Double-Descent Curves in Neural Networks: A New Perspective Using Gaussian Processes

Ouns El Harzli[1] *, **Bernardo Cuenca Grau**[1], **Guillermo Valle-Pérez** [2], **Ard A. Louis** [2]

[1]Department of Computer Science, University of Oxford
[2]Rudolf Peierls Centre for Theoretical Physics, University of Oxford

Abstract

Double-descent curves in neural networks describe the phenomenon that the generalisation error initially descends with increasing parameters, then grows after reaching an optimal number of parameters which is less than the number of data points, but then descends again in the overparameterized regime. In this paper, we use techniques from random matrix theory to characterize the spectral distribution of the empirical feature covariance matrix as a width-dependent perturbation of the spectrum of the neural network Gaussian process (NNGP) kernel, thus establishing a novel connection between the NNGP literature and the random matrix theory literature in the context of neural networks. Our analytical expressions allow us to explore the generalisation behavior of the corresponding kernel and GP regression. Furthermore, they offer a new interpretation of double-descent in terms of the discrepancy between the width-dependent empirical kernel and the width-independent NNGP kernel.

Introduction

Deep learning has achieved unparalleled success across a wide range of tasks (Krizhevsky, Sutskever, and Hinton 2012; Hannun et al. 2014; LeCun, Bengio, and Hinton 2015; Schmidhuber 2015). Surprisingly, however, the best-performing deep neural networks (DNNs) operate in a highly over-parametrised regime, where the number of parameters in the model is much larger than the number of training examples (Simonyan and Zisserman 2014). This appears to violate the conventional statistical wisdom of bias-variance trade-off which predicts that, in order to avoid overfitting and obtain the best possible generalisation, the number of parameters should be lower than the number of training examples (Mohri, Rostamizadeh, and A.Talwalkar 2012; Vapnik 1995; Shalev-Shwartz and Ben-David 2014).

The generalisation error of DNNs as a function of the number of parameters has been studied empirically (Belkin et al. 2019; Nakkiran et al. 2021), and it has been observed that it follows a double-descent curve instead of the classical U-shaped curve characteristic of the bias-variance trade-off. Specifically, for a fixed number of training examples, the generalisation error increases as the number of parameters approaches the number of training examples and, past this so-called *interpolation threshold*, it starts decreasing again finding its global minimum when the number of parameters goes to infinity. Understanding these surprising observations at a more fundamental level is an important step towards tackling the deeper question as to *why* DNNs generalise so well in practice (Geiger, Petrini, and Wyart 2020).

A number of mathematical frameworks for explaining the double-descent phenomenon in a variety of DNN architectures have been proposed (Mei and Montanari 2019; Hastie et al. 2019; Advani, Saxe, and Sompolinsky 2020; Liu, Liao, and Suykens 2020). Assuming a teacher-student setting (Seung and Sompolinsky 1992), these works derive an analytical expression for the generalisation error as a function of the ratios $\gamma = \frac{n}{N}$ between the number n of training examples and the width N of the neural network and $\psi = \frac{n}{d}$ between the number of training examples and the dimension d of the input. For a fixed value of ψ and varying γ, the generalisation error derived in these approaches follows a double descent curve; furthermore, the value of the generalisation error for a given γ and ψ is obtained as a limit where n, N and d go to infinity while γ and ψ remain constant. These works typically assume that all layers except the last one remain untrained, and thus coincide with random features models (Liao, Couillet, and Mahoney 2020; Gerace et al. 2020).

A (largely orthogonal) line of research has studied the equivalence between infinitely-wide neural networks with random weights and Gaussian processes with a particular covariance function (Lee et al. 2018; Matthews et al. 2018) called the *NNGP (or conjugate) kernel*. In the limit of infinite width, the class of functions obtained by choosing the network weights randomly converges in distribution to a Gaussian process, where the structure of the covariance function is defined inductively on the number of layers. The random feature map corresponding to an NNGP is, however, not explicitly known and thus its study falls within the realm of the nonparametric kernel literature (Györfi et al. 2002; Berthier, Bach, and Gaillard 2020).

