Chapter 1

Introduction

In the last decade, machine learning and artificial intelligence (AI) have experienced a surge in research activity and investment, largely due to the subfield of deep learning. Deep learning and AI have a long and interdisciplinary history, interacting with fields from neuroscience to physics. In fact, it is argued that the current success in deep learning is mostly due to a recent increase in the amount of data and compute capabilities, while many of the ideas and questions are much older. In 1995, the influential statistician Leo Brieman wrote an article titled "Reflections After Refereeing Papers for NIPS" [Breiman, 1995] in which he poses some of the main open questions of the day in machine learning theory:

- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?
- When should one stop the backpropagation and use the current parameters?

Even though these questions were written 25 years ago, they are still open and are some of the main topics studied by recent papers in the theory of deep neural networks. In that article, Brieman also expresses his opinions on the value of theory, and how achieving understanding of new phenomena, like machine learning systems, relies on much more than mathematically rigorous theorems. He proposes inquiry techniques such as: "mathematical heuristics, simplified analogies (like the Ising Model), simulations, comparisons of methodologies, devising new tools, theorems where useful (rare!), and shunning panaceas" [Breiman, 1995].

Recently, in a commentary in Nature Physics, Lenka Zdeborová [Zdeborová, 2020] builds on this idea to propose that physicists should play a key role in the quest to

understand deep learning. Physics typically employs a diversity of inquery techniques as Brieman suggests, while maintaining a high standard of scientific rigour. The study of machine learning systems is particularly connected to the study of complex systems and emergence in physics. In both of these areas, it is often the case that the microscopic rules of a system are known precisely, but the macroscopic behaviour is still surprising, and one would like to find a simple explanation of it. A mathematical theorem is only one approach at tackling this problem. In physics, a more common approach relies on the scientific method: performing carefully designed experiments to rule out possible hypotheses. This approach may not give quite the same level of confidence as a mathematical theorem, but it can still provide insight into observed phenomena, produce non-trivial and correct predictions, and may also inspire new theorems.

1.0.1 Early history of artificial neural networks

The history of artificial neural networks (ANNs) can be traced to the classic paper McCulloch and Pitts [1943] in which they introduced a model of networks of neurons based on simplifying assumptions of real physiological neurons. These neurons had binary 0 or 1 activations, and discrete "weights" (corresponding to number of "synapses"). McCulloch and Pitts didn't propose an explicit training mechanism for these networks. The first learning algorithm for neural networks was proposed by Hebb [1949], based on the idea that "neurons that fire together wire together". Interestingly a year earlier, Alan Turing, considered one of the founding fathers of AI, had published a report [Turing, 1948] in which he introduced a learning machine similar to ANNs, and speculated about how these could learn. Frank Rosenblatt proposed in 1959 a model which he called the perceptron [Rosenblatt, 1958], consisting of a network of neurons like that of McCulloch and Pitts, and a new training algorithm, inspired by the work of Hebb, and other literature on "brain models" [Ashby, 1952, Hayek, 1952, Uttley, 1956]. He considered neural networks with hidden layers, but only the last layer was trained, and the parameters corresponding to biases in the neurons only, rather than weights in the synapses. His model was simple enough that it allowed theoretical analysis of its training and generalization learning curves [Rosenblatt, 1961, Joseph, 1960, Lumb, 1959]. Rosenblatt's paper was very influential and inspired several works on extensions to the perceptron.

It was quickly recognized that multiple layers allowed for more expressivity and better generalization [Rosenblatt, 1961, Bryan, 1963]. However, training parameters across layers remained a challenge [Rosenblatt, 1961]. Some early attempts at training

