The Abdus Salam International Centre for Theoretical Physics

United Nations Educational, Scientific and Cultural Organization

IAEA International Atomic Energy Agency

Youth in High-Dimensions: Recent Progress in Machine Learning, High-Dimensional Statistics and Inference | (SMR 3719)

27 Jun 2022 - 30 Jun 2022
ICTP, Trieste, Italy

P01 - ALIDOU Abdou Majeed

Relative Entropy, Compression Algorithms, and Applications

P02 - ARPINO Gabriel

Computational Hardness of Sparse Linear Regression

P03 - BARBIER Damien

Compressed sensing with l0-norm: statistical mechanics analysis and recovering algorithm

P04 - BODIN Antoine Philippe Michel

Model, sample, and epoch-wise descents: exact solution of gradient flow in the random feature model

P05 - CAGNETTA Francesco

Exploring the compositional structure of images with deep convolutional kernel methods

P06 - CAMILLI Francesco

Bayesian inference in a mismatched setting: a spin-glass model with Mattis interaction

P07 - CARBONE Ginevra

Resilience of Bayesian Layer-Wise Explanations under Adversarial Attacks

P08 - CARETTI Federico

Localisation of Mitosis Transition

P09 - CHAPARRO AMARO Roberto Oscar

Comparison between NN and RF techniques for hybrid en- ergy γ-ray reconstruction in simulation data of the 55 Imaging Air- Cherenkov Telescopes HAWC's Eye at High Altitude

P10 - CUTURELLO Francesca

Determining the impact of remote homology detection on protein structure predictions by Deep Learning models

P11 - DE MORAIS GOULART José Henrique

A random matrix perspective on the spiked rank-one tensor model

P12 - DONHAUSER Konstantin

Fast rates for noisy interpolation require rethinking the effects of inductive bias

P13 - ERBA Vittorio

Optimal denoising of rotationally invariant rectangular matrices

P14 - GERACE Federica

Gaussian Universality of Linear Classiers with Random Labels in High-Dimension

P15 - IACOANGELI Alfredo

A knowledge-based machine learning approach to gene prioritisation in amyotrophic lateral sclerosis

P16 - INGROSSO Alessandro

Data-driven emergence of convolutional structure in neural networks

P17 - KIZILDAG Can Eren

A Curious Case of Symmetric Binary Perceptron Model: Algorithms and Barriers

P18 - LIANG Shansuo

Industrial Problems

P19 - LOUVARIS Michail

Universality of the least singular value and singular vector delocalization for levy non-symmetric matrices

P20 - LUCIBELLO Carlo

Deep Learning via Message Passing

P21 - MACOCCO Iuri

Topological analysis of genomics sequences shows that evolutive pressure acts on a low-dimensional manifold

P22 - MAILLARD Antoine

Perturbative construction of mean-field equations in extensive-rank matrix factorization and denoising

P23 - MAINALI Nischal

Neural Receptive field as Gaussian Process

P24 - MISIAKIEWICZ Jakub Theodor

The merged-staircase property: a necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks

P25 - PAPANIKOLAOU Nikolaos

Consensus from group interactions: An adaptive voter model on hypergraphs

P26 - PETRINI Leonardo

How learning features can lead to over-fitting in neural nets

P27 - REFINETTI Maria

The dynamics of representation learning in shallow, non-linear autoencoders

P28 - ROMANOV Elad

On the Role of Channel Capacity in Learning Gaussian Mixture Models

P29 - SAGLIETTI Luca

Inducing bias is simpler than you think

P30 - SALGADO CORRADO Olaf Ariel

Who initiated the division of the Karate Club? Detecting communities' most representative nodes through graphical models

P31 - SARAO MANNELLI Stefano

Maslow's Hammer for Catastrophic Forgetting: Node Re-Use vs Node Activation

P32 - SCHÖNSBERG Francesca

May subjective sensory experiences be the result of a neural reservoir?

P33 - SCLOCCHI Antonio

High-dimensional optimization under nonconvex excluded volume constraints

P34 - STEPHAN Théo Ludovic

Phase diagram of Stochastic Gradient Descent in high-dimensional two-layer neural networks

P35 - SZEKELY Eszter

Data-driven separation of two-layer neural networks and random features

P36 - TAKAHASHI Takashi

Sharp Asymptotics of Self-training with Linear Classifier

P37 - UDOMBOSO Godwin Christopher

Image-Based Algorithm for Vehicle Class Prediction in the University Of Ibadan, Nigeria

P38 - WU Yi

High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation

P39 - XIE Rongrong

A random energy approach to deep learning

P40 - ZHANG Alex Chen Yi

No poster

P41 - ZHU Yizhe

Non-backtracking spectral clustering in sparse hypergraphs

P01

Relative Entropy, Compression Algorithms, and Applications

Abdou Majeed ALIDOU

May 12, 2022

This work is concerned with the important problem of extracting relevant information from sequences of data. It explains key concepts and results from information theory such as entropy, conditional entropy, mutual information, relative entropy, and Gibbs' inequality. We present some lossless compression algorithms including Symbol Codes (Huffman coding), Arithmetic Codes, and the popular Lempel-Ziv algorithm, as well as some important results such as the Asymptotic Equipartition Principle and Shannon Source Coding Theorem.

The powerful concept of relative entropy is then interpreted in terms of compression algorithms. Indeed, the relative entropy between two distributions X and Y can be estimated by comparing the performance of some compression algorithms on a sequence A+b, with their performance on a sequence B+b, where A and B are long texts taken respectively from the distributions X and Y, and b is a small text taken from the distribution Y. That technique can be used to distinguish two sequences of data from different sources.

Finally, we discuss some of the key applications of such algorithms to the problem of language recognition and authorship attribution.

P02

Computational Hardness of Sparse Linear Regression

When it comes to statistical inference in high-dimensions, the sparsity assumption has become the scientist's saving grace. It formalizes the a priori belief that only a few parameters among many are significant for the inference task at hand. This assumption has become ubiquitous in modern technological developments, from robot localization to MRI brain scanning, and has provided a solution to the otherwise uncompromising problem of inferring more parameters than we have data points. There is growing evidence, however, that such gains in information imply hardness of computation. Consider the problem of recovering a hidden binary k-sparse p-dimensional vector x from n noisy linear observations $y = \lambda Ax+z$ where $A_{i,j} \sim N(0,1)$, $z_i \sim N(0,1)$, $\lambda > 0$, which we call Sparse High-Dimensional Linear Regression (SHLR). A detection variant of the problem consists of distinguishing the pair (A,y) generated from the SHLR model from a null model where (A, y) are drawn independently from Gaussian distributions, which we call Sparse High-Dimensional Linear Detection (SHLD). Indeed, in the contexts of SHLR and SHLD, we bring novel rigorous evidence towards the existence of a statistical price to pay for computational efficiency through the study of low-degree polynomials. Our results indicate the presence of an algorithmic obstruction to solving SHLD and SHLR efficiently within a certain achievable regime.

P03

Compressed sensing with l0-norm: statistical mechanics analysis and recovering algorithm

Compressed (or compressive) sensing (CS) is a framework detailing the reconstruction of a sparse N-dimensional vector encoding a signal from a lower-dimensional feature vector, whose components represent a set of measurements on the signal. With this work we propose to focus on CS with l0-norm minimization, meaning that we will study the solutions for the measurement protocol (i.e. N-dimensional vectors giving back the correct feature vector) with maximum sparsity. In particular we frame this set-up as statistical mechanics problem, defining a cost function with tuning parameters which take into account both the measurement process and the l0-norm minimization. We numerically show that the solutions to this problem form clusters with a 1-step replica symmetry breaking structure and highlight the presence of two regimes depending on the signal sparsity and the compression rate. We also propose an algorithm based on survey propagation that achieve signal recovery for sufficiently low compression rate. Numerical analyzes show that this algorithm offers better performance than the well-known l1-norm message passing algorithm.

P04

Model, sample, and epoch-wise descents: exact solution of gradient flow in the random feature model

Recent evidence has shown the existence of a so-called double-descent and even triple-descent behavior for the generalization error of deep-learning models. This important phenomenon commonly appears in implemented neural network architectures, and also seems to emerge in epoch-wise curves during the training process. A recent line of research has highlighted that random matrix tools can be used to obtain precise analytical asymptotics of the generalization (and training) errors of the random feature model. In this contribution, we analyze the whole temporal behavior of the generalization and training errors under gradient flow for the random feature model. We show that in the asymptotic limit of large system size the full time-evolution path of both errors can be calculated analytically. This allows us to observe how the double and triple descents develop over time, if and when early stopping is an option, and also observe time-wise descent structures. Our techniques are based on Cauchy complex integral representations of the errors together with recent random matrix methods based on linear pencils.

P05

Locality defeats the curse of dimensionality in convolutional teacher-student scenarios

Alessandro Favero *
Institute of Physics
École Polytechnique Fédérale de Lausanne
alessandro.favero@epfl.ch

Francesco Cagnetta *
Institute of Physics
École Polytechnique Fédérale de Lausanne
francesco.cagnetta@epfl.ch

Matthieu Wyart
Institute of Physics
École Polytechnique Fédérale de Lausanne
matthieu.wyart@epfl.ch

Abstract

Convolutional neural networks perform a local and translationally-invariant treatment of the data: quantifying which of these two aspects is central to their success remains a challenge. We study this problem within a teacher-student framework for kernel regression, using 'convolutional' kernels inspired by the neural tangent kernel of simple convolutional architectures of given filter size. Using heuristic methods from physics, we find in the ridgeless case that locality is key in determining the learning curve exponent β (that relates the test error $\epsilon_t \sim P^{-\beta}$ to the size of the training set P), whereas translational invariance is not. In particular, if the filter size of the teacher t is smaller than that of the student s, β is a function of s only and does not depend on the input dimension. We confirm our predictions on β empirically. We conclude by proving, under a natural universality assumption, that performing kernel regression with a ridge that decreases with the size of the training set leads to similar learning curve exponents to those we obtain in the ridgeless case.

1 Introduction

Deep Convolutional Neural Networks (CNNs) are widely recognised as the engine of the latest successes of deep learning methods, yet such a success is surprising. Indeed, any supervised learning model suffers *in principle* from the curse of dimensionality: under minimal assumptions on the function to be learnt, achieving a fixed target generalisation error ϵ requires a number of training samples P which grows exponentially with the dimensionality d of input data [1], i.e. $\epsilon(P) \sim P^{-1/d}$. Nonetheless, empirical evidence shows that the curse of dimensionality is beaten *in practice* [2, 3, 4], with

$$\epsilon(P) \sim P^{-\beta}, \quad \beta \gg 1/d. \tag{1}$$

CNNs, in particular, achieve excellent performances on high-dimensional tasks such as image classification on ImageNet with state-of-the-art architectures, for which $\beta \approx [0.3, 0.5]$ [2]. Natural data must then possess additional structures that make them learnable. A classical idea [5] ascribes the success of recognition systems to the compositionality of data, i.e. the fact that objects are made of features, themselves made of sub-features [6, 7, 8]. In this view, the locality of CNNs plays a key role for their performance, as supported by empirical observations [9]. Yet, there is no clear

*Equal contribution.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

analytical understanding of the relationship between the compositionality of the data and learning curves.

In order to study this relationship quantitively, we introduce a teacher-student framework for kernel regression, where the function to be learnt takes one of the following two forms:

$$f^{LC}(\boldsymbol{x}) = \sum_{i \in \mathcal{P}} g_i(\boldsymbol{x}_i), \quad f^{CN}(\boldsymbol{x}) = \sum_{i \in \mathcal{P}} g(\boldsymbol{x}_i). \tag{2}$$

Here, $\boldsymbol{x}$ is a d-dimensional input and $\boldsymbol{x}_i$ denotes the i-th t-dimensional patch of $\boldsymbol{x}$, $\boldsymbol{x}_i = (x_i, \dots, x_{i+t-1})$. i ranges in a subset $\mathcal{P}$ of $\{1, \dots, d\}$. The g_i's and g are random functions of t variables whose smoothness is controlled by some exponent α_t. Such functions model the local nature of certain datasets and can be generated, for example, by randomly-initialised one-hidden-layer neural networks: f^{LC} corresponds to a *locally connected* network (LCN) [10, 11], in which the input is split into lower-dimensional patches before being processed, whereas a network enforcing invariance with respect to shifts of the input patches via weight sharing can be described by f^{CN}. In such cases t would be the filter size of the network. Our goal is to compute the asymptotic decay of the error of a student kernel performing regression on such data, and to relate the corresponding exponent β to the locality of the target function. The student kernel corresponds to a prior on the true function of the form described by Eq. (2), except that the filter size s and its prior α_s on the smoothness of the g functions can differ from those of the target function. Such students include overparametrised one-hidden-layer neural networks operating in the *lazy training regime* [12, 13, 14, 15, 16].

1.1 Our contributions

We consider a teacher-student framework for kernel regression, where the target function has one of the forms in Eq. (2), where the g_i's and g are Gaussian random fields of given covariance. Target functions are characterised by the dimensionality t of the g functions—the *filter size*—and a smoothness exponent α_t, such that $\alpha_t > 2n$ implies that typical target functions are at least n times differentiable. Kernel regression is performed by *local* or *convolutional* student kernels, having filter size s and a prior on the target smoothness characterised by another exponent $\alpha_s > 0$. Our main contributions follow:

- We use recent results based on the replica method of statistical physics on the generalisation error of kernel methods [17, 18, 19] to estimate the exponent β. We find that $\beta = \alpha_t/s$ if $t \leq s$ and $\alpha_t \leq 2(\alpha_s + s)$. This approach is non-rigorous, but it can be proven if data are sampled on a lattice [4] and corresponds to a provable lower bound on the error when teacher and student are equal [20].
- In particular, we find the same exponent for students with a prior on the shift invariance of the target function and students without this prior, implying that the curse of dimensionality is beaten due to locality and not shift invariance.
- We confirm systematically our predictions by performing kernel ridgeless regression numerically for various t, s and embedding dimension d.
- We use the recent framework of [21] and a natural Gaussian universality assumption to prove a rigorous estimate of β in the case where the ridge decreases with the size of the training set. The estimate of β depends again on s and not on d, demonstrating that the curse of dimensionality can indeed be beaten by using local filters on such compositional data.

