On the $O(\frac{\sqrt{d}}{T^{1/4}})$ Convergence Rate of AdamW Measured by ℓ_1 Norm

Huan Li
Nankai University

Yiming Dong
Peking University

Zhouchen Lin
Peking University

Introduction

Consider nonconvex optimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$$

where d is the dimension, and it can be extremely large

$$d = 1.75 \times 10^{11} \text{ in GPT-3}$$

Adaptive Moment Estimation with Decoupled Weight Decay (AdamW)

- The default optimizer for training large language models
- However, its convergence behavior is not well-understood

Algorithm 1 AdamW

Hyper parameters: $\eta, \theta, \beta, \lambda, \varepsilon$
Initialize $\mathbf{x}^1$, $\mathbf{m}^0 = 0$, $\mathbf{v}^0 = 0$
for $k = 1, 2, \cdots, K$ **do**
  $\mathbf{g}^k = \text{GradOracle}(\mathbf{x}^k)$
  $\mathbf{m}^k = \theta \mathbf{m}^{k-1} + (1-\theta)\mathbf{g}^k$
  $\mathbf{v}^k = \beta \mathbf{v}^{k-1} + (1-\beta)(\mathbf{g}^k)^{\odot 2}$
  $\mathbf{x}^{k+1} = (1-\lambda\eta)\mathbf{x}^k - \frac{\eta}{\sqrt{\mathbf{v}^k + \varepsilon}} \odot \mathbf{m}^k$
end for

Contributions

We prove the following convergence rate for AdamW measured by ℓ_1 norm

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{x}^k)\|_1\right] \leq O\left(\frac{\sqrt{d}}{K^{1/4}}\sqrt[4]{\sigma_s^2 L(f(\mathbf{x}^1)-f^*)}+\sqrt{\frac{dL(f(\mathbf{x}^1)-f^*)}{K}}\right)$$

and $\|\mathbf{x}^k\|_\infty < \frac{1}{\lambda}$ for all iterates. It can be considered to be analogous to the following optimal convergence rate of SGD

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{x}^k)\|_2\right] \leq O\left(\frac{1}{K^{1/4}}\sqrt[4]{\sigma_s^2 L(f(\mathbf{x}^1)-f^*)}\right)$$

in the ideal case of $\|\nabla f(\mathbf{x})\|_1 = \Theta(\sqrt{d})\|\nabla f(\mathbf{x})\|_2$

Assumptions

- Smoothness: $\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| \leq L\|\mathbf{y} - \mathbf{x}\|$

- Unbiased estimator: $\mathbb{E}\left[\mathbf{g}^k\right] = \nabla f(\mathbf{x}^k)$

- Coordinate-wise bounded noise variance: $\mathbb{E}\left[|\mathbf{g}_i^k - \nabla_i f(\mathbf{x}^k)|^2\right] \leq \sigma_i^2$

Denoting $\sigma_s^2 = \sum_{i=1}^d \sigma_i^2$, we have $\mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|^2\right] \leq \sigma_s^2$

Theorem

Suppose that the three assumptions hold. Define $\hat{\sigma}_s^2 = \max\left\{\sigma_s^2, \frac{L(f(\mathbf{x}^1)-f^*)}{K\gamma^2}\right\}$ with any constant $\gamma \in (0,1]$. Let

$$1-\theta = \sqrt{\frac{L(f(\mathbf{x}^1)-f^*)}{K\hat{\sigma}_s^2}}, \quad \theta \le \beta \le \sqrt{\theta}, \quad \eta = \sqrt{\frac{f(\mathbf{x}^1)-f^*}{4KdL}},$$

$$\varepsilon = \frac{\hat{\sigma}_s^2}{d}, \quad \lambda \le \frac{\sqrt{d}}{\sqrt{72}K^{3/4}}\sqrt[4]{\frac{L^3}{\hat{\sigma}_s^2(f(\mathbf{x}^1)-f^*)}}, \quad \|\mathbf{x}^1\|_\infty \le \sqrt{\frac{K(f(\mathbf{x}^1)-f^*)}{dL}}.$$

Then for AdamW, we have $\|\mathbf{x}^k\|_\infty < \frac{1}{\lambda}$ for all $k = 1, 2, \cdots, K$ and

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\|\nabla f(\mathbf{x}^k)\|_1\right] \le \frac{8\sqrt{d}}{K^{1/4}}\sqrt[4]{\hat{\sigma}_s^2 L(f(\mathbf{x}^1)-f^*)} + 30\sqrt{\frac{dL(f(\mathbf{x}^1)-f^*)}{K}}.$$