several layers include the Gamba networks [Gamba, 1961, Palmieri and Sanna, 1960], and networks of Adaline neurons [Widrow and Hoff, 1960, Widrow, 1962]. The Adaline neurons [Widrow and Hoff, 1960] are interesting in that they were inspired by earlier work on switching networks by Shannon [1938] and on adaptive neuron models by Von Neumann [1956] and Mattson [1959], Mattson, they are described in a way which is closer to the modern interpretation of the "perceptron", and they were trained with a version of gradient descent (in the case of no hidden layers). In 1965, Ivakhenko et al. published a training algorithm for adjusting all the weights of multi-layer neural networks, which they later used to successfully train deep neural networks (DNNs) of up to 8 layers [Ivakhnenko and Lapa, 1965]. This work appears to have been unknown to Minsky and Papert when they published their influential book on perceptrons 4 years later [Marvin and Seymour, 1969]. In this book they analyzed the theoretical limitations of a class of perceptrons in which the features feeding to the trained linear classifier were assumed to be local. They commented on other classes of perceptrons including the Gamba networks, and more general multilayer perceptrons, speculating that these extensions are probably not promising, although they could not yet offer proof of this due to the difficulty in analysing these systems. This book is claimed to have caused a huge drop in funding for work on neural networks, and shifted most focus on AI research to systems based on logical rules, deduction, and explicit heuristics [Olazaran, 1996, Cohen]. The high expectation placed on this approach, combined with the comparatively poor results, led to the first "AI winter" during the decade of the 70s [Howe, 2007]. In the 80s, there was renewed interest in the rule-based approach to AI with the development of expert systems, but these systems also proved too brittle, resulting in a second AI winter at the end of the 80s [Stuart et al., 2003].

Despite the sharp decrease in funding for research into neural networks and related "connectionist" approaches during these two decades, some important advances were made that laid the foundations for later theoretical and applied developments. On the theoretical side, William Little proposed in 1974 an analogy between the Ising model of spin glasses and neural networks [Little, 1974]. We also saw the development of neural networks with emergent collective behaviour that allowed to recover a previously seen pattern, from a subpart [Amari, 1977, Kohonen, 1972]. This type of "associative" or "content-addressable" memory inspired the work of Hopfield [1982], which popularized the Ising-like model of Little, and in turn inspired an increasing number of physicists to work on the theory of neural networks [Sompolinsky, 1988]. In 1989, expressivity theorems were developed [Hornik et al., 1989], based on earlier

work by Kolmogorov [1957], which generalized the older results on approximating Boolean functions to real-valued functions. On the more applied side, the 70s and 80s saw the development of backpropagation [Linnainmaa, 1970, Werbos, 1974, 1982, Parker, 1985, Rumelhart et al., 1985], and convolutions [Fukushima, 1979, LeCun et al., 1989]. These developments ushered in a big resurgence of work on neural networks during the 90s, including recurrent neural networks (RNNs) and increasingly performant architectures and algorithms, that ultimately led to the current 'AI spring' based on deep learning, starting in the early 2010s. We refer the reader to the articles Schmidhuber [2015], LeCun et al. [2015] for a good review of these developments, focused on practical developments.

1.0.2 Learning theory and ANNs

The theory of neural networks is tightly linked to the theory of machine learning. The later traditionally encompasses three main subareas: supervised learning, unsupervised learning, and reinforcement learning. We will focus on supervised learning because it is arguably the most studied and applied of the three areas, and holds many of the foundational questions which are relevant throughout all of machine learning. As we mentioned earlier, work on predicting the learning properties of neural networks goes back at least to the original paper on perceptrons [Rosenblatt, 1958]. Rosenblatt formalized "learning" as equivalent to "generalization", which he defines as being able to predict correct responses to inputs not seen during training. He distinguished that from the ability to correctly predict responses to inputs in the training set (for which it was given the correct answer, or at least a reinforcing signal to learn the correct answer). This separation of the learning problem into fitting the training data (an optimization problem), and generalizing to new data (an statistical problem) is still used as the basis of most analyses of supervised learning – see the modern formalization we present in **??**.

