Understanding interpolation in machine learning (with a quick overview of other current hot related topics for the analysis of deep neural networks)

Stéphane Chrétien

University of Lyon 2
ERIC Laboratory

Mathematics of Machine Learning LMS-Bath Symposium,
3-7 August 2020, University of Bath

Introduction

Modern machine learning and statistics deal with the problem of learning from data:

given a training dataset (y_i, x_i) $i \in I$ where

$x_i \in \mathbb{R}^d$ is the input
$y_i \in \mathbb{R}$ is the output,

one seeks a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ from a certain function class $\mathcal{F}$ that has good prediction performance on test data $(y_t, x_t))$, $t \in T$, i.e. which has small testing error

$$\sum_{t \in T} \ell(y_t, f(x_t)) \tag{1}$$

This problem is of fundamental significance and finds applications in numerous scenarios.

For instance, in image recognition,

the input x corresponds to the raw image
the output y is the image category

and the goal is to find a mapping f that can classify new images with acceptable accuracy.

Decades of research efforts in statistical machine learning have been devoted to developing methods to find f efficiently with provable guarantees.

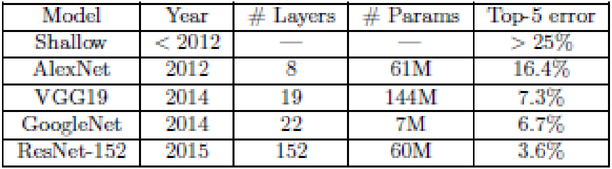
- Prominent examples include
  - linear classifiers (e.g., linear / logistic regression, linear discriminant analysis),
  - kernel methods (e.g., support vector machines),
  - tree-based methods (e.g., decision trees, random forests),
  - nonparametric regression (e.g., nearest neighbors, local kernel smoothing), etc.
- Roughly speaking, each aforementioned method corresponds to a different function class $\mathcal{F}$ from which the final classifier f is chosen.

- Deep learning, in its simplest form, consists in looking for functions of the form

$$\mathcal{F} = \left\{ f(x, \theta) = W_L(\sigma_L(W_{L-1}(\sigma_{L-1}(\cdots \sigma_2(W_1(x))))) \right\}.$$

where σ_l is a non-linear function which applies componentwise and W_l is an affine operator, $l = 1, \dots, L$.

- Evolution of the performances over the last 7 years . . .

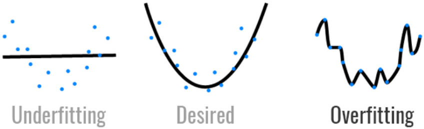


Model	Year	# Layers	# Params	Top-5 error
Shallow	< 2012	—	—	> 25%
AlexNet	2012	8	61M	16.4%
VGG19	2014	19	144M	7.3%
GoogleNet	2014	22	7M	6.7%
ResNet-152	2015	152	60M	3.6%

- It is widely acknowledged that two indispensable factors contribute to the success of deep learning, namely
    - huge datasets that often contain millions of samples and
    - immense computing power resulting from clusters of graphics processing units (GPUs).
- Admittedly, these resources are only recently available.

- However, these two alone are not sufficient to explain the mystery of deep learning:
    - over-parametrization: the number of parameters in state-of-the-art models is very often much larger than the sample size,

        ← which might make them prone to overfitting,



and

- nonconvexity does not seem to be a problem: even with the help of GPUs, training deep learning models is still NP-hard in the worst case due to the highly nonconvex loss function to minimize.

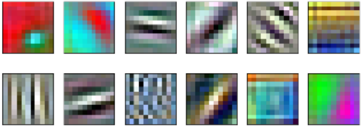
  Nevertheless, standard incremental algorithms (Stochastic Gradient Descent, etc) often reach good minimisers of the Empirical Risk

- A lot remains to be understood ! ...

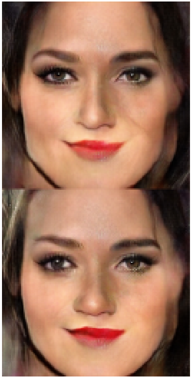
- Deep learning is able to approximate complicated nonlinear maps through composing many simple nonlinear functions.

- The motivation for the multilayer architecture is that there are different levels of features and the layers might be able to properly account for these different levels independently.

- Here, we sample and visualize weights from a pre-trained AlexNet model.



This can be used to generate new images using for instance, Generative Adversarial Networks



Why overparametrise ?

- It is often observed that depth helps efficiently extract features from the dataset, whereas
- (recent studies found that *increasing both depth and width in a shallow model leads to very nice continuous limits, where PDE tools can be put to work...*)

What consequences does overparametrisation have on learning ?

- In deep neural networks, *over-parametrization* usually entails existence of many local minimisers with potentially different statistical performance.

- Common practice advises to runs stochastic gradient descent with random initialization and converges to parameters with *very good practical prediction accuracy*.

Why is this simple approach actually often working ?

The goal of current research is to resolve these paradoxes !

We will now survey some of the recent results from the literature and the most intriguing mysteries of deep learning

- expressivity
- generalisation bounds (PAC + compression)
- optimisation algorithms
- interpolation

Expressivity

- Recent works have been devoted to the approximation accuracy of deep neural networks for various measures of the error (expressivity)

- Some notable works include
  * the approximation results of Yarotsky
  * the nonlinear approximation analysis of RELU networks by Daubechies, Devore Foucart, Hanin and Petrova
  * the approximation of analytic maps by Weinan E and Wang.
  * approximation of functions in Sobolev spaces by Guhring, Kutyniok and Petersen
  * the definition of new approximation spaces (Barron spaces) by Weinan E. et al. (which play the same role as Besov spaces for nonlinear approximation with wavelet bases)
  * etc

We will use the following theorem from Guhring, Kutyniok and Petersen.

Theorem

Let $k \in \mathbb{N}_{\geq 2}, 1 \leq p \leq \infty, B > 0,$ *and* $0 \leq s \leq 1$. *Then, there exists a constant* $c = c(d, p, k, B, s) > 0$ *with the following properties: for any* $\varepsilon \in (0, 1/2)$, *for any* $f\colon (0,1)^d \mapsto \mathbb{R}$ *in the ball of radius* B *in* $W^{k,p}$, *there exists a vector of weights* W *and an associated neural network* f_W *such that*

$$\|f_W - f\|_{W^{s,p}((0,1)^d)} \leq \varepsilon$$

Theorem

and

(i) the number L of layers is bounded by

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

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

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

Generalisation and compression

Really interesting work by Bartlett, Barron and others ...

Theorem (B., Foster, Telgarsky, 2017)

With high probability over n training examples $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{\pm 1\}$, every f_W with $R_W \leq r$ has

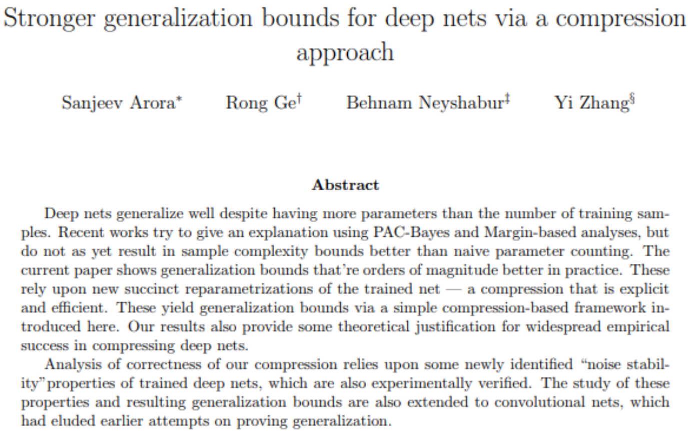
$$\Pr(\operatorname{sign}(f(X)) \neq Y) \leq \frac{1}{n}\sum_{i=1}^{n} 1[Y_i f(X_i) \leq \gamma] + \tilde{O}\left(\frac{rL}{\gamma\sqrt{n}}\right).$$

Here, f_W is computed in a network with L layers and parameters $W_1, \ldots, W_L$:

$$f_W(x) := \sigma_L(W_L \sigma_{L-1}(W_{L-1} \cdots \sigma_1(W_1 x) \cdots)),$$

where the σ_i are 1-Lipschitz, and we measure the scale of f_W using a product of norms of the matrices W_i,
for example, $r := \prod_{i=1}^{L} \|W_i\|_* \left(\sum_{i=1}^{L} \frac{\|W_i\|_{2,1}^{2/3}}{\|W_i\|_*^{2/3}}\right)^{3/2}$.