Specially, when $\sigma_s^2 \le \frac{L(f(\mathbf{x}^1)-f^*)}{K\gamma^2}$, we have

$$1-\theta = \gamma, \quad \theta \le \beta \le \sqrt{\theta}, \quad \eta = \sqrt{\frac{f(\mathbf{x}^1)-f^*}{4KdL}}, \quad \varepsilon = \frac{L(f(\mathbf{x}^1)-f^*)}{dK\gamma^2},$$

$$\lambda \le \sqrt{\frac{dL\gamma}{72K(f(\mathbf{x}^1)-f^*)}}, \quad \|\mathbf{x}^1\|_\infty \le \sqrt{\frac{K(f(\mathbf{x}^1)-f^*)}{dL}}, \quad \|\mathbf{x}^k\|_\infty < \frac{1}{\lambda},$$

and accordingly $$\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\|\nabla f(\mathbf{x}^k)\|_1\right] \le 38\sqrt{\frac{dL(f(\mathbf{x}^1)-f^*)}{K\gamma}}.$$

Discussions

Optimality of Our Convergence Rate

- Optimal with respect to $K, \sigma_s, L, f(\mathbf{x}^1) - f^*$
- We have empirically confirmed $\|\nabla f(\mathbf{x})\|_1 = \Theta(\sqrt{d})\|\nabla f(\mathbf{x})\|_2$ on real deep neural networks training, demonstrating that our rate can be considered to be analogous to the optimal convergence rate of SGD
- Jiang et al . [COLT2025] established the $O\left(\sqrt[4]{\frac{d\|\boldsymbol{\sigma}\|_1^2 L(f(\mathbf{x}^1)-f^*)}{K}} + \sqrt{\frac{dL(f(\mathbf{x}^1)-f^*)}{K}}\right)$ lower bound for SGD when measuring the gradients by ℓ_1 norm. This lower bound precisely aligns with our convergence rate when $\|\boldsymbol{\sigma}\|_1 \approx \sqrt{d}\|\boldsymbol{\sigma}\|_2 = \sqrt{d}\sigma_s$

Discussions

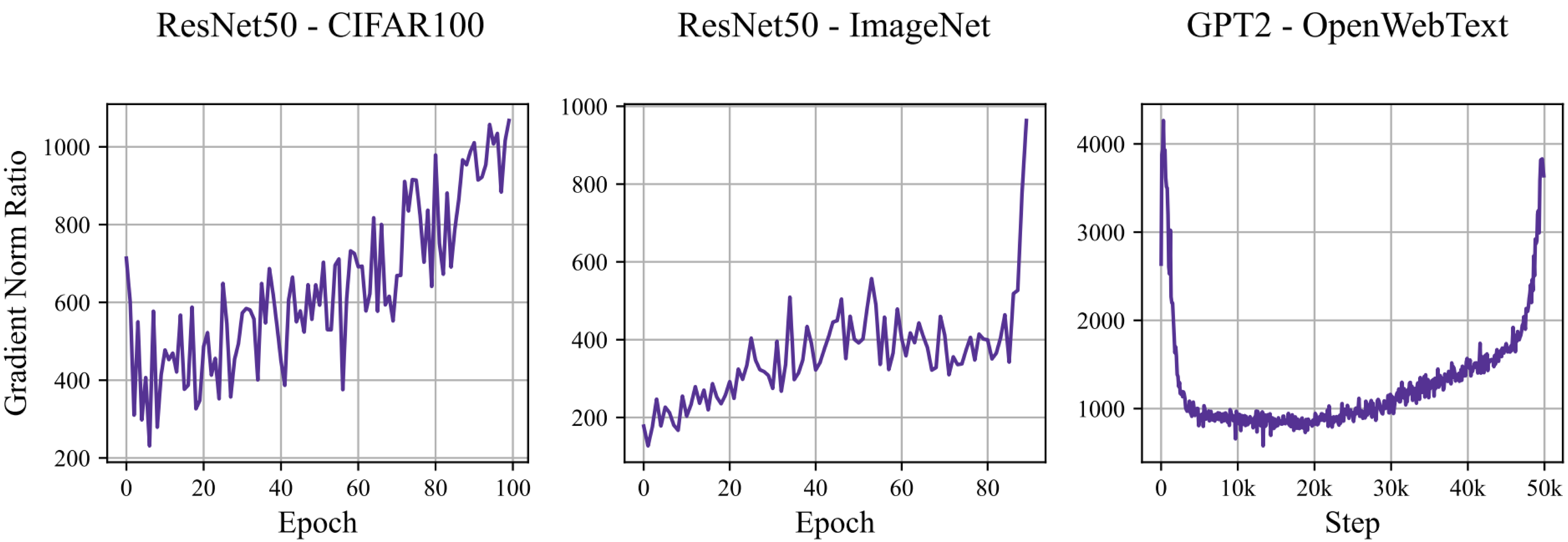


Illustration of $\|\nabla f(\mathbf{x}^k)\|_1 = \Theta(\sqrt{d})\|\nabla f(\mathbf{x}^k)\|_2$ for AdamW over epochs/steps. The gradient norm ratio shows $\frac{\|\nabla f(\mathbf{x}^k)\|_1}{\|\nabla f(\mathbf{x}^k)\|_2}$, and $\sqrt{d} = 4868$, 5060, and 11136, respectively.

