ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Benign overfitting of fully connected Deep Nets: A Sobolev space viewpoint

Emmanuel Caron [1] and Stéphane Chrétien[2]

1- Laboratoire de Mathématiques d'Avignon,
Université d'Avignon,
Avignon, France

2- Université Lumière-Lyon-II,
69676 Bron Cedex, France

Abstract. Deep neural nets have undergone tremendous improvements in the last decade, which revolutionised the field of machine learning in a broad and lasting manner, achieving unprecedented performance in such diverse fields as image analysis, point cloud registration, natural language processing and model free control. On the theoretical side, understanding the underpinnings of deep learning remains a formidable challenge, despite impressive breakthroughs in the last decade. One particularly interesting new prospect is the analysis of the double descent phenomenon described in Belkin et al. [2019], a counter-intuitive theory bringing new insight on the performance of learning systems in the greatly over-parametrised regime.The list of contribution to the understanding of the double descent paradigm has grown substantially in the last two years, but all available results in the literature mainly focus on the linear and the kernel setups. In the present paper, we study the overparametrised part of the double descent curve introduced in Belkin et al. [2019] and propose a new approach to the study of benign overfitting in the setting of learning Sobolev maps.

1 Introduction

This paper aims to study the theoretical performance of deep neural networks in the overparametrised regime. Understanding the striking success of deep learning in so many applications has remained a formidable challenge since the early days of the deep learning revolution. One of the very counter-intuitive phenomena encountered in the field of deep learning is the fact that zero training error can does not preclude good generalisation on unobserved data. Most recent results in machine learning theory address this puzzling phenomenon via the theory of random matrices for kernel methods, and therefore struggle to resolve the interpolation/generalisation paradox for general deep learning architectures. In this paper, we propose a novel analysis of the empirical risk minimisation with the lens of perturbation theory and provide rigourous justifications for this conundrum in generalisation theory.

Let us now describe our statistical model and survey the recent approaches for the study of generalisation for overparametrised networks.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

1.1 Statistical setup

1.1.1 The observation model

Let $Z_i = (X_i, Y_i)$ in $\mathbb{R}^d \times \mathbb{R}$, $i = 1, \ldots, n$ be observations drawn from the following model

$$Y_i = f^*(X_i) + \varepsilon_i$$

$i = 1, \ldots, n$, where we assume that the vectors X_i are random and i.i.d., taking values in $\mathbb{R}^d$ and the noise vector $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_n]^t$ is sub-Gaussian, with sub-Gaussian constant denoted by K_ϵ. The goal of machine learning is to estimate f^* based on the observation $Z_1, \ldots, Z_n$.

1.1.2 Empirical risk minimisation

The estimation of f^* will be based on restricting the search to a subset $\mathcal{F}$ of functions of a Banach space $\mathcal{B}$. The estimator will be chosen in the set of stationary points of the empirical version of the risk $R : \mathcal{F} \to \mathbb{R}$ defined by

$$R(f) = \mathbb{E}\left[\ell(Y, f(X))\right],$$

where $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ satisfies $\ell(y, y) = 0$ for all $y \in \mathbb{R}$ and $\ell(y, \cdot)$: $\mathbb{R} \mapsto \mathbb{R}$ is a strictly convex twice continuously differentiable nonnegative function. Let $\hat{R}_n(f)$ denote the empirical risk defined by

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)).$$

Then, the Empirical Risk Minimizer $\hat{f}^{ERM}$ will be a solution to

$$\hat{f}^{ERM} \in \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}_n(f). \tag{1}$$

1.2 Generalisation of deep networks

Understanding generalisation is one of the greatest challenges in deep learning. Brute force use of standard complexities from statistical learning theory often lead to vacuous bounds. Some recent approaches, based on compression Arora et al. [2018], Baykal et al. [2018], etc or PAC-Bayes techniques Laviolette [2017], Guedj [2019] have recently proved useful for deep networks Dziugaite and Roy [2017], Neyshabur et al. [2017]. The dependency on the number of layers was discussed in Golowich et al. [2018] and refined results were recently obtained in Wei and Ma [2019]. Relationships with kernels, called tangent kernels, were discovered two years ago in Jacot et al. [2018], which opens the door to new refined generalisation guarantees in certain regimes; but see also the discussion in Chizat et al. [2019]. In the present paper, we take a different, perturbative, route to the analysis of the generalisation problem.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