New trends involve compression showing that intrinsic dimension of the deep networks is not as large as we think !



Stronger generalization bounds for deep nets via a compression approach

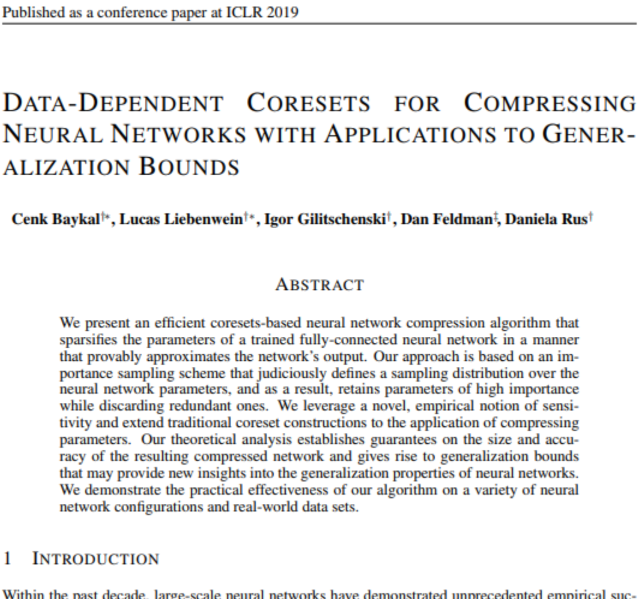
Sanjeev Arora[*] Rong Ge[†] Behnam Neyshabur[‡] Yi Zhang[§]

Abstract

Deep nets generalize well despite having more parameters than the number of training samples. Recent works try to give an explanation using PAC-Bayes and Margin-based analyses, but do not as yet result in sample complexity bounds better than naive parameter counting. The current paper shows generalization bounds that're orders of magnitude better in practice. These rely upon new succinct reparametrizations of the trained net — a compression that is explicit and efficient. These yield generalization bounds via a simple compression-based framework introduced here. Our results also provide some theoretical justification for widespread empirical success in compressing deep nets.

Analysis of correctness of our compression relies upon some newly identified "noise stability"properties of trained deep nets, which are also experimentally verified. The study of these properties and resulting generalization bounds are also extended to convolutional nets, which had eluded earlier attempts on proving generalization.

even the theory of coresets was used in order to compress !



Published as a conference paper at ICLR 2019

Data-Dependent Coresets for Compressing Neural Networks with Applications to Generalization Bounds

Cenk Baykal[†*], Lucas Liebenwein[†*], Igor Gilitschenski[†], Dan Feldman[‡], Daniela Rus[†]

Abstract

We present an efficient coresets-based neural network compression algorithm that sparsifies the parameters of a trained fully-connected neural network in a manner that provably approximates the network's output. Our approach is based on an importance sampling scheme that judiciously defines a sampling distribution over the neural network parameters, and as a result, retains parameters of high importance while discarding redundant ones. We leverage a novel, empirical notion of sensitivity and extend traditional coreset constructions to the application of compressing parameters. Our theoretical analysis establishes guarantees on the size and accuracy of the resulting compressed network and gives rise to generalization bounds that may provide new insights into the generalization properties of neural networks. We demonstrate the practical effectiveness of our algorithm on a variety of neural network configurations and real-world data sets.

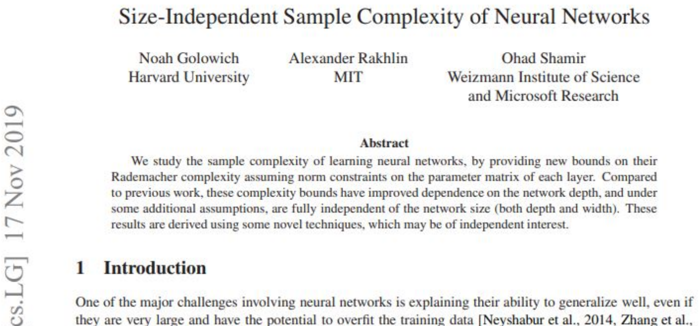
1 Introduction

Within the past decade, large-scale neural networks have demonstrated unprecedented empirical suc-

One of the main issues using the traditional approaches was to obtain bounds which do not blow up as the number of layers increases

$\Rightarrow$ this would go against empirical findings

Noah Golowich, Alexander Rakhlin, Ohad Shamir recently solved the problem using Rademacher complexity in a careful way



Size-Independent Sample Complexity of Neural Networks

Noah Golowich
Harvard University

Alexander Rakhlin
MIT

Ohad Shamir
Weizmann Institute of Science
and Microsoft Research

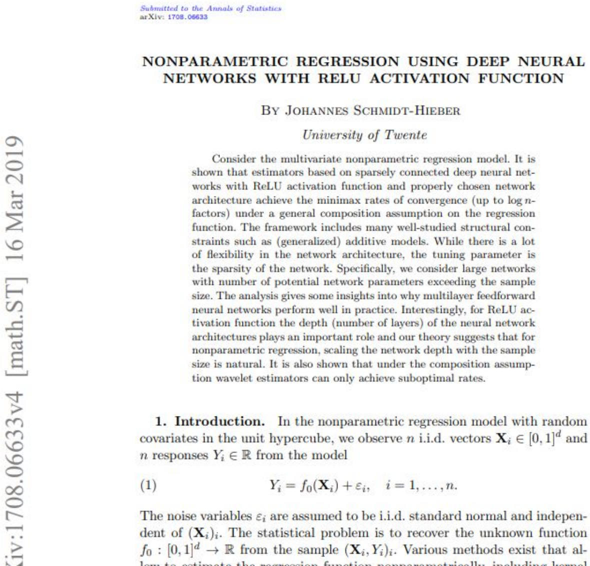
Abstract

We study the sample complexity of learning neural networks, by providing new bounds on their Rademacher complexity assuming norm constraints on the parameter matrix of each layer. Compared to previous work, these complexity bounds have improved dependence on the network depth, and under some additional assumptions, are fully independent of the network size (both depth and width). These results are derived using some novel techniques, which may be of independent interest.

1 Introduction

One of the major challenges involving neural networks is explaining their ability to generalize well, even if they are very large and have the potential to overfit the training data [Neyshabur et al., 2014, Zhang et al.,

Interesting non-parametric statistical oriented results for composition functions using sparsity are given in



Submitted to the Annals of Statistics
arXiv: 1708.06633

NONPARAMETRIC REGRESSION USING DEEP NEURAL NETWORKS WITH RELU ACTIVATION FUNCTION

By Johannes Schmidt-Hieber

University of Twente

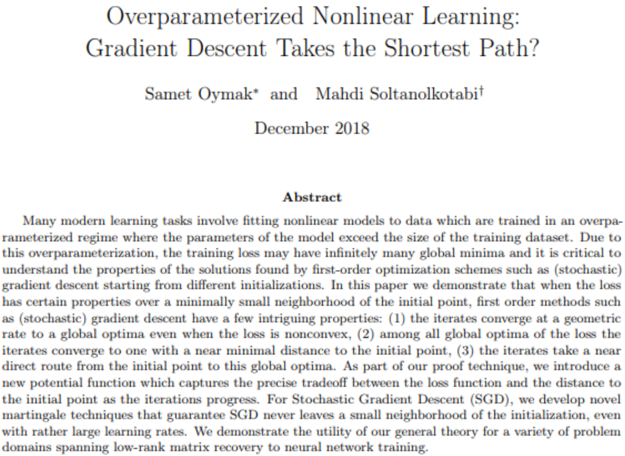
Consider the multivariate nonparametric regression model. It is shown that estimators based on sparsely connected deep neural networks with ReLU activation function and properly chosen network architecture achieve the minimax rates of convergence (up to $\log n$-factors) under a general composition assumption on the regression function. The framework includes many well-studied structural constraints such as (generalized) additive models. While there is a lot of flexibility in the network architecture, the tuning parameter is the sparsity of the network. Specifically, we consider large networks with number of potential network parameters exceeding the sample size. The analysis gives some insights into why multilayer feedforward neural networks perform well in practice. Interestingly, for ReLU activation function the depth (number of layers) of the neural network architectures plays an important role and our theory suggests that for nonparametric regression, scaling the network depth with the sample size is natural. It is also shown that under the composition assumption wavelet estimators can only achieve suboptimal rates.

