Adam revisited: a weighted past gradients perspective

Hui ZHONG, Zaiyi CHEN, Chuan QIN, Zai HUANG, Vincent W. ZHENG, Tong XU, Enhong CHEN

Frontiers of Computer Science, DOI: 10.1007/s11704-019-8457-x

Problems & Ideas

- Problems of the non-convergence of Adam
  - Give a better solution to fix non-convergence issues of Adam
  - Existing fix solutions can be further improved
- Ideas: Weighted Adaptive Algorithm
  - Weighted Adaptive Gradient Method Framework (WAGMF)
  - Applying the linear growing weighting strategy to WAGMF.

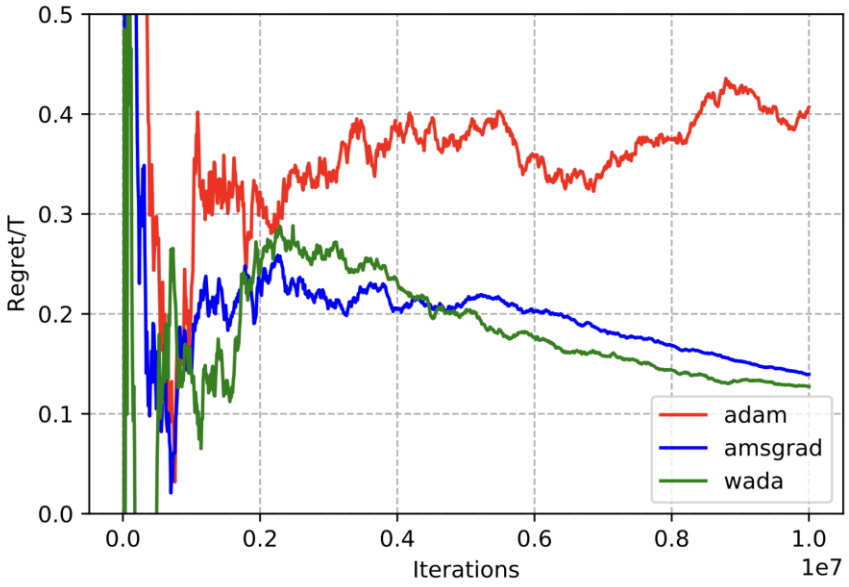

Figure 1. The non-convergence of Adam. As we can see, the average of the Regret doesn't converge to 0 for Adam.

Main Contributions

- We provide a convergence analysis of WADA, and demonstrate that it can achieve following weighted data-dependent regret bound

$$R(T) \leqslant \frac{D_\infty^2}{2(1-\beta_1)} \sum_{i=1}^{d} \sqrt[4]{\sum_{j=1}^{T} j \cdot g_{j,i}^2} + \frac{\beta_1 D_\infty^2 \sqrt{G_\infty}}{2(1-\beta_1)(1-\lambda)^2} + \frac{\alpha d G_\infty}{(1-\beta_1)^2} \sum_{i=1}^{d} \sqrt[4]{\sum_{j=1}^{T} j \cdot g_{j,i}^2}.$$

- The experimental results demonstrate that our methods outperform ADAM and its variants on training convex problems and deep neural networks.

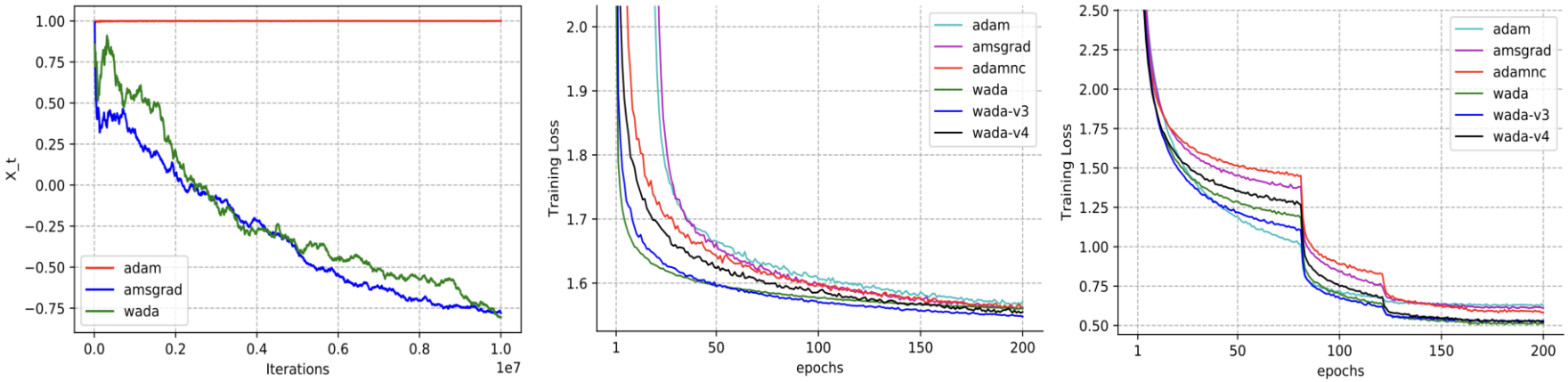

Figure 2. Figures for the convergence of wada, the training loss on Softmax, the training loss on ResNet.