In this paper, we establish a novel connection between both of these orthogonal lines of research. Given a fully-connected neural architecture defined by the input dimension, the layer widths, the activation function, and the distribution of the weights, we derive an analytical expression for a *width-dependent NNGP kernel* which generalises the *em-*

*Corresponding author: ouns.elharzli@new.ox.ac.uk

pirical covariance matrices by (Mei and Montanari 2019) to a kernel function. We then exploit elements of random matrix theory (Wigner 1955; Livan, Novaes, and Vivo 2018) to analytically compute, given a set of training examples, the spectral distribution of (the covariance matrix associated to) our width-dependent NNGP kernel. Although similar expressions have been computed in prior work (Louart, Liao, and Couillet 2018; Fan and Wang 2020), ours is unique in that it is given as a function of the spectral distribution of the NNGP kernel. As a result, our expression enables a new interpretation of the width-dependent spectral distribution as a perturbation of the width-independent spectral distribution that decreases as γ tends to zero so that, in the limit of infinite wdith, the width-dependent spectral distribution converges to that of an NNGP. The kernel function and the the analytical formula for the spectral distribution that we propose allow us compute the generalisation error of both GP and kernel regression as a function of γ and ψ. Similarly to (Mei and Montanari 2019; Hastie et al. 2019; Advani, Saxe, and Sompolinsky 2020; Bordelon, Canatar, and Pehlevan 2020; Canatar, Bordelon, and Pehlevan 2021) the value of the generalisation error at γ and ψ is computed as a limit where n, N and d go to infinity while the ratios γ and ψ remain constant. Furthermore, for a fixed ψ and varying γ, the generalisation error exhibits double descent behaviour. Our generalisation error expression is novel in that it isolates the term that contributes to the double-descent behavior.

Our approach requires only mild assumptions on the neural architecture and the data generating process. The target function for learning is assumed to have a bounded second moment with respect to the data distribution and we only require mild regularity assumptions (measurability and Lipschitzianity) for the activation functions, which are satisfied by all commonly-used activations functions.

Our results thus provide a new interpretation to the double-descent phenomenon where the behaviour of the generalisation error is governed by the discrepancy between the width-dependent empirical kernel characterising the network's architecture and the width-independent NNGP kernel of the limit Gaussian process.

Preliminaries

We next introduce the basic concepts underpinning our technical results. Throughout the paper, we denote matrices by bold uppercase letters and vectors by bold lowercase letters.

Random matrix theory Random matrix theory (Wigner 1955; Livan, Novaes, and Vivo 2018) is the study of the spectral distributions of large matrices whose elements are random variables. The spectral measure F_n of a given matrix with eigenvalues λ_i is a measure over $x \in \mathbb{R}$ given by $F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i}(x)$, where $\delta_{\lambda_i}(x)$ is the Dirac measure at an eigenvalue λ_i. When the matrix is random, the spectral measure becomes a random measure, often referred to as the empirical spectral distribution. We study weak convergences (convergences in distribution) of spectral measures to nonrandom measures (Geronimo and Hill 2002). A useful tool to manipulate spectral measures is the Stieltjes transform; for a measure F supported on the real interval I, the Stieltjes transform is given as follows for each $z \in \mathbb{C} - I$: $S_F(z) = \int_I \frac{dF(\lambda)}{\lambda - z}$. There is a one-to-one correspondence between measures and their Stieltjes transforms, as per the inversion formula: $F(x) = \lim_{y\to 0^+} S_F(x+iy)$ for $x \in I \setminus \{0\}$. Pointwise convergence in their Stieltjes transforms ensures weak convergence of measures.