1.2 Related work

Several recent works study the role of the compositional structure of data [6, 22, 23]. When such structure is hierarchical, deep convolutional networks can be much more expressive than shallow ones [6, 24, 7]. Concerning training, [25] shows that both convolutional and locally-connected networks can achieve a target generalisation error in polynomial time, whereas fully-connected networks cannot, for a class of functions which depend only on s consecutive bits of the d-dimensional input, with $s = \mathcal{O}(\log d)$. In [8] the effects of the architecture's locality are studied from a kernel perspective, using a class of deep convolutional kernels introduced in [26, 27] and characterising their Reproducing Kernel Hilbert Space (RKHS). In general, belonging to the RKHS ensures favourable bounds on performance and, for isotropic kernels, is a constraint on the function smoothness that becomes

stringent in large d. For local functions, the corresponding constraint on smoothness is governed by the filter size s and not d [8]. Lastly, a recent work shows that weight sharing, in the absence of locality, leads to a mild improvement of the generalisation error of shift-invariant kernels [28].

By contrast, our work focuses on computing non-trivial training curve exponents in a setup where the locality and shift-invariance priors of the kernel can differ from those of the class of functions being learnt. In our setup, the latter are in general not in the RKHS of the kernel[2]. Technically, our result that the size of the student filter s controls the learning curve (and not that of the teacher t) relates to the fact that kernels are not able to detect data anisotropy (the fact that the function depends only on a subset of the coordinates) in worst-case settings [30] nor in the typical case for Gaussian fields [31].

2 Setup

Kernel ridge regression Kernel ridge regression is a method to learn a target function $f^* : \mathbb{R}^d \to \mathbb{R}$ from P observations $\{(\boldsymbol{x}^\mu, f^*(\boldsymbol{x}^\mu))\}_{\mu=1}^P$, where the inputs $\boldsymbol{x}^\mu$ are i.i.d. random variables distributed according to a certain measure $p\left(d^d x\right)$ on $\mathbb{R}^d$. Let K be a positive-definite kernel and $\mathcal{H}$ the corresponding Reproducing Kernel Hilbert Space (RKHS). The kernel ridge regression estimator f of the target function f^* is defined as

$$f = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \left\{ \frac{1}{P} \sum_{\mu=1}^{P} \left(f(\boldsymbol{x}^\mu) - f^*(\boldsymbol{x}^\mu) \right)^2 + \lambda \left\| f \right\|_{\mathcal{H}}^2 \right\}, \tag{3}$$

where $\| \cdot \|_{\mathcal{H}}$ denotes the RKHS norm and λ is the ridge parameter. The limit $\lambda \to 0^+$ is known as the ridgeless case and corresponds to the solution with minimum RKHS norm that interpolates the P observations. Eq. (3) is a convex optimisation problem, having the unique solution

$$f(\boldsymbol{x}) = \frac{1}{P} \sum_{\mu,\nu=1}^{P} K(\boldsymbol{x}, \boldsymbol{x}^\mu) \left(\left(\frac{1}{P} \mathbb{K}_P + \lambda \mathbb{I}_P \right)^{-1} \right)_{\mu,\nu} f^*(\boldsymbol{x}^\nu), \tag{4}$$

where $\mathbb{K}_P$ is the *Gram matrix* defined as $(\mathbb{K}_P)_{\mu\nu} = K(\boldsymbol{x}^\mu, \boldsymbol{x}^\nu)$, and $\mathbb{I}_P$ denotes the P-dimensional identity matrix. Our goal is to compute the generalisation error, which we define as the expectation of the mean squared error over the data distribution $p\left(d^d x\right)$, averaged over an ensemble of target functions f^*, i.e

$$\epsilon(P) = \mathbb{E}_{\boldsymbol{x}, f^*} \left[\left(f(\boldsymbol{x}) - f^*(\boldsymbol{x}) \right)^2 \right]. \tag{5}$$

The error ϵ depends on the number of samples P through the predictor of Eq. (4) and we refer to the graph of $\epsilon(P)$ as *learning curve*.

Statistical mechanics of generalisation in kernel regression The theoretical understanding of generalisation is still an open problem. A few recent works [17, 21, 18] relate the generalisation error ϵ to the decomposition of the target function in the eigenbasis of the kernel. A positive-definite kernel K can indeed be written, by Mercer's theorem, in terms of its eigenvalues $\{\lambda_\rho\}$ and eigenfunctions $\{\phi_\rho\}$:

$$K(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\rho=1}^{\infty} \lambda_\rho \phi_\rho(\boldsymbol{x}) \overline{\phi_\rho(\boldsymbol{y})}, \quad \int p\left(d^d y\right) K(\boldsymbol{x}, \boldsymbol{y}) \phi_\rho(\boldsymbol{y}) = \lambda_\rho \phi_\rho(\boldsymbol{x}). \tag{6}$$

In [17, 21, 18] it is shown that, when the target function can be written in terms of the kernel eigenbasis,

$$f^*(\boldsymbol{x}) = \sum_{\rho} c_\rho \phi_\rho(\boldsymbol{x}), \tag{7}$$

the error ϵ can also be cast as a sum of modal contributions, $\epsilon = \sum_\rho \epsilon_\rho$. The details of the general formulation are summarised in Appendix A. Here we present an intuitive limiting case, obtained in the ridgeless limit $\lambda \to 0^+$, when $\lambda_\rho \sim \rho^{-a}$ for large ρ, and $\mathbb{E}[|c_\rho|^2] \sim \rho^{-b}$ with $2a > b - 1$, that is

$$\epsilon(P) \sim \sum_{\rho > P} \mathbb{E}[|c_\rho|^2] \equiv \mathcal{B}(P), \tag{8}$$

[2] A Gaussian field of covariance K is never in the RKHS of the kernel K, see e.g. [29].

with $\sim$ denoting asymptotic equivalence for large P. Eq. (8) indicates that, given P examples, the generalisation error can be estimated as the tail sum of the power in the target function past the first P modes of the kernel, which we denote as $\mathcal{B}(P)$. Although the general modal decomposition cannot be proven rigorously in the ridgeless limit [21, 19], additional results are available when the target functions are Gaussian random fields with covariance specified by a teacher kernel:

- Eq. (8) can be proven rigorously [4] if teacher and student are isotropic kernels and the input points $\boldsymbol{x}^\mu$ are sampled on the lattice $\mathbb{Z}^d$, i.e. all the elements of each input sequence are integer multiples of an arbitrary unit;
- If teacher and student coincide then $\mathbb{E}[|c_\rho|^2]$ equals the ρ-th eigenvalue λ_ρ and (see e.g. [20]) $\epsilon(P) \geq \mathcal{B}(P)$, i.e. the estimate of Eq. (8) is a lower bound.

3 Kernels for local and convolutional teacher-student scenarios

In this section we introduce convolutional and local kernels that will be used as teachers, i.e. to generate different ensembles of target functions f^* with controlled smoothness and degree of locality, and as student kernels. We motivate our choice by considering one-hidden-layer neural networks with simple local and convolutional architectures. Because of the relationship between our kernels and the Neural Tangent Kernel [12] of the aforementioned architectures, our framework encompasses regression with simple overparametrised networks trained in the lazy regime [16]. For the sake of clarity we limit the discussion to inputs which are sequences in $\mathbb{R}^d$, i.e. $\boldsymbol{x} = (x_1, \dots, x_d)$. Extension to higher-order tensorial inputs such as images $\boldsymbol{X} \in \mathbb{R}^{d\times d}$ is straightforward. To avoid dealing with the boundaries of the sequence we identify x_{i+d} with x_i for all $i = 1, \dots, d$.

Definition 3.1 (one-hidden-layer CNN)**.** *A one-hidden-layer convolutional network with H hidden neurons and average pooling is defined as follows,*

$$f^{CNN}(\boldsymbol{x}) = \frac{1}{\sqrt{H}} \sum_{h=1}^{H} a_h \frac{1}{|\mathcal{P}|} \sum_{i\in\mathcal{P}} \sigma(\boldsymbol{w}_h \cdot \boldsymbol{x}_i), \tag{9}$$

where $\boldsymbol{x} \in \mathbb{R}^d$ is the input, H is the width, σ a nonlinear activation function, $\mathcal{P} \subseteq \{1, \dots, d\}$ is a set of patch indices and $|\mathcal{P}|$ its cardinality. For all $i \in \mathcal{P}$, $\boldsymbol{x}_i$ is an s-dimensional patch of $\boldsymbol{x}$. For all $h = 1, \dots, H$, $\boldsymbol{w}_h \in \mathbb{R}^s$ is a filter with filter size s, $a_h \in \mathbb{R}$ is a scalar weight. The dot $\cdot$ denotes the standard Euclidean scalar product.

In the network defined above, a d-dimensional input sequence $\boldsymbol{x}$ is first mapped to s-dimensional *patches* $\boldsymbol{x}_i$, which are ordered subsequences of the input. Comparing each patch to a filter $\boldsymbol{w}_h$ and applying the activation function σ leads to a $|\mathcal{P}|$-dimensional hidden representation which is equivariant for shifts of the input. The summation over the patch index i promotes this equivariance to full invariance, leading to a model which is both local and shift-invariant as f^{CN} in Eq. (2). A model which is only local, as f^{LC} in Eq. (2), can be obtained by lifting the constraint of weight-sharing, which forces, for each $h = 1, \dots, H$, the same filter $\boldsymbol{w}_h$ to apply to all patches $\boldsymbol{x}_i$.

Definition 3.2 (one-hidden-layer LCN)**.** *In the notation of Definition 3.1, a one-hidden-layer locally-connected network with H hidden neurons is defined as follows,*

$$f^{LCN}(\boldsymbol{x}) = \frac{1}{\sqrt{H}} \sum_{h=1}^{H} \frac{1}{\sqrt{|\mathcal{P}|}} \sum_{i\in\mathcal{P}} a_{h,i} \sigma(\boldsymbol{w}_{h,i} \cdot \boldsymbol{x}_i), \tag{10}$$

For all $i \in \mathcal{P}$ and $h = 1, \dots, H$: $\boldsymbol{x}_i$ is an s-dimensional patch of $\boldsymbol{x}$, $\boldsymbol{w}_{h,i} \in \mathbb{R}^s$ is a filter with filter size s, $a_{h,i} \in \mathbb{R}$ is a scalar weight.

Notice that the definition above reduces to that of a fully-connected network when the filter size is set to the input dimension, $s = d$, and $\mathcal{P} = \{1\}$. With the target functions taking one of the two forms in Eq. (2), our framework contains the case where the observations are generated by neural networks such as (3.1) and (3.2). Let us now introduce the neural tangent kernels of such architectures.

Definition 3.3 (Neural Tangent Kernel)**.** *Given a neural network function $f(\boldsymbol{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N)$ denotes the complete set of parameters and N the total number of parameters, the Neural Tangent Kernel (NTK) is defined as [12]*

$$\Theta_N(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) = \sum_{n=1}^{N} \partial_{\theta_n} f(\boldsymbol{x}, \boldsymbol{\theta}) \partial_{\theta_n} f(\boldsymbol{y}, \boldsymbol{\theta}), \tag{11}$$

where ∂_{θ_n} denotes partial derivation w.r.t. the n-th parameter θ_n.

For one-hidden-layer networks with random, $\mathcal{O}(1)$-variance Gaussian initialisation of all the weights, and normalisation by $\sqrt{H}$ as in (3.1) and (3.2), the NTK converges to a deterministic limit $\Theta(\boldsymbol{x}, \boldsymbol{y})$ as $N \propto H \to \infty$ [12]. Furthermore, training $f(\boldsymbol{x}, \boldsymbol{\theta}) - f(\boldsymbol{x}, \boldsymbol{\theta}_0)$, with $\boldsymbol{\theta}_0$ denoting the network parameters at initialisation, under gradient descent on the mean squared error is equivalent to performing ridgeless regression with kernel $\Theta(\boldsymbol{x}, \boldsymbol{y})$ [12]. The following lemmas relate the NTK of convolutional and local architectures acting on d-dimensional inputs to that of a fully-connected architecture acting on s-dimensional inputs. Both lemmas are proved in Appendix B.

Lemma 3.1. *Call Θ^{FC} the NTK of a fully-connected network function acting on s-dimensional inputs and Θ^{CN} the NTK of a convolutional network function (3.1) with filter size s acting on d-dimensional inputs. Then*

$$\Theta^{CN}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{|\mathcal{P}|^2} \sum_{i,j \in \mathcal{P}} \Theta^{FC}(\boldsymbol{x}_i, \boldsymbol{y}_j) \tag{12}$$

As the functions in Eq. (2), Θ^{CN} is written as a combination of lower-dimensional constituent kernels Θ^{FC} acting on patches, and the dimensionality of the constituent kernel coincides with the filter size of the corresponding network. This observation extends to local kernels, via

Lemma 3.2. *Call Θ^{LC} the NTK of a locally-connected network function (3.2) with filter size s acting on d-dimensional inputs. Then*

$$\Theta^{LC}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \Theta^{FC}(\boldsymbol{x}_i, \boldsymbol{y}_i) \tag{13}$$

Following the general structure of Eq. (12) and Eq. (13), we introduce convolutional (K^{CN}) and local (K^{LC}) student and teacher kernels, defined as sums of lower-dimensional constituent kernels C,

$$K^{CN}(\boldsymbol{x}, \boldsymbol{y}) = |\mathcal{P}|^{-2} \sum_{i,j \in \mathcal{P}} C(\boldsymbol{x}_i, \boldsymbol{y}_j), \tag{14a}$$

$$K^{LC}(\boldsymbol{x}, \boldsymbol{y}) = |\mathcal{P}|^{-1} \sum_{i \in \mathcal{P}} C(\boldsymbol{x}_i, \boldsymbol{y}_i). \tag{14b}$$

The kernels in Eq. (14) are characterised by the dimensionality of the constituent kernel C, or *filter size* s (for the student, or t for the teacher) and the nonanalytic behaviour of C when the two arguments approach, i.e. $C(\boldsymbol{x}_i, \boldsymbol{y}_j) \sim \|\boldsymbol{x}_i - \boldsymbol{y}_j\|^{\alpha_s}$ (for the student, or $\|\boldsymbol{x}_i - \boldsymbol{y}_j\|^{\alpha_t}$ for the teacher) plus analytic contributions, with $\alpha_{s/t} \neq 2m$ for $m \in \mathbb{N}$. Using the kernels in Eq. (14) as covariances allows us to generate random target functions with the desired degree of locality t (as in Eq. (2)), which can also be invariant for shifts of the patches. Having a student kernel as in Eq. (14) results in an estimator f also having the form displayed in Eq. (2), with a different filter size with respect to the target function. The α's control the smoothness of these functions as, if $\alpha > 2n \in \mathbb{N}$, then the functions are at least n times differentiable in the mean-square sense.

