Mijnheer de Rector Magnificus, zeer gewaardeerde toehoorders, dear colleagues and guests,

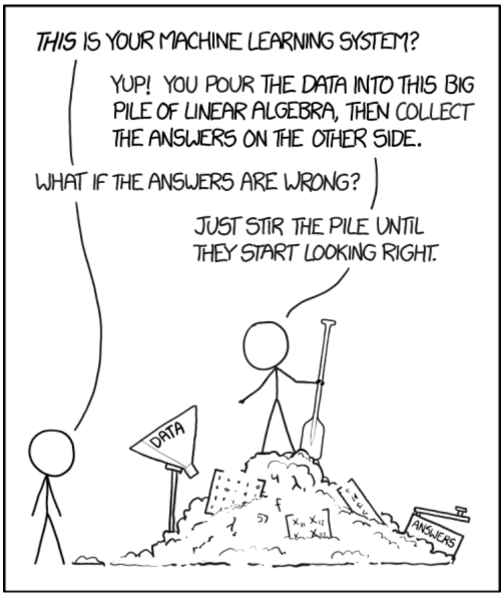
machine learning and in particular the search for a theory of machine learning are currently considered as one of the biggest trends in science. In my talk, I will survey the state of the art and describe the challenges that we face in the current development of a mathematical foundation of machine learning. But before I come to that, I want to go back to the origins of mathematics to highlight why we need mathematics and theoretical foundations.

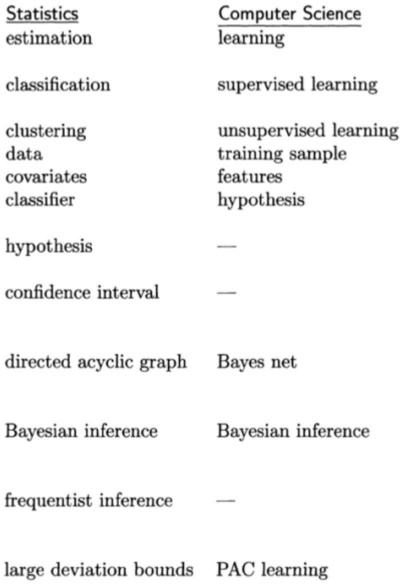
One of the beginnings of mathematics is the Babylonian period (around 1700 BC). Driven by real-world problems, a lot of mathematical formulae were found during this time. However, there was no concept of a mathematical proof yet on how to relate and derive results from each other. Essentially, Babylonian mathematics is a pile of disconnected formulae. Some results are only approximately correct, for instance, the number pi is set to 3.

In the old testament there is this story about the tower of Babel. The story is that people wanted to build a tower to the sky. During the construction work, they started to talk in different languages leading to an enormous confusion and because of this they had to give up and the tower was never completed. Maybe this story can also be viewed as a metaphor for the less rigorous scientific style during the Babylonian period. If science only creates a pile of disconnected results, without a common structure such as a proper scientific reasoning, then different terminologies, different scientific languages will emerge and errors are introduced that are hard to detect and to get rid of, leading to considerable confusion and potentially causing the collapse of a field.

The proper way to structure mathematics -as it is done until today- was subsequently developed during the ancient Greek period. One of the important inventions is that mathematical objects get rigorous definitions. Instead of disconnected results, mathematical statements are connected by logical arguments, more advanced results being derived from simpler statements. We can think of the structure of modern mathematics as a reversed tower where we start with the top representing the simplest possible statements that cannot be reduced any further; axioms, in the mathematical language. From the axioms, we then derive subsequently more and more complex statements until we arrive at the mathematics that can be used to solve complex real-world problems.

With my research interests, I sit somewhere in the middle of this tower, well separated from the applications. One might still call it the ivory tower of science, where more abstract mathematical theory is developed. Although the ivory tower of science has a negative connotation it fulfils an important role namely to provide a high level understanding that can be subsequently used to design methods to tackle real world problems.


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS



Statistics	Computer Science
estimation	learning
classification	supervised learning
clustering	unsupervised learning
data	training sample
covariates	features
classifier	hypothesis
hypothesis	—
confidence interval	—
directed acyclic graph	Bayes net
Bayesian inference	Bayesian inference
frequentist inference	—
large deviation bounds	PAC learning

xkcd.com

[2], page xi

Figure 1: A cartoon on machine learning and a small dictionary from a standard statistics textbook translating between computer science and statistics terminology.

