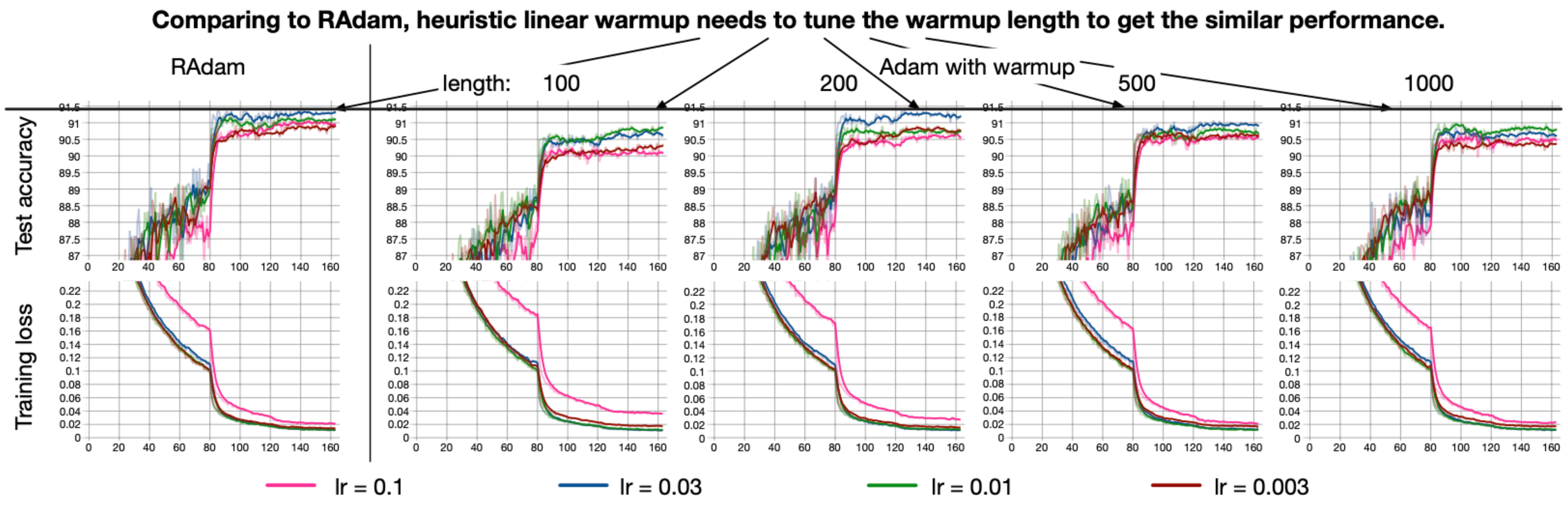
Figure 6: Performance of RAdam, Adam and SGD with different learning rates on CIFAR10. X-axis is the number of epochs.

- RAdam achieves consistent model performance, while Adam and SGD are shown to be sensitive to the learning rate

Comparing to heuristic warmup

Table 3: BLEU score on Neural Machine Translation.

Method	IWSLT'14 DE-EN	IWSLT'14 EN-DE	WMT'16 EN-DE
Adam with warmup	34.66 ± 0.014	28.56 ± 0.067	27.03
RAdam	34.76 ± 0.003	28.48 ± 0.054	27.27



- RAdam achieves similar performance to that of pervious SOTA (Adam with warmup)
- Adam with warmup is relatively more sensitive to the choice of learning rate
- Whereas, RAdam is robust, but controls the warmup behavior

Conclusion

- **Explored the underlying principle** of the effectiveness of the **warmup heuristic** used for adaptive optimization algorithms
- Identified that due to the limited amount of samples in the early stage of model training, t**he adaptive learning rate has an undesirably large variance** and can cause the model to converge to suspicious/ bad local optima
- The paper provide both **empirical and theoretical evidence** to support the hypothesis and **proposed new variant of Adam**
- In future work, we plan to apply RAdam to other applications such as Named Entity Recognition

Putting the world's medical data to work

hello@vuno.co

Simulated verification

- First order approximation
- Scaled Inverse Chi-Square Distribution Assumption
  - Assume gradients accords to normal distribution with zero mean
  - Assume square of psi accords to scaled inverse chi-square distribution
    - To derive variance of psi based on the similarity between the EMA and SMA