1. Introduction. In the nonparametric regression model with random covariates in the unit hypercube, we observe n i.i.d. vectors $\mathbf{X}_i \in [0,1]^d$ and n responses $Y_i \in \mathbb{R}$ from the model

$$Y_i = f_0(\mathbf{X}_i) + \varepsilon_i, \quad i = 1, \dots, n. \tag{1}$$

The noise variables ε_i are assumed to be i.i.d. standard normal and independent of $(\mathbf{X}_i)_i$. The statistical problem is to recover the unknown function $f_0 : [0,1]^d \to \mathbb{R}$ from the sample $(\mathbf{X}_i, Y_i)_i$. Various methods exist that al-

Where do gradient methods end up landing in this wild lanscape ?



Overparameterized Nonlinear Learning: Gradient Descent Takes the Shortest Path?

Samet Oymak* and Mahdi Soltanolkotabi†

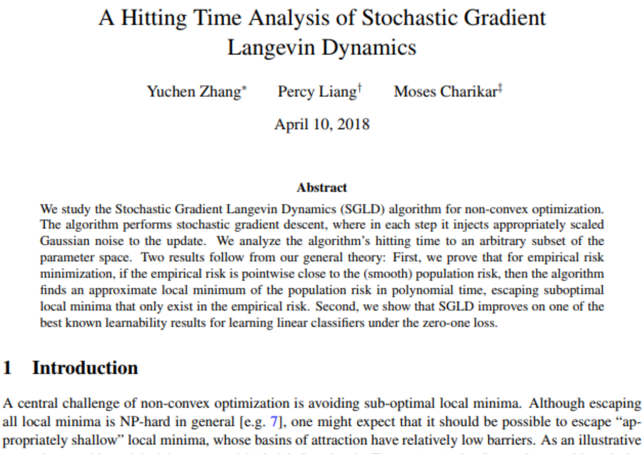
December 2018

Abstract

Many modern learning tasks involve fitting nonlinear models to data which are trained in an overparameterized regime where the parameters of the model exceed the size of the training dataset. Due to this overparameterization, the training loss may have infinitely many global minima and it is critical to understand the properties of the solutions found by first-order optimization schemes such as (stochastic) gradient descent starting from different initializations. In this paper we demonstrate that when the loss has certain properties over a minimally small neighborhood of the initial point, first order methods such as (stochastic) gradient descent have a few intriguing properties: (1) the iterates converge at a geometric rate to a global optima even when the loss is nonconvex, (2) among all global optima of the loss the iterates converge to one with a near minimal distance to the initial point, (3) the iterates take a near direct route from the initial point to this global optima. As part of our proof technique, we introduce a new potential function which captures the precise tradeoff between the loss function and the distance to the initial point as the iterations progress. For Stochastic Gradient Descent (SGD), we develop novel martingale techniques that guarantee SGD never leaves a small neighborhood of the initialization, even with rather large learning rates. We demonstrate the utility of our general theory for a variety of problem domains spanning low-rank matrix recovery to neural network training.

Stochastic Gradient Langevin Dynamics (SGLD) avoids spurious local minimisers !

Proved in the work of Charikar et al. for the Langevin approximation



A Hitting Time Analysis of Stochastic Gradient Langevin Dynamics

Yuchen Zhang* Percy Liang† Moses Charikar‡

April 10, 2018

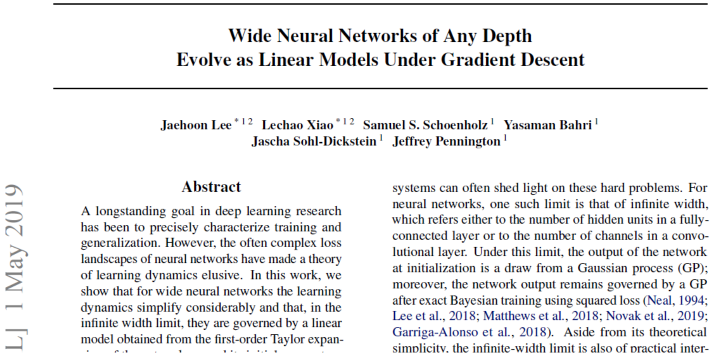
Abstract

We study the Stochastic Gradient Langevin Dynamics (SGLD) algorithm for non-convex optimization. The algorithm performs stochastic gradient descent, where in each step it injects appropriately scaled Gaussian noise to the update. We analyze the algorithm's hitting time to an arbitrary subset of the parameter space. Two results follow from our general theory: First, we prove that for empirical risk minimization, if the empirical risk is pointwise close to the (smooth) population risk, then the algorithm finds an approximate local minimum of the population risk in polynomial time, escaping suboptimal local minima that only exist in the empirical risk. Second, we show that SGLD improves on one of the best known learnability results for learning linear classifiers under the zero-one loss.

1 Introduction

A central challenge of non-convex optimization is avoiding sub-optimal local minima. Although escaping all local minima is NP-hard in general [e.g. 7], one might expect that it should be possible to escape "appropriately shallow" local minima, whose basins of attraction have relatively low barriers. As an illustrative

Width helps simplify the analysis !



Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent

Jaehoon Lee[*12] Lechao Xiao[*12] Samuel S. Schoenholz[1] Yasaman Bahri[1]
Jascha Sohl-Dickstein[1] Jeffrey Pennington[1]

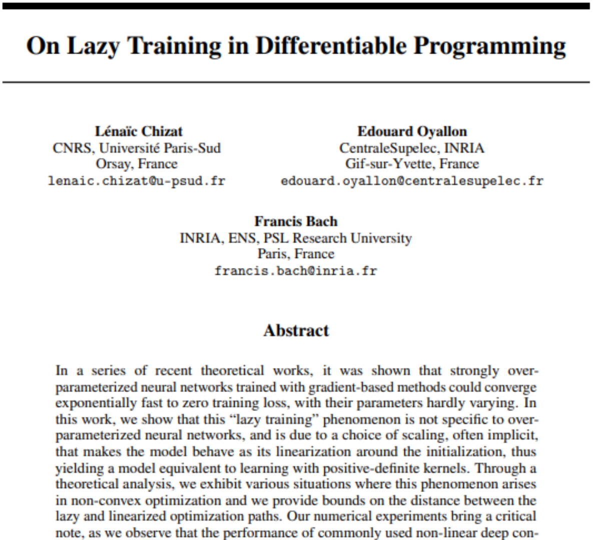
L] 1 May 2019

Abstract

A longstanding goal in deep learning research has been to precisely characterize training and generalization. However, the often complex loss landscapes of neural networks have made a theory of learning dynamics elusive. In this work, we show that for wide neural networks the learning dynamics simplify considerably and that, in the infinite width limit, they are governed by a linear model obtained from the first-order Taylor expan-

systems can often shed light on these hard problems. For neural networks, one such limit is that of infinite width, which refers either to the number of hidden units in a fully-connected layer or to the number of channels in a convolutional layer. Under this limit, the output of the network at initialization is a draw from a Gaussian process (GP); moreover, the network output remains governed by a GP after exact Bayesian training using squared loss (Neal, 1994; Lee et al., 2018; Matthews et al., 2018; Novak et al., 2019; Garriga-Alonso et al., 2018). Aside from its theoretical simplicity, the infinite-width limit is also of practical inter-

But recent work of Chizat, Oyallon and Bach showed that deep neural networks then reach in a tangent kernel regime . . .



