Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup

Sebastian Goldt[1], Madhu S. Advani[2], Andrew M. Saxe[3]
Florent Krzakala[4], Lenka Zdeborová[1]
[1] Institut de Physique Théorique, CNRS, CEA, Université Paris-Saclay, Saclay, France
[2] Center for Brain Science, Harvard University, Cambridge, MA 02138, USA
[3] Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom
[4] Laboratoire de Physique Statistique, Sorbonne Universités,
Université Pierre et Marie Curie Paris 6, Ecole Normale Supérieure, 75005 Paris, France

Abstract

Deep neural networks achieve stellar generalisation even when they have enough parameters to easily fit all their training data. We study this phenomenon by analysing the dynamics and the performance of over-parameterised two-layer neural networks in the teacher-student setup, where one network, the student, is trained on data generated by another network, called the teacher. We show how the dynamics of stochastic gradient descent (SGD) is captured by a set of differential equations and prove that this description is asymptotically exact in the limit of large inputs. Using this framework, we calculate the final generalisation error of student networks that have more parameters than their teachers. We find that the final generalisation error of the student increases with network size when training only the first layer, but stays constant or even decreases with size when training both layers. We show that these different behaviours have their root in the different solutions SGD finds for different activation functions. Our results indicate that achieving good generalisation in neural networks goes beyond the properties of SGD alone and depends on the interplay of at least the algorithm, the model architecture, and the data set.

Deep neural networks behind state-of-the-art results in image classification and other domains have one thing in common: their size. In many applications, the free parameters of these models outnumber the samples in their training set by up to two orders of magnitude[1,2]. Statistical learning theory suggests that such heavily over-parameterised networks generalise poorly without further regularisation[3–9], yet empirical studies consistently find that increasing the size of networks to the point where they can easily fit their training data and beyond does not impede their ability to generalise well, even without any explicit regularisation[10–12]. Resolving this paradox is arguably one of the big challenges in the theory of deep learning.

One tentative explanation for the success of large networks has focused on the properties of stochastic gradient descent (SGD), the algorithm routinely used to train these networks. In particular, it has been proposed that SGD has an implicit regularisation mechanism that ensures that solutions found by SGD generalise well irrespective of the number of parameters involved, for models as diverse as (over-parameterised) neural networks[10,13], logistic regression[14] and matrix factorisation models[15,16].

In this paper, we analyse the dynamics of one-pass (or online) SGD in two-layer neural networks. We focus in particular on the influence of over-parameterisation on the final generalisation error. We use the teacher-student framework[17,18], where a training data set is generated by feeding random inputs through a two-layer neural network with M hidden units called the *teacher*. Another neural network, the *student*, is then trained using SGD on that data set. The generalisation error is defined as the mean

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.