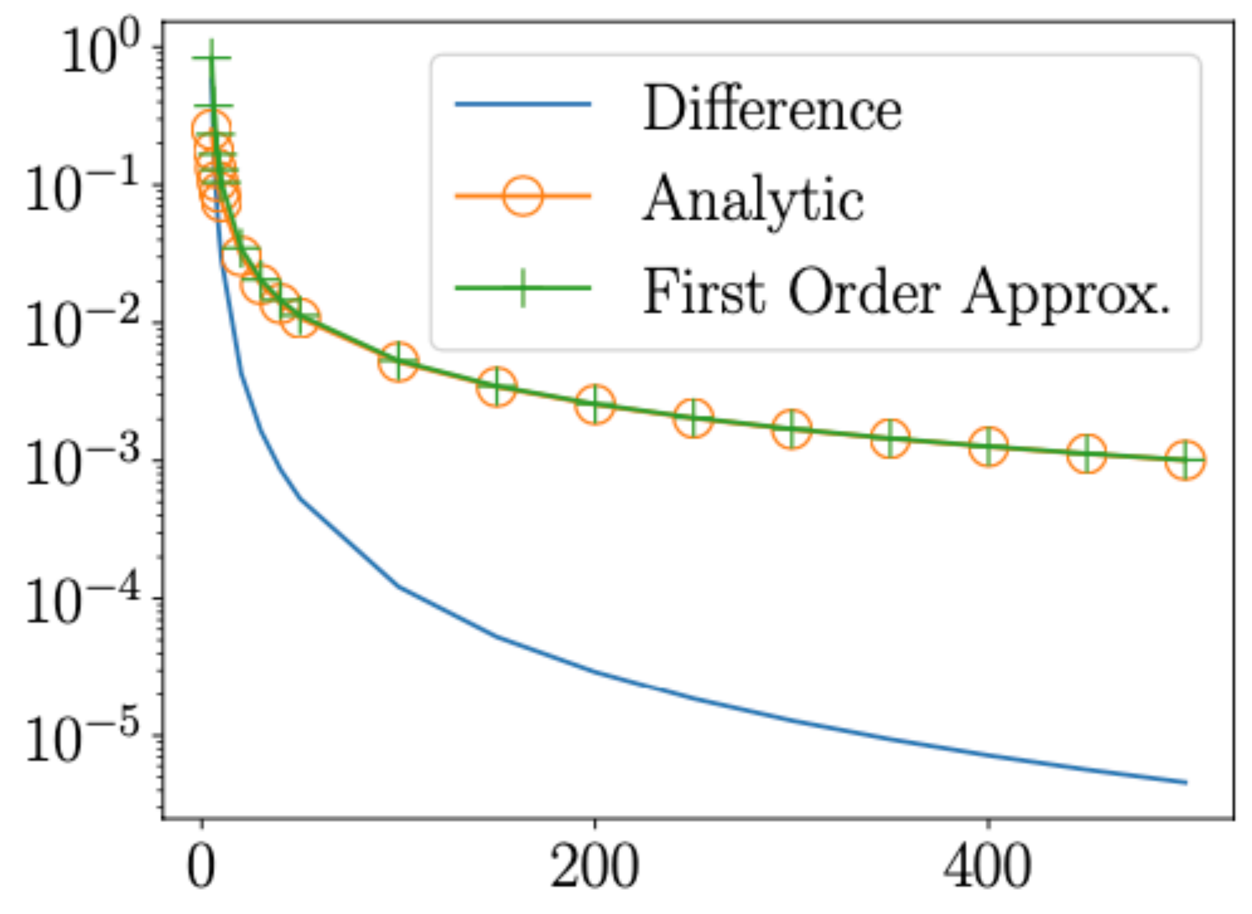


Figure 8: The value of Equation 4, Equation 6 and their difference (calculated as the absolute difference value). The x-axis is ρ and the y-axis is the variance in the log scale.

