Analysis of Stochastic Gradient Descent for Learning Linear Neural Networks

Abstract—**In this work we analyze stochastic gradient descent (SGD) for learning deep linear neural networks. We use an analytical approach that combines SGD iterates and gradient flow trajectories base on stochastic approximation theory. Then establish the almost sure boundedness of SGD iterates and its convergence guarantee for learning deep linear neural networks. Most studies on the analysis of SGD for nonconvex problem have entirely focused on convergence property which only indicate that the second moment of the loss function gradient tend to zero [11]–[14]. Our study demonstrates the convergence of SGD to a critical point of the square loss almost surely for learning deep linear neural networks.**

EXTENDED ABSTRACT

The remarkable performance of deep learning (DL) models strongly relies on its capacity to best tune its parameters during the learning process. In general, this learning process essentially consists of optimizing an associated cost function composed of the training data over the model parameters, via (stochastic) gradient descent or any of its variants. This then result in an optimization problem of a highly non-convex cost function which is challenging to analyze.

Nevertheless, practitioners obtain impressive results by training and deploy DL models in a vast amount of applications. Besides, these impressive results of DL models, there is still lack of theoretical explanation that completely describe the process that enable its models to achieve spectacular results. This gap between the theory behind DL models and practical DL has made its models to be observed, so far, as a black box. Consequently, a lot of effort is being devoted to shed light in this regard. Two main directions of contributions were developed to address the issue of describing the training process of DL models namely; works analyzing gradient flow (GF) (which is GD with infinitesimal step-sizes) for learning deep linear neural networks and works analyzing (S)GD with practical step-sizes for learning deep linear neural networks. Linear neural networks here refer to neural networks with linear activation function (in our work we used identity). Even with a linear activation function, the corresponding cost function still remains highly non-convex due to the overparameterized matrix of the network.

The analysis of GF for learning deep linear neural networks has provided intuitive properties on its optimization process [2], [3], [6]. For instance, the work in [2] suggests that overparameterization caused by depth leads gradient descent for training a deep linear neural networks to behave as if it were training a shallow network using a specific preconditioning scheme at the same time. Authors of [6] proved that if the number of instances in the training set is greater than the dimensions of the input data and the dimensions of the input data is in turn greater than the dimension of the output layer, GF converges to a global minimum for almost every initial condition. In addition, their results require the minimum dimension of the hidden layers to be greater or equal to the dimension of the output layer. The study in [3] significantly extends previous convergence results and provides a more general dimension setting. However, these works on GF are still far from practice, given that they are all base on infinitesimal step sizes and do not provide any concrete informations on realistic step sizes.

The preliminary work that addresses GD with realistic step sizes for learning a linear neural networks is provided in [4]. This work precisely studies GD initialized with identity for learning a deep linear residual network (a particular subclass of neural networks where layers dimensions are all equal) over whitened data. It shows that, if the initial value of the cost function is sufficiently close to a global minimum, or a global minimum is attained when the product of all layers is positive definite then GD converges to a global minimum at a linear rate. Furthermore, authors of [1] introduced approximate balanced initialization then extended the results in [4] into a setting that enables the input, output and hidden dimensions to take any other values that could even allows the minimum intermediate layer dimension to be greater or equal to the minimum between the first and last hidden layer dimension. Besides, they proved the convergence of GD with a constant step-size using a deficiency margin condition which indicates that the target matrix should be closed to the parameterized matrix of the network at initialization. GD run with constant step-sizes also converges to a global optimum for learning linear neural networks with a near-zero initialization [8]. Authors of [8] precisely devised their convergence guarantee on a single variable regression problem. Moreover, the work in [17] complements and generalizes previous results of GD for learning linear neural networks. It carefully analyzes GD dynamic for almost all initialization and takes into account both constant step-sizes with decreasing step-sizes. This work establishes the global convergence of GD to a minimizer of the square loss without the deficient margin condition and not necessarily over whitened data. Our study focus on the extension of [17] from GD to SGD.

In practice, computing the full gradient is time demanding for very large-scale problems. Hence, practitioners use SGD (or any of its variants) which only involve a small number of randomly selected data points in the computation of the gradient and is, therefore, less costly. Our work makes further steps ahead in closing the gap between theoritical DL and practical DL by directly analyzing SGD for learning linear neural networks. In contrast to GF trajectory for learning neural networks whose bound has been established and convergence properties are clearly proved in the literature, tangibly established SGD (GF stochastic approximation) iterates bound and convergence for learning neural networks are still open questions. Indeed, some results in literature of online learning algorithms such as SGD assume that its iterates are bounded [7], [15]. Preliminaries results on stochastic approximation provides the convergence of SGD under the assumption that its supremum over iteration is bounded [15]. Other results in the state of the art circumvent the boundedness assumption on SGD iterates by using a condition that requires the gradient of the cost function to be Lipschitz [10], [11], [13], [16]. However, SGD sequence can escape to infinity and would, therefore, need a careful analysis. In addition, the Lipschitz condition on cost function gradient is not a fair condition in the sense that parameterization makes it impossible to be satisfied with linear neural networks cost function (such as square loss for instance).