1.Background

Let me now discuss the current state of machine learning. Nowadays most scientific disciplines are quantitative and draw conclusions from data. To deal with the enormous amount of available data and all the new data structures, various new data analysis tools have been proposed. Machine learning denotes a number of methods that have been derived to analyze large dataset with a complicated latent structure. These methods are designed based on intuition, imitating processes in nature and by empirical search over many possible configurations. Most of the newly developed data analysis tools are essentially just combinations of complicated formulae or algorithms. This is also what is highlighted in the cartoon above. We often do not understand how these methods work and why they work. Most experts would probably agree that at this moment machine learning rather resembles the Babylonian mathematics.

And similar as in the metaphor with the tower of Babel, that I was alluding to earlier, we speak different languages. Different scientific communities have developed their own languages, their own scientific terminologies. Although we all work on the same problems we have trouble understanding researchers from other scientific communities. As an example, the small dictionary in Figure 1, taken from a standard statistics textbook, shows that even the most elementary scientific terms are denoted differently by different communities or the same word (for instance hypothesis) can have different meanings. Having different languages obviously causes considerable confusion and contributes to the fact that machine learning -which is mostly formulated in computer science terminology - is an extremely chaotic field. Another factor that makes machine learning very hard to oversee is that it is so rapidly growing, with tens of thousands of articles being uploaded on the internet every year. Most of the claims in these preprints are never checked by a proper scientific peer review process. The major hope is that a mathematical theory would unify different approaches reducing methods to key concepts and presenting the whole field in a coherent way. Moreover, it should lead to a more high-level understanding on how and why these methods work and provide a common ground on which we can compare different procedures theoretically.

2. Prediction

Machine learning is an umbrella term for a collection of methods that have been developed by computer scientists and that all have one goal: to make predictions. Prediction essentially means that we have to make a guess. To come up with such a guess is exactly what machine learning can do. These methods are often extremely complex with many parameters and hyperparameters and perform surprisingly well if enough data are available. Machine learning is now implemented for a variety of complex tasks that all can be cast into a prediction problem. To mention a few examples, this includes autonomous systems such as self-driving cars, speech recognition, machine translation, game playing and many others. Machine learning has therefore many relevant applications and if we think of mathematics as a reversed tower, it can be located at the upper end of the mathematics tower. But - as I mentioned before - it lacks the mathematical foundation. In my research, I want to fill the gap in the theory that is closest to my previous expertise. My approach is to study machine learning from a more abstract point of view by interpreting machine learning methods as statistical methods. By doing so, we can build on many powerful results from mathematical statistics for a theoretical analysis.

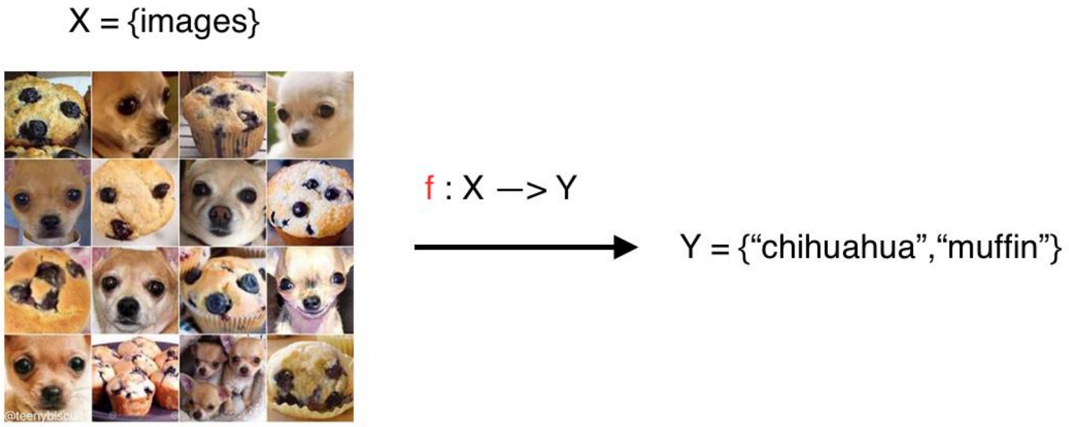


Figure 2: Illustration of the mathematical formulation underlying object detection on images. The aim of a machine learning procedure is to recover the unknown function f. Images taken from @teenybiscuit

What is now the mathematical problem that I then want to solve? To explain the idea, let us consider object detection of images. We want to train a machine to see. To see as the human eye can see things. This means if we give an image to a machine, it should be able to recognize the object on this image. We can simplify that if we give a machine images of two objects only, say muffins and chihuahuas, which look surprisingly similar. For humans this is very simple - if we look at the images in Figure 2, we immediately know which images display chihuahuas and which images display muffins. But for a computer to see what is on an image is an extremely hard problem and only because of machine learning there has been a lot of progress recently. In mathematical terms, we can state the problem as follows. On a computer a colored pixel is just a number and therefore an image is just a collection of numbers. We can now suppose that there exists a function - that is unknown to us - such that whenever we stick all the pixel values in it, it will return the true label- here muffin or chihuahua (see illustration in Figure 2).