Since the 80s, theory of supervised learning neural networks began developing in two main branches. On the one hand, there was the work following the approach from statistical physics [Hopfield, 1982, Sompolinsky, 1988, Tishby, 1995, Smieja, 1989, Wallace, 1987], which was often less rigorous, and based on extensive experiments, as well as focusing on "average-case" analysis. On the other hand, the development of learning theory as a branch of statistics by Vapnik and Chervonenkis [Vapnik, 1968, Vapnik and Chervonenkis, 1974, Vapnik, 1995], and of theoretical computer science by Leslie Valiant [Valiant, 1984], took a very rigorous approach, mostly based on theorems, and "worst-case" analysis with minimal assumptions. Vapnik and Chervonenkis' theory

gave rise to the important concept of VC dimension, as measure of the 'richness' of a family of classifiers, while Valiant introduced probably approximately correct (PAC) learning, which formalized the concept of 'generalization', using frequentist statistics ideas to minimize the number of assumptions. Already in a workshop on the theory of generalization in NIPS 1992, it was recognized that these two communities, as well as others, should find more effective ways to communicate and exchange ideas [Wolpert, 1995]. Although the situation has improved today, we see from the article by Zdeborova with which we opened the thesis [Zdeborová, 2020], that better collaboration between the different disciplines is still a live issue today.

In 1989, Levin, Tishby and Solla introduced a probabilistic view of learning based on a formal analogy between Bayesian inference and statistical physics Tishby et al. [1989], Levin et al. [1990]. This framework led David MacKay in 1992 to propose a way to do model selection on neural networks by selecting the solutions where the error surface was flatter, which he associated with larger Bayesian evidence MacKay [1992]. MacKay found that "empirically, the correlation between the evidence and generalisation is often good. But a theoretical connection between the two is not yet established". Two years later, Hochretier and Schimdhuber developed algorithms to find flat minima of ANNs, and justified their performance with the minimum description length (MDL) principle, and the observation that real-world problems tended to be "simple" and are far from uniformly distributed [Hochreiter and Schmidhuber, 1994, 1995, 1997]. However, a mathematically rigorous result that could explain this correlation in some cases was first proposed by John Shawe-Taylor in his PAC analysis of a Bayesian estimator Shawe-Taylor and Williamson [1997] (based on an extension to the structural risk minimization framework [Shawe-Taylor et al., 1998] of Vapnik and Chervonenkis). A similar but more general result was introduced in 1998 by McAllister with his PAC-Bayes theorems [McAllester, 1998]. As we will see in **??**, both flatness of minima, and PAC-Bayes play a big role in recent literature on deep learning theory Keskar et al. [2016], Neyshabur et al. [2017], Bartlett et al. [2017], Dziugaite and Roy [2017], Zhou et al. [2018].

The work of MacKay on Bayesian priors [MacKay, 1992] also inspired another line of work starting with Neal [1994]. In this paper, he switched focus from the prior over parameters to the prior *over functions*, and he showed that for of a one-hidden-layer ANN, the later approached a Gaussian process when the number of hidden neurons grew to infinity. The result of Neal was recently extended to DNNs of many layers [Lee et al., 2017], and forms the foundation of the mean field theory (MFT) of DNNs, which gives analytical results on the infinite-width limit of the prior over functions for

DNNs. The MFT of DNNs is reviewed in **??**. This theory has inspired a lot of recent work on deep learning theory **?????????????**.

Neal and MacKay's Bayesian approach was a more principled way to avoid 'overfiting' and helped popularize the use of Gaussian processes (GPs) in the late 90s and early 2000s [Rasmussen, 2004]. During this same time, support vector machines (SVMs) were developed by Vapnik and others [Boser et al., 1992, Drucker et al., 1997] as a way to perform classification and regression, which came with rigorous generalization guarantees based on VC dimension theory Vapnik [1995]. In part because of their theoretical foundations, GPs and SVMs became increasingly popular in the 2000s, under the common name of "kernel machines" Rasmussen [2004].

This changed in the early 2010s, when a series of ANNs with many layers started winning several international machine learning competitions, and producing breakthroughs in classic AI tasks like computer vision [Cireşan et al., 2011, Krizhevsky et al., 2012], playing games [Mnih et al., 2013], and speech recognition [Hannun et al., 2014]. Because these ANNs have many layers, they are called 'deep', and their use is called 'deep learning'. The success of deep learning has kept growing and now finds applications in a great many academic and applied disciplines [Pierson and Gashler, 2017, Miotto et al., 2018, Yannakakis and Togelius, 2018, Guest et al., 2018, Foster, 2019, Robila and Robila, 2019].

