NEURAL INFORMATION PROCESSING SYSTEMS

MBZUAI

JOHNS HOPKINS UNIVERSITY

INNOPOLIS UNIVERSITY

Remove that Square Root: A New Efficient Scale-Invariant Version of AdaGrad

Sayantan Choudhury, Nazarii Tupitsa, Nicolas Loizou,
Samuel Horvarth, Martin Takac, Eduard Gorbunov

- We are interested in solving the unconstrained minimization problem

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

- These problems appear in computer vision, natural language processing and other machine-learning tasks.

- The standard algorithm for solving this problem is stochastic gradient descent (SGD):

$$w_{t+1} = w_t - \nu_t g_t$$

- However, tuning the step size is a significant issue when training large models. Therefore, we need adaptive methods.
- AdaGrad step size:

$$\nu_t = \frac{\beta}{\sqrt{\operatorname{diag}\left(\Delta I + \sum_{\tau=0}^{t} g_\tau g_\tau^\top\right)}}$$

- AdaGrad is not scale-invariant and suffers when data is poorly scaled.
- We propose a scale-invariant algorithm, KATE (by removing the square root from the denominator), which uses step size:

$$\nu_t = \frac{\beta m_t}{\operatorname{diag}\left(\Delta I + \sum_{\tau=0}^{t} g_\tau g_\tau^\top\right)}$$

Require: Initial point $w_0 \in \mathbb{R}^d$, step size $\beta > 0, \eta \in \mathbb{R}^d_+$ and $b_{-1}, m_{-1} = 0$.

1: **for** $t = 0, 1, ..., T$ **do**
2: Compute $g_t \in \mathbb{R}^d$ such that $\mathbb{E}[g_t] = \nabla f(w_t)$.
3: $b_t^2 = b_{t-1}^2 + g_t^2$
4: $m_t^2 = m_{t-1}^2 + \eta g_t^2 + \frac{g_t^2}{b_t^2}$
5: $w_{t+1} = w_t - \frac{\beta m_t}{b_t^2} g_t$

Scan me!

- We rigorously prove that KATE is scale-invariant for solving generalised linear models (GLMs).

$$\min_{w\in\mathbb{R}^d} f(w) = \frac{1}{n}\sum_{i=1}^{n} \varphi_i\left(x_i^\top w\right)$$

Linear: $\varphi_i(z) = (z - y_i)^2$

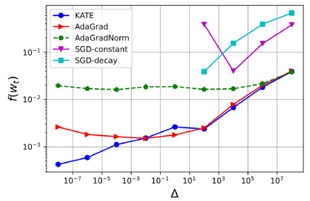
Logistic: $\varphi_i(z) = \log\left(1 + \exp(-y_i z)\right)$

- Convergence Guarantees:

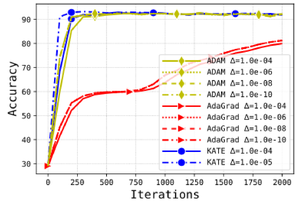
Algorithm	Convergence rate	Scale invariance
AdaGradNorm (Ward et al., 2020)	$\mathcal{O}\left(\log T/\sqrt{T}\right)$	✗
AdaGrad (Défossez et al., 2020)	$\mathcal{O}\left(\log T/\sqrt{T}\right)$	✗
Adam (Défossez et al., 2020)	$\mathcal{O}\left(\log T/\sqrt{T}\right)$	✗
KATE (this work)	$\mathcal{O}\left(\log T/\sqrt{T}\right)$	✓

- Numerical Experiments:

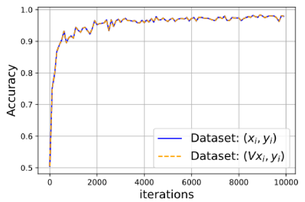
Robustness of KATE



Training ResNet18 on CIFAR10



Scale-Invariance of KATE



BERT Finetuning on Emotions