If we would know the function, we could write a computer program and running this program the computer could perfectly distinguish these two objects. Any method that teaches the computer to recognize objects on images needs to reconstruct the function f from data. To illustrate the problem, suppose that all these pictures would consist of one pixel only. The unknown function f would map then one pixel value to the label. Suppose that small pixel values are chihuahuas and large pixel values are muffins (see Figure 3 for an illustration).

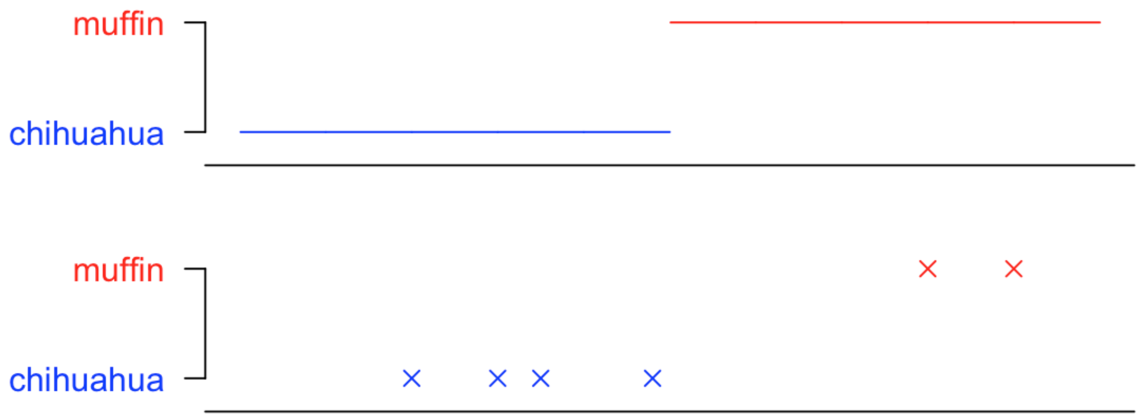


Figure 3: An example of a unknown function f that maps pixel values into labels (top). Correctly labeled training data correspond to points on this function. This means that in practice we know the function f only at some points.

The thing is that we do not know this function and all what we have are a few examples. This means, in practice we know the function values at some points. The problem is now that if someone gives us a new image for which the label is unknown. Then, we have to come up with a guess what the label is. This means we need to make a prediction of the function value at that point.

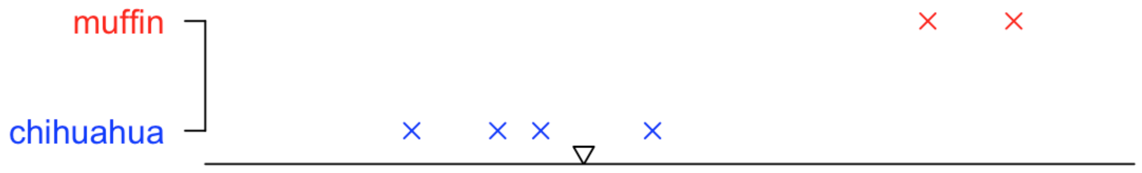


Figure 4: The prediction problem is to predict the label corresponding to a newly observed pixel value (black triangle).

Suppose we get to see the new pixel value in Figure 4. As there are examples in our dataset with slightly lower and a slightly larger pixel value which all resulted in a chihuahua, it is clear that the best guess would be to say that the corresponding label is also a chihuahua.

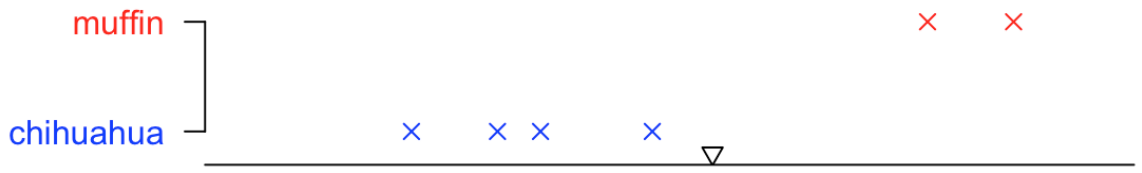


Figure 5: A pixel value that is hard to classify.

