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.

squared error between teacher and student outputs, averaged over all of input space. We will focus on student networks that have a larger number of hidden units $K \geq M$ than their teacher. This means that the student can express much more complex functions than the teacher function they have to learn; the students are thus over-parameterised with respect to the generative model of the training data in a way that is simple to quantify. We find this definition of over-parameterisation cleaner in our setting than the oft-used comparison of the number of parameters in the model with the number of samples in the training set, which is not well justified for non-linear functions. Furthermore, these two numbers surely cannot fully capture the complexity of the function learned in practical applications.

The teacher-student framework is also interesting in the wake of the need to understand the effectiveness of neural networks and the limitations of the classical approaches to generalisation[11]. Traditional approaches to learning and generalisation are data agnostic and seek worst-case type bounds[19]. On the other hand, there has been a considerable body of theoretical work calculating the generalisation ability of neural networks for data arising from a probabilistic model, particularly within the framework of statistical mechanics[17,18,20–22]. Revisiting and extending the results that have emerged from this perspective is currently experiencing a surge of interest[23–28].

In this work we consider two-layer networks with a large input layer and a finite, but arbitrary, number of hidden neurons. Other limits of two-layer neural networks have received a lot of attention recently. A series of papers[29–32] studied the mean-field limit of two-layer networks, where the number of neurons in the hidden layer is very large, and proved various general properties of SGD based on a description in terms of a limiting partial differential equation. Another set of works, operating in a different limit, have shown that infinitely wide over-parameterised neural networks trained with gradient-based methods effectively solve a kernel regression[33–38], without any feature learning. Both the mean-field and the kernel regime crucially rely on having an infinite number of nodes in the hidden layer, and the performance of the networks strongly depends on the detailed scaling used[38,39]. Furthermore, a very wide hidden layer makes it hard to have a student that is larger than the teacher in a quantifiable way. This leads us to consider the opposite limit of large input dimension and finite number of hidden units.

Our **main contributions** are as follows:

(i) The dynamics of SGD (online) learning by two-layer neural networks in the teacher-student setup was studied in a series of classic papers[40–44] from the statistical physics community, leading to a heuristic derivation of a set of coupled ordinary differential equations (ODE) that describe the *typical* time-evolution of the generalisation error. *We provide a rigorous foundation of the ODE approach to analysing the generalisation dynamics in the limit of large input size by proving their correctness.*

(ii) These works focused on training only the first layer, mainly in the case where the teacher network has the same number of hidden units and the student network, $K = M$. *We generalise their analysis to the case where the student's expressivity is considerably larger than that of the teacher* in order to investigate the *over-parameterised regime* $K > M$.

(iii) We provide a detailed analysis of the dynamics of learning and of the generalisation when only the first layer is trained. We derive a reduced set of coupled ODE that describes the generalisation dynamics for any $K \geq M$ and obtain analytical expressions for the asymptotic generalisation error of networks with linear and sigmoidal activation functions. Crucially, we find that with all other parameters equal, the final generalisation error *increases* with the size of the student network. In this case, SGD alone thus does not seem to be enough to regularise larger student networks.

(iv) We finally analyse the dynamics when learning both layers. We give an analytical expression for the final generalisation error of sigmoidal networks and find evidence that suggests that SGD finds solutions which amount to performing an effective model average, thus improving the generalisation error upon over-parameterisation. In linear and ReLU networks, we experimentally find that the generalisation error does change as a function of K when training both layers. However, there exist student networks with better performance that are fixed points of the SGD dynamics, but are not reached when starting SGD from initial conditions with small, random weights.

Crucially, we find this range of different behaviours while keeping the training algorithm (SGD) the same, changing only the activation functions of the networks and the parts of the network that are trained. Our results clearly indicate that the implicit regularisation of neural networks in our setting goes beyond the properties of SGD alone. Instead, a full understanding of the generalisation properties of even very simple neural networks requires taking into account the interplay of at least