1.3 Intrinsic dimension of the data and relative distance between the samples

Many recent works have studied the intrinsic dimension of various data sets arising in machine learning Costa et al. [2005], Hein and Audibert [2005], Amsaleg et al. [2015], Ansuini et al. [2019], Mezard [2020]. Manifold learning is one way of decribing the data space. One often considers that the data lives on a metric measure spaces $(\mathcal{X}, dist, \mu)$, where a set $\mathcal{X}$ is equipped with both a metric $dist$ and a measure μ. We assume that $\mu(\mathcal{X}) = 1$ as in Clarkson [2006]. The counting measure $|A|$, can thus be used for estimation purposes using concentration of measure tools, when the data are independently identically distributed on $\mathcal{X}$ with distribution μ. The information dimension is closely related to the pointwise dimension $\alpha_\mu(x)$ for $x \in \mathcal{X}$, also known as the local dimension or Hölder exponent which is defined as

$$\alpha_\mu(x) = \lim_{\epsilon \to 0} \frac{\log \mu(B(x, \epsilon))}{\log \epsilon}$$

It was shown in Cutler and Dawson [1989] that for all $i = 1, \ldots, n$,

$$\min_{i'=1}^{n} \|x - X_{i'}\|_2 = n^{-1/\alpha_\mu(x) + o(1)}$$

as $n \to \infty$. Similar observations were made by Camastra [2003], Clarkson [2006], [Mezard, 2020, 50:08/1:34:37]. In the sequel, we will make the following assumption.

Assumption 1 *There holds*

$$\min_{i,i'=1}^{n} \|X_i - X_{i'}\|_2 \geq c n^{-1/\nu} \tag{2}$$

with probability larger than or equal to $1 - \delta$*, for some positive constants* c*,* ν *and for* $\delta \in (0, 1)$*.*

The Holder exponent ν is usually interpreted as a surrogate for the intrinsic dimension of the data manifold. E.g., this intrinsic dimension was estimated to be less than 20 for the MNIST dataset in Hein and Audibert [2005].

2 Main results

2.1 A general bound for the ERM in a Banach space

Our first contribution is the following general error bound for minimisers of the empirical risk in Banach spaces.

Theorem 1 *Assume that* f^* *belongs to a family* $\mathcal{F}$ *of functions in a Banach space* $\mathcal{B}$*. Set* ℓ *to be the* ℓ_2^2 *loss, i.e.* $\ell(y, z) = \frac{1}{2}(y - z)^2$ *for all* y*,* z *in* $\mathbb{R}$*. Let*

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Assumption 1 hold. Let ψ denote the bump function

$$\psi(x) = \begin{cases} \exp\left(1 - \frac{1}{1-\|x\|_2^2}\right) & \text{if } \|x\|_2^2 \leq 1, \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

and let $\psi_\sigma = \psi(\cdot/\sigma)$. Take any $\sigma \leq cn^{-1/\nu}$ such that the ball in $\mathcal{B}$ centered at f^* with radius $6K_\epsilon\, n\|\psi_\sigma\|_{\mathcal{B}}$ is included in $\mathcal{F}$. Then, with probability larger than $1-\delta$, there exists a mapping $\hat{f}^{ERM}\colon \mathbb{R}^d \mapsto \mathbb{R}$ which solves the empirical risk minimisation problem (1) and which lies at a distance at most $6K_\epsilon\, n\|\psi_\sigma\|_{\mathcal{B}}$ from the neural network f^*.

2.2 Application to deep neural networks

We will concentrate on the case of $\mathcal{B}$ being equal to the Sobolev space $\mathcal{W}^{k,p}(\mathcal{D})$ on a compact domain $\mathcal{D} \subset (0,1)^d$ of $\mathbb{R}^d$.

2.2.1 Deep networks

Traditional feedforward architectures implement a map as a sequence of affine-linear transformations, denoted by $T_l : \mathbb{R}^{N_{l-1}} \to \mathbb{R}^{N_l}$, $l = 1, \ldots, L$, followed by a componentwise application of a non-linear function, denoted by $\varrho : \mathbb{R} \to \mathbb{R}$ and called activation function. The parameters defining the affine transformations T_l are referred to as weights and will be denoted by W. Let

$$f_l : \mathbb{R}^{N_{l-1}} \to \mathbb{R}^{N_l},\ x \mapsto \varrho\left(T_l(x)\right)$$

for $l = 1, \ldots, L-1$,

$$f_L : \mathbb{R}^{N_{L-1}} \to \mathbb{R}^{N_L},\ x \mapsto T_L(x)$$

and define

$$f_W = f_L \circ f_{L-1} \circ \cdots \circ f_1.$$

Let $\mathcal{F}$ denote the subset of $\mathcal{W}^{k,p}$ consisting of deep neural networks.