But if we observe a new pixel value in the setup of Figure 5, the situation is less clear. Because there is a chihuahua image with a very similar pixel value we would say this is a chihuahua. But here we make a mistake because if we compare it to the true function f displayed in Figure 3, it displays in fact a muffin. From this example, one can see that it is particularly hard to come up with a good guess in regions where the function changes. It is also clear that we have to make mistakes because there is just not enough information in the available data. If we have more data though, we have more information about the unknown function and will make therefore less errors. For instance, for the same new pixel value as in Figure 5, we would have it correctly classified if we would have had much more examples of correctly classified images (Figure 6).

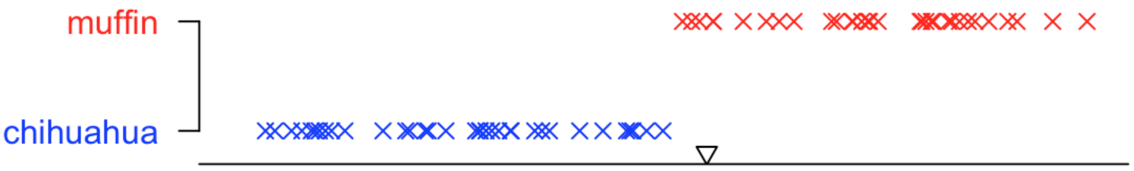


Figure 6: Same as Figure 5 but with more test images. Based on this data, any ad hoc prediction rule would predict that the label of the new pixel value (black triangle) would be muffin. This shows that the availability of more labeled images should lead to less mistakes of the method.

The mathematical problem that we ask is the following. Given that we have n examples (labeled images), what is the expected error that a given machine learning method will make. What we are mostly interested in the dependence of the error on the number of data points. We have just seen that more data means more information and should result in fewer errors. What I work on is to find the speed at which this error decreases as the sample size becomes large. If we can establish such a result, we have some theoretical guarantees for the performance of the considered method.

One should also say that of course nobody is interested in distinguishing muffins from chihuahuas. As I mentioned before there are many highly relevant applications and for the image detection problem, there are numerous medical applications such as for instance automatic detection of cancer cells. Also for the self-driving car, object recognition is a crucial element as the car has to know what the objects are that surrounds it. As with any tool, one can of course also abuse it and there are also applications that are ethically problematic. Obviously, tracking and automatic surveillance of citizens is of course highly problematic.

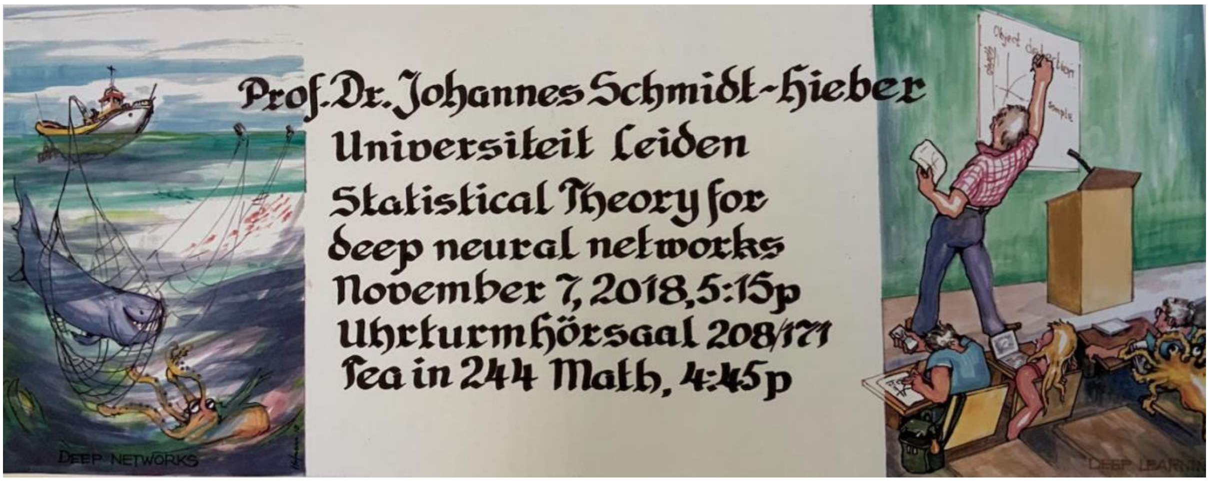


Figure 7: Poster for a talk on deep networks at the University of Darmstadt. On the left side, there is an artist impression of a deep network. Deep networks have, however, nothing to do with fishing. Instead they are inspired by the human brain.

3. Deep Networks

As mentioned before machine learning is a collection of many methods. The currently most important ones relate to so called deep networks. Deep networks are inspired by imitating the brain. Human brains are extremely good in distinguishing objects on images. It is natural that by building a method that has an underlying structure inspired by the brain and that learns to adapt all the free parameters in a similar way as a child learns that this machine will then also perform as good as the human brain.