On Lazy Training in Differentiable Programming

Lénaïc Chizat
CNRS, Université Paris-Sud
Orsay, France
lenaic.chizat@u-psud.fr

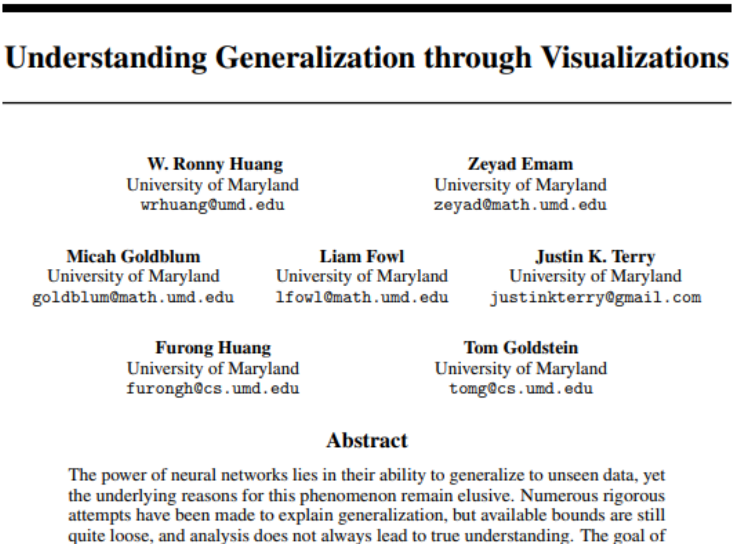
Edouard Oyallon
CentraleSupelec, INRIA
Gif-sur-Yvette, France
edouard.oyallon@centralesupelec.fr

Francis Bach
INRIA, ENS, PSL Research University
Paris, France
francis.bach@inria.fr

Abstract

In a series of recent theoretical works, it was shown that strongly overparameterized neural networks trained with gradient-based methods could converge exponentially fast to zero training loss, with their parameters hardly varying. In this work, we show that this "lazy training" phenomenon is not specific to overparameterized neural networks, and is due to a choice of scaling, often implicit, that makes the model behave as its linearization around the initialization, thus yielding a model equivalent to learning with positive-definite kernels. Through a theoretical analysis, we exhibit various situations where this phenomenon arises in non-convex optimization and we provide bounds on the distance between the lazy and linearized optimization paths. Our numerical experiments bring a critical note, as we observe that the performance of commonly used non-linear deep con-

The importance of being flat !



Understanding Generalization through Visualizations

W. Ronny Huang
University of Maryland
wrhuang@umd.edu

Zeyad Emam
University of Maryland
zeyad@math.umd.edu

Micah Goldblum
University of Maryland
goldblum@math.umd.edu

Liam Fowl
University of Maryland
lfowl@math.umd.edu

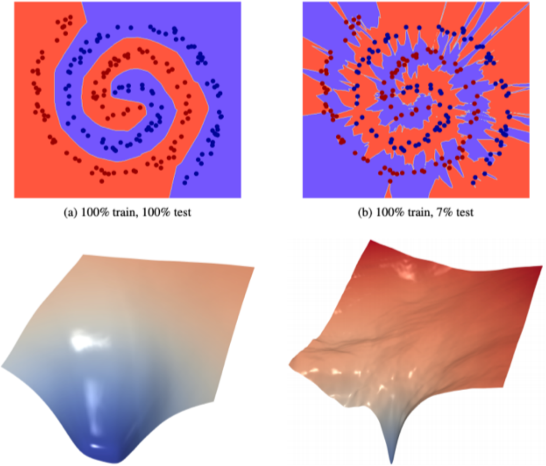
Justin K. Terry
University of Maryland
justinkterry@gmail.com

Furong Huang
University of Maryland
furongh@cs.umd.edu

Tom Goldstein
University of Maryland
tomg@cs.umd.edu

Abstract

The power of neural networks lies in their ability to generalize to unseen data, yet the underlying reasons for this phenomenon remain elusive. Numerous rigorous attempts have been made to explain generalization, but available bounds are still quite loose, and analysis does not always lead to true understanding. The goal of



(a) 100% train, 100% test

(b) 100% train, 7% test

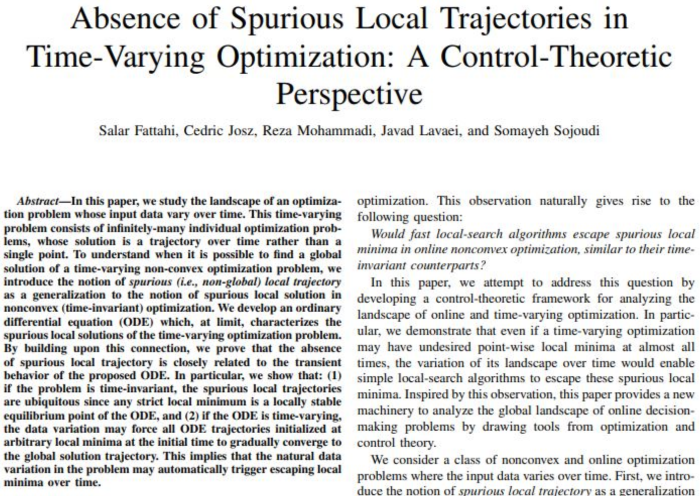
(c) Minimizer of network in (a) above

(d) Minimizer of network in (b) above

Figure: Flat minimisers are easier to reach and have better generalisation properties (empirical)

Yet another fascinating direction . . .

Recent work of Lavaei and co-authors shows that incremental methods can avoid spurious minimisers



Absence of Spurious Local Trajectories in Time-Varying Optimization: A Control-Theoretic Perspective

Salar Fattahi, Cedric Josz, Reza Mohammadi, Javad Lavaei, and Somayeh Sojoudi

***Abstract*—In this paper, we study the landscape of an optimization problem whose input data vary over time. This time-varying problem consists of infinitely-many individual optimization problems, whose solution is a trajectory over time rather than a single point. To understand when it is possible to find a global solution of a time-varying non-convex optimization problem, we introduce the notion of *spurious (i.e., non-global) local trajectory* as a generalization to the notion of spurious local solution in nonconvex (time-invariant) optimization. We develop an ordinary differential equation (ODE) which, at limit, characterizes the spurious local solutions of the time-varying optimization problem. By building upon this connection, we prove that the absence of spurious local trajectory is closely related to the transient behavior of the proposed ODE. In particular, we show that: (1) if the problem is time-invariant, the spurious local trajectories are ubiquitous since any strict local minimum is a locally stable equilibrium point of the ODE, and (2) if the ODE is time-varying, the data variation may force all ODE trajectories initialized at arbitrary local minima at the initial time to gradually converge to the global solution trajectory. This implies that the natural data variation in the problem may automatically trigger escaping local minima over time.**

optimization. This observation naturally gives rise to the following question:

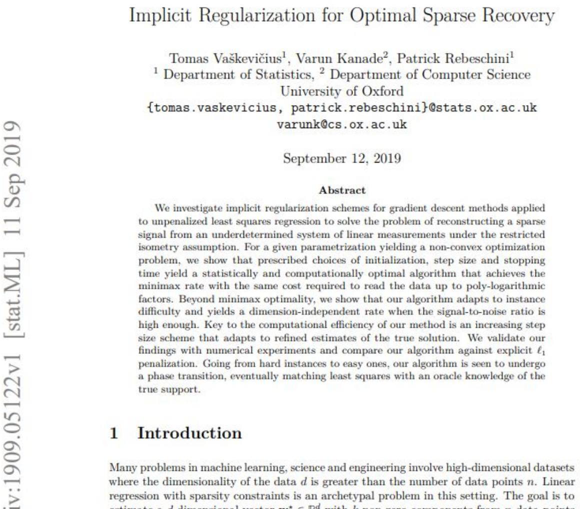
Would fast local-search algorithms escape spurious local minima in online nonconvex optimization, similar to their time-invariant counterparts?

In this paper, we attempt to address this question by developing a control-theoretic framework for analyzing the landscape of online and time-varying optimization. In particular, we demonstrate that even if a time-varying optimization may have undesired point-wise local minima at almost all times, the variation of its landscape over time would enable simple local-search algorithms to escape these spurious local minima. Inspired by this observation, this paper provides a new machinery to analyze the global landscape of online decision-making problems by drawing tools from optimization and control theory.