A notable example of such constituent kernels is the NTK of ReLU networks Θ^{FC}, which presents a cusp at the origin corresponding to $\alpha_s = 1$ [32]. In addition, in the $H \to \infty$ limit, a network initialised with random weights converges to a Gaussian process [33, 34, 35]. For networks with ReLU activations, the covariance kernel of such process has nonanalytic behaviour with $\alpha_t = 3$ [36].

3.1 Mercer's decomposition of local and convolutional kernels

We now turn to describing how the eigendecomposition of the constituent kernel C induces an eigendecomposition of convolutional and local kernels. We work under the following assumptions,

i) The constituent kernel $C(\boldsymbol{x}, \boldsymbol{y})$ on $\mathbb{R}^s \times \mathbb{R}^s$ admits the following Mercer's decomposition,

$$C(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\rho=1}^{\infty} \lambda_\rho \phi_\rho(\boldsymbol{x}) \phi_\rho(\boldsymbol{y}), \tag{15}$$

with (ordered) eigenvalues λ_ρ and eigenfunctions ϕ_ρ such that, with $p^{(s)}(d^s x)$ denoting the s-dimensional patch measure, $\phi_1(\boldsymbol{x}) = 1 \; \forall \boldsymbol{x}$ and $\int p^{(s)}(d^s x) \phi_\rho(\boldsymbol{x}) = 0$ for all $\rho > 1$;

$ii)$ Convolutional and local kernels from Eq. (14) have *nonoverlapping* patches, i.e. d is an integer multiple of s and $\mathcal{P} = \{1 + n \times s \,|\, n = 1, \dots, d/s\}$ with $|\mathcal{P}| = d/s$;

$iii)$ The s-dimensional marginals on patches of the d-dimensional input measure $p^{(d)}(d^d x)$ are all identical and equal to $p^{(s)}(d^s x)$.

We stress here that the request of nonoverlapping patches in assumption $ii)$ can be relaxed at the price of further assumptions, i.e. $C(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{C}(\boldsymbol{x} - \boldsymbol{y})$ and data distributed uniformly on the torus, so that C is diagonalised in Fourier space. The resulting eigendecompositions are qualitatively similar to those described in this section (details in Appendix C). Let us also remark that assumptions $i)$ and $iii)$—together with all the assumptions on the data distribution that might follow—are technical in nature and required only to carry out the Mercer's decomposition analytically. We believe that the main results of this paper hold under much more general conditions, namely the support of the distribution being truly d-dimensional—such that the distance between neighbouring points in a collection of P data points scales as $P^{-1/d}$—and the distribution itself decaying rapidly away from the mean or having compact support. Our experiments, discussed in Section 5, support this hypothesis.

Lemma 3.3 (Spectra of convolutional kernels)**.** *Let K^{CN} be a convolutional kernel defined as in Eq. (14a), with a constituent kernel C satisfying assumptions $i)$, $ii)$ and $iii)$ above. Then K^{CN} admits the following Mercer's decomposition,*

$$K^{CN}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\rho=1}^{\infty} \Lambda_\rho \Phi_\rho(\boldsymbol{x}) \overline{\Phi_\rho(\boldsymbol{y})}, \tag{16}$$

with eigenvalues and eigenfunctions

$$\Lambda_1 = \lambda_1,\ \Phi_1(\boldsymbol{x}) = 1;\ \Lambda_\rho = \frac{s}{d}\lambda_\rho,\ \Phi_\rho(\boldsymbol{x}) = \sqrt{\frac{s}{d}} \sum_{i \in \mathcal{P}} \phi_\rho(\boldsymbol{x}_i) \text{ for } \rho > 1. \tag{17}$$

Lemma 3.4 (Spectra of local kernels)**.** *Let K^{LC} be a local kernel defined as in Eq. (14b), with a constituent kernel C satisfying assumptions $i)$, $ii)$ and $iii)$ above. Then K^{LC} admits the following Mercer's decomposition,*

$$K^{LC}(\boldsymbol{x}, \boldsymbol{y}) = \Lambda_1 \Phi_1(\boldsymbol{x}) + \sum_{\rho > 1}^{\infty} \sum_{i \in \mathcal{P}} \Lambda_{\rho,i} \Phi_{\rho,i}(\boldsymbol{x}) \overline{\Phi_{\rho,i}(\boldsymbol{y})}, \tag{18}$$

with eigenvalues and eigenfunctions ($\forall i \in \mathcal{P}$)

$$\Lambda_1 = \lambda_1,\ \Phi_1(\boldsymbol{x}) = 1;\ \Lambda_{\rho,i} = \frac{s}{d}\lambda_\rho,\ \Phi_{\rho,i}(\boldsymbol{x}) = \phi_\rho(\boldsymbol{x}_i) \text{ for } \rho > 1. \tag{19}$$

Under assumptions $i)$, $ii)$ and $iii)$ above, lemmas 3.3 and 3.4 follow from the definitions of convolutional and local kernels and the eigendecompositions of the constituents (see Appendix C for a proof of the lemmas and generalisation to kernels with overlapping patches). In the next section, we explore the consequences of these results for the asymptotics of learning curves.

4 Asymptotic learning curves for ridgeless regression

In what follows, we consider explicitly translationally-invariant constituent kernels $C(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathcal{C}(\boldsymbol{x}_i - \boldsymbol{x}_j)$ and a d-dimensional data distribution $p(d^d x)$ which is uniform on the torus, so that all lower-dimensional marginals are also uniform on lower-dimensional tori. Under these conditions, all results of Section 3 can be extended to kernels with overlapping patches ($\mathcal{P} = \{1, \dots, d\}$), so that the main results of this paper apply to nonoverlapping as well as overlapping-patches kernels. Furthermore, Mercer's decomposition Eq. (15) can be written in Fourier space [37], with s-dimensional plane waves $\phi^{(s)}_{\boldsymbol{k}}(\boldsymbol{x}) = e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$ as eigenfunctions and the eigenvalues coinciding with the Fourier transform of $\mathcal{C}$. Furthermore, for kernels with filter size s (or t) and positive smoothness exponent α_s (or α_t), the eigenvalues decay with a power $-(s + \alpha_s)$ (or $-(t + \alpha_t)$) of the modulus of the wavevector $k = \sqrt{\boldsymbol{k} \cdot \boldsymbol{k}}$ [38]. In this setting, we obtain our main result:

Theorem 4.1. *Let K_T be a d-dimensional convolutional kernel with a translationally-invariant t-dimensional constituent and leading nonanalyticity at the origin controlled by the exponent $\alpha_t > 0$. Let K_S be a d-dimensional convolutional or local student kernel with a translationally-invariant s-dimensional constituent, and with a nonanalyticity at the origin controlled by the exponent $\alpha_s > 0$. Assume, in addition, that if the kernels have overlapping patches then $s \geq t$, whereas if the kernels have nonoverlapping patches s is an integer multiple of t; and that data are uniformly distributed on a d-dimensional torus. Then, the following asymptotic equivalence holds in the limit $P \to \infty$,*

$$\mathcal{B}(P) \sim P^{-\beta}, \quad \beta = \alpha_t / s.$$

Theorem 4.1, together with Eq. (8) and the additional assumption $\alpha_t \leq 2(\alpha_s + s)$, yields the following expression for the learning curves asymptotics,

$$\epsilon(P) \sim P^{-\beta}, \quad \beta = \alpha_t / s. \tag{20}$$

As β is independent of the embedding dimension d, we conclude that the curse of dimensionality is beaten when a convolutional target is learnt with a convolutional or local kernel. In fact, Eq. (20) indicates that there is no asymptotic advantage in using a convolutional rather than local student when learning a convolutional task, confirming the picture that locality, not weight sharing, is the main source of the convolutional architecture's performances [6]. In Appendix D we show that the generalization error of a local student learning convolutional teacher decays as

$$\epsilon(P) \sim \left(\frac{P}{|\mathcal{P}|} \right)^{-\beta}, \quad \beta = \alpha_t / s. \tag{21}$$

Eq. (21) implies that including weight sharing only amounts to a rescaling of P by a factor $|\mathcal{P}|$—the size of the translation group over patches—recovering the result obtained in [28]. Intuitively, a local student will need $|\mathcal{P}|$ times more points than a convolutional student to learn the target with comparable accuracy, since it has to learn the same local function in all the possible $|\mathcal{P}|$ locations. The predictions in Eq. (20) and Eq. (21) are confirmed empirically, as discussed in Section 5 and Appendix G. Let us mention in particular that, although our predictions are valid only asymptotically, they hold already in the range $P \sim 10^2 - 10^3$, consistently with the number of training points typically used in applications.

Theorem 4.1 is proven in Appendix D and extended to the case of a local teacher and local student in Appendix E. Here we sketch the proof for the nonoverlapping case, which begins with the calculation of the variance of the coefficients of the target function in the student basis. By indexing the coefficients with the s-dimensional wavevectors $\boldsymbol{k}$,

$$\begin{aligned} \mathbb{E}[|c_{\boldsymbol{k}}|^2] &= \int_{[0,1]^d} d^d x \Phi_{\boldsymbol{k}}(\boldsymbol{x}) \int_{[0,1]^d} d^d y \overline{\Phi_{\boldsymbol{k}}}(\boldsymbol{y}) \mathbb{E}[f^*(\boldsymbol{x}) f^*(\boldsymbol{y})] \\ &= \int_{[0,1]^d} d^d x \Phi_{\boldsymbol{k}}(\boldsymbol{x}) \int_{[0,1]^d} d^d y \overline{\Phi_{\boldsymbol{k}}}(\boldsymbol{y}) K_T(\boldsymbol{x}, \boldsymbol{y}). \end{aligned} \tag{22}$$

If the size of teacher and student coincide, $s = t$, teacher and student have the same eigenfunctions. Thus, using the eigenvalue equation Eq. (6) of the teacher yields $\mathbb{E}[|c_{\boldsymbol{k}}|^2] \sim k^{-(\alpha_t + t)} = k^{-(\alpha_t + s)}$. After ranking eigenvalues by k, with multiplicity k^{s-1} from all the wavevectors having the same modulus k, one has

$$\mathcal{B}(P) = \sum_{\left\{ \boldsymbol{k} | k > P^{1/s} \right\}} k^{-(\alpha_t + s)} \sim \int_{P^{1/s}}^{\infty} dk k^{s-1} k^{-(\alpha_t + s)} \sim P^{-\frac{\alpha_t}{s}}. \tag{23}$$

When the filter size of the teacher t is lowered, some of the coefficients $\mathbb{E}[|c_{\boldsymbol{k}}|^2]$ vanish. As the target function becomes a composition of t-dimensional constituents, the only non-zero coefficients are found for $\boldsymbol{k}$'s which lie in some t-dimensional subspaces of the s-dimensional Fourier space. These subspaces correspond to the $\boldsymbol{k}$ having at most a patch of t consecutive non-vanishing components. In other words, $\mathbb{E}[|c_{\boldsymbol{k}}|^2]$ is finite only if $\boldsymbol{k}$ is effectively t-dimensional and the integral on the right-hand side of Eq. (23) becomes t-dimensional, thus

$$\mathcal{B}(P) \sim \int_{P^{1/s}}^{\infty} dk k^{t-1} k^{-(\alpha_t + t)} \sim P^{-\frac{\alpha_t}{s}}. \tag{24}$$

If the teacher patches are not contained in the student ones, the target cannot be represented with a combination of student eigenfunctions, hence the error asymptotes to a finite value when $P \to \infty$.

5 Empirical learning curves for ridgeless regression

This section investigates numerically the asymptotic behaviour of the learning curves for our teacher-student framework. We consider different combinations of convolutional and local teachers and students with overlapping patches and Laplacian constituent kernels, i.e. $\mathcal{C}(\boldsymbol{x}_i - \boldsymbol{x}_j) = e^{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|}$. In order to test the robustness of our results to the data distribution, data are uniformly generated in the hypercube $[0,1]^d$ (results in Fig. 1) or on a d-hypersphere (results in Appendix G). Fig. 1 shows learning curves for both convolutional (left panels) and local (right panels) students learning a convolutional target function. The results in the case of a local teacher are presented in Appendix G, and display no qualitative differences.