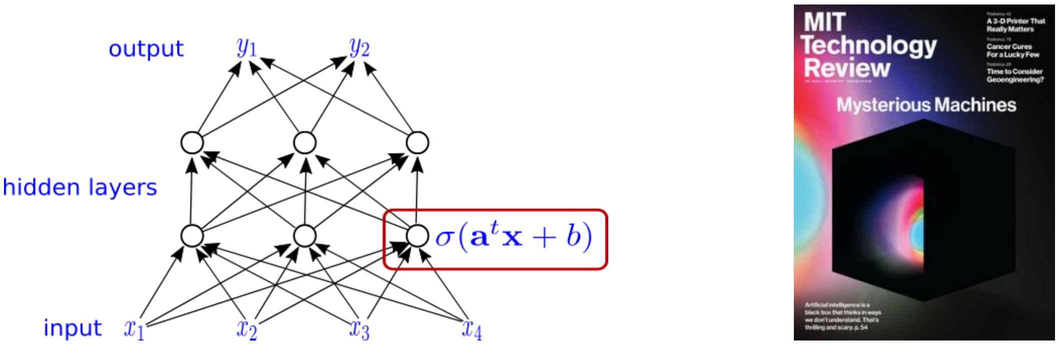


Figure 8: Graph representation of a neural network and interpretation as a black box method.

Deep networks are mathematical functions that can be represented by a graph, see Figure 8. In this graph representation we see the similarity with the brain – we see neurons and connections between neurons. In the deep network graph representation, the nodes are arranged in layers where every of these nodes stands for a simple mathematical operation. A deep network just means that a neural network with many of these layers are used. A deeper network can compute more complex functions. On the contrary, the width of the network describes how many of the computational units are in one of the layers and it roughly means how much the network can memorize. As in the human brain there are many free parameters that are learned once some examples are provided. This is also called deep learning.

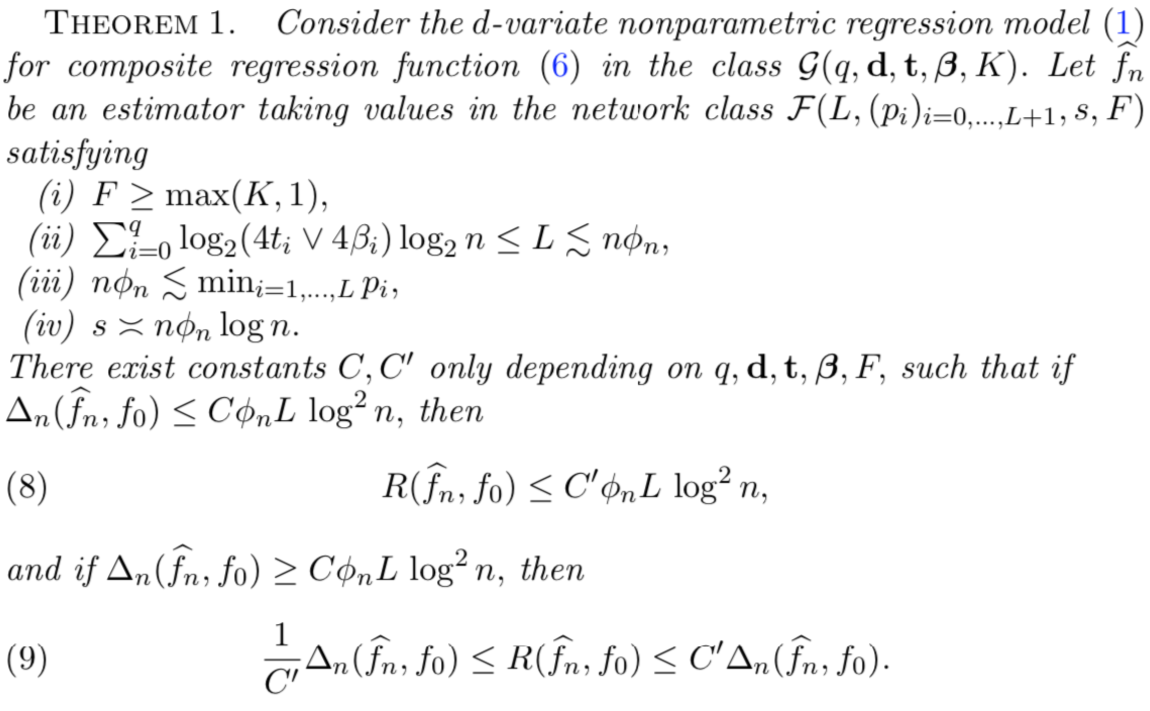
The network drawn in Figure 8 only has a few neurons, but modern network architectures can have millions of neurons and free parameters. After the network is adjusted to the test images for which we know the labels, all these millions of parameters are assigned to some values. In contrast to classical statistical methods where the parameters have an interpretations as mean, variance or regression coefficient, these millions of numbers have no meaning to us, we do not know why they are assigned to these values. It just happens. Because we understand so little about the role of the parameters deep learning is commonly referred to as a black box method.