2.2.2 Main result

Our main result concerning deep networks is the following

Theorem 2 *Set ℓ to be the ℓ_2^2 loss, i.e. $\ell(y,z) = \frac{1}{2}(y-z)^2$ for all y, z in $\mathbb{R}$. Let Assumption 1 hold. Assume that $\|f^*\|_{\mathcal{W}^{k,p}(\mathcal{D})} \leq B$ for some $k \in \mathbb{N}$ and $p \in [1, +\infty]$. Let ψ denote the bump function (3) with $\sigma \leq \frac{1}{2}cn^{-1/\nu}$ and let $\psi_\sigma = \psi(\cdot/\sigma)$. Assume that d, p, ν and n are such that*

$$3K_\epsilon n^{1-d/(\nu p)} 2\sqrt{C}\; C_\psi\; \pi^{d/4}\,(6k)^{3k} \left(\frac{d^7}{4\exp(6)}\right)^{k/2} \leq B. \tag{4}$$

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Then, for any $s \in [0,1]$*, there exists a positive constant* $C = C(d,k,B,s)$ *such that with probability at least* $1-\exp(-n)$*, the class* $\mathcal{F}$ *of neural networks parametrised by*

(i) the number L *of layers is bounded by*

$$L \leq c\ \log_2\left(\rho^{-k/(k-s)}\right)$$

(ii) the number $d + \sum_{l=1}^{L} N_l$ *of neurons is bounded by*

$$d + \sum_{l=1}^{L} N_l \leq c\ \rho^{-d/(k-s)} \cdot \log_2\left(\rho^{-k/(k-s)}\right).$$

contains at least one neural network $f_{\hat{W}}\colon \mathbb{R}^d \mapsto \mathbb{R}$ *with* $\mathcal{W}^{s,2}(\mathcal{D})$*-distance at most*

$$K_\epsilon n^{1-d/(2\nu)} 2\sqrt{C}\ C_\psi\ \pi^{d/4}\,(6k)^{3k}\left(\frac{d^7}{4\exp(6)}\right)^{k/2} \tag{5}$$

to the solution set of the empirical risk minimisation problem (1) *over the Sobolev class* $\mathcal{W}^{k,2}(\mathcal{D})$ *and such that the estimation error is bounded by with*

$$\left\|f_{\hat{W}} - f^*\right\|_{L^2(\mathcal{D})} \leq \rho + 3K_\epsilon n^{1-d/(2\nu)} 2\sqrt{C}\ C_\psi\ \pi^{d/4}\,(6k)^{3k}\left(\frac{d^7}{4\exp(6)}\right)^{k/2}. \tag{6}$$

Theorem 2 shows the existence of an almost ERM minimising deep neural network over a Sobolev class for Sobolev map estimation, with controlled complexity. Our result shows that the error bound improves as the intrinsic dimensionality ν decreases, and that the influence of ν is exponential. Our conclusion is that deep networks natural adapts to the intrinsic complexity of the data.

3 Perspectives

We are currently extending our results to a deeper understanding of the double descent phenomenon for general deep networks using our new approach.

References

Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E Houle, Ken-ichi Kawarabayashi, and Michael Nett. Estimating local intrinsic dimensionality. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 29–38, 2015.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 6109–6119, 2019.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. *arXiv preprint arXiv:1802.05296*, 2018.

Cenk Baykal, Lucas Liebenwein, Igor Gilitschenski, Dan Feldman, and Daniela Rus. Data-dependent coresets for compressing neural networks with applications to generalization bounds. *arXiv preprint arXiv:1804.05345*, 2018.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Francesco Camastra. Data dimensionality estimation methods: a survey. *Pattern recognition*, 36(12):2945–2954, 2003.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2933–2943, 2019.

Kenneth L Clarkson. Nearest-neighbor searching and metric space dimensions. *Nearest-neighbor methods for learning and vision: theory and practice*, pages 15–59, 2006.

Jose A Costa, Abhishek Girotra, and AO Hero. Estimating local intrinsic dimension with k-nearest neighbor graphs. In *IEEE/SP 13th Workshop on Statistical Signal Processing, 2005*, pages 417–422. IEEE, 2005.

Colleen Diane Cutler and Donald Andrew Dawson. Estimation of dimension for spatially distributed data and related limit theorems. *Journal of multivariate analysis*, 28(1):115–148, 1989.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299, 2018.

Benjamin Guedj. A primer on pac-bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019.

Matthias Hein and Jean-Yves Audibert. Intrinsic dimensionality estimation of submanifolds in rd. In *Proceedings of the 22nd international conference on Machine learning*, pages 289–296, 2005.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

François Laviolette. A tutorial on pac-bayesian theory. In *Talk at the NIPS 2017 Workshop:(Almost)*, volume 50, 2017.

Marc Mezard. Artificial intelligence: success, limits, myths and threats. statistical physics and statistical inference (lecture 3). video link, 2020.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

Colin Wei and Tengyu Ma. Improved sample complexities for deep neural networks and robust classification via an all-layer margin. In *International Conference on Learning Representations*, 2019.