In the following, we always refer to Fig. 1. Panels A and B show that, with $\alpha_t = \alpha_s = 1$, our prediction $\beta = 1/s$ holds independently of the embedding dimension d. Furthermore, notice that fixing the dimension d and the teacher filter size t, the generalisation errors of a convolutional and a local student with the same filter size differ only by a multiplicative constant independent of P. Indeed, the shift-invariant nature of the convolutional student only results in a pre-asymptotic correction to our estimate of the generalisation error $\mathcal{B}(P)$. In Appendix G, we check that this multiplicative constant corresponds to rescaling P by the number of patches, as predicted in Section 4. Panels C and D show learning curves for several values of s and fixed t. The curse of dimensionality is recovered when the size of the student filters coincides with the input dimension, both for local and convolutional students. Finally, panels E and F show learning curves for fixed t and s being smaller than, equal to or larger than t. We stress that, when $s < t$ the student kernel cannot reproduce the target function, hence the error does not decrease by increasing P. Further details on the experiments are provided in Appendix G, together with learning curves for data distributed uniformly on the unit sphere $\mathbb{S}^{d-1}$ and for regression with the actual analytical and empirical NTKs of one-hidden-layer convolutional networks. It is worthwhile to notice that experiments are always in excellent agreement with our predictions, despite using data distributions that are out of the hypotheses of Theorem 4.1. Indeed, for regression with the actual NTK even the assumption of translationally-invariant constituents is violated. Moreover, we report the learning curves of local kernels on the CIFAR-10 dataset showing that smaller filter sizes correspond to faster decays even for real and anisotropic data distributions, in agreement with the picture emerging from our synthetic model.

6 Asymptotics of learning curves with decreasing ridge

We now prove an upper bound for the exponent β implying that the curse of dimensionality is beaten by a local or convolutional kernel learning a convolutional target (as in Eq. (2)), using the framework developed in [21] and a natural universality assumption on the kernel eigenfunctions. It is worth noticing that this framework does not require the target function to be generated by a teacher kernel. Proofs are presented in Appendix F. Let $\mathcal{D}(\Lambda)$ denote the density of eigenvalues of the student kernel, $\mathcal{D}(\Lambda) = \sum_\rho \delta(\Lambda - \Lambda_\rho)$, with $\delta(x)$ denoting Dirac delta function. Having a random target function with coefficients c_ρ in the kernel eigenbasis having variance $\mathbb{E}[|c_\rho|^2]$, one can define the following reduced density (with respect to the teacher):

$$\mathcal{D}_T(\Lambda) = \sum_{\{\rho | \mathbb{E}[|c_\rho|^2] > 0\}} \delta(\Lambda - \Lambda_\rho) \tag{25}$$

$\mathcal{D}_T(\Lambda)$ counts eigenvalues for which the target has a non-zero variance, such that:

$$\sum_\rho \mathbb{E}[|c_\rho|^2] = \int d\Lambda \mathcal{D}_T(\Lambda) c^2(\Lambda), \tag{26}$$

where the function $c(\Lambda)$ is defined by $c^2(\Lambda_\rho) = \mathbb{E}[|c_\rho|^2]$ for all ρ such that $\mathbb{E}[|c_\rho|^2] > 0$. The following theorem then follows from the results of [21].

Theorem 6.1. *Let us consider a positive-definite kernel K with eigenvalues Λ_ρ, $\sum_\rho \Lambda_\rho < \infty$, and eigenfunctions Φ_ρ learning a (random) target function f^* in kernel ridge regression (Eq. (3)) with ridge λ from P observations $f^*(\boldsymbol{x}^\mu)$, with $\boldsymbol{x}^\mu \in \mathbb{R}^d$ drawn from a certain probability distribution. Let us denote with $\mathcal{D}_T(\Lambda)$ the reduced density of kernel eigenvalues with respect to the target and $\epsilon(\lambda, P)$ the generalisation error and also assume that*

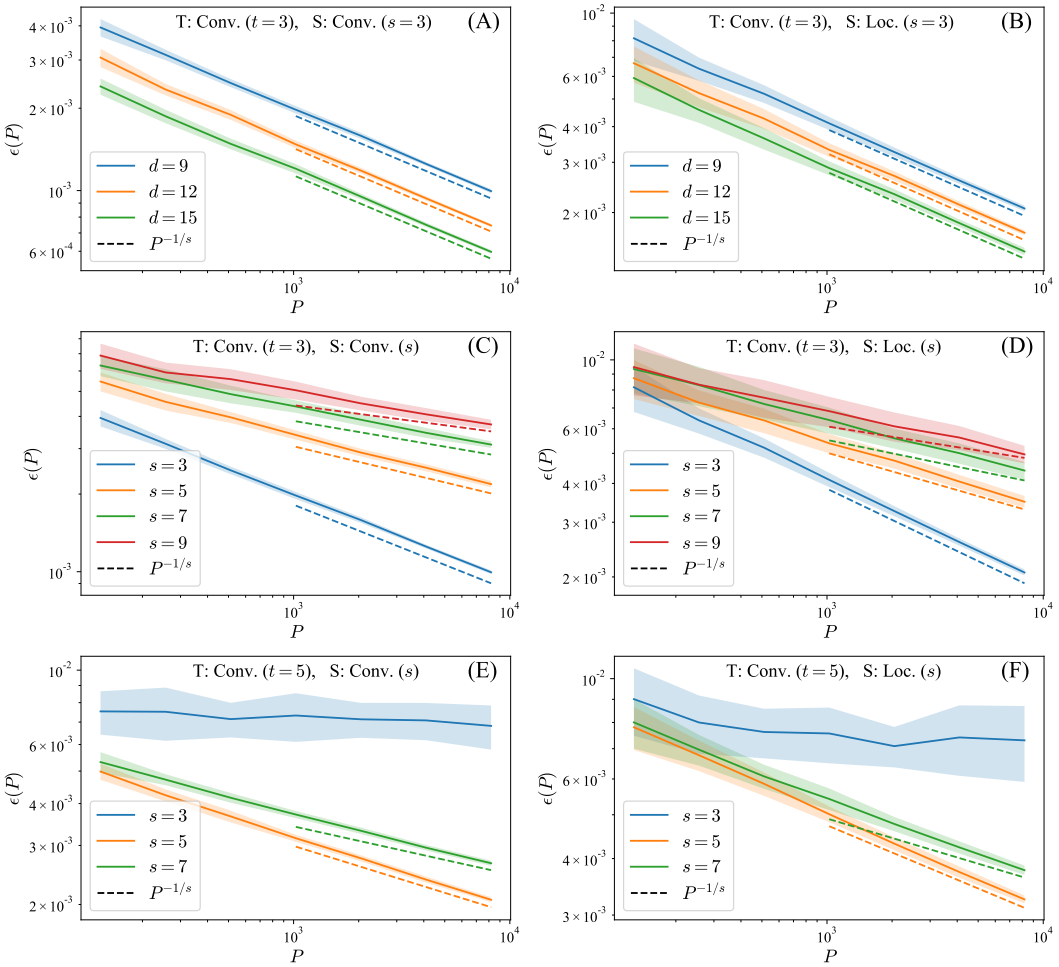


Figure 1: Learning curves for different combinations of convolutional teachers with convolutional (left panels) and local (right panels) students. The teacher and student filter sizes are denoted with t and s respectively. Data are sampled uniformly in the hypercube $[0,1]^d$, with $d=9$ if not specified otherwise. Solid lines are the results of numerical experiments averaged over 128 realisations and the shaded areas represent the empirical standard deviations. The predicted scalings are shown by dashed lines. All the panels are discussed in Section 5, while additional details on experiments are reported in Appendix G, together with additional experiments.

i) *For any P-tuple of indices $\rho_1, \dots, \rho_P$, the vector $(\Phi_{\rho_1}(\boldsymbol{x}^1), \dots, \Phi_{\rho_P}(\boldsymbol{x}^P))$ is a Gaussian random vector;*

ii) *The target function can be written in the kernel eigenbasis with coefficients c_ρ and $c^2(\Lambda_\rho) = \mathbb{E}[|c_\rho|^2]$, with $\mathcal{D}_T(\Lambda) \sim \Lambda^{-(1+r)}$, $c^2(\Lambda) \sim \Lambda^q$ asymptotically for small Λ and $r>0$, $r<q<r+2$;*

Then the following equivalence holds in the joint $P \to \infty$ and $\lambda \to 0$ limit with $1/(\lambda\sqrt{P}) \to 0$:

$$\epsilon(\lambda, P) \sim \sum_{\{\rho | \Lambda_\rho < \lambda\}} \mathbb{E}[|c_\rho|^2] = \int_0^\lambda d\Lambda \mathcal{D}_T(\Lambda) c^2(\Lambda). \tag{27}$$

Note that the assumption $i)$ of the theorem on the Gaussianity of the eigenbasis does not hold in our setup where the Φ_ρ's are plane waves. However, the random variables $\Phi_\rho(\boldsymbol{x}^\mu)$ have a probability density with compact support. It is thus natural to assume that a Gaussian universality assumption holds, i.e. that Theorem 6.1 applies to our problem. With this assumption, we obtain the following

Corollary 6.1.1. *Performing kernel ridge regression in a teacher-student scenario with smoothness exponents α_t (teacher) and α_s (student), with ridge $\lambda \sim P^{-\gamma}$ and $0 < \gamma < 1/2$, under the joint hypotheses of Theorem 4.1 and Theorem 6.1, the exponent governing the asymptotic scaling of the generalisation error with P is given by:*

$$\beta = \gamma \frac{\alpha_t}{\alpha_s + s}, \tag{28}$$

which does not vanish in the limit $d \to \infty$. Furthermore, Eq. (28) depends on s and not on t as the prediction of Eq. (20).

7 Conclusions and future work

Our work shows that, even in large dimension d, a function can be learnt efficiently if it can be expressed as a sum of constituent functions each depending on a smaller number of variables t, by performing regression with a kernel that entails such a compositional structure with s-dimensional constituents. The learning curve exponent is then independent of d and governed by s if $s \geq t$, optimal for $s = t$ and null if $s < t$.

In the context of image classification, this result relates to the "Bag of Words" viewpoint. Consider for example two-dimensional images consisting of M features of t adjacent pixels, and that different classes correspond to distinct subsets of (possibly shared) features. If features can be located anywhere, then data lie on a $2M$-dimensional manifold. On the one hand, we expect a one-hidden-layer convolutional network with filter size $s \geq t$ to learn well with a learning curve exponent governed by s and independent of M. On the other hand, a fully-connected network would suffer from the curse of dimensionality for large M.

Our work does not consider that the compositional structure of real data is hierarchical, with large features that consist of smaller sub-features. It is intuitively clear that depth and locality taken together are well-suited for such data structure [8, 6]. Extending the present teacher-student framework to this case would offer valuable quantitative insights into the question of how many data are required to learn such tasks.

Acknowledgements

We thank Alberto Bietti, Stefano Spigler, Antonio Sclocchi, Leonardo Petrini, Mario Geiger, and Umberto Maria Tomasini for helpful discussions. This work was supported by a grant from the Simons Foundation (#454953 Matthieu Wyart).

References

[1] Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with lipschitz functions. *Journal of Machine Learning Research*, 5(Jun):669–695, 2004.

[2] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[4] Stefano Spigler, Mario Geiger, and Matthieu Wyart. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124001, December 2020. Publisher: IOP Publishing.

[5] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

[6] Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

[7] Arturo Deza, Qianli Liao, Andrzej Banburski, and Tomaso Poggio. Hierarchically compositional tasks and deep convolutional networks, 2020.

[8] Alberto Bietti. On approximation in deep convolutional networks: a kernel perspective, 2021.

[9] Behnam Neyshabur. Towards learning convolutions from scratch. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8078–8088. Curran Associates, Inc., 2020.

[10] Kunihiko Fukushima. Cognitron: A self-organizing multilayered neural network. *Biological cybernetics*, 20(3):121–136, 1975.

[11] Yann LeCun et al. Generalization and network design strategies. *Connectionism in perspective*, 19:143–155, 1989.

[12] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 8580–8589, USA, 2018. Curran Associates Inc.

[13] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

[14] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. In *Advances in Neural Information Processing Systems 32*, pages 8572–8583. Curran Associates, Inc., 2019.

[15] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On Exact Computation with an Infinitely Wide Neural Net. In *Advances in Neural Information Processing Systems 32*, pages 8141–8150. Curran Associates, Inc., 2019.

[16] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2937–2947, 2019.

[17] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum Dependent Learning Curves in Kernel Regression and Wide Neural Networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, November 2020. ISSN: 2640-3498.

[18] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1):1–12, 2021.

[19] Bruno Loureiro, Cédric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Capturing the learning curves of generic features maps for realistic data sets with a teacher-student model. *arXiv preprint arXiv:2102.08127*, 2021.

[20] Charles A Micchelli and Grace Wahba. Design problems for optimal surface interpolation. Technical report, WISCONSIN UNIV-MADISON DEPT OF STATISTICS, 1979.

[21] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clement Hongler, and Franck Gabriel. Kernel alignment risk estimator: Risk prediction from training data. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15568–15578. Curran Associates, Inc., 2020.

[22] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning*, pages 2747–2755. PMLR, 2018.

[23] H. H. Zhou, Y. Xiong, and V. Singh. Building bayesian neural networks with blocks: On structure, interpretability and uncertainty. *arXiv preprint, arXiv:1806.03563*, 2018.

[24] Tommaso Poggio, Andrzej Banburski, and Qianli Liao. Theoretical issues in deep networks. *Proceedings of the National Academy of Sciences*, 117(48):30039–30045, 2020.

[25] Eran Malach and Shai Shalev-Shwartz. Computational separation between convolutional and fully-connected networks. In *International Conference on Learning Representations*, 2021.

[26] Julien Mairal. End-to-end kernel learning with supervised convolutional kernel networks. *arXiv preprint arXiv:1605.06265*, 2016.

[27] Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. *arXiv preprint arXiv:1905.12173*, 2019.

[28] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models, 2021.

[29] Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

[30] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[31] Jonas Paccolat, Stefano Spigler, and Matthieu Wyart. How isotropic kernels perform on simple invariants. *Machine Learning: Science and Technology*, 2(2):025020, March 2021. Publisher: IOP Publishing.

[32] Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, and Basri Ronen. On the similarity between the laplace and neural tangent kernels. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1451–1461. Curran Associates, Inc., 2020.

[33] Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996.

[34] Christopher KI Williams. Computing with infinite networks. *Advances in neural information processing systems*, pages 295–301, 1997.