This surge in interest and research revived several theoretical questions including those posed by Brieman in 1995. One of the most influential articles in this regard was Zhang et al. [2017] which showed that modern neural networks with millions of parameters have a huge capacity to fit even random labellings of millions of images, and that this implied that existing theory (mostly based on VC dimension) was not able to explain their generalization. In particular, VC dimension theory did not take into account that a model could fail to generalize for some data sets while generalizing well for others, which is precisely what Zhang et al. found about DNNs. The varied approaches that have been tried to solve this problem are presented in a unified manner in **??**.

1.0.3 Machine learning and algorithmic information theory

Another line of work that will be relevant for this thesis applies ideas from algorithmic information theory (AIT) to machine learning. AIT studies a notion of information content for an object defined as the length of the shortest program producing a representation of that object. This quantity is called the Kolmogorov complexity of the object, after Andrei Kolmogorov who proposed it in 1965 and developed a

theory analogous to Shannon's information theory based on it [Kolmogorov, 1965]. In 1964, Solomonoff introduced the related concept of universal a priori probability, which argues that a good prior to use in inductive inference is the prior that assigns probabilities to hypotheses based on their probability that they are produced by a Turing machine which is fed random inputs [Solomonoff, 1964]. In the 1970s, Levin made significant contributions to the field, including the coding theorem [Zvonkin and Levin, 1970, Ming and Vitányi, 2014] which shows a tight connection between the universal prior and Kolmogorov complexity. With regards to applying AIT to machine learning, Hutter has made many contributions [Hutter, 2004], including proofs that an ideal algorithm which is perfectly biased towards simple functions will generalize when the true function is simple [Lattimore and Hutter, 2013]. This result is the AIT version of a similar result based on PAC learning by Blumer et al. [1987], where it was shown that algorithms perfectly biased towards functions with short codes (for any fixed but arbitrary code on the set of possible functions) will generalize in the PAC sense[1] when the target function also has a short code. Recently, a version of the Levin's coding theorem for computable functions [Gács, 1988, Dingle et al., 2018] was applied to a variety of real-world input-output maps and it was found that many systems show a bias towards simple outputs in a way which follows the coding theorem's predictions [Dingle et al., 2018].

1.0.4 Towards a new theory of generalisation for ANNs

In this thesis, I will explore the question of why deep neural networks generalize. A lot of recent work has explored this question, but the question is still considered largely open [Kawaguchi et al., 2017, Poggio et al., 2018, Neyshabur et al., 2017, Jiang et al., 2019, 2020]. I will combine theoretical tools from PAC-Bayes, the infinite-width MFT of DNNs, and AIT, with an experiment approach, to develop a consistent picture of why DNNs generalize.

Generalization is linked to *inductive bias*. A learning algorithm usually tries to find a function that fits the data (referred to as *empirical risk minimization*). Any further mechanism which makes choosing one function more likely than other is called an inductive bias[2]. Statistical learning theory establishes that an inductive bias is

[1] PAC learning guarantees generalization error less than a certain value, with at least a certain probability, under samples of the training set. See **??** and Shalev-Shwartz and Ben-David [2014]

[2] A general learning algorithm may not have a clean distinction between a "fitting the data" component and an "inductive bias" component. Bayesian algorithms, however, do offer such a clean distinction: "fiting the data" is formally identified with the likelihood part of the posterior and the inductive bias is identified with the prior.

necessary for generalization, which can be illustrated by considering what would happen in the absence of any inductive bias. An algorithm that chooses any function that fits the data (from the set of all possible functions[3]) with equal probability, will have very poor generalization, as its output on any new example is equivalent to a uniform random guess.

In Valle-Pérez et al. [2018], we introduced the notion of the *parameter-function map* (PF map) defined as the map between the parameters of a neural network and the function that the neural network with those parameters expresses. It is formally defined in **??**. We empirically found that this map showed a bias towards simple functions, of the same form as the *simplicity bias* found in other maps [Dingle et al., 2018]. We suggest the following hypothesis: that this bias may be the main source of inductive bias for DNNs, and support this with other experiments comparing the PF map bias of DNN architectures with different generalization.