We consider a class of nonconvex and online optimization problems where the input data varies over time. First, we introduce the notion of *spurious local trajectory* as a generalization

Interactions between optimisation and regularisation

- Implicit regularisation of stochastic gradient *(based on older seminal ideas of Rosasco and others)*



Implicit Regularization for Optimal Sparse Recovery

Tomas Vaškevičius[1], Varun Kanade[2], Patrick Rebeschini[1]
[1] Department of Statistics, [2] Department of Computer Science
University of Oxford
{tomas.vaskevicius, patrick.rebeschini}@stats.ox.ac.uk
varunk@cs.ox.ac.uk

September 12, 2019

Abstract

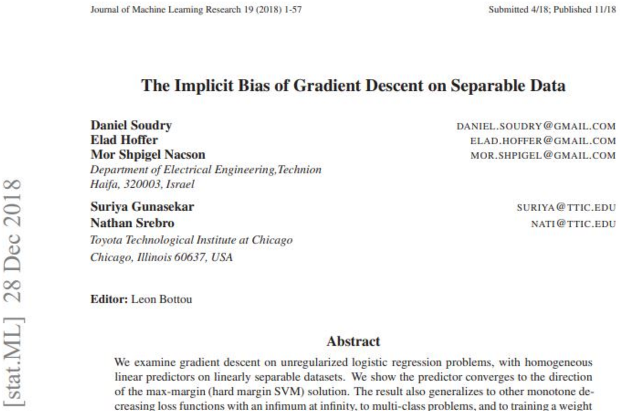
We investigate implicit regularization schemes for gradient descent methods applied to unpenalized least squares regression to solve the problem of reconstructing a sparse signal from an underdetermined system of linear measurements under the restricted isometry assumption. For a given parametrization yielding a non-convex optimization problem, we show that prescribed choices of initialization, step size and stopping time yield a statistically and computationally optimal algorithm that achieves the minimax rate with the same cost required to read the data up to poly-logarithmic factors. Beyond minimax optimality, we show that our algorithm adapts to instance difficulty and yields a dimension-independent rate when the signal-to-noise ratio is high enough. Key to the computational efficiency of our method is an increasing step size scheme that adapts to refined estimates of the true solution. We validate our findings with numerical experiments and compare our algorithm against explicit ℓ_1 penalization. Going from hard instances to easy ones, our algorithm is seen to undergo a phase transition, eventually matching least squares with an oracle knowledge of the true support.

1 Introduction

Many problems in machine learning, science and engineering involve high-dimensional datasets where the dimensionality of the data d is greater than the number of data points n. Linear regression with sparsity constraints is an archetypal problem in this setting. The goal is to

Interactions between optimisation and regularisation

- Implicit bias of stochastic gradient
- for linearly separable data, training a linear classifier with gradient descent on the logistic loss, or any loss with an exponential tail, implicitly leads to a max-margin linear classifier for the ℓ_2-norm



Journal of Machine Learning Research 19 (2018) 1-57 Submitted 4/18; Published 11/18

The Implicit Bias of Gradient Descent on Separable Data

Daniel Soudry DANIEL.SOUDRY@GMAIL.COM
Elad Hoffer ELAD.HOFFER@GMAIL.COM
Mor Shpigel Nacson MOR.SHPIGEL@GMAIL.COM
Department of Electrical Engineering,Technion
Haifa, 320003, Israel

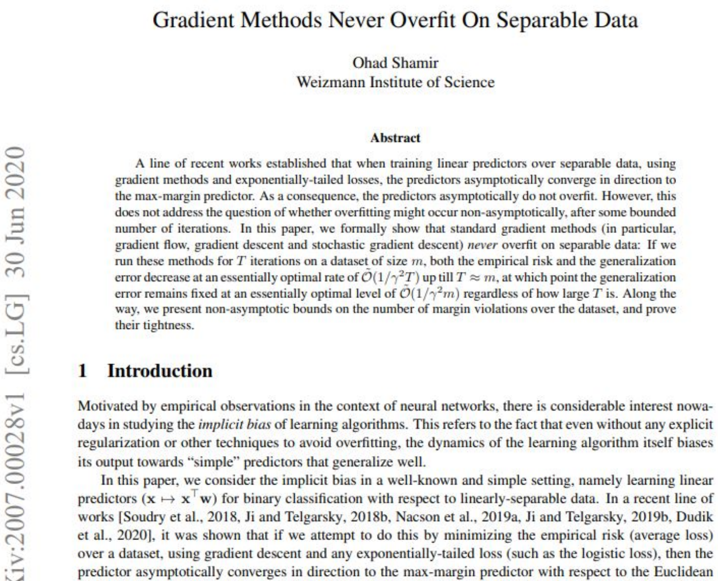
Suriya Gunasekar SURIYA@TTIC.EDU
Nathan Srebro NATI@TTIC.EDU
Toyota Technological Institute at Chicago
Chicago, Illinois 60637, USA

Editor: Leon Bottou

Abstract

We examine gradient descent on unregularized logistic regression problems, with homogeneous linear predictors on linearly separable datasets. We show the predictor converges to the direction of the max-margin (hard margin SVM) solution. The result also generalizes to other monotone decreasing loss functions with an infimum at infinity, to multi-class problems, and to training a weight

Interactions between optimisation and regularisation



Gradient Methods Never Overfit On Separable Data

Ohad Shamir
Weizmann Institute of Science

Abstract

A line of recent works established that when training linear predictors over separable data, using gradient methods and exponentially-tailed losses, the predictors asymptotically converge in direction to the max-margin predictor. As a consequence, the predictors asymptotically do not overfit. However, this does not address the question of whether overfitting might occur non-asymptotically, after some bounded number of iterations. In this paper, we formally show that standard gradient methods (in particular, gradient flow, gradient descent and stochastic gradient descent) *never* overfit on separable data: If we run these methods for T iterations on a dataset of size m, both the empirical risk and the generalization error decrease at an essentially optimal rate of $\tilde{\mathcal{O}}(1/\gamma^2 T)$ up till $T \approx m$, at which point the generalization error remains fixed at an essentially optimal level of $\tilde{\mathcal{O}}(1/\gamma^2 m)$ regardless of how large T is. Along the way, we present non-asymptotic bounds on the number of margin violations over the dataset, and prove their tightness.

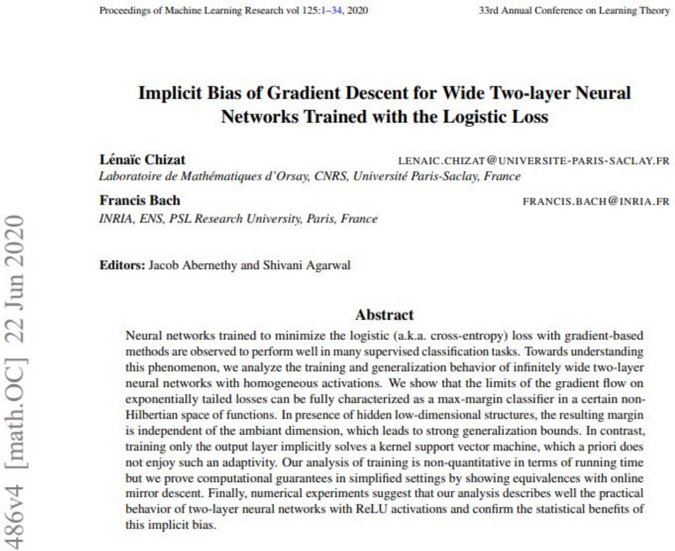
1 Introduction

Motivated by empirical observations in the context of neural networks, there is considerable interest nowadays in studying the *implicit bias* of learning algorithms. This refers to the fact that even without any explicit regularization or other techniques to avoid overfitting, the dynamics of the learning algorithm itself biases its output towards "simple" predictors that generalize well.

In this paper, we consider the implicit bias in a well-known and simple setting, namely learning linear predictors ($\mathbf{x} \mapsto \mathbf{x}^\top \mathbf{w}$) for binary classification with respect to linearly-separable data. In a recent line of works [Soudry et al., 2018, Ji and Telgarsky, 2018b, Nacson et al., 2019a, Ji and Telgarsky, 2019b, Dudik et al., 2020], it was shown that if we attempt to do this by minimizing the empirical risk (average loss) over a dataset, using gradient descent and any exponentially-tailed loss (such as the logistic loss), then the predictor asymptotically converges in direction to the max-margin predictor with respect to the Euclidean

Interactions between optimisation and regularisation

Implicit bias of stochastic gradient for two layer networks



Proceedings of Machine Learning Research vol 125:1–34, 2020 33rd Annual Conference on Learning Theory

Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks Trained with the Logistic Loss

Lénaïc Chizat LENAIC.CHIZAT@UNIVERSITE-PARIS-SACLAY.FR
Laboratoire de Mathématiques d'Orsay, CNRS, Université Paris-Saclay, France

Francis Bach FRANCIS.BACH@INRIA.FR
INRIA, ENS, PSL Research University, Paris, France