[35] 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.

[36] Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

[37] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[38] Harold Widom. Asymptotic behavior of the eigenvalues of certain integral equations. ii. *Archive for Rational Mechanics and Analysis*, 17(3):215–229, 1964.

[39] Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[40] Marc Potters and Jean-Philippe Bouchaud. *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists*. Cambridge University Press, 2020.

[41] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Implicit regularization of random feature models. In *International Conference on Machine Learning*, pages 4631–4640, 2020.

P06

Bayesian inference in a mismatched setting: a spin-glass model with Mattis interaction

The poster will focus on the rank-one matrix estimation problem under Gaussian additive noise when the statistician assumes a prior on the ground truth signal, Rademacher for simplicity, that does not match the real one. As a consequence, the setting is not Bayes-optimal and the Nishimori identities break down. The statistician's posterior is a Boltzmann-Gibbs measure whose Hamiltonian is that of a Sherrington-Kirkpatrick (SK) model with an added Mattis interaction. Thanks to the Parisi solution of the SK model the cross entropy of the true and the statistician's evidences, closely related to the free energy, is rigorously expressed in terms of a variational principle over two order parameters: the Parisi overlap distribution and the Mattis magnetization. The phase diagram of the specific mismatch between Rademacher and Gaussian priors is analyzed in detail and shown to contain a glassy region.

P07

Resilience of Bayesian Layer-Wise Explanations under Adversarial Attacks

We consider the problem of the stability of saliency-based explanations of Neural Network predictions under adversarial attacks in a classification task. Saliency interpretations of deterministic Neural Networks are remarkably brittle even when the attacks fail, i.e. for attacks that do not change the classification label. We empirically show that interpretations provided by Bayesian Neural Networks are considerably more stable under adversarial perturbations of the inputs and even under direct attacks to the explanations. By leveraging recent results, we also provide a theoretical explanation of this result in terms of the geometry of the data manifold. Additionally, we discuss the stability of the interpretations of high level representations of the inputs in the internal layers of a Network. Our results demonstrate that Bayesian methods, in addition to being more robust to adversarial attacks, have the potential to provide more stable and interpretable assessments of Neural Network predictions.

P08

Localisation of Mitosis Transition

A recent study by Doimo, Glielmo, Goldt, Laio, 2021 has shown that in the last layer of a sufficiently wide neural network duplication of information happens, possibly related to benign overfitting. To do so, they showed that, after the network is properly trained, the error committed by using only "chunks" (that are, a certain number of random neurons) in the last layer follows two different regimes, separated by a phase transition. In the first regime, the error (or better, the difference between the error and the least possible error committed by the full network) decreases steeply as the chunk size increases, while in the second regime it decreases as N^(-1/2), with N the size of the chunks. The goal of my project is to understand better the dependence of the transition point of the input complexity and the architecture of the network, testing the results they obtained with artificial data and smaller networks to understand what are the requirements for the phenomenon to happen.

P09

Title: Comparison between NN and RF techniques for hybrid energy γ-ray reconstruction in simulation data of the 55 Imaging Air-Cherenkov Telescopes HAWC's Eye at High Altitude

Abstract

We present the result of the application of supervised regression ML algorithms such as Neuronal Network (NN) and the Random Forest regressor (RF), in the energy reconstruction of simulated Monte Carlo data of compact light-weight Imaging-Air-Cherenkov Telescope (IACT) (HAWC's Eye) and High-Altitude-Water-Cherenkov (HAWC) observatory of γ-showers in CORSIKA-9600 and MARS, using an array of 55 telescopes, in which a set of features showed a better hybrid reconstruction of the energy ($3 \sim 5\ TeV$).

In the processing, the dimensionality is reduced using feature selection, in which are reduced from 48 to 9 features. Some cuts were applied to the data, evaluating and choosing the distribution of data and the features that improve better the performance of the training in the energy reconstruction. The performance of the total energy reconstruction is measured using the coefficient of determination (r^2). Besides, the correlation and bias plots are compared with the σ and the Root Mean Square (RMS).

After several experiments, we reduced the optimal amount of shower entries (0.08-1.5)% of the total data to the training and testing. The parameters were optimized and adjusted as appropriate. For the case of NN, it was searched different net topologies and parameters, selecting the Relu activation function, the Adam optimizer lr=0.001 and 800 epochs for the training. In the case of RF, it was adjusted the max depth in each leaf using 1000 trees (comparing with previous tests). Also, the K-fold method was applied with ten splits with low σ (NN 0.0024 and RF 0.0043). Finally, RF showed better generalization than NN at distribution testing changes.

P10

Determining the impact of remote homology detection on protein structure predictions by Deep Learning models

Deep Learning models unveil unprecedented ability of predicting protein contacts and structures only relying on homologous sequences information in Multiple Sequence Alignments (MSA). In particular, self-supervised models exploit the attention mechanism to capture long-range correlations between positions in alignments, encoding both co-evolutionary and phylogenetic signals in high-dimensional embedding spaces. The influence of remote homology detection strategies on the accuracy of MSA-Transformer and AlphaFold2 predictions is estimated by comparing inference results on sequences classified with different methods. Both models are trained on alignments built via iterative pairwise comparison of profile Hidden Markov Models (HMM) of a query sequence against large databases of primary sequences. In a similar fashion, for Pfam-A alignments the query HMM profile is built on manually curated seed alignments. Differently, the unsupervised DPCfam algorithm identifies protein domains classifying pairwise aligned sequences through Density Peak Clustering technique. MSA-transformer contact prediction accuracy is similar on DPCfam and Pfam-A alignments but outperforms the two methods on the HMM-based alignments, suggesting a dependency of the model on the training MSA format. This is confirmed by the boost in performances if aligning DPCfam domains sequences with the training hhblits procedure and if fine-tuning the model parameters on such alignments. On the other hand, AlphaFold2 lDDT scores indicate that custom MSAs are better suited for domains structure prediction with respect to HMM-based alignments. Inferences on DPCfam domains return the most reliable structures, showing that such unsupervised method captures evolutionary relationships among protein domains with accuracy comparable to that of the manually curated Pfam-A.

P11

A random matrix perspective on the spiked rank-one tensor model

The task of recovering a low-rank tensor from noisy observations is at the heart of various methods used for information extraction in signal processing, data analysis and machine learning. While it is generally quite hard to analyze the performance of such methods, substantial progress has been recently achieved in the large-dimensional setting, thanks in large part to fairly advanced results and tools borrowed from statistical physics. In particular, sharp results were derived in the case of a deterministic rank-one symmetric tensor corrupted by symmetric Gaussian noise, unveiling an abrupt, discontinuous phase transition in the performance of maximum likelihood estimation as the signal-to-noise ratio grows. The random landscape of this maximum likelihood problem has also been thoroughly studied, shedding light on geometric phase transitions that take place and explain the aforementioned discontinuity. In this work, we connect these results to the notion of tensor eigenpairs, which are by definition critical points of the maximum likelihood problem. A simple but crucial observation is that each eigenpair of a tensor is also an eigenpair of a matrix obtained from that tensor by a contraction with the concerned eigenvector. As we argue, this link opens the door to the use of standard tools from random matrix theory, leading to an alternative and more elementary way of reaching some of the same predictions that had been obtained with the statistical physics machinery, while also providing interesting additional insights.

P12

Fast rates for noisy interpolation require rethinking the effects of inductive bias

Good generalization performance on high-dimensional data crucially hinges on a simple structure of the ground truth and a corresponding strong inductive bias of the estimator. Even though this intuition is valid for regularized models, in this paper we caution against a strong inductive bias for interpolation in the presence of noise: Our results suggest that, while a stronger inductive bias encourages a simpler structure that is more aligned with the ground truth, it also increases the detrimental effect of noise. Specifically, for both linear regression and classification with a sparse ground truth, we prove that minimum lp-norm and maximum lp-margin interpolators achieve fast polynomial rates up to order 1/n for p > 1 compared to a logarithmic rate for p = 1. Finally, we provide experimental evidence that this trade-off may also play a crucial role in understanding non-linear interpolating models used in practice.

P13

Optimal denoising of rotationally invariant rectangular matrices

In this work we consider denoising of large rectangular matrices: given a noisy observation of a signal matrix, what is the best way of recovering the signal matrix itself? For Gaussian noise and rotationally-invariant signal priors, we completely characterize the optimal denoiser and its performance in the high-dimensional limit, in which the size of the signal matrix goes to infinity with fixed aspects ratio, and under the Bayes optimal setting, that is when the statistician knows how the signal and the observations were generated. Our results generalise previous works that considered only symmetric matrices to the more general case of non-symmetric and rectangular ones. We explore analytically and numerically a particular choice of factorized signal prior that models cross-covariance matrices and the matrix factorization problem. As a byproduct of our analysis, we provide an explicit asymptotic evaluation of the rectangular Harish-Chandra-Itzykson-Zuber integral in a special case. See https://arxiv.org/abs/2203.07752

P14

Gaussian Universality of Linear Classiers with Random Labels in High-Dimension

While classical in many theoretical seings, the assumption of Gaussian i.i.d. inputs is oen perceived as a strong limitation in the analysis of high-dimensional learning. In this study, we redeem this line of work in the case of generalized linear classication with random labels. Our main contribution is a rigorous proof that data coming from a range of generative models in high-dimensions have the same minimum training loss as Gaussian data with corresponding data covariance. In particular, our theorem covers data created by an arbitrary mixture of homogeneous Gaussian clouds, as well as multi-modal generative neural networks. In the limit of vanishing regularization, we further demonstrate that the training loss is independent of the data covariance. Finally, we show that this universality property is observed in practice with real datasets and random labels.

P15

A knowledge-based machine learning approach to gene prioritisation in amyotrophic lateral sclerosis

Amyotrophic lateral sclerosis is a neurodegenerative disease of the upper and lower motor neurons resulting in death from neuromuscular respiratory failure, typically within two to five years of first symptoms. Several rare disruptive gene variants have been associated with ALS and are responsible for about 15% of all cases. Although our knowledge of the genetic landscape of this disease is improving, it remains limited. Machine learning models trained on the available protein–protein interaction and phenotype-genotype association data can use our current knowledge of the disease genetics for the prediction of novel candidate genes. Here, we describe a knowledge-based machine learning method for this purpose. We trained our model on protein–protein interaction data from IntAct, gene function annotation from Gene Ontology, and known disease-gene associations from DisGeNet. Using several sets of known ALS genes from public databases and a manual review as input, we generated a list of new candidate genes for each input set. We investigated the relevance of the predicted genes in ALS by using the available summary statistics from the largest ALS genome-wide association study and by performing functional and phenotype enrichment analysis. The predicted sets were enriched for genes associated with other neurodegenerative diseases known to overlap with ALS genetically and phenotypically, as well as for biological processes associated with the disease. Moreover, using ALS genes from ClinVar and our manual review as input, the predicted sets were enriched for ALS-associated genes (ClinVar p = 0.038 and manual review p = 0.060) when used for gene prioritisation

P16

Data-driven emergence of convolutional structure in neural networks

Exploiting invariances in the inputs is crucial for constructing efficient representations and accurate predictions in neural circuits. In neuroscience, translation invariance is at the heart of models of the visual system, while convolutional neural networks designed to exploit translation invariance triggered the first wave of deep learning successes. While the hallmark of convolutions, namely localised receptive fields that tile the input space, can be implemented with fully-connected neural networks, learning convolutions directly from inputs in a fully-connected network has so far proven elusive. Here, we show how initially fully-connected neural networks solving a discrimination task can learn a convolutional structure directly from their inputs, resulting in localised, space-tiling receptive fields. We find that both translation invariance and non-trivial higher-order statistics are needed to learn convolutions from scratch. We provide an analytical and numerical characterisation of the pattern-formation mechanism responsible for this phenomenon in a simple model, which results in an unexpected link between receptive field formation and the tensor decomposition of higher-order input correlations.

P17

A Curious Case of Symmetric Binary Perceptron Model: Algorithms and Barriers

David Gamarnik [1] Eren C. Kızıldağ [2] Will Perkins [3] Changji Xu [4]

[1]MIT ORC [2]MIT EECS & LIDS [3]University of Illinois at Chicago [4]Harvard CMSA

Symmetric Binary Perceptron (SBP)

Setup: Fix $\kappa, \alpha > 0$, set $M = \lfloor n\alpha \rfloor \in \mathbb{N}$. Generate i.i.d. $X_i \overset{d}{=} \mathcal{N}(0, I_n)$, $1 \le i \le M$. Define

$$S_\alpha(\kappa) = \bigcap_{1 \le i \le M} \left\{\sigma \in \mathcal{B}_n : |\langle \sigma, X_i \rangle| \le \kappa\sqrt{n}\right\} = \left\{\sigma \in \mathcal{B}_n : \|\mathcal{M}\sigma\|_\infty \le \kappa\sqrt{n}\right\},$$

where $\mathcal{B}_n = \{-1, 1\}^n$ and $\mathcal{M} \in \mathbb{R}^{M \times n}$ is the matrix of **disorder** with rows $X_1, \dots, X_M \in \mathbb{R}^n$.

Algorithmic Goal: Find a $\sigma \in S_\alpha(\kappa)$ in **polynomial-time** whenever $S_\alpha(\kappa) \neq \varnothing$ (whp).

Motivation

Neural Networks: Toy one-layer neural network (Wendel'62, Cover'65).

- Patterns $X_i \in \mathbb{R}^n$ to be **stored**.
- **Storage:** Find $\sigma \in \mathcal{B}_n$ "consistent" with X_i's: $\langle \sigma, X_i \rangle \ge 0$.

Constraint Satisfaction Problems: X_i rules out certain $\sigma \in \mathcal{B}_n$. **Constraint Density:** $\alpha = M/n$.

Discrepancy Theory: Given $\mathcal{M} \in \mathbb{R}^{M \times n}$, explore its discrepancy $\min_{\sigma \in \mathcal{B}_n} \|\mathcal{M}\sigma\|_\infty$.