The black box is often highlighted in the extensive public media news coverage about machine learning, see Figure 8. Then it is said that deep networks are scary because we build black box machines that we do not understand anymore ourselves. Although I agree in principle, the word black box is slightly exaggerating because we can look into the system, we see all these numbers. The problem is that they do not have any meaning to us. It is maybe better to compare them to a Dutch croquet. As deep networks, Dutch croquet are great and nobody seems to know what is inside. If you want to know what a croquet consist of, you can open them and look inside. It is therefore not a black box. But opening doesn't help because if one looks inside, the filling is something that is very hard to tell what it is and also real Dutch people do not really know.

When I started to think about a theoretical foundation some years ago, there was a general belief that modern machine learning is too complex that one can still prove theorems. One argument is that in order to understand the outcome of such a Dutch croquette method, you first need to understand what all these millions of individual parameters mean and which values they are assigned to for a given dataset. If you do not understand the complex interior behavior, nothing can be said about the theoretical properties of the outcome.

This seems to be a very logical argumentation, but it is not true. Here the translation into a statistics problem becomes crucial since there is some machinery in theoretical statistics that one can use and that can lead to theoretical bounds on the error rates of what we have identified as a croquette method. To apply this machinery, instead of a full understanding of what happens in the interior, one only needs to know two things. The first one is that there should be one parameter assignment that gives a good outcome. This part of the proof is completely independent of the data. Secondly, one needs to know something about the space of functions that can be generated by deep network functions. Both of these properties are not easy to show but they still can be derived for deep networks.

What can we now prove about deep networks? Actually we can give a rigorous prove that deep neural networks can have the smallest possible error rate that is in an information theoretic sense achievable by any method. To show you how such a mathematical theorem looks like, the full statement is reproduced below. For more information we refer to the article [1].



THEOREM 1. *Consider the d-variate nonparametric regression model* (1) *for composite regression function* (6) *in the class* $\mathcal{G}(q, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, K)$. *Let* $\widehat{f}_n$ *be an estimator taking values in the network class* $\mathcal{F}(L, (p_i)_{i=0,\ldots,L+1}, s, F)$ *satisfying*

(i) $F \geq \max(K, 1)$,

(ii) $\sum_{i=0}^{q} \log_2(4t_i \vee 4\beta_i) \log_2 n \leq L \lesssim n\phi_n$,

(iii) $n\phi_n \lesssim \min_{i=1,\ldots,L} p_i$,

(iv) $s \asymp n\phi_n \log n$.

There exist constants C, C' *only depending on* $q, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, F$, *such that if* $\Delta_n(\widehat{f}_n, f_0) \leq C\phi_n L \log^2 n$, *then*

$$R(\widehat{f}_n, f_0) \leq C'\phi_n L \log^2 n, \tag{8}$$

and if $\Delta_n(\widehat{f}_n, f_0) \geq C\phi_n L \log^2 n$, *then*

$$\frac{1}{C'}\Delta_n(\widehat{f}_n, f_0) \leq R(\widehat{f}_n, f_0) \leq C'\Delta_n(\widehat{f}_n, f_0). \tag{9}$$

Something that is also hidden in this statement is that deep networks perform particularly well if the underlying relationship has some hierarchical structure. What do we mean with that? This is best explained with a heuristic. Text has for instance a hierarchical structure. To generate text, we can first generate letters, then we combine letters into words, then words into sentences and so on. With this example one can well see how an object decomposes into several abstraction layers. The visual cortex in the human brain works in a very similar way by extracting more and more abstract information before we finally know what we see. Such composition structures occur everywhere and seem also to be present in the cases where deep networks outperform other methods.

What is maybe even more interesting is that we can give a mathematical proof which shows that so called wavelet methods – which are typically performing very well – will have a much larger error than deep networks under such a hierarchical composition structure.

These results are not the end of the story. As some restrictive assumptions are needed, it is rather a starting point and I hope to extend the theory with my group in Twente over the coming years.

4. Challenges for a theory of machine learning

At the end of my talk, I want to outline some challenges that have to be addressed by any serious attempt to build an exhaustive theory of machine learning.