To further investigate the hypothesis, we adopt the MFT of DNNs, also known as Neural Network Gaussian processes (NNGPs). This theory allows one to perform Bayesian inference with models which are equivalent to infinite-width DNNs, and with a i.i.d. prior $P(\theta)$ for the parameters of the network (typically Guassian). We summarize the MFT of DNNs and NNGPs in **??**. We argue that because this prior is essentially uniform in high dimensions, the prior on function space $P(f)$ is mainly determined by the PF map, and empirically show that the function space prior is not very sensitive to the choice of parameter space prior.

The literature on NNGPs has shown that NNGPs show similar generalization performance to their corresponding finite-width DNNs trained with stochastic gradient descent (SGD). This implies that the inductive bias in $P(f)$ is *enough* to explain the generalization of DNNs. In Valle-Pérez et al. [2018], we hypothesized that this was because SGD may be sampling the parameter space close to uniformly (and thus close to Bayesian inference), and we gave extensive evidence in support of this hypothesis in Mingard et al. [2020]. This offers empirical justification for studying Bayesian DNNs, even though most DNNs in practice are trained with SGD (or variants thereof).

In Valle-Pérez et al. [2018], we also applied PAC-Bayes theory with the NNGP prior, to connect the generalization error to the Bayesian evidence (as originally found by MacKay [1992]). This offers quantitative insight into the statistical origins of generalization in DNNs, and allows us to estimate the error using only the training set. A method for predicting generalization based only on the training set allows one to train the algorithm on all data available, as no test set is required to estimate

[3]Note that in this work a restriction to the hypothesis class is referred to as an inductive bias too.

performance. Using all the data is particularly useful when the amount of data is small. Therefore the PAC-Bayes prediction may be useful when performing model selection, or neural architecture search [[]], on certain low-data regimes. However, we think a more promising direction is the study of learning curves, as we will see next. Learning curves is the name given to the behaviour of the average generalization error versus the number of samples, where the average is typically over draws of the training set.

In Valle-Perez and Louis, we extended the PAC-Bayes theorem to give high-probability predictions to the actual error, rather than the expected error (see **??**), and we prove a result that shows that under some weak conditions, the Bayesian evidence should display a 'learning curve' with the same exponent as the true learning curve. We experimentally demonstrate this over a large range of modern DNN architectures and datasets, demonstrating that our mean field PAC-Bayes theory has the best predictive power for generalization error of the different proposed theories. We also discuss its limitations, which come mainly from its use of some uncontrolled approximations, and the significant computational expense in computing the bound for large datasets sizes.

The experiments of Mingard et al. [2020] and Valle-Perez and Louis together offer substantial evidence for the hypothesis originally proposed in Valle-Pérez et al. [2018] that the PF map is the main source of inductive bias of DNNs, and represent one of the main results of this thesis. On the theoretical side, and building on this hypothesis, we offer a quantitative framework for predicting generalization and learning curves based on extensions to the PAC-Bayes theory for DNNs. We also analyzed some intriguing properties of the PF map of the simplest neural network, the perceptron, proving that it has a very specific form of bias towards low entropy functions [Mingard et al., 2019]. In Valle-Pérez et al. [2018], we also considered the deeper question of why the inductive bias leads to generalization. Fundamentally, generalization is expected when the inductive bias is similar to the target functions which we want the networks to learn (an statement made rigorous by the PAC-Bayes theorem). Characterizing the distribution of target functions found in real-world problems is a difficult task, but it is generally agreed that problems on which AI is applied are "simple" or "structured" in a way that can be at least approximately captured by the notion of Kolmogorov complexity [Schmidhuber, 1997, Bengio et al., 2007, Lin et al., 2017]. Therefore, we argue, that the inductive bias of the PF map is good because it is also biased towards simple functions.

