Convergence of Optimizers Without Bounded Gradient Assumption

Yufeng (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 Proposal and Selected Papers

We plan to survey and analyze five recent papers on convergence issues without bounded gradient assumption of optimization algorithms including Root Mean Squared Propagation (RMSProp) [3], SGD [4], SGD with momentum (SGDM) [5], Adaptive Moment Estimation (Adam) [6], and Adaptive Subgradient (AdaGrad) [7], comparing and contrasting their approaches, methodologies, and findings, and discussing their broader implications for the convergence problem and the general field of optimizers.

As a non-momentum specificity of Adam [10], RMSProp [8] is known for performing well empirically, despite theoretical analyses of it suggesting divergence even for simple complex functions. In their ICRL 2021 paper, Shi et al. [3] examined a counter-example presented by Reddi et al. [9] in 2018 in their study of the convergence of Adam, discovering a gap in their analysis of convergence for large β_2 parameter. They then ran simulations based on this gap. By plotting whether convergence was achieved for different choices of β_2 for RMSprop, they found that in general, there was a continuous and nontrivial curve sloping upwards, from divergence to convergence, suggesting their conjecture that RMSprop converges for large enough β_2. This paper is notable and interesting as it is the first to prove the convergence of RMSprop without the bounded gradient assumption.

In their AISTATS 2021 paper, Gower et al. [4] showed that SGD converges at a rate of $\mathcal{O}(1/\sqrt{k})$ on Quasar (Strongly) Convex functions and proved linear convergence to a neighborhood for functions satisfying the Polyak-Łojasiewicz conditions without any bounded gradient assumption, instead relying on the expected residual (ER) condition for support. The significance of this study resides in the provided insights on the complexity of minibatching and determination of optimal batch sizes, as well as the demonstration that for models interpolating the training data, the ER condition could be discarded and still give state-of-the-art results.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

The third paper we chose is interesting for a similar reason: it is the first convergence guarantee for Multistage SGDM without uniformly bounded gradient assumption by Liu et al. [5]—Multistage refers to applying a constant stepsize which is then dropped by a constant factor to encourage fine-tuning of training, and the momentum weight is either kept unchanged or gradually increased. Moreover, although SGDM has an upper hand against SGD empirically, the theoretical understanding of momentum in the stochastic case is far from complete, given that previous analyses of SGDM either "provide worse convergence bounds than those of SGD, or assume Lipschitz or quadratic objectives, which fail to hold in practice." This work demonstrated that SGDM has the same convergence bound as SGD for both strongly convex and nonconvex functions without uniformly bounded gradient assumption.

In their NeurIPS 2021 paper, Huang et al. [6] proposed a faster and universal framework for adaptive gradients that includes most already existing adaptive gradient forms of optimizers. Under the current framework of adaptive gradients, the adaptive learning rate varies between individual algorithms, e.g. Adam vs. AdaGrad-Norm [11]. Not only does the proposed framework have a universal adaptive learning rate, it also guarantees convergence without assuming boundedness of the gradient. Furthermore, they experimentally validate the outperformance of their algorithm against existing adaptive algorithms, demonstrating the signifance of this proposition as it is both more generalizable theoretically and empirically better in performance.

Finally, last paper we look at is a very recent study from Faw et al. [7] on convergence rates of AdaGrad-Norm[11]. This work holds a particular significance in their demonstration that neither the bounded gradient assumption nor the bounded variance assumption are necessary in AdaGrad-Norm[11], thus adaptive gradient methods converge in much broader situations than previously understood.

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.

[3] Naichen Shi, Dawei Li, Mingyi Hong, and Ruoyu Sun. RMSprop converges with proper hyper-parameter. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3UDSdyIcBDA.

[4] Robert M. Gower, Othmane Sebbouh, and Nicolas Loizou. Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation, 2020. URL https://arxiv.org/abs/2006.10311.

[5] Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18261–18271. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d3f5d4de09ea19461dab00590df91e4f-Paper.pdf.

[6] Feihu Huang, Junyi Li, and Heng Huang. Super-adam: Faster and universal framework of adaptive gradients, 2021. URL https://arxiv.org/abs/2106.08208.

[7] Matthew Faw, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward. The power of adaptivity in sgd: Self-tuning step sizes with unbounded gradients and affine variance, 2022. URL https://arxiv.org/abs/2202.05791.

[8] Tijmen Tieleman, Geoffrey Hinton, et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[9] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *CoRR*, abs/1904.09237, 2019. URL http://arxiv.org/abs/1904.09237.

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

[11] Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes, 2018.