Discussions

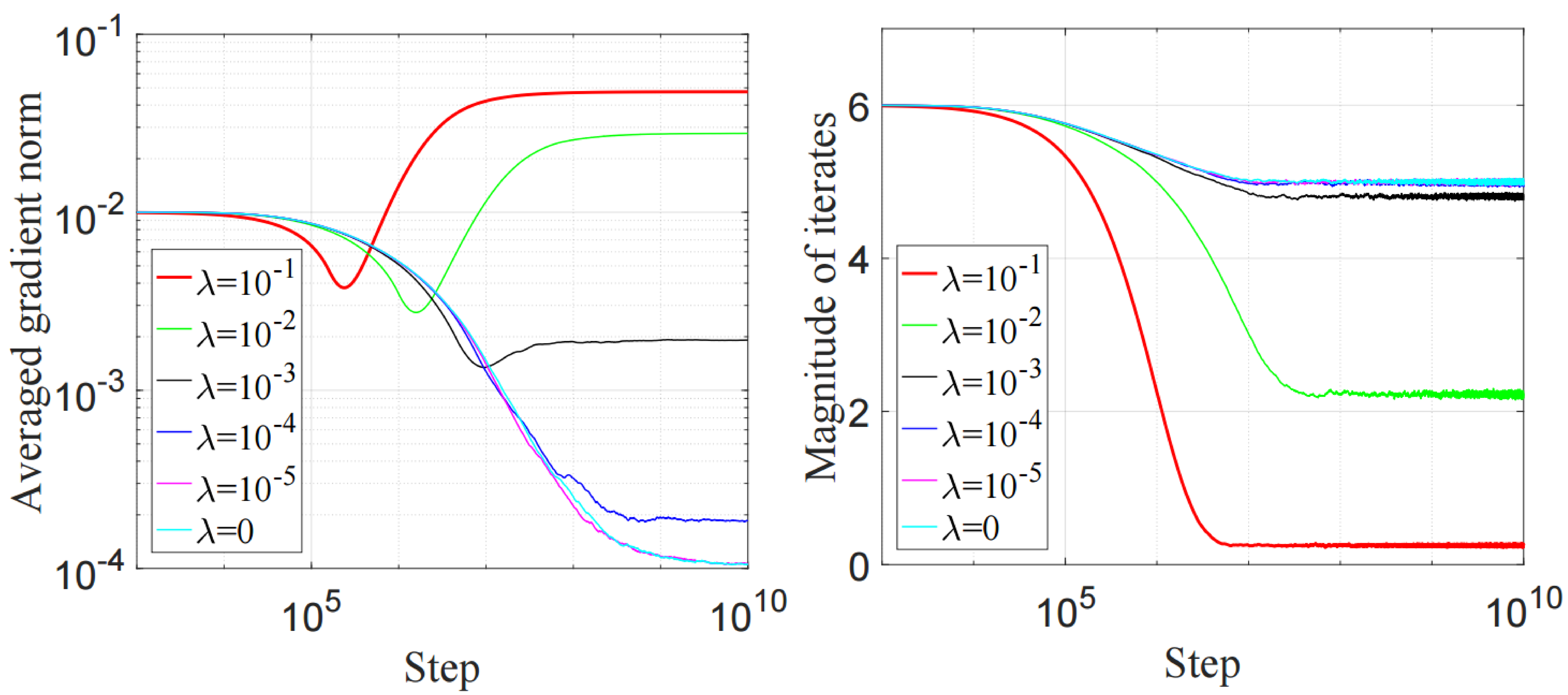
Reasonable Weight Decay Parameter

- We require $\lambda \leq \frac{\sqrt{d}}{\sqrt{72}K^{3/4}}\sqrt[4]{\frac{L^3}{\hat{\sigma}_s^2(f(\mathbf{x}^1)-f^*)}}$
- In modern deep neural networks, the dimension d is extremely large, making $\frac{\sqrt{d}}{K^{3/4}}$ almost certainly exceed 0.01, which is the default setting of λ in PyTorch
- In our experiments, we observe $(K,d)=(39100, 2.37\times 10^7), (28080, 2.56\times 10^7),$ and $(50000, 1.24\times 10^8)$, resulting in $\frac{\sqrt{d}}{K^{3/4}}\approx 1.75, 2.33,$ and 3.33, respectively
- We empirically show that large values of λ exceeding a certain threshold may cause AdamW neither to converge to the minimum solution nor to a KKT point