The thesis is organized as follows. In **??**, we show that the PF map is biased towards simple functions for a simple DNN implementing Boolean functions. We

explore the effect of changing the complexity measure, parameter distribution and number of layers. We also link the simplicity bias to some effects that the target function complexity has on learning. In **??**, we cover the theoretical and experimental results about the PF map of the perceptron. We prove that it is biased towards low entropy (high class imbalance) functions, and extend the result to infinite-width mutilayer perceptrons with ReLU nonlinearity. We also show empirical results on the effect of bias. In **??**, we frame existing works on generalization in deep learning under a unified perspective, propose our new PAC-Bayes mean-field theory of generalization, and prove some theoretical results regarding its tightness, based on the asymptotic behaviour of learning curves. In **??**, we present the results from extensive experiments testing the predictive power of the PAC-Bayes theory on a large range of architectures and datasets. In **??**, we present extensive empirical evidence for the applicability of our theory by showing that DNNs trained with standard optimization algorithms have a similar inductive bias to Bayesian neural networks. Finally, in **??**, the significance of the results and future directions are discussed.

Bibliography

S-I Amari. Neural theory of association and concept-formation. *Biological cybernetics*, 26(3):175–185, 1977.

VV Ashby. R.(1952) design for a brain, 1952.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

Yoshua Bengio, Yann LeCun, et al. Scaling learning algorithms towards ai. *Large-scale kernel machines*, 34(5):1–41, 2007.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Occam's razor. *Information processing letters*, 24(6):377–380, 1987.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

Leo Breiman. Reflections after refereeing papers for nips. *The Mathematics of Generalization*, pages 11–15, 1995.

James S Bryan. Experiments in adaptive pattern recognition. *IEEE Transactions on Military Electronics*, (2 & 3):174–179, 1963.

Dan C Cireşan, Ueli Meier, Jonathan Masci, Luca M Gambardella, and Jürgen Schmidhuber. High-performance neural networks for visual object classification. *arXiv preprint arXiv:1102.0183*, 2011.

Harvey A Cohen. The perceptron controversy. URL `http://harveycohen.net/image/perceptron.html`.

Kamaludin Dingle, Chico Q Camargo, and Ard A Louis. Input–output maps are strongly biased towards simple outputs. *Nature communications*, 9(1):761, 2018.

Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161, 1997.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*, 2017. URL `http://auai.org/uai2017/proceedings/papers/173.pdf`.

David Foster. *Generative deep learning: teaching machines to paint, write, compose, and play*. O'Reilly Media, 2019.

Kunihiko Fukushima. Neural network model for a mechanism of pattern recognition unaffected by shift in position-neocognitron. *IEICE Technical Report, A*, 62(10): 658–665, 1979.

Peter Gács. *Lecture notes on descriptional complexity and randomness*. Citeseer, 1988.

A Gamba. Optimum performance of learning machines. *PROCEEDINGS OF THE INSTITUTE OF RADIO ENGINEERS*, 49(1):349, 1961.

Dan Guest, Kyle Cranmer, and Daniel Whiteson. Deep learning and its application to lhc physics. *Annual Review of Nuclear and Particle Science*, 68:161–181, 2018.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

Friedrich August Hayek. *The sensory order: An inquiry into the foundations of theoretical psychology*. University of Chicago Press, 1952.

Donald Olding Hebb. *The organization of behavior: a neuropsychological theory*. J. Wiley; Chapman & Hall, 1949.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minimum search finds simple nets. Technical report, 1994. Technical Report FKI-200-94.

Sepp Hochreiter and Jürgen Schmidhuber. Simplifying neural nets by discovering flat minima. In *Advances in neural information processing systems*, pages 529–536, 1995.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1): 1–42, 1997.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8): 2554–2558, 1982.

Kurt Hornik, Maxwell Stinchcombe, Halbert White, et al. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Jim Howe. Artificial intelligence at edinburgh university: A perspective. *Archived from the original on*, 17, 2007.

Marcus Hutter. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media, 2004.

AG Ivakhnenko and VG Lapa. Cybernetic predictive devices, 1965.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

Yiding Jiang, Pierre Foret, Scott Yak, Daniel M. Roy, Hossein Mobahi, Gintare Karolina Dziugaite, Samy Bengio, Suriya Gunasekar, Isabelle Guyon, and Behnam Neyshabur. Neurips 2020 competition: Predicting generalization in deep learning, 2020.