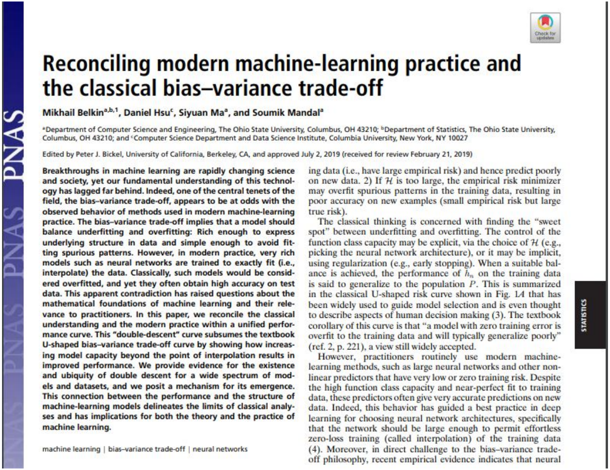
Editors: Jacob Abernethy and Shivani Agarwal

Abstract

Neural networks trained to minimize the logistic (a.k.a. cross-entropy) loss with gradient-based methods are observed to perform well in many supervised classification tasks. Towards understanding this phenomenon, we analyze the training and generalization behavior of infinitely wide two-layer neural networks with homogeneous activations. We show that the limits of the gradient flow on exponentially tailed losses can be fully characterized as a max-margin classifier in a certain non-Hilbertian space of functions. In presence of hidden low-dimensional structures, the resulting margin is independent of the ambiant dimension, which leads to strong generalization bounds. In contrast, training only the output layer implicitly solves a kernel support vector machine, which a priori does not enjoy such an adaptivity. Our analysis of training is non-quantitative in terms of running time but we prove computational guarantees in simplified settings by showing equivalences with online mirror descent. Finally, numerical experiments suggest that our analysis describes well the practical behavior of two-layer neural networks with ReLU activations and confirm the statistical benefits of this implicit bias.

Interpolation

- Overparametrisation works despite contradicting the intuition that "overfitting makes no sense"



Reconciling modern machine-learning practice and the classical bias–variance trade-off

Mikhail Belkin[a,b,1], Daniel Hsu[c], Siyuan Ma[a], and Soumik Mandal[a]

[a]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210; [b]Department of Statistics, The Ohio State University, Columbus, OH 43210; and [c]Computer Science Department and Data Science Institute, Columbia University, New York, NY 10027

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved July 2, 2019 (received for review February 21, 2019)

Breakthroughs in machine learning are rapidly changing science and society, yet our fundamental understanding of this technology has lagged far behind. Indeed, one of the central tenets of the field, the bias–variance trade-off, appears to be at odds with the observed behavior of methods used in modern machine-learning practice. The bias–variance trade-off implies that a model should balance underfitting and overfitting: Rich enough to express underlying structure in data and simple enough to avoid fitting spurious patterns. However, in modern practice, very rich models such as neural networks are trained to exactly fit (i.e., interpolate) the data. Classically, such models would be considered overfitted, and yet they often obtain high accuracy on test data. This apparent contradiction has raised questions about the mathematical foundations of machine learning and their relevance to practitioners. In this paper, we reconcile the classical understanding and the modern practice within a unified performance curve. This "double-descent" curve subsumes the textbook U-shaped bias–variance trade-off curve by showing how increasing model capacity beyond the point of interpolation results in improved performance. We provide evidence for the existence and ubiquity of double descent for a wide spectrum of models and datasets, and we posit a mechanism for its emergence. This connection between the performance and the structure of machine-learning models delineates the limits of classical analyses and has implications for both the theory and the practice of machine learning.

machine learning | bias–variance trade-off | neural networks

ing data (i.e., have large empirical risk) and hence predict poorly on new data. 2) If $\mathcal{H}$ is too large, the empirical risk minimizer may overfit spurious patterns in the training data, resulting in poor accuracy on new examples (small empirical risk but large true risk).

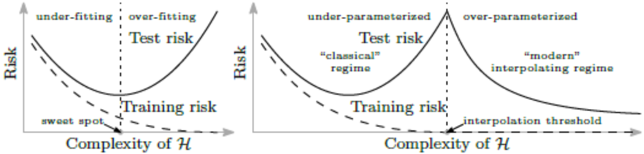
The classical thinking is concerned with finding the "sweet spot" between underfitting and overfitting. The control of the function class capacity may be explicit, via the choice of $\mathcal{H}$ (e.g., picking the neural network architecture), or it may be implicit, using regularization (e.g., early stopping). When a suitable balance is achieved, the performance of h_n on the training data is said to generalize to the population P. This is summarized in the classical U-shaped risk curve shown in Fig. 1*A* that has been widely used to guide model selection and is even thought to describe aspects of human decision making (3). The textbook corollary of this curve is that "a model with zero training error is overfit to the training data and will typically generalize poorly" (ref. 2, p. 221), a view still widely accepted.

However, practitioners routinely use modern machine-learning methods, such as large neural networks and other nonlinear predictors that have very low or zero training risk. Despite the high function class capacity and near-perfect fit to training data, these predictors often give very accurate predictions on new data. Indeed, this behavior has guided a best practice in deep learning for choosing neural network architectures, specifically that the network should be large enough to permit effortless zero-loss training (called interpolation) of the training data (4). Moreover, in direct challenge to the bias–variance trade-off philosophy, recent empirical evidence indicates that neural

STATISTICS

This topic is called "interpolation"

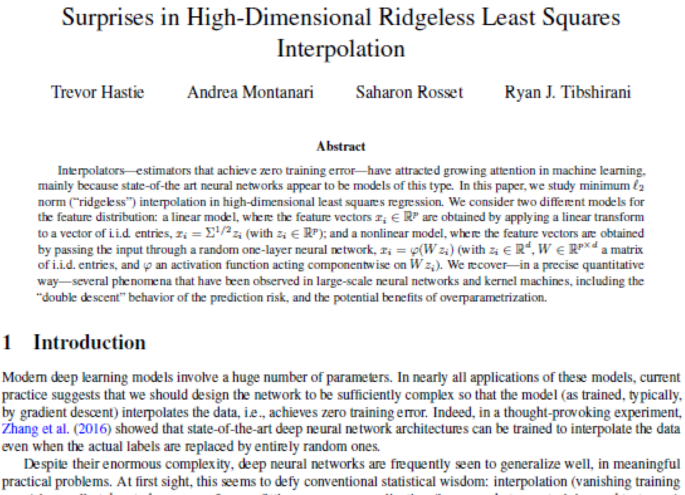
- Belkin et al. introduced the "double descent curve"



(a) U-shaped "bias-variance" risk curve

(b) "double descent" risk curve

Montanari et al. resolved this paradox ... for the linear model !
(Uses a lot of random matrix theory in the asymptotic regime)



Surprises in High-Dimensional Ridgeless Least Squares Interpolation

Trevor Hastie Andrea Montanari Saharon Rosset Ryan J. Tibshirani

Abstract

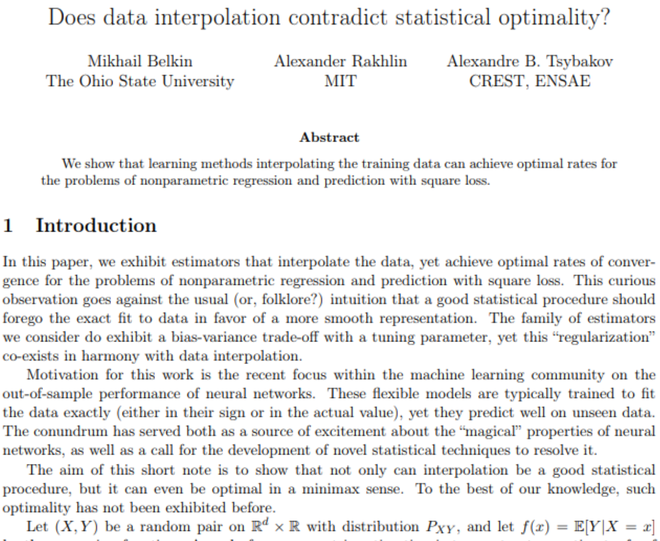
Interpolators—estimators that achieve zero training error—have attracted growing attention in machine learning, mainly because state-of-the art neural networks appear to be models of this type. In this paper, we study minimum ℓ_2 norm ("ridgeless") interpolation in high-dimensional least squares regression. We consider two different models for the feature distribution: a linear model, where the feature vectors $x_i \in \mathbb{R}^p$ are obtained by applying a linear transform to a vector of i.i.d. entries, $x_i = \Sigma^{1/2} z_i$ (with $z_i \in \mathbb{R}^p$); and a nonlinear model, where the feature vectors are obtained by passing the input through a random one-layer neural network, $x_i = \varphi(W z_i)$ (with $z_i \in \mathbb{R}^d$, $W \in \mathbb{R}^{p \times d}$ a matrix of i.i.d. entries, and φ an activation function acting componentwise on $W z_i$). We recover—in a precise quantitative way—several phenomena that have been observed in large-scale neural networks and kernel machines, including the "double descent" behavior of the prediction risk, and the potential benefits of overparametrization.