Existential and Algorithmic Guarantees

Sharp Phase Transition. Let $\alpha_c(\kappa) = -1/\log_2 \mathbb{P}\left[|\mathcal{N}(0,1)| \le \kappa\right]$. Perkins-Xu'21, Abbe-Li-Sly'21:

$$S_\alpha(\kappa) \neq \varnothing \text{ (whp) if } \alpha < \alpha_c(\kappa). \quad S_\alpha(\kappa) = \varnothing \text{ (whp) if } \alpha > \alpha_c(\kappa).$$

Algorithmic (Polynomial-Time). Bansal-Spencer'20: for $\alpha = O(\kappa^2)$, outputs a $\sigma_{\mathrm{ALG}} \in S_\alpha(\kappa)$ (whp).

A *Statistical-to-Computational Gap*

Gap between existential guarantee and the best polynomial-time algorithmic guarantee.

Most pronounced for $\kappa \to 0$:

- $S_\alpha(\kappa) \neq \varnothing$ (whp) iff $\alpha < -1/\log_2 \kappa$. Algorithms exist for $\alpha = O(\kappa^2)$.
- A striking gap: $-1/\log_2 \kappa$ vs κ^2.

Source of this gap/hardness?

Extreme Clustering and Freezing

Also known as Frozen 1-RSB in physics. For **any** $0 < \alpha < \alpha_c(\kappa)$:

- **Typical** solutions of SBP are isolated (whp). Distance to nearest solution is $\Theta(n)$.
- Suggests algorithmic hardness (Achlioptas & Coja-Oghlan'08).

A Conundrum: Extreme clustering/freezing coexist with polynomial-time algorithms.

Study of Statistical-to-Computational Gap

Common feature in many algorithmic problems in high-dimensional statistics & random combinatorial structures: *Random k-SAT, optimization over random graphs, p-spin model, number partitioning...*

Average-Case Problems: No analogue of worst-case theory (such as $P \neq NP$).

Rigorous Evidences of Hardness: *low-degree methods, reductions from the planted clique, failure of MCMC, failure of BP/AMP, SoS/SQ lower bounds,...*

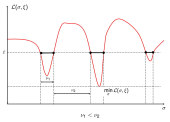
Overlap Gap Property (OGP)

Another approach from *spin glass theory*: **Overlap Gap Property (OGP).**

- Generic optimization problem with random instance ξ: $\min_{\sigma \in \Theta} \mathcal{L}(\sigma, \xi)$.
- (Informally) OGP for **energy** $\mathcal{E}$ if $\exists 0 < \nu_1 < \nu_2$ s.t. w.h.p. over ξ, $\forall \sigma_1, \sigma_2 \in \Theta$,

$$\mathcal{L}(\sigma_j, \xi) \le \mathcal{E} \implies \text{distance}(\sigma_1, \sigma_2) < \nu_1 \quad \text{or} \quad \text{distance}(\sigma_1, \sigma_2) > \nu_2.$$

- Any two **near optimal** σ_1, σ_2 are either *too similar or too dissimilar*.



First algorithmic implication: Finding maximum independent set in $\mathbb{G}_d(n)$. (Gamarnik-Sudan'13).

Problems with OGP: Many, *random k-SAT, p−spin model, number partitioning...*

OGP as a Provable Barrier to Algorithms: *WALKSAT, local algorithms, stable algorithms, low-degree polynomials, approximate message passing (AMP), MCMC, low-depth circuits, QAOA...*

Landscape Results: Presence of OGP

Consider i.i.d. $\mathcal{M}_i \in \mathbb{R}^{M \times n}, 0 \le i \le m$, each with i.i.d. $\mathcal{N}(0,1)$ entries. Interpolate:

$$\mathcal{M}_i(\tau) = \cos(\tau)\mathcal{M}_0 + \sin(\tau)\mathcal{M}_i \in \mathbb{R}^{M \times n}, \quad \tau \in [0, \pi/2], \quad 1 \le i \le m.$$

Fix $\kappa > 0$. SBP exhibits **Ensemble** $m-$**OGP** with $(m, \beta, \eta, \mathcal{I})$, if for any $\sigma_1, \dots, \sigma_m \in \mathcal{B}_n$ with

$$\|\mathcal{M}_i(\tau_i)\sigma_i\|_\infty \le \kappa\sqrt{n}, \quad \tau_i \in \mathcal{I}, \quad 1 \le i \le m,$$

there exists $1 \le i < j \le m$ such that $n^{-1}\langle \sigma_i, \sigma_j \rangle \notin (\beta - \eta, \beta)$.

$m-$**tuples:** Hardness for broader range of parameters (i.e. lower threshold for α).

Ensemble: Correlated instances. Rule out any sufficiently stable algorithm.

Small κ regime, $\kappa \to 0$: Statistical-to-Computational Gap is most pronounced.

Theorem. $\forall \kappa > 0$ small and $\mathcal{I} \subset [0, \pi/2]$ with $|\mathcal{I}| \le \exp(O(n))$, there exists $m \in \mathbb{N}$ and $1 > \beta > \eta > 0$ such that the SBP exhibits (whp) the Ensemble $m-$OGP with $(m, \beta, \eta, \mathcal{I})$ for $\alpha = \Omega(\kappa^2 \log \frac{1}{\kappa})$.

- Nearly **tight:** Matches algorithmic κ^2 threshold up to $\log \frac{1}{\kappa}$ factor.
- $\beta \gg \eta$: no equidistant $m-$tuples each satisfying constraint $\mathcal{M}_i(\tau_i)$, $1 \le i \le m$.

Large κ regime: Set $\kappa = 1$, $\alpha_c(\kappa) \approx 1.8158$. Thus $S_\alpha(\kappa) \neq \varnothing$ (whp) iff $\alpha < 1.8158$.

Theorem. Let $\kappa = 1$. $\exists 0 < \beta_2, \beta_3, \eta_2, \eta_3 < 1$ (where $\beta_i > \eta_i$) such that the following holds whp:

- SBP exhibits Ensemble 2−OGP with $(2, \beta_2, \eta_2, \mathcal{I})$ for $\alpha \ge 1.71$.
- SBP exhibits Ensemble 3−OGP with $(3, \beta_3, \eta_3, \mathcal{I})$ for $\alpha \ge 1.67$.

Algorithmic Hardness Results

Algorithm $\mathcal{A} : \mathbb{R}^{M \times n} \to \mathcal{B}_n$, potentially randomized.

Stable Algorithms. Informally, $\mathcal{A}$ is stable if small change in X yields small change in $\mathcal{A}(X)$.

Success:

$$\mathbb{P}\left[\left\|\mathcal{M}\mathcal{A}(\mathcal{M})\right\|_\infty \le \kappa\sqrt{n}\right] \ge 1 - p_f.$$

Stability: $\exists \rho \in (0, 1]$ such that for i.i.d. $\mathcal{M}, \overline{\mathcal{M}} \in \mathbb{R}^{M \times n}$ with $\mathrm{Cov}(\mathcal{M}_{ij}, \overline{\mathcal{M}}_{ij}) = \rho$

$$\mathbb{P}\left[d_H\Big(\mathcal{A}(\mathcal{M}), \mathcal{A}(\overline{\mathcal{M}})\Big) \le f + L\|\mathcal{M} - \overline{\mathcal{M}}\|_F\right] \ge 1 - p_{\mathrm{st}}.$$

AMP and low-degree polynomials are stable (Gamarnik-Jagannath-Wein'20).

Question: *"Are known efficient algorithms for perceptron models stable?"*

Theorem. Kim-Roche algorithm (Kim-Roche'98) for the asymmetric perceptron is stable.

$m-$OGP $\implies$ **Failure of Stable Algorithms.**

Theorem. Stable algorithms fail to find a solution for the SBP for $\alpha = \Omega(\kappa^2 \log \frac{1}{\kappa})$.

Rule out $p_f, p_{\mathrm{st}} = O(1)$. **No** need for **high-probability** guarantee.

Proof Idea. By contradiction. Suppose $\exists \mathcal{A}$.

- m-OGP: a structure occurs with *vanishing probability*.
- Run $\mathcal{A}$ on correlated instances. Show that w.p. > 0, *forbidden* structure occurs.
- Uses Ramsey Theory (Gamarnik-Kızıldağ'21).

Failure of Online Algorithms for High Densities:

Columns of $\mathcal{M}$: $\mathcal{C}_1, \dots, \mathcal{C}_n \in \mathbb{R}^M$. $\mathcal{A}$ is **online** if $\exists f_t$ s.t. $\sigma_t = f_t(\mathcal{C}_i : 1 \le i \le t)$ for $1 \le t \le n$.

Theorem. $\exists c > 0$ such that for $\alpha \ge \alpha_c(\kappa) - \epsilon$, there is no **online** $\mathcal{A}$ for SBP.

Future Directions

Algorithmic Threshold: Let $\alpha_m(\kappa)$ be the smallest density such that for some $0 < \eta < \beta < 1$, SBP exhibits (whp) Ensemble $m-$OGP with $(m, \beta, \eta, \{0\})$ for $\alpha \ge \alpha_m(\kappa)$. Define

$$\alpha_\infty^*(\kappa) \triangleq \lim_{m \to \infty} \alpha_m(\kappa).$$

Conjecture. $\alpha_\infty^*(\kappa)$ marks the true algorithmic threshold of SBP.

- Bansal-Spencer algorithm is likely optimal (up to logarithmic factors).
- $\log \frac{1}{\kappa}$ factor? More delicate structure (Wein'20, Bresler-Huang'21, Huang-Sellke'21).

Stability of Other Algorithms: *"Is Bansal-Spencer algorithm stable? Other discrepancy algorithms?"*

Asymmetric Perceptron: Many open problems.

- Existence/Location of sharp phase transition point. *Krauth-Mézard (89) prediction.*
- Rigorously verifying Frozen 1-RSB picture.
- OGP and failure of stable algorithms.

More Enthusiastic Questions on OGP.

- Largest class of algorithms ruled out by OGP: Includes stable algorithms, MCMC, etc.
- Counterexample to OGP: Is there a model where efficient algorithms coexist with OGP?

arXiv:2203.15667

kizildag@mit.edu

P18

Industrial Problems

Some industrial practical problems can be discussed in details

P19

Universality of the least singular value and singular vector delocalization for levy non-symmetric matrices

In this paper we consider N×N matrices D_{n} with i.i.d. entries all following an a−stable law divided by N^{1/a}. We prove that the least singular value of D_{N}, multiplied by N, tends to the same law as in the Gaussian case, for almost all a∈(0,2). This is proven by considering the symmetrization of the matrix D_{N} and using a version of the three step strategy, a well known strategy in the random matrix theory literature. In order to apply the three step strategy, we also prove an isotropic local law for the symmetrization of matrices after slightly perturbing them by a Gaussian matrix with a similar structure. The isotropic local law is proven for a general class of matrices that satisfy some regularity assumption. We also prove the complete delocalization for the left and right singular vectors of DN at small energy, i.e., for energies at a small interval around 0.

P20

Deep Learning via Message Passing

Message-passing algorithms based on the Belief Propagation (BP) equations constitute a well-known distributed computational scheme. It is exact on tree-like graphical models and has also proven to be effective in many problems defined on graphs with loops (from inference to optimization, from signal processing to clustering). The BP-based scheme is fundamentally different from stochastic gradient descent (SGD), on which the current success of deep networks is based. In this paper, we present and adapt to mini-batch training on GPUs a family of BP-based message-passing algorithms with a reinforcement field that biases distributions towards locally entropic solutions. These algorithms are capable of training multi-layer neural networks with discrete weights and activations with performance comparable to SGD-inspired heuristics (BinaryNet) and are naturally well-adapted to continual learning. Furthermore, using these algorithms to estimate the marginals of the weights allows us to make approximate Bayesian predictions that have higher accuracy than point-wise solutions.

P21

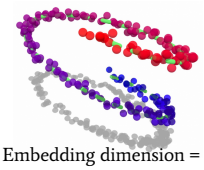
Topological analysis of genomics sequences shows that evolutive pressure acts on a low-dimensional manifold

Iuri Macocco[1], Aldo Glielmo[1, 2], Jacopo Grilli[3] and Alessandro Laio[1]

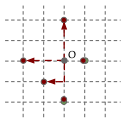
[1)] *International School for Advanced Studies (SISSA), Via Bonomea 265, 34136 Trieste, Italy*
[2)] *Banca d'Italia, Italy*
[3)] *The Abdus Salam International Centre for Theoretical Physics (ICTP), Strada Costiera 11, 34014 Trieste, Italy*

Real-world data are often defined by a very large number of features, but are effectively contained in a manifold which can be described, at least locally, by a relatively small number of coordinates. Such number is called Intrinsic Dimension (ID).

All present estimators have been formulated in spaces where distances can vary continuously. However, many systems are characterised by discrete features and discrete distances.