RD Joseph. On predicting perceptron performance. In *PROCEEDINGS OF THE INSTITUTE OF RADIO ENGINEERS*, volume 48, pages 398–398, 1960.

Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *CoRR*, abs/1710.05468, 2017. URL `http://arxiv.org/abs/1710.05468`.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *CoRR*, abs/1609.04836, 2016. URL `http://arxiv.org/abs/1609.04836`.

Teuvo Kohonen. Correlation matrix memories. *IEEE transactions on computers*, 100 (4):353–359, 1972.

Andrei Nikolaevich Kolmogorov. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. In *Doklady Akademii Nauk*, volume 114, pages 953–956. Russian Academy of Sciences, 1957.

Andrei Nikolaevich Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information and Transmission*, 1(1):3–11, 1965.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Tor Lattimore and Marcus Hutter. No free lunch versus occams razor in supervised learning. In *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence*, pages 223–235. Springer, 2013.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436, 2015.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

Esther Levin, Naftali Tishby, and Sara A Solla. A statistical approach to learning and generalization in layered neural networks. *Proceedings of the IEEE*, 78(10): 1568–1574, 1990.

Henry W Lin, Max Tegmark, and David Rolnick. Why does deep and cheap learning work so well? *Journal of Statistical Physics*, 168(6):1223–1247, 2017.

Seppo Linnainmaa. The representation of the cumulative rounding error of an algorithm as a taylor expansion of the local rounding errors. *Master's Thesis (in Finnish), Univ. Helsinki*, pages 6–7, 1970.

William A Little. The existence of persistent states in the brain. *Mathematical biosciences*, 19(1-2):101–120, 1974.

Dale Raymond Lumb. An evaluation of the perceptron theory. 1959.

David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

Minsky Marvin and A Papert Seymour. Perceptrons, 1969.

RA Mattson. A self-organizing logical system. 1959 eastern joint computer conf. *Conv. Ree. Inst. Radio Engrs, NY.*

Richard Lewis Mattson. *The design and analysis of an adaptive system for statistical classification.* PhD thesis, Massachusetts Institute of Technology, Department of Electrical Engineering, 1959.

David A McAllester. Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234. ACM, 1998.

Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

LI Ming and Paul MB Vitányi. Kolmogorov complexity and its applications. *Algorithms and Complexity*, 1:187, 2014.

Chris Mingard, Joar Skalse, Guillermo Valle-Pérez, David Martínez-Rubio, Vladimir Mikulik, and Ard A Louis. Neural networks are a priori biased towards boolean functions with low entropy. *arXiv preprint arXiv:1909.11522*, 2019.

Chris Mingard, Guillermo Valle-Pérez, Joar Skalse, and Ard A Louis. Is sgd a bayesian sampler? well, almost. *arXiv preprint arXiv:2006.15191*, 2020.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Radford M Neal. Priors for infinite networks (tech. rep. no. crg-tr-94-1). *University of Toronto*, 1994.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5949–5958, 2017.

Mikel Olazaran. A sociological study of the official history of the perceptrons controversy. *Social Studies of Science*, 26(3):611–659, 1996.

G Palmieri and R Sanna. *Automatic Probabilistic Programmer: Analizer for Pattern Recognition*. Feltrinelli, 1960.

DB Parker. Learning-logic (tr-47). *Center for Computational Research in Economics and Management Science. MIT-Press, Cambridge, Mass*, 8, 1985.

Harry A Pierson and Michael S Gashler. Deep learning in robotics: a review of recent research. *Advanced Robotics*, 31(16):821–835, 2017.

Tomaso Poggio, Kenji Kawaguchi, Qianli Liao, Brando Miranda, Lorenzo Rosasco, Xavier Boix, Jack Hidary, and Hrushikesh Mhaskar. Theory of deep learning iii: the non-overfitting puzzle. Technical report, CBMM memo 073, 2018.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

Mihaela Robila and Stefan A Robila. Applications of artificial intelligence methodologies to behavioral and social sciences. *Journal of Child and Family Studies*, pages 1–13, 2019.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Frank Rosenblatt. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY, 1961.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

Jürgen Schmidhuber. Discovering neural nets with low kolmogorov complexity and high generalization capability. *Neural Networks*, 10(5):857–873, 1997.