1 Introduction

Modern deep learning models involve a huge number of parameters. In nearly all applications of these models, current practice suggests that we should design the network to be sufficiently complex so that the model (as trained, typically, by gradient descent) interpolates the data, i.e., achieves zero training error. Indeed, in a thought-provoking experiment, Zhang et al. (2016) showed that state-of-the-art deep neural network architectures can be trained to interpolate the data even when the actual labels are replaced by entirely random ones.

Despite their enormous complexity, deep neural networks are frequently seen to generalize well, in meaningful practical problems. At first sight, this seems to defy conventional statistical wisdom: interpolation (vanishing training

Other models in non-parametric regression have been addressed by Belkin, Rakhlin and Tsybakov



Does data interpolation contradict statistical optimality?

Mikhail Belkin
The Ohio State University

Alexander Rakhlin
MIT

Alexandre B. Tsybakov
CREST, ENSAE

Abstract

We show that learning methods interpolating the training data can achieve optimal rates for the problems of nonparametric regression and prediction with square loss.

1 Introduction

In this paper, we exhibit estimators that interpolate the data, yet achieve optimal rates of convergence for the problems of nonparametric regression and prediction with square loss. This curious observation goes against the usual (or, folklore?) intuition that a good statistical procedure should forego the exact fit to data in favor of a more smooth representation. The family of estimators we consider do exhibit a bias-variance trade-off with a tuning parameter, yet this "regularization" co-exists in harmony with data interpolation.

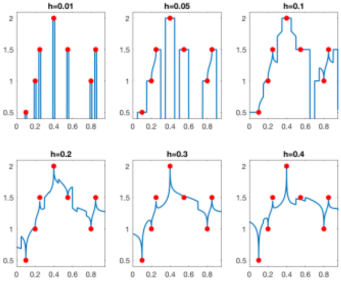
Motivation for this work is the recent focus within the machine learning community on the out-of-sample performance of neural networks. These flexible models are typically trained to fit the data exactly (either in their sign or in the actual value), yet they predict well on unseen data. The conundrum has served both as a source of excitement about the "magical" properties of neural networks, as well as a call for the development of novel statistical techniques to resolve it.

The aim of this short note is to show that not only can interpolation be a good statistical procedure, but it can even be optimal in a minimax sense. To the best of our knowledge, such optimality has not been exhibited before.

Let (X, Y) be a random pair on $\mathbb{R}^d \times \mathbb{R}$ with distribution P_{XY}, and let $f(x) = \mathbb{E}[Y|X = x]$

The Nadaraya-Watson estimator for a singular kernel K is defined as

$$f_n(x) = \begin{cases} Y_i & \text{if } x = X_i \text{ for some } i = 1, \ldots, n, \\ 0 & \text{if } \sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) = 0, \\ \frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)} & \text{otherwise.} \end{cases}$$



Interpolation with $K(u) = \|u\|^{-a}\mathbf{I}\{\|u\| \leq 1\}$, $a = 0.49$, and various values of h.

Figure: Singular Kernel estimators that interpolate !

A simple analysis of interpolation

Mathematical Model

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

$$Y_i = f^*(X_i) + \varepsilon_i \tag{2}$$

$i = 1, \ldots, n$, 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 is to estimate f^* based on the observation $Z_1, \ldots, Z_n$.

The estimation of f^* will be based on restricting the search to a subset $\mathcal{F}$ of functions of a Banach space $\mathcal{B}$.

In order to generalise, the estimator should 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

$$\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)). \tag{3}$$

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

$$\hat{f}^{ERM} \in \mathrm{argmin}_{f\in\mathcal{F}}\ \hat{R}_n(f). \tag{4}$$

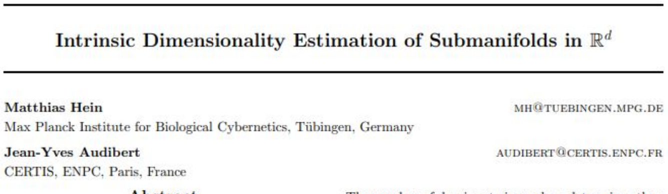
Assumption

The sample satisfies the following separation

$$\min_{i,i'=1}^{n} \|X_i - X_{i'}\|_2 \geq cn^{-1/\nu} \tag{5}$$

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 .



Intrinsic Dimensionality Estimation of Submanifolds in $\mathbb{R}^d$

Matthias Hein MH@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Jean-Yves Audibert AUDIBERT@CERTIS.ENPC.FR
CERTIS, ENPC, Paris, France

An handy result from Neuberger about the distance of the solution of a zero finding problem, i.e. consisting in solving

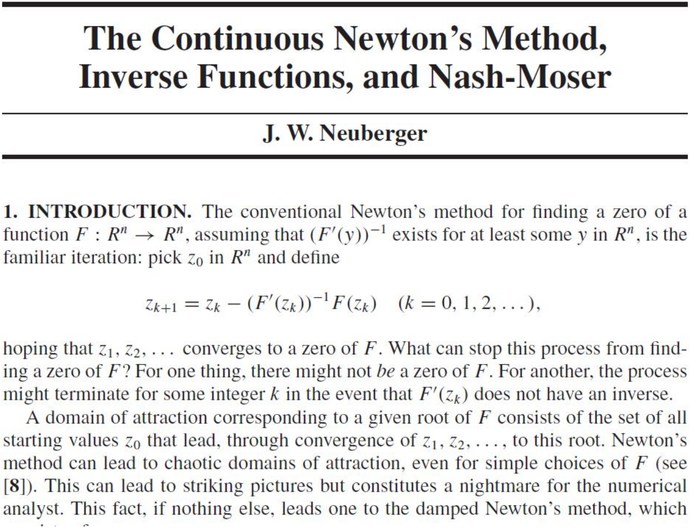
$$F(\hat{f}) = 0$$

to a given initial guess f^*

An handy result from Neuberger about the distance of the solution of a zero finding problem, i.e. consisting in solving

$F(\hat{f}) = 0$ (corresponding to first order optimality condition for the ERM !)

to a given initial guess f^*



The Continuous Newton's Method, Inverse Functions, and Nash-Moser

J. W. Neuberger

1. INTRODUCTION. The conventional Newton's method for finding a zero of a function $F : R^n \to R^n$, assuming that $(F'(y))^{-1}$ exists for at least some y in R^n, is the familiar iteration: pick z_0 in R^n and define

$$z_{k+1} = z_k - (F'(z_k))^{-1} F(z_k) \quad (k = 0, 1, 2, \dots),$$

hoping that $z_1, z_2, \dots$ converges to a zero of F. What can stop this process from finding a zero of F? For one thing, there might not *be* a zero of F. For another, the process might terminate for some integer k in the event that $F'(z_k)$ does not have an inverse.

A domain of attraction corresponding to a given root of F consists of the set of all starting values z_0 that lead, through convergence of $z_1, z_2, \dots$, to this root. Newton's method can lead to chaotic domains of attraction, even for simple choices of F (see [**8**]). This can lead to striking pictures but constitutes a nightmare for the numerical analyst. This fact, if nothing else, leads one to the damped Newton's method, which

Theorem (Neuberger's theorem)

Suppose that $\mathcal{B}$, $\mathcal{J}$, and $\mathcal{K}$ are three Banach spaces and that $\mathcal{B}$ is compactly embedded in $\mathcal{J}$.
Suppose that $F : \mathcal{B} \to \mathcal{K}$ is continuous with respect to the topologies of $\mathcal{J}$ and $\mathcal{K}$.
Suppose that $f \in \mathcal{B}$, that $r > 0$, and that for each g in $B_r(f)$, there is an h in $\bar{B}_r(0)$ such that

$$\lim_{t \to 0^+} \frac{1}{t}(F(g + th) - F(g)) = -F(f).$$

Then there is $\hat{f}$ in $\bar{B}_r(f)$ such that $F(\hat{f}) = 0$.

