Risk bounds for deep learning

Bos, J.M.

Chapter 5

General discussion

In this thesis, risk bounds for deep learning have been established in various settings. The central aim was to use statistical theory to obtain new insights into the performance of deep neural networks. Chapter 2 showed that deep neural networks can achieve optimal convergence rates under the (truncated) cross-entropy risk for the conditional class probabilities in the classification model. Furthermore, this chapter includes approaches to deal with the unboundedness of the cross-entropy loss for conditional class probabilities near zero. The used approaches are truncation and the small-value bound assumption. This last bound controls the probability that the conditional probabilities are close to zero. In Chapter 3 a method was studied that transforms the unsupervised density estimation problem into a supervised regression problem. In this way, convergence rates were obtained using existing results for regression. These rates show that deep neural networks can exploit a compositional structure to partly circumvent the curse of dimensionality. Furthermore, it was demonstrated that different existing density models indeed satisfy the compositional structure assumption. Chapter 4 considered an optimization method motivated by biological networks: forward gradient descent. It was shown that the extra randomness in forward gradient descent leads to a convergence rate in the linear regression model that is a dimension-dependent factor $d\log(d)$ slower than the optimal rate that can be achieved by gradient descent.

These findings rely on certain assumptions. This chapter discusses some of these underlying assumptions in more detail, relates them to existing literature on neural networks and discusses whether and how these assumptions can be adapted to extend the results in this thesis.

5.1 Statistical theory and training of neural networks

In Chapters 2 and 3 the risk bounds depend on the assumption that the estimator has empirical risk close to the risk of an empirical risk minimizer. The analysis of empirical risk minimizers without specifying how to obtain them is standard in statistical literature on risk bounds for deep neural networks. Examples of this approach include [67, 74, 76, 134]. In practice it is non-trivial to compute such an estimator. One additional issue is that the constraints on the deep neural network classes in Chapters 2 and 3 do not necessarily match the network structures considered in the deep learning literature. Most importantly, overparametrized neural networks are studied in practice because they can be trained relatively easily and successfully by simple gradient methods [10, 15]. But such overparametrized neural networks do not match the neural network classes studied in this thesis.

On the other hand, Chapter 4 considers forward gradient descent in the linear regression model. In this case the training method is the focus of the analysis, including an explicit (theoretical) learning rate. When the relevant properties of the covariance matrix Σ are known, the theory provides all the information required to run the method. This in contrast to the training of neural networks, as done in the simulation study in Chapter 3, where various (training) parameters must be chosen before the neural networks can be trained properly. The limitation here is that the results of Chapter 4 are for the linear regression model, a setting that is much easier to deal with than deep neural networks. There exists (optimization) literature on the complexity of stochastic gradient descent and zero-order methods that expands results for those methods to more general strongly convex-optimization problems, [115, 133]. This suggests the possibility for further research extending the results in Chapter 4 to general convex problems. The key challenge is to deal simultaneously with the randomness from the data and the additional randomness introduced by forward gradient descent. As training deep neural networks is a non-convex optimization problem it remains unclear if it is feasible to extend the analysis to the deep neural networks considered in Chapters 2 and 3.

5.2 Model assumptions

In this thesis various assumptions on the target function are imposed. In Chapter 2 it is assumed that the conditional class probabilities are β-Hölder smooth. In Chapter 3 it is assumed that the densities have a compositional structure, where each function in the composition is in some Hölder-smoothness class. The main motivation behind the choice for these smoothness assumptions is that this makes comparison with existing risk bounds in the literature possible, as convergence of the risk under these assumptions has been widely studied. The compositional structure in Chapter 3, as well as the possible inclusion of a compositional structure as discussed in Chapter 2, are motivated by existing results for regression [62, 72, 11, 127, 75]. In these works, it is shown that deep neural networks can circumvent the curse of dimensionality under compositional structure assumptions. This provides a possible explanation for the observed good performance on high-dimensional input problems of deep neural networks in practice.

For image classification there exists a related assumption, the hierarchical max-pooling model considered in [74, 76]. This compositional model is tailored to the image classification task in combination with convolutional neural networks. The principal idea behind this model is that the question: "contains the image a prespecified object?", can be answered by estimating the probability that this is true for subparts of the image and then taking the maximum of the probabilities over the subparts.

A different kind of model assumption is based on the observation that in many practical datasets the data seem to lie around a low dimensional manifold. In [103] it is shown for the regression problem that if the data are scattered around a lower dimensional manifold, then deep neural networks can exploit this to obtain convergence rates that depend on the intrinsic manifold dimension instead of the full dimension of the input space. For this result it is also assumed that the regression function is Hölder-smooth. This paper includes a numerical estimation of the intrinsic dimension of the MNIST and CIFAR-10 benchmark datasets, showing that these datasets indeed have an intrinsic dimension that is much smaller than their full dimension. An assumption that is closer related to the composition structure assumption in this thesis is the assumption of local low dimensionality studied in [73]. The idea of the local low dimensionality assumption is that the function locally only depends on very few of its components. Under this assumption it is shown that in the regression problem the bounds depend on the local dimensionality instead of the full input dimension. These works [103, 73] suggest that it should be possible to combine the idea of the data lying around a lower dimensional manifold with the results for the classification and density estimation models studied in this thesis. How to combine the composition and manifold assumptions in a manner that is realistic for practical datasets and the exact effects of such a combination on the risk bounds is an avenue for further research.