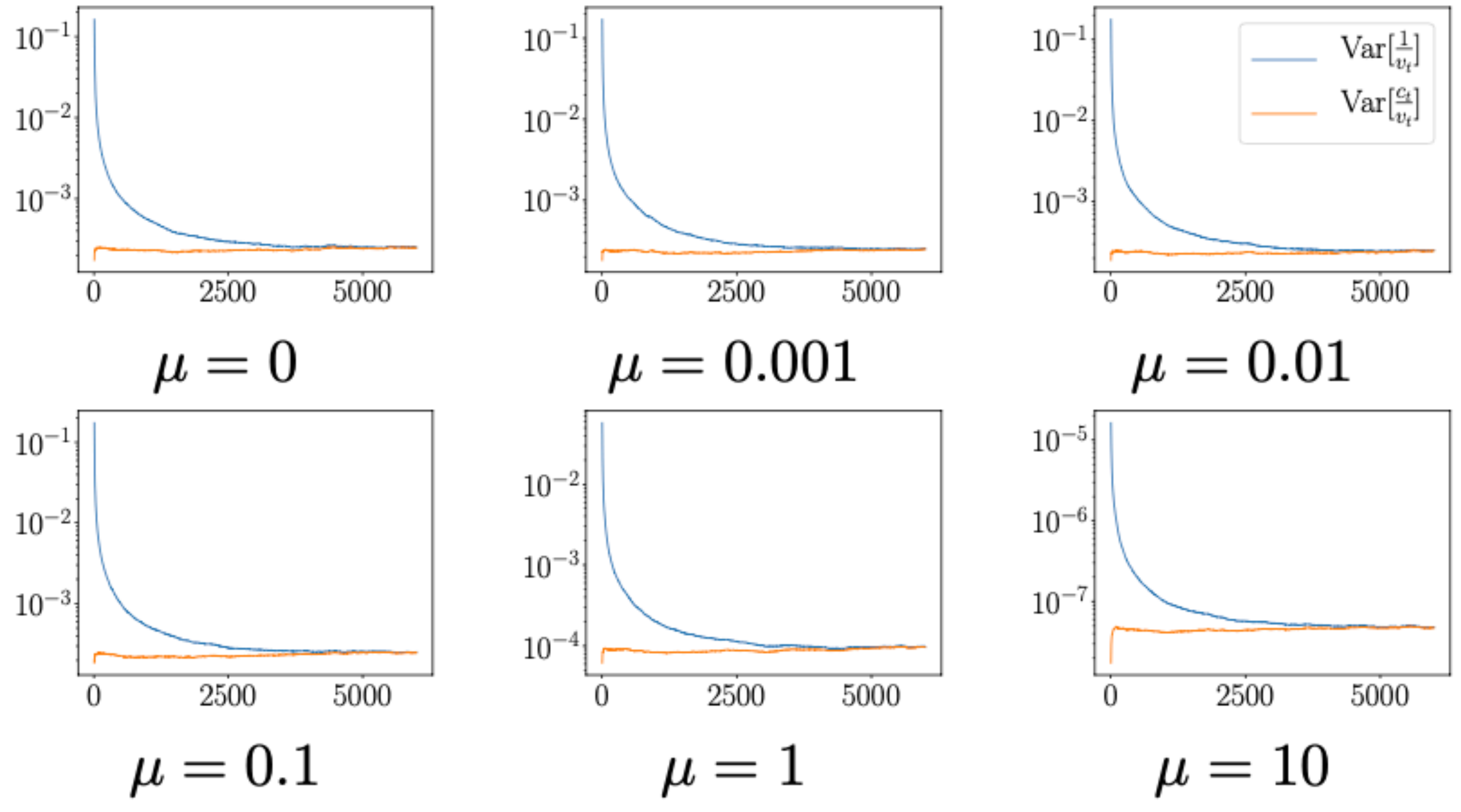


Figure 9: The simulation of $\mathrm{Var}[\frac{1}{v_t}]$ and $\mathrm{Var}[\frac{c_t}{v_t}]$. The x-axis is iteration number (the simulation starts from 5) and the y-axis is the variance in the log scale.

Rectified Adaptive learning rate

- Estimation of p

Exponential moving average (EMA)

Simple moving average (SMA)

where $f(t, \beta_2)$ is the length of the SMA which allows the SMA has the same "center of mass" with the EMA. In other words, $f(t, \beta_2)$ satisfies:

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

By solving this equation, we have: $f(t, \beta_2) = \frac{2}{1-\beta_2} - 1 - \frac{2t\beta_2^t}{1-\beta_2^t}$. In the previous section, we assume: $\frac{1-\beta_2^t}{(1-\beta_2)\sum_{i=1}^{t}\beta_2^{t-i}g_i^2} \sim \text{Scale-inv-}\mathcal{X}^2(\rho, \frac{1}{\sigma^2})$. Here, since $g_i \sim \mathcal{N}(0, \sigma^2)$, we have $\frac{\sum_{i=1}^{f(t,\beta_2)} g_{t+1-i}^2}{f(t,\beta_2)} \sim \text{Scale-inv-}\mathcal{X}^2(f(t,\beta_2), \frac{1}{\sigma^2})$. Thus, Equation $\boxed{5}$ views $\text{Scale-inv-}\mathcal{X}^2(f(t,\beta_2), \frac{1}{\sigma^2})$ as an approximation to $\text{Scale-inv-}\mathcal{X}^2(\rho, \frac{1}{\sigma^2})$. Therefore, we treat $f(t, \beta_2)$ as an estimation of ρ. For ease of notation, we mark $f(t, \beta_2)$ as ρ_t. Also, we record $\frac{2}{1-\beta_2} - 1$ as ρ_∞ (maximum length of the approximated SMA), due to the inequality $f(t, \beta_2) \leq \lim_{t\to\infty} f(t, \beta_2) = \frac{2}{1-\beta_2} - 1$.