Embedding dimension = 3
Intrinsic dimension = 1



Examples:

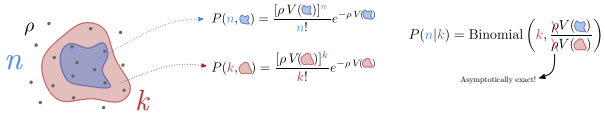
GAAGGTCTTCGGAT
GATGCCTTTCGGAA
GACGGCCTTCGGGT
genomics strands

Problem: discrete features lead to two main issues

- Distances degeneracy
- No Scaling $N \nsim r^d$

Construction of the Intrinsic Dimension for Discrete Dataset (I3D) estimator

1) Space Poisson processes [1]



2) Measure of discrete volumes [2]

$$V_o(t,d) = \sum_{l=0}^{d} \binom{d}{l}\binom{t-l+d}{d}$$

3) Set t_1 and t_2 → Compute n_i and k_i → Write the likelihood for all datapoints

$$\mathcal{L}\{d|(n_i,k_i)\} = \prod_i^N \binom{k_i}{n_i} \left(\frac{V_o(t_1,d)}{V_o(t_2,d)}\right)^{n_i} \left(1 - \frac{V_o(t_1,d)}{V_o(t_2,d)}\right)^{k_i - n_i}$$

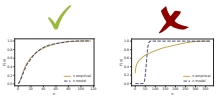
4) Infer the ID

$$\frac{\partial}{\partial d} \ln \mathcal{L}\{d\,|\,(n_i,k_i)\} = \frac{V_o(t_1,d)}{V_o(t_2,d)} - \frac{\langle n \rangle}{\langle k \rangle} = 0 \;\rightarrow \hat{d}$$

5) Model validation

empirical distribution $P(n)$

model distribution $P(n) = \sum_k P(k)\text{Binomial}(k, \frac{V_o(t_1,\hat{d})}{V_o(t_2,\hat{d})})$

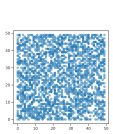


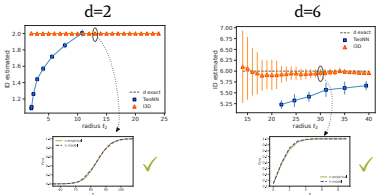
6) ID scaling by varying t_1 and t_2

Performance on toy models against continuous ID estimator [3]

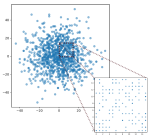
The lattice structure, together with the high density and the consequent possible overlap of some points, prevents other estimators from providing a precise estimate, which is obtained only under aggressive decimation of the dataset. The I3D estimator, instead, returns accurate values for the ID at all scales.

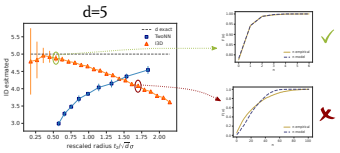
Uniform lattice distribution





Gaussian lattice distribution

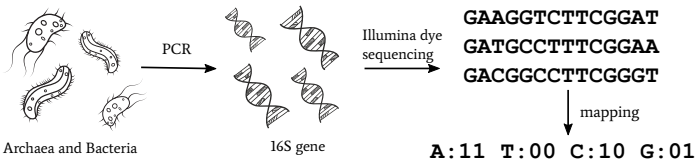


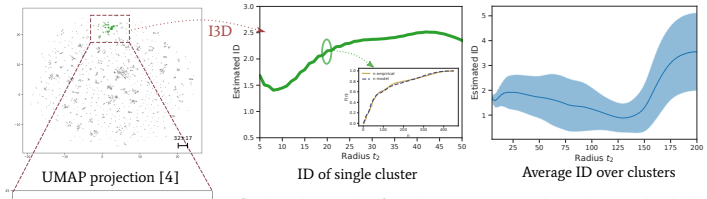


The evolutionary relationships among biological species, based upon similarities and differences in their genetic characteristics, are often represented and studied through phylogenetic trees. However, such trees are not necessarily correct and relative inferred relationship are not unquestionably true. Here we propose a complementary method that provides effective information about evolution and selection based directly on the sequences, without the need of constructing phylogenetic trees

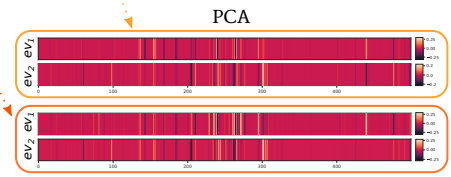
Topological analysis of prokaryotic 16S gene

Prokaryotic 16S rRNA sequences are widely used in environmental microbiology and molecular evolution as reliable markers for the taxonomic classification and phylogenetic analysis of microbes, due to its slow rates of evolution





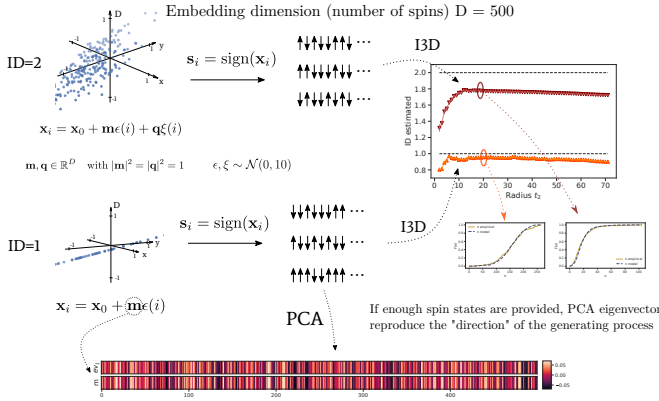
Such a low value for the ID suggests that, despite the high dimensionality of sequences' space, evolution effectively operates in a low-dimensional space. Qualitatively, an ID of ~2 on a scale of ~20 means that if one considers all the sequences differing by approximately 20 mutations from a given sequence, these mutations cannot be considered independent one from each other, but are correlated in such a way that approximately 18 degrees of freedom are effectively forbidden.



The "direction" of these correlated mutation can be, at least approximately, measured by performing PCA. The eigenvectors do not change significantly on different distance range, indicating that, consistently with the low value of the ID, the data manifold on this scale can be approximately described by a two-dimensional plane. The components of eigenvectors can be qualitatively interpreted as proportional to the mutation probabilities of the associated nucleotides for a collective mutation process.

Behaviour on controlled-ID spin dataset

In order to better grasp the meaning of PCA on genomics data, we performed the same procedure on artificial, controlled-ID spin ensembles



References

[1] Moltchanov, D. Distance distributions in random networks. Ad Hoc Networks10, (2012)
[2] Beck, M. & Robins, S. Computing the continuous discretely: integer-point enumeration in polyhedra. Choice Rev.
[3] Facco, E., D'Errico, M., Rodriguez, A. & Laio, A. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. Sci. Reports (2017).
[4] Mcinnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2020). 1802.03426v3.

Contact: imacocco@sissa.it

DADApy
SISSA
ICTP

P22

Perturbative construction of mean-field equations in extensive-rank matrix factorization and denoising

Factorization of matrices where the rank of the two factors diverges linearly with their sizes has many applications in diverse areas such as unsupervised representation learning, dictionary learning or sparse coding. We consider a setting where the two factors are generated from known component-wise independent prior distributions, and the statistician observes a (possibly noisy) component-wise function of their matrix product. In the limit where the dimensions of the matrices tend to infinity, but their ratios remain fixed, we expect to be able to derive closed form expressions for the optimal mean squared error on the estimation of the two factors. However, this remains a very involved mathematical and algorithmic problem. A related, but simpler, problem is extensive-rank matrix denoising, where one aims to reconstruct a matrix with extensive but usually small rank from noisy measurements. In this paper, we approach both these problems using high-temperature expansions at fixed order parameters. This allows to clarify how previous attempts at solving these problems failed at finding an asymptotically exact solution. We provide a systematic way to derive the corrections to these existing approximations, taking into account the structure of correlations particular to the problem. Finally, we illustrate our approach in detail on the case of extensive-rank matrix denoising. We compare our results with known optimal rotationally-invariant estimators, and show how exact asymptotic calculations of the minimal error can be performed using extensive-rank matrix integrals.

P23

Neural Receptive field as Gaussian Process

Neural receptive field such as place fields exhibit interesting transition when representing small to large space. I develop a framework of understanding neural activity as a function of it's input as a Gaussian Process and use related notions to characterise the distribution of receptive fields and show that it is in good agreement with empirical observations.

P24

The merged-staircase property: a necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks

It is currently known how to characterize functions that neural networks can learn with SGD for two extremal parametrizations: neural networks in the linear regime, and neural networks with no structural constraints. However, for the main parametrization of interest —non-linear but regular networks— no tight characterization has yet been achieved, despite significant developments. We take a step in this direction by considering depth-2 neural networks trained by SGD in the mean-field regime. We consider functions on binary inputs that depend on a latent low-dimensional subspace (i.e., small number of coordinates). This regime is of interest since it is poorly understood how neural networks routinely tackle high-dimensional datasets and adapt to latent low-dimensional structure without suffering from the curse of dimensionality. Accordingly, we study SGD-learnability with O(d) sample complexity in a large ambient dimension d. Our main results characterize a hierarchical property — the merged-staircase property— that is both necessary and nearly sufficient for learning in this setting. We further show that non-linear training is necessary: for this class of functions, linear methods on any feature map (e.g., the NTK) are not capable of learning efficiently. The key tools are a new “dimension-free” dynamics approximation result that applies to functions defined on a latent space of low-dimension, a proof of global convergence based on polynomial identity testing, and an improvement of lower bounds against linear methods for non-almost orthogonal functions.

P25

Consensus from group interactions: An adaptive voter model on hypergraphs

We study the effect of group interactions on the emergence of consensus in a spin system. Agents with discrete opinions {0,1} form groups. They can change their opinion based on their group's influence (voter dynamics), but groups can also split and merge (adaptation). In a hypergraph, these groups are represented by hyperedges of different sizes. The heterogeneity of group sizes is controlled by a parameter β. To study the impact of β on reaching consensus, we provide extensive computer simulations and compare them with an analytic approach for the dynamics of the average magnetization. We find that group interactions amplify small initial opinion biases, accelerate the formation of consensus and lead to a drift of the average magnetization. The conservation of the initial magnetization, known for basic voter models, is no longer obtained.

P26

How learning features can lead to over-fitting in neural nets

Understanding why neural networks can learn data in high dimensions remains a challenge. It has been proposed that they do so by adapting to features in the data that are relevant for the task, effectively reducing the input dimension and making the problem tractable. Indeed, in fully-connected networks, learning features is beneficial when the task is insensitive to linear directions in input space and this setting is theoretically well understood. However, when looking at real data (image classification), it is empirically found that learning features is beneficial for modern convolutional architectures, but not for simple fully-connected ones. How to rationalize the drawbacks of feature learning? In our work, we argue that the lazy regime (in which features are not learned) can be advantageous if the target function is smooth enough along certain input-space directions. We prove it on a class of target functions where we show that the asymptotic decay of the generalization error with the number of training points is faster in the lazy regime, when compared to feature-learning. We give empirical evidence that this same phenomenon takes place in the context of image classification by studying the deformation stability of the predictor in both lazy and feature regimes. We conclude by arguing that the benefits of adaptivity in the feature regime may be counterbalanced by the drawbacks of over-fitting, depending on the data symmetries and architecture choice.

P27

The dynamics of representation learning in shallow, non-linear autoencoders

Autoencoders are the simplest neural network for unsupervised learning, and thus an ideal framework for studying feature learning. While a detailed understanding of the dynamics of linear autoencoders has recently been obtained, the study of non-linear autoencoders has been hindered by the technical diculty of handling training data with non-trivial correlations – a fundamental prerequisite for feature extraction. Here, we study the dynamics of feature learning in non-linear, shallow autoencoders. We derive a set of asymptotically exact equations that describe the generalisation dynamics of autoencoders trained with stochastic gradient descent (SGD) in the limit of high-dimensional inputs. These equations reveal that autoencoders learn the leading principal components of their inputs sequentially. An analysis of the long-time dynamics explains the failure of sigmoidal autoencoders to learn with tied weights, and highlights the importance of training the bias in ReLU autoencoders. Building on previous results for linear networks, we analyse a modification of the vanilla SGD algorithm which allows learning of the exact principal components. Finally, we show that our equations accurately describe the generalisation dynamics of non-linear autoencoders on realistic datasets such as CIFAR10.

P28

On the Role of Channel Capacity in Learning Gaussian Mixture Models

Elad Romanov [*1], Tamir Bendory [†2], and Or Ordentlich [‡1]

[1]School of Computer Science and Engineering, The Hebrew University of Jerusalem
[2]School of Electrical Engineering, Tel Aviv University

Abstract

This paper studies the sample complexity of learning the k unknown centers of a balanced Gaussian mixture model (GMM) in $\mathbb{R}^d$ with spherical covariance matrix $\sigma^2\boldsymbol{I}$. In particular, we are interested in the following question: what is the maximal noise level σ^2, for which the sample complexity is essentially the same as when estimating the centers from labeled measurements? To that end, we restrict attention to a Bayesian formulation of the problem, where the centers are uniformly distributed on the sphere $\sqrt{d}\mathcal{S}^{d-1}$. Our main results characterize the *exact noise threshold* σ^2 below which the GMM learning problem, in the large system limit $d, k \to \infty$, is as easy as learning from labeled observations, and above which it is substantially harder. The threshold occurs at $\frac{\log k}{d} = \frac{1}{2}\log\left(1+\frac{1}{\sigma^2}\right)$, which is the capacity of the additive white Gaussian noise (AWGN) channel. Thinking of the set of k centers as a code, this noise threshold can be interpreted as the largest noise level for which the error probability of the code over the AWGN channel is small. Previous works on the GMM learning problem have identified the *minimum distance* between the centers as a key parameter in determining the statistical difficulty of learning the corresponding GMM. While our results are only proved for GMMs whose centers are uniformly distributed over the sphere, they hint that perhaps it is the decoding error probability associated with the center constellation as a channel code that determines the statistical difficulty of learning the corresponding GMM, rather than just the minimum distance.

[*]elad.romanov@gmail.com
[†]bendory@tauex.tau.ac.il
[‡]or.ordentlich@mail.huji.ac.il

P29

Inducing bias is simpler than you think

Machine learning systems are nowadays involved in almost every aspect of our life, given their flexibility and the abundance of available training data. However, increasing evidence shows that blind applications of these tools might incur in negative societal impact. Cultural biases against marginalised communities are often reflected in the very data used at training, and may be perpetuated or even enlarged by the learned models. The harmful effect is often disproportional and impacts some communities more severely compared to the rest. This paper proposes a high-dimensional theoretical model of imbalance, amenable of analytic treatment through the tools of statistical physics. The parametric control over the structural properties of the data allow an in depth study of the multiple bias-inducing factors at play. By extensively exploring different learning settings and parameters regions, we identify the regimes in which the data imbalances may severely impact the under-represented communities. On the other hand, we also trace a positive transfer effect between different communities, especially in the low sampling regime. This suggests that mixing data with different statistical properties is not necessarily harmful if the model is made aware of them. Finally, we discuss the issue of defining appropriate bias mitigation techniques, showing how the standard fairness assessment metrics often make incompatible requirements on the learned models. We also propose an interpretable and robust mitigation strategy, based on the introduction of coupled learning models that specialise on the different communities represented in the data.