Work in [9] presents a general structure of online learning algorithms and indicates that addressing these algorithms with stochastic approximations theory produces desired convergence result. It used stochastic approximations theory to devise the convergence of SGD for minimizing a convex loss function. In our work, we use an approach that combines SGD iterates and GF trajectories base on stochastic approximation theory. Then develop the almost sure boundedness of SGD iterates and its convergence guarantee for training deep linear neural networks. Giving that works of GF have provided comprehensive analysis, involving GF in the analysis of SGD is very helpful in the sense that some of GF's properties would complement the analysis of SGD and, therefore, contribute in overcoming the major difficulties that arise when addressing SGD alone. This analytical approach was introduced in [5]. More precisely, this work defined a continuous time process via an interpolation of a discrete time stochastic process. Then involved it in the study to determine the almost sure dynamics of the discrete stochastic process (as a stochastic approximation process of a semi-flow) with decreasing step-sizes. In our work, we define such continuous time process by interpolating SGD iterates for learning linear neural networks then use it to address the long term behavior of SGD through stochastic approximation theory.

The great benefit in training DL models with SGD algorithm in practice has resulted its convergence properties to become a central topic of research [10]–[14], [16]. Study done in [16] examined the almost sure convergence of SGD for non-convex cost function partly based on a similar analytical approach to our work. This study assumes that the sublevels of the objective function are bounded. It also uses the smoothness assumption for cost function which is generally used in the SGD literature [12]–[14], [18] and basically means the gradient of the objective function is Lipschitz. The potential consequence of this condition is that it provides a relation which ensure a decrease of loss function iterates and, therefore, constitute a plausible argument for convergence. Unfortunately, loss functions such as square loss of linear neural networks does not satisfy these two conditions through SGD algorithm due to the factorized matrix of the network. In contrast, we use approximate balanced initialization of the weight matrices and elaborate a decrease relation of loss function iterates in expectation by extending previous results of [17]. Most studies on the analysis of SGD for nonconvex problem have entirely focused on convergence property which only indicate that the second moment of the loss function gradient tend to zero [11]–[14]. Besides, we are not aware of a single work that concretely provides the convergence properties of SGD for learning deep linear neural networks. Our study demonstrates the convergence of SGD to critical a point of the square loss for learning deep linear neural networks.

References

[1] S. Arora, N. Cohen, N. Golowich, and W. Hu. A convergence analysis of gradient descent for deep linear neural networks, 2018.

[2] S. Arora, N. Cohen, and E. Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. *Preprint. https://arxiv.org/abs/1802.06509*, 2018.

[3] B. Bah, H. Rauhut, U. Terstiege, and M. Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *arXiv preprint arXiv:1910.05505*, 2019.

[4] P. Bartlett, D. Helmbold, and P. Long. Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks. In *International conference on machine learning*, pages 521–530. PMLR, 2018.

[5] M. Benaïm and M. W. Hirsch. Asymptotic pseudotrajectories and chain recurrent flows, with applications. *Journal of Dynamics and Differential Equations*, 8(1):141–176, 1996.

[6] Y. Chitour, Z. Liao, and R. Couillet. A geometric approach of gradient descent algorithms in neural networks. *arXiv preprint arXiv:1811.03568*, 2018.

[7] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.

[8] O. Elkabetz and N. Cohen. Continuous vs. discrete optimization of deep neural networks. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[9] L. eon Bottou. Online learning and stochastic approximations. *Onlinelearning in neural networks*, 17(9):142, 1998.

[10] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pages 797–842. PMLR, 2015.

[11] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[12] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

[13] A. Khaled and P. Richtárik. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2023. Survey Certification.

[14] Y. Lei, T. Hu, G. Li, and K. Tang. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE transactions on neural networks and learning systems*, 31(10):4394–4400, 2019.

[15] L. Ljung. Analysis of recursive stochastic algorithms. *IEEE transactions on automatic control*, 22(4):551–575, 1977.

[16] P. Mertikopoulos, N. Hallak, A. Kavis, and V. Cevher. On the almost sure convergence of stochastic gradient descent in non-convex problems. *Advances in Neural Information Processing Systems*, 33:1117–1128, 2020.

[17] G. M. Nguegnang, H. Rauhut, and U. Terstiege. Convergence of gradient descent for learning linear neural networks. *arXiv preprint arXiv:2108.02040*, 2021.

[18] S. Vlaski and A. H. Sayed. Second-order guarantees of stochastic gradient descent in nonconvex optimization. *IEEE Transactions on Automatic Control*, 67:6489–6504, 2019.