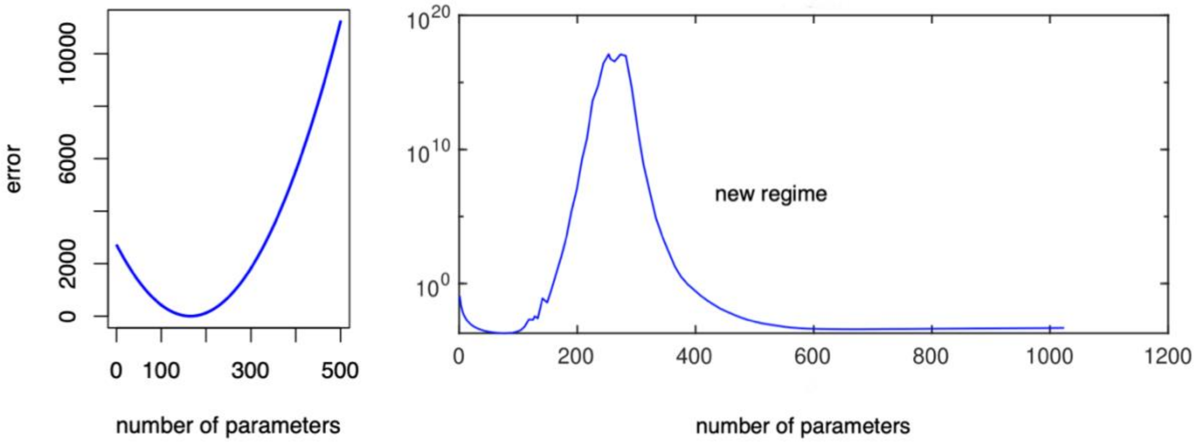


Figure 9: Classical bias/variance trade-off (left) and double descent phenomenon (right).

One thing that fascinates me about this field is that completely new phenomena are observed empirically that contradict our classical understanding. For instance, something that we cover in all statistics courses is the so called bias-variance trade-off. This says that if one plots the performance of a method in dependence on the number of parameters, we obtain a curve that looks a bit like a "U" according to the classical theory. Methods with too few parameters are not flexible enough and lead to large error rates of the methods. Increasing the number of parameters improves the method first. But at some point the method becomes worse again and as we add additional parameters the error explodes. This yields then this U curve and one conclusion from that is that too many parameters lead to poor performance. This is how we teach the bias-variance trade-off and how it is explained in the textbooks.

For machine learning method, a new phenomenon occurs. For extremely large number of parameters, the error decreases again contradicting our intuition. Figure 9 shows an example of this behavior. On the left side of the plot, we recover this U-shaped curve, but then instead of going up, as predicted by the classical theory, the performance improves again. When the first articles on the double descent phenomenon appeared a year ago, the setups were quite artificial. During this summer, the community met during a two month meeting in Berkeley and we had a lot of intense discussion. My point was that this phenomenon is very specific and the authors argued that it occurs in much more generality. At some point we agreed on a setting where they believed double descent occurs and I was convinced the phenomenon will not be present. We then run a simulation study and the plot on the right of Figure 9 is what we got. It clearly shows the double descent and thus, I was completely wrong. Although I still do not have a good intuition yet I am now convinced that the double-descent is indeed a much more general phenomenon.

This new regime is an instance for a phenomenon that challenges the classical statistical theory. Because of such new phenomena, one might be tempted to compare the current situation in machine learning with physics at the beginning of the 20th century. This was a period where several new phenomena were observed in experiments that could not be explained anymore by classical mechanics and finally led to a much more profound understanding of the physical world.

As explained before, for the results so far, we can make a mathematical statement about the outcome without understanding of what happens in the deep network. To give an explanation of the output of a machine learning method, we need, however, to be able to say more about what happens in the interior. The European data protection guideline ensures that everyone has the "right to explanation". This means that if a machine learning method is used to determine whether someone goes to jail or not or to determine the credit score of a customer, this customer has the right to know how the machine came to the conclusion. This makes many powerful machine learning methods such as deep networks useless because we do not know how they come to their conclusion. There is a conjecture that there could be a trade-off between interpretability and performance of a method. This conjecture states that the output of very accurate methods are hard to interpret and methods that allow for better interpretation necessarily have to give up a bit on performance. To prove or disprove such a conjecture requires of course that we can formalize what interpretability means in mathematical terms. This seems to be already a hard problem.

