Convergence of Optimizers Without Bounded Gradient Assumption

Roger Wang, Eric Xia
rogerwyf@uw.edu, ericxia@uw.edu

1 Introduction

The primary goal of most deep learning models is to minimize the loss function. Optimizers are crucial for the task of updating model weights such that the model will actually converge to a minimum in a computationally efficient manner instead of overshooting or moving away from the minimum. While existing optimizers are intuitively straightforward in convex learning, in non-convex settings they (notably for Adam-type adaptive gradient methods) often require the assumption on the boundedness of gradients for achieving convergence.

For example, for an objective function $f(x)$ satisfying Polyak-Łojasiewicz conditions [1], convergence rates of $O(1/T)$ were established for Stochastic Gradient Descent (SGD) under the assumption that $\mathbb{E}[||\nabla f_i(x_k)||^2] \leq C^2$ for all x_k and some C, where f_i typically represents the fit on an individual training sample [2].

While convenient, the imposition of assumptions on the boundedness of gradients can be difficult to verify in practical settings, hence in recent years there has been a trend in research efforts towards analyzing existing optimization methods and proposing novel methods without such assumptions.

2 General Assumptions

Preprint. Under review.

3 Stochastic Gradient Descent for Structured Nonconvex Functions

3.1 Previous research & Motivation

3.2 Additional Assumptions

3.3 Main Result

3.4 Key Approaches & Insights

3.5 Discussion

4 Stochastic Gradient Descent with Momentum

4.1 Previous research & Motivation

4.2 Additional Assumptions

4.3 Main Result

4.4 Key Approaches & Insights

4.5 Discussion

5 AdaGrad

5.1 Previous research & Motivation

5.2 Additional Assumptions

5.3 Main Result

5.4 Key Approaches & Insights

5.5 Discussion

6 RMSProp

6.1 Previous research & Motivation

6.2 Additional Assumptions

6.3 Main Result

6.4 Key Approaches & Insights

6.5 Discussion

7 Adam

7.1 Previous research & Motivation

7.2 Additional Assumptions

7.3 Main Result

7.4 Key Approaches & Insights

7.5 Discussion

8 Addition Dicussion & Comments

9 Future Research

References

[1] B.T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(63)90382-3. URL https://www.sciencedirect.com/science/article/pii/0041555363903823.

[2] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. *CoRR*, abs/1608.04636, 2016. URL http://arxiv.org/abs/1608.04636.