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Claude E Shannon. A symbolic analysis of relay and switching circuits. *Electrical Engineering*, 57(12):713–723, 1938.

John Shawe-Taylor and Robert C Williamson. A pac analysis of a bayesian estimator. In *Annual Workshop on Computational Learning Theory: Proceedings of the tenth annual conference on Computational learning theory*, volume 6, pages 2–9, 1997.

John Shawe-Taylor, Peter L Bartlett, Robert C Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE transactions on Information Theory*, 44(5):1926–1940, 1998.

Frank J Smieja. Learning and generalization in feed-forward neural networks (neural networks, backpropagation learning algorithm). 1989.

Ray J Solomonoff. A formal theory of inductive inference. part i. *Information and control*, 7(1):1–22, 1964.

Haim Sompolinsky. Statistical mechanics of neural networks. *Physics Today*, 41(21): 70–80, 1988.

Russell Stuart, Norvig Peter, et al. Artificial intelligence: a modern approach, 2003.

Naftali Tishby. Statistical physics models of supervised learning. In *The Mathematics Of Generalization*, pages 215–242. Addison-Wesley Longman Publishing Co., Inc., 1995.

Naftali Tishby, Esther Levin, and Sara A Solla. Consistent inference of probabilities in layered networks: Predictions and generalization. In *International Joint Conference on Neural Networks*, volume 2, pages 403–409, 1989.

Alan Mathison Turing. Intelligent machinery, 1948.

AM Uttley. Conditional probability machines and conditioned reflexes. inautomata studies. eds. ce shannon and j. mccarthy, 1956.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11): 1134–1142, 1984.

Guillermo Valle-Perez and Ard A. Louis. Generalization theory of deep learning.

Guillermo Valle-Pérez, Chico Q Camargo, and Ard A Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. *arXiv preprint arXiv:1805.08522*, 2018.

Vladimir Vapnik. On the uniform convergence of relative frequencies of events to their probabilities. In *Doklady Akademii Nauk USSR*, volume 181, pages 781–787, 1968.

Vladimir Vapnik and Alexey Chervonenkis. Theory of pattern recognition, 1974.

Vladimir N Vapnik. The nature of statistical learning theory, 1995.

John Von Neumann. Probabilistic logics and the synthesis of reliable organisms from unreliable components. *Automata studies*, 34:43–98, 1956.

DJ Wallace. Neural network models: a physicist's primer. *Computational Physics (SUSSP 32)*, pages 168–211, 1987.

Paul J Werbos. Applications of advances in nonlinear sensitivity analysis. In *System modeling and optimization*, pages 762–770. Springer, 1982.

PJ Werbos. *Beyond regression: new tools for prediction and analysis in the behavioral sciences, Harvard University*. PhD thesis, Masters Thesis, 1974.

Bernard Widrow. Generalization and information storage in network of ada-line'neurons'. *Self-organizing systems-1962*, pages 435–462, 1962.

Bernard Widrow and Marcian E Hoff. Adaptive switching circuits. Technical report, Stanford Univ Ca Stanford Electronics Labs, 1960.

David H Wolpert. Mathematics of generalization: Proceedings: Sfi-cnls workshop on formal approaches to supervised learning (1992: Santa fe, nm), 1995.

Georgios N Yannakakis and Julian Togelius. *Artificial intelligence and games*, volume 2. Springer, 2018.

Lenka Zdeborová. Understanding deep learning is also a job for physicists. *Nature Physics*, pages 1–3, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. URL `https://arxiv.org/abs/1611.03530`.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P Adams, and Peter Orbanz. Non-vacuous generalization bounds at the imagenet scale: a pac-bayesian compression approach. 2018.

Alexander K Zvonkin and Leonid A Levin. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Mathematical Surveys*, 25(6):83, 1970.