Another challenge is to develop a mathematical theory of human learning. I have mentioned before that deep networks are inspired by the brain and deep learning tries to mimic the way a brain learns. Although deep networks and brains differ in important aspects there is still the hope that a theoretical foundation of deep learning translates into a mathematical theory of the human brain with the potential to lead to a more profound understanding of psychological phenomena.

One of the biggest challenges is to predict the development of machine learning over the coming years. In 2019, we live in a world were machine learning is believed to have unlimited potential, were countries and companies invest billions into its developments and were everyone is driven by the fear of missing out what is believed to be a technological revolution. At the moment everyone seems to be euphoric. I think that there are many warning signs that are overlooked and that lead to the fact that I am more pessimistic.

Historically, there have been already two artificial intelligence hypes in the past, in the 50ies/60ies and in the 80ies. Both of them started with some breakthroughs and ended because of a large gap between the outrageous expectations concerning the development of the field and the much more modest real progress.

Even the highly tuned, state of the art machine learning procedures have many drawbacks, which limit their applicability. Known issues are that these methods are very instable and easy to fool. For instance they perform typically poorly under a slight change of the underlying scenario, a problem that could not be resolved yet, despite tremendous efforts. If a method cannot cope with changes in the input, this can have severe effects. The financial crisis ten years ago brought the world economy close to a collapse because major banks applied mathematical formulae for asset pricing which did not work anymore under the scenario of falling house prices. Since machine learning performs poorly if a new situation arises, there is the real danger that these methods would also fail spectacularly under a similar change of the underlying scenario.

After the financial crisis, mathematics was criticized for inventing all these new financial products that were so complicated that no one would understand them anymore, [3]. As a consequence, financial mathematics experienced a sharp decline. Machine learning is even more extreme as we understand less. As researchers working in this field, we have the duty to mention the risks. This year, I was invited to participate in the advisory board for the Dutch AI agenda, where a lot of expertise was gathered. During the discussions, I pushed for more emphasis on the risks and warning signs. The Dutch AI agenda will be released soon, but the latest version that I have seen is, does unfortunately not contain anything on this and sketches a future that I believe is far too optimistic.

So far I have been working on results that show that in certain scenarios deep networks are in a sense optimal. One of the main challenges for my future research is to find a more complete theoretical description of the strength and weaknesses of these methods. For that one needs of course to prove theorems in cases where deep networks work but also for scenarios where deep networks fail. To show when a statistical method does not work is mathematically even more demanding. It is worth mentioning that the first neural network hype in the 50ies/60ies ended because of a negative mathematical result, which showed that an important function was not representable by the back then considered network structures. This shows again the importance which mathematics plays in this field.

5. Conclusions

For the next years, I plan to continue to work along these challenges. The math department at the University of Twente is a perfect environment to conduct this research and it is a privilege to work here. Among others, there are already strong groups in optimization and computer science on the campus providing a perfect academic embedding of my research. With my research area, I hope to bring in something new and to contribute to the flourishing of the university. In the first months after my appointment, we have already started to revive the statistics group and to contribute to the statistics education on the campus.

As it is common for such an inaugural speech, I would like to thank the people who have supported me throughout my professional life. First and most importantly, there is my wife Maria and our families – some of them living not far from here. I also want to thank the math department and the faculty in particular the head of our department Stephan van Gils and our dean Joost Kok for investing a lot of their time and energy in the new statistics group. I owe a lot of thanks to my many collaborators as well as all my current and previous colleagues. In particular, I want to thank my PhD advisor and mentor Axel Munk who introduced me to mathematical statistics during my undergraduate studies. Finally, I have to thank the statistics community in the Netherlands. Instead of competing, we closely collaborate via many joint initiatives. This is what makes us strong and also attracts many foreigners - such as myself – to work in the Netherlands and to join the community.

One of the main points in my talk was to compare machine learning methods to Dutch croquettes. But then, I talked only about demystifying machine learning methods and never mentioned what we know about the content of a croquette. I have spent hours trying to unravel this mystery. There is a book on croquettes but it is not very helpful. In fact, it is incomplete and got published only after the author -who tried to finish it for many years – passed away. Even Wikipedia only provides a very vague and lengthy description. We know very little about the functioning of deep networks and machine learning. But it seems that we even know less about croquettes.

Thank you very much for your attention. Ik heb gezegd.

References:

[1] J. Schmidt-Hieber (2019). *Nonparametric regression using deep neural networks with ReLU activation function.* To appear as a discussion article in the Annals of Statistics.

[2] L. Wasserman (2004). *All of Statistics.* Springer Texts in Statistics.

[3] D. Bikramjit, P. Embrechts, V. Fasen (2013). *Four theorems and a financial crisis*. Internat. J. Approx. Reason. 54, 701-716.