For $r > 0$ and u in $\mathcal{B}$, $B_r(u)$ and $\bar{B}_r(u)$ will denote the open and closed balls in $\mathcal{B}$, respectively, with center u and radius r.

We recall that $f \in \mathcal{F}$, and $d' \in \mathcal{B}$ such that $\mathcal{F} \subset \mathcal{B}$. Let us compute the directional derivative of $\hat{R}_n$

$$\begin{aligned} D\hat{R}_n(f) \cdot h' &= \lim_{t \to 0} \frac{\hat{R}_n(f + th') - \hat{R}_n(f)}{t} \\ &= \lim_{t \to 0} \frac{\frac{1}{n} \ \sum_{i=1}^n \ \ell(Y_i, f(X_i) + th'(X_i)) - \ell(Y_i, f(X_i))}{t} \end{aligned}$$

$$= \lim_{t \to 0} \frac{\frac{1}{n} \sum_{i=1}^n \ \partial_2 \ \ell(Y_i, f(X_i)) \ th'(X_i) + c \ \partial_2^2 \ \ell(Y_i, f(X_i) \ t^2 h'^2(X_i)}{t}$$

with $c \in [0, 1]$, and thus

$$D\hat{R}_n(f) \cdot h' = \frac{1}{n} \sum_{i=1}^n \partial_2 \ell(Y_i, f(X_i)) h'(X_i).$$

In the same spirit, we get

$$D^2\hat{R}_n(f)\cdot(h',h)=\frac{1}{n}\sum_{i=1}^{n}\partial_2^2\,\ell(Y_i,f(X_i))\;h'(X_i)h(X_i).$$

Based on these computations, Neuberger's theorem resorts to obtaining a bound on the norm of an appropriate solution d' to the following linear system

$$\frac{1}{n}\sum_{i=1}^{n} \partial_2^2\, \ell(Y_i, f(X_i))\, h'(X_i)h(X_i) = -\frac{1}{n}\, \sum_{i=1}^{n}\, \partial_2\ell(Y_i, f^*(X_i))h'(X_i)$$

for all $f \in B_r(f^*)$ and for all $h' \in \mathcal{B}$.

Let ψ denote the bump function

$$\psi(x) = \begin{cases} \exp\left(1 - \frac{1}{1-\|x\|_2^2}\right) & \text{if } \|x\|_2^2 \leq 1, \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

and let $\psi_\sigma = \psi(\cdot/\sigma)$.

We will decouple the problem and first solve it in a Sobolev space, and then approximate the solution by a deep neural network using the Guhring, Kutyniok and Petersen theorem . . .

Theorem

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. Take any $\sigma \leq cn^{-1/\nu}$ such that the ball in $\mathcal{B}$ centered at f^* with radius $6K_\epsilon \; n\|\psi_\sigma\|_{\mathcal{B}} \subset \mathcal{F}$.

Then, with high probability, there exists a mapping $\hat{f}^{ERM} : \mathbb{R}^d \mapsto \mathbb{R}$ which is a stationary point of the empirical risk minimisation problem and which lies at a distance at most

$$6K_\epsilon \; n\|\psi_\sigma\|_{\mathcal{B}}$$

from f^*.

Our main result is the following

Theorem

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 $\|f^*\|_{\mathcal{W}^{k,p}} \leq B$ for some $k \in \mathbb{N}$ and $p \in [1,+\infty]$. Assume that d, p, ν and n are such that $6K_\epsilon n^{1/2-d/(\nu p)}\|\psi\|_{\mathcal{W}^{k,p}} \leq B$.

Then, for any $s \in [0,1]$, with high probability, there exists a deep neural network $f_{\hat{W}}$: $\mathbb{R}^d \mapsto \mathbb{R}$ with

$$\left\|f_{\hat{W}} - f^*\right\|_{L^p(\mathcal{D})} \leq CK_\epsilon n^{1-d/(\nu p)}\|\psi\|_{\mathcal{W}^{k,p}(\mathcal{D})}$$

for some positive constant $C = C(d,p,k,B,s)$

Theorem

and with

(i) a number L of layers upper bounded by

$$L \leq c \; \log_2\left(\left(CK_\epsilon n^{1-d/(\nu p)}\|\psi\|_{\mathcal{W}^{k,p}(\mathcal{D})}\right)^{-k/(k-s)}\right)$$

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

$$d + \sum_{l=1}^{L} N_l \leq c \; \left(CK_\epsilon n^{1-d/(\nu p)}\|\psi\|_{\mathcal{W}^{k,p}(\mathcal{D})}\right)^{-d/(k-s)} \cdot \log_2\left(\left(CK_\epsilon n^{1-d/(\nu p)}\|\psi\|_{\mathcal{W}^{k,p}(\mathcal{D})}\right)^{-k/(k-s)}\right)$$

which approximately solves the empirical risk minimisation problem's first order optimality conditions over the Sobolev class $\mathcal{W}^{k,p}(\mathcal{D})$ with $\mathcal{W}^{s,p}(\mathcal{D})$-distance at most $K_\epsilon n^{1-d(\nu p)}\|\psi\|_{\mathcal{W}^{k,p}(\mathcal{D})}$ to its solution set.

Sketch of the proof

Notice that for all $f \in B_s(f_{W^*})$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ell}{\partial_2}(Y_i, f(X_i))\ h'(X_i) = -\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))\ \ h'(X_i),$$

and that

$$\frac{1}{n}\sum_{i=1}^{n}\ \frac{\partial^2 \ell}{\partial_2^2}(Y_i, f(X_i))\ h'(X_i)h(X_i) = \frac{1}{n}\sum_{i=1}^{n} h'(X_i)h(X_i).$$

Then, using the fact that ℓ is the ℓ_2^2 loss, Neuberger's condition reads

$$\frac{1}{n}\sum_{i=1}^{n} h'(X_i)h(X_i) = \frac{1}{n}\sum_{i=1}^{n}\ h'(X_i)\left(Y_i - f_{W^*}(X_i)\right).$$

One possible solution can be obtained by setting

$$h(X_i) = Y_i - f_{W^*}(X_i) = \varepsilon_i$$

$i = 1, \ldots, n$, i.e. using a noise interpolating solution (*reminiscent of the work by Belkin, Rakhlin and Tsybakov on singular kernels previously mentioned*).

One simple option is to take

$$h(x) = \sum_{i=1}^{n} \epsilon_i \ \psi \left(\frac{x - X_i}{\sigma} \right)$$

where $\psi : \mathbb{R}^p \to \mathbb{R}$ is a kernel function and $\sigma > 0$ is a bandwidth.

Let

$$\psi_\sigma = \psi\left(\cdot/\sigma\right).$$

Now, observe that, based on Assumption 1, the functions $\psi((x - X_i)/\sigma)$, and their successive derivatives up to k, $i = 1, \ldots, n$, have disjoint supports for with probability larger than or equal to $1 - \delta$ as long as $\sigma \leq cn^{-1/\nu}$.
We thus obtain that

$$\|h\|_{\mathcal{B}} = \|\epsilon\|_1 \ \|\psi_\sigma\|_{\mathcal{B}}$$

Moreover, as is well known for subGaussian vectors, the norm is controlled by

$$\|\epsilon\|_2 \leq 6K_\epsilon n.$$

with probability at least $1 - \exp(-n)$, combining the conclusion of Theorem 4 follows from Neuberger's Theorem 3.

The proof for the deep neural network case is completed by using the approximation result of Guhring, Kutyniok and Petersen.

The number of layers may have to increase logarithmically with the number of samples

The total number of parameters blows up polynomially in the number of samples and exponentially in the dimension of the problem

Conclusion and perspectives

- This simple exercice in using quantitative zero finding theorems such as Neubergers shows that we can easily prove results that do not blow up with the number of layers with interpolating networks
- We can easily study local minimisers as well using the same technique

We would need to explore approximation theory in unusual/non standard directions:

- improve the Guhring, Kutyniok and Petersen theorem by introducing the constraint that the network be a flat minimiser
    - This would explain that Stochastic Gradient methods can find the correct approximation with large probability (?)

Some papers:

A finite sample analysis of the double descent phenomenon for ridge function estimation, Emmanuel Caron and Stephane Chretien: arXiv preprint arXiv:2007.12882

The double descent phenomenon for Deep Nets, Emmanuel Caron and Stephane Chretien (soon on Arxiv)