P30

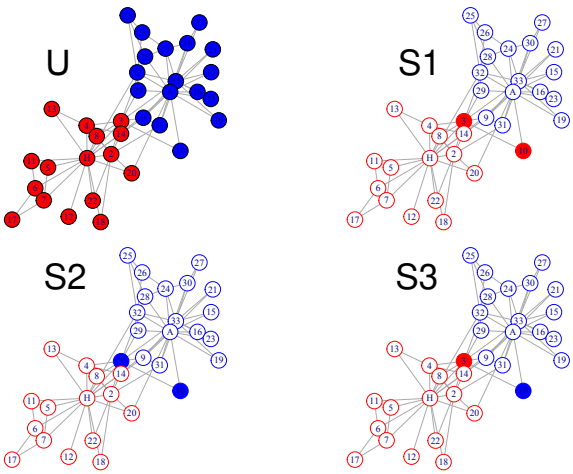


Figure: Partition observed by Zachary (U, upper left) and the three most (and equally) probable states (S1, S2 and S3), after fixing node A and H to be in different groups. Nodes 3 and 10 (with filled color in S1, S2 and S3) may be associated with A or H depending on the state S.

P31

Maslow's Hammer for Catastrophic Forgetting: Node Re-Use vs Node Activation

Continual learning - learning new tasks in sequence while maintaining performance on old tasks - remains particularly challenging for artificial neural networks. Surprisingly, the amount of forgetting does not increase with the dissimilarity between the learned tasks, but appears to be worst in an intermediate similarity regime. In this paper we theoretically analyse both a synthetic teacher-student framework and a real data setup to provide an explanation of this phenomenon that we name Maslow's hammer hypothesis. Our analysis reveals the presence of a trade-off between node activation and node re-use that results in worst forgetting in the intermediate regime. Using this understanding we reinterpret popular algorithmic interventions for catastrophic interference in terms of this trade-off, and identify the regimes in which they are most effective.

P32

May subjective sensory experiences be the result of a neural reservoir?

Subjective sensory experience, such as the tactile or auditory perception of a presented stimulus, is known to be history dependent. In particular contraction bias, i.e. the attraction of the perception towards the center of the distribution of the stimuli observed in the past, discovered over a century ago through behavioral experiments, is currently being studied also from a neurophysiological point of view. Reservoir neural networks, as LSM/ESN, which one can think as the recurrent neural circuits from which experimenters record from, offer a putative computational mechanisms which could underlay the shading memory phenomena observed in the experiments. In this poster we present a working hypothesis over the possibility of using reservoir neural networks to model sensory perception.

P33

High-dimensional optimization under nonconvex excluded volume constraints

We consider high-dimensional random optimization problems where the dynamical variables are subjected to nonconvex excluded volume constraints. We focus on the case in which the cost function is a simple quadratic cost and the excluded volume constraints are modeled by a perceptron constraint satisfaction problem. We show that depending on the density of constraints, one can have different situations. If the number of constraints is small, one typically has a phase where the ground state of the cost function is unique and sits on the boundary of the island of configurations allowed by the constraints. In this case, there is a hypostatic number of marginally satisfied constraints. If the number of constraints is increased one enters a glassy phase where the cost function has many local minima sitting again on the boundary of the regions of allowed configurations. At the phase transition point, the total number of marginally satisfied constraints becomes equal to the number of degrees of freedom in the problem and therefore we say that these minima are isostatic. We conjecture that by increasing further the constraints the system stays isostatic up to the point where the volume of available phase space shrinks to zero. We derive our results using the replica method and we also analyze a dynamical algorithm, the Karush-Kuhn-Tucker algorithm, through dynamical mean-field theory and we show how to recover the results of the replica approach in the replica symmetric phase (see Sclocchi and Urbani 2022).

P34

Phase diagram of Stochastic Gradient Descent in high-dimensional two-layer neural networks

Despite the non-convex optimization landscape, over-parametrized shallow networks are able to achieve global convergence under gradient descent. The picture can be radically different for narrow networks, which tend to get stuck in badly-generalizing local minima. Here we investigate the cross-over between these two regimes in the high-dimensional setting, and in particular investigate the connection between the so-called mean-field/hydrodynamic regime and the seminal approach of Saad & Solla. Focusing on the case of Gaussian data, we study the interplay between the learning rate, the time scale, and the number of hidden units in the high-dimensional dynamics of stochastic gradient descent (SGD). Our work builds on a deterministic description of SGD in high-dimensions from statistical physics, which we extend and for which we provide rigorous convergence rates.

P35

Data-driven separation of two-layer neural networks and random features

We compare shallow (two-layer) neural networks and random features approximations of kernel-ridge regression on tasks that rely on the information contained in the higher-order statistics of the inputs. We design synthetic models of data where we control the relative importance of higher-order cumulants and study in which settings end-to-end trained networks achieve better performance than random features. We further study the features of data that neural networks and kernels fit.

P36

Sharp Asymptotics of Self-training with Linear Classifier

Self-training (ST) is a straightforward and standard approach in semi-supervised learning, successfully applied to many machine learning problems. ST uses the model itself to give predictions on unlabeled data and then refines the model by fitting to these labels using supervised learning methods. This refinement step is iterated several times. The performance of ST strongly depends on the supervised learning method used in the refinement step and the nature of the given data; hence a general performance guarantee from a concise theory may become loose in a concrete setup. However, the theoretical methods that sharply predict how the performance of ST depends on various details for each learning scenario are scarce. This paper develops a novel theoretical framework for sharply characterizing the generalization abilities of the models trained by ST using the non-rigorous replica method of statistical physics. In particular, we consider the ST of the linear model that minimizes the ridge-regularized cross-entropy loss when the data is generated from a two-component Gaussian mixture. Consequently, we show that the generalization performance of ST in each iteration is sharply characterized by a small finite number of variables, which satisfy a set of deterministic self-consistent equations. Numerically solving these self-consistent equations, we find that ST's generalization performance approaches the supervised learning method with a simple regularization schedule when the label bias is small, and a moderately large number of iterations are used.

P37

Image-Based Algorithm for Vehicle Class Prediction in the University Of Ibadan, Nigeria

This study employed image based features to study the patterns in the vehicles that access the premises of the University of Ibadan, as well as to recognize them. Image processing techniques, such as the Red-Green-Blue scale to grayscale conversion and the Histogram Equalization were used to remove color differences and to improve the quality of the images respectively. The two-dimensional discrete wavelet decomposition technique was used to extract features (sub-bands) up to the third level of decomposition. The Multivariate Adaptive Regression Splines (MARS) was used for the predictive recognition. The algorithm recognized two optimal clusters in the vehicles. The algorithm was also able to recognize about 85 percent and 94 percent of cars in entry and exit. It also did fairly well in recognizing SUVs and Space-Buses, but very poorly on other classes of vehicles. However, being more specific about the identity of the vehicles, increasing the features and increasing the sample size will increase the recognizing ability of the algorithm.

P38

High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation

Jimmy Ba[1], Murat A. Erdogdu[1], Taiji Suzuki[2], Zhichao Wang[3], Denny Wu[1], Greg Yang[4]

[1]University of Toronto and Vector Institute. [2]University of Tokyo and RIKEN AIP.
[3]University of California, San Diego. [4]Microsoft Research.

We consider the following fully-connected two-layer neural network (NN) with N neurons,

$$f_{\text{NN}}(\boldsymbol{x}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{x}, \boldsymbol{w}_i\rangle) = \frac{1}{\sqrt{N}}\boldsymbol{a}^\top \sigma(\boldsymbol{W}^\top \boldsymbol{x}),$$

where $\boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{W} \in \mathbb{R}^{d\times N}, \boldsymbol{a} \in \mathbb{R}^N$, σ is a nonlinearity applied entry-wise, and the training objective is to minimize the empirical risk computed on n training points. Our analysis will be made in the *proportional asymptotic limit*, i.e., the number of training data n, the input dimensionality d, and the number of features (neurons) N jointly tend to infinity. In this regime, the performance of *random features* (RF) kernel regression (where $\boldsymbol{W}$ is randomly initialized and only $\boldsymbol{a}$ is learned) has been precisely characterized [LLC18, MM19].

While these asymptotic analyses reveal interesting phenomena also present in practical deep learning, RF and kernel models do not fully explain the empirical success of neural networks: one crucial advantage of deep learning is the *ability to learn useful features* that "adapt" to the learning problem [Suz18, GMMM19]. In our setting this adaptivity arises from the learning of first-layer weights $\boldsymbol{W}$, which can make a difference even in the "early phase" of gradient descent training. Our goal is to rigorously characterize the presence of feature learning in the proportional limit, and demonstrate its improvement over the initial (fixed) kernel.

Our Results. We investigate arguably the most simplified scenario of feature learning: how *the first gradient step* on the first-layer parameters $\boldsymbol{W}$ impacts the representation of the two-layer NN. Specifically, we consider the regression setting with the MSE loss, and a student-teacher model in the proportional asymptotic limit; we precisely characterize the prediction risk of the kernel ridge regression estimator on top of the first-layer CK feature $\boldsymbol{x} \to \sigma(\boldsymbol{W}^\top \boldsymbol{x})$, before and after the gradient step with learning rate η.

When the target function (teacher) f^* is a single-index model, it is known that the prediction risk of a large class of RF/kernel ridge regression estimators is lower-bounded by the L^2-norm of the "nonlinear" component the teacher $\|\mathsf{P}_{>1} f^*\|_{L^2}^2$, i.e., they can only learn *linear* functions on the input [BMR21]. After taking one gradient step on $\boldsymbol{W}$, we compute the CK ridge estimator using separate training data, and compare its prediction risk against this linear lower bound. Our analysis will be made under the following two choices of learning rate scaling (see Figure 1):

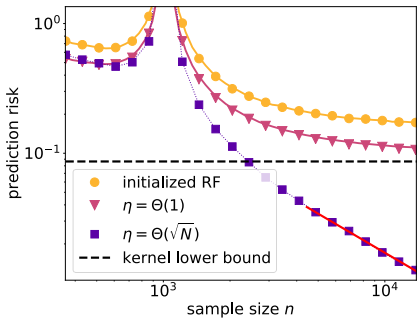


Figure 1: Prediction risk of ridge regression on trained features (erf) after one gradient step. Dots represent empirical simulations and solid lines are predicted asymptotic values; red line indicates $\Theta(d/n)$ rate.

- **Small lr:** $\eta = \Theta(1)$. We extend the *Gaussian Equivalence Theorem* (GET) in [HL20] to the feature map trained via multiple gradient descent steps on $\boldsymbol{W}$; this allows us to precisely characterize the prediction risk using random matrix theoretical tools. We prove that after one gradient step, the ridge regression estimator on the learned CK features already exhibits nontrivial improvement over the initial RF ridge model, but it remains in the "linear regime" and cannot outperform the best linear estimator on the input.
- **Large lr:** $\eta = \Theta(\sqrt{N})$. As for larger learning rate that coincides with the *maximal update parameterization* in [YH20], we prove that for certain target functions f^*, the kernel ridge regression estimator after one feature learning step can achieve lower risk than the linear lower bound $\|\mathsf{P}_{>1} f^*\|_{L^2}^2$, and thus outperform a wide range of RF and kernel models.

References

[BMR21] Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin, *Deep learning: a statistical viewpoint*, arXiv preprint arXiv:2103.09177 (2021).

[GMMM19] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari, *Limitations of lazy training of two-layers neural networks*, arXiv preprint arXiv:1906.08899 (2019).

[HL20] Hong Hu and Yue M Lu, *Universality laws for high-dimensional learning with random features*, arXiv preprint arXiv:2009.07669 (2020).

[LLC18] Cosme Louart, Zhenyu Liao, and Romain Couillet, *A random matrix approach to neural networks*, The Annals of Applied Probability **28** (2018), no. 2, 1190–1248.

[MM19] Song Mei and Andrea Montanari, *The generalization error of random features regression: Precise asymptotics and double descent curve*, arXiv preprint arXiv:1908.05355 (2019).

[Suz18] Taiji Suzuki, *Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality*, arXiv preprint arXiv:1810.08033 (2018).

[YH20] Greg Yang and Edward J Hu, *Feature learning in infinite-width neural networks*, arXiv preprint arXiv:2011.14522 (2020).

P39

A random energy approach to deep learning

We study a generic ensemble of Deep Belief Networks which is parametrized by the distribution of energy levels of the hidden states of each layer. We show that, within a random energy approach, statistical dependence can propagate from the visible to deep layers only if each layer is tuned close to the critical point during learning. As a consequence, efficiently trained learning machines are characterised by a broad distribution of energy levels. The analysis of Deep Belief Networks and Restricted Boltzmann Machines on different datasets confirms these conclusions.

P40

No poster

No poster

P41

Non-backtracking spectral clustering in sparse hypergraphs

The stochastic block model has been one of the most fruitful research topics in community detection and clustering. Recently, community detection on hypergraphs has become an important topic in higher-order network analysis. We consider the detection problem in a sparse random tensor model called the hypergraph stochastic block model (HSBM). We prove that a spectral method based on the non-backtracking operator for hypergraphs works with high probability down to the generalized Kesten-Stigum detection threshold conjectured by Angelini et al (2015). We characterize the spectrum of the non-backtracking operator for the sparse random hypergraph and provide an efficient dimension reduction procedure using the Ihara-Bass formula for hypergraphs. As a result, the community detection problem can be reduced to an eigenvector problem of a non-normal matrix constructed from the adjacency matrix and the degree matrix of the hypergraph. Based on joint work with Ludovic Stephan (EPFL).