Discussions

$$f(x)=\frac{(x-x^*)^2}{200},\ \text{with}\ g(x)=\begin{cases} x-x^*-1, & \text{with probability } p=0.1,\\ -\frac{1}{10}(x-x^*-\frac{10}{9}), & \text{with probability } 1-p. \end{cases}$$

we set $K=10^{10}, \theta=1-\frac{1}{\sqrt{K}}, \beta=\sqrt{\theta}, \eta=\frac{1}{\sqrt{K}}, \varepsilon=10^{-10}, \mathrm{m}^0=0, v^0=0, x^1=x^*+1$ with $x^*=5$, and test $\lambda=\{10^{-1},10^{-2},10^{-3},10^{-4},10^{-5},0\}$. We observe that AdamW fails to converge to x^* when $\lambda=\{10^{-1},10^{-2},10^{-3}\}$.



Illustrations of $\frac{1}{k}\sum_{t=1}^{k}|\nabla f(x^t)|$ (left) and x^k (right) over steps on the toy example

Discussions

Small ε Setting

- We set $\varepsilon = \frac{\hat{\sigma}_s^2}{d} = \max\left\{\frac{\sigma_s^2}{d}, \frac{L(f(\mathbf{x}^1)-f^*)}{dK\gamma^2}\right\}$, which remains small due to extremely large d and modest σ_s^2
- We empirically show that $\frac{\sigma_s^2}{d} \approx 10^{-7}, 10^{-9},$ and 10^{-10} in our experiments

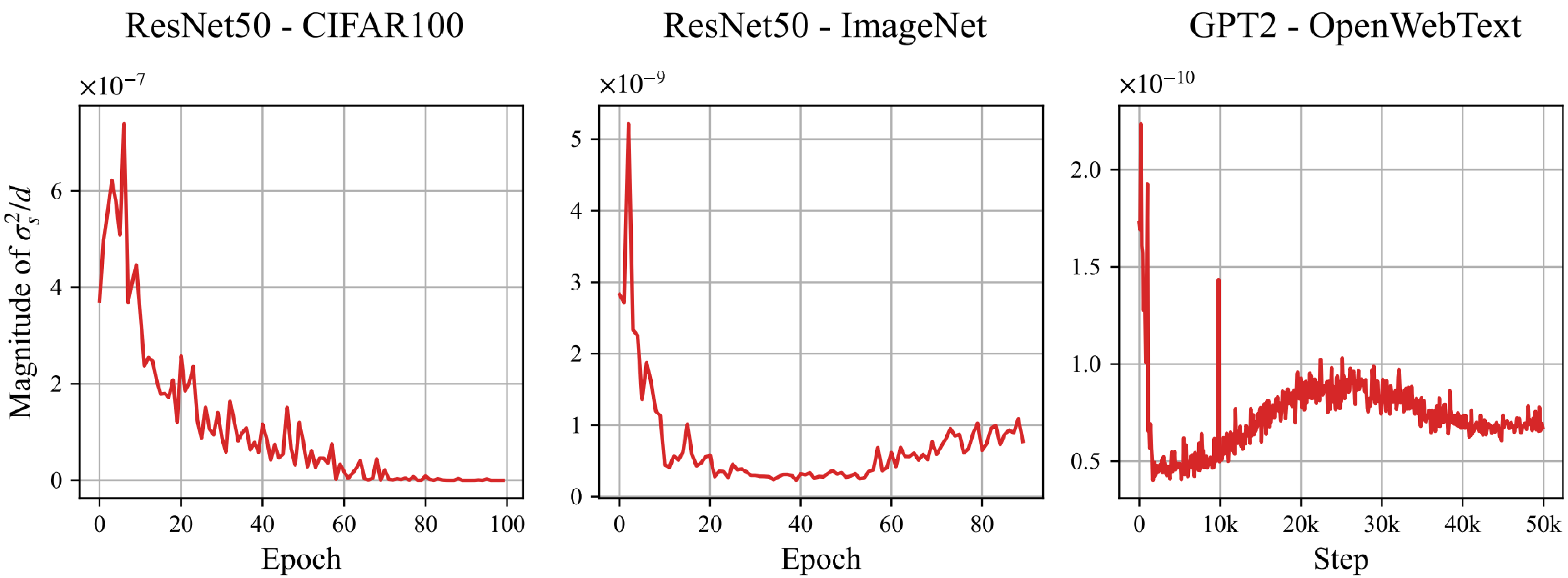


Illustration of small $\frac{\sigma_s^2}{d}$ over epochs/steps. The magnitude σ_s^2 is approximated by $\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|^2$ for AdamW without taking expectation, and $d = 2.37 \times 10^7$, 2.56×10^7 , and 1.24×10^8 , respectively.

Thanks for listening!