We will rely on a famous result in random matrix theory. Consider $\mathbf{X} \in \mathbb{R}^{N\times n}$, a random matrix with i.i.d. entries drawn from $\mathcal{N}(0, \frac{1}{N})$ and Ψ a nonrandom positive semi-definite matrix. Suppose that Ψ has a limiting spectral measure μ, and let $n, N \to \infty$ with fixed ratio $\gamma := \frac{n}{N}$, then the random matrix $\mathbf{\Psi}^{1/2}\mathbf{X}^T\mathbf{X}\mathbf{\Psi}^{1/2}$ has a limiting nonrandom spectral measure $\rho^{\gamma}_{MP} \boxtimes \mu$. The measure $\rho^{\gamma}_{MP} \boxtimes \mu$ is defined by its Stieltjes transform S, which solves the Marchenko-Pastur fixed-point equation (Marchenko and Pastur 1967):

$$S(z) = \int \frac{1}{x\left(1-\gamma-\gamma z S(z)\right) - z} d\mu(x). \tag{1}$$

The measure $\rho^{\gamma}_{MP} \boxtimes \mu$ is called the Marchenko-Pastur map of μ. In the particular case $\mathbf{\Psi} = \mathbf{I}_n$, μ is the Dirac measure at 1, and one recovers the Marchenko-Pastur distribution ρ^{γ}_{MP}.

Neural network Gaussian processes A Gaussian process f over a space $\mathbb{R}^d$ is a random scalar field such that its evaluation at any collection of finitely many points $(f(x_1), ..., f(x_n))$ follows a multivariate Gaussian. A Gaussian process is determined by a mean function $\mu : \mathbb{R}^d \mapsto \mathbb{R}$, and a covariance function $K : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$, which describe respectively the mean of the Gaussian distribution at each point and the covariance between the Gaussians at any two points. For Gaussian processes, the covariance function is a kernel, i.e. a positive semi-definite symmetric function (Rasmussen and Williams 2006). We note $f \sim \mathcal{GP}(\mu, K)$.

We consider a random fully-connected neural network (FCN) with zero bias as in (Mei and Montanari 2019)):

$$\mathbf{x}^l := \phi(\mathbf{h}^l) \qquad \mathbf{h}^l := \mathbf{W}^l\mathbf{x}^{l-1} \qquad \forall l \in 1, ..., L \tag{2}$$

where $N_0 := d$ is the dimension of the input space, $x^0 \in \mathbb{R}^d$ is an arbitrary input, N_l is the width of the l-th layer, $\mathbf{h}^l := \mathbf{h}^l(\mathbf{x_0})$ is the preactivation of the l-th layer, the weight matrices $\mathbf{W}^l \in \mathbb{R}^{N_l \times N_{l-1}}$ have entries drawn i.i.d. from a Gaussian distribution $\mathcal{N}(0, \frac{1}{N_{l-1}})$, and ϕ is an arbitrary nonlinear activation function acting componentwise.

Applying successively the central limit theorem to each layer, the infinite-width limit of (2) yields a Gaussian process, called the *Neural Network Gaussian Process (NNGP)*. More precisely, if we let $N_0, ..., N_{L-1} \to \infty$, the $\mathbf{h}^L_i \sim \mathcal{GP}(\mu^L, K^L)$ are independent and defined inductively by layers as follows for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ and each $l \in 1, ..., L$:

$$\mu^l(\mathbf{x}) = 0 \qquad K^{\phi,0}(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}' \tag{3}$$

$$\begin{aligned} &\mathbf{h}^{l-1}_i \sim \mathcal{GP}(\mu^{l-1}, K^{\phi,l-1}) \\ &K^{\phi,l}(\mathbf{x}, \mathbf{x}') = \mathbb{E}_h\left(\phi(\mathbf{h}^{l-1}_i(\mathbf{x}))\phi(\mathbf{h}^{l-1}_i(\mathbf{x}'))\right) \end{aligned} \tag{4}$$

The covariance function $K^{\phi,L}$ is called the *NNGP kernel* or conjugate kernel (Daniely, Frostig, and Singer 2016), which is determined by the network depth L and the activation function ϕ; when these are clear, we simply denote it as K.

There is potential for many subtleties in the way the infinite-width limits are approached (Matthews et al. 2018), and we follow the approach in (Lee et al. 2018) where infinite limits taken sequentially. The recursive formulae for the NNGP kernel have also been determined by (Poole et al. 2016) in the context of mean-field theory of random neural networks. We will use some of their techniques in our proofs.

The equivalence between randomly-initialised neural networks with infinite width and Gaussian processes with particular covariance functions is a general result which has been first established in (Neal 1994) and revisited recently in (Lee et al. 2018; Matthews et al. 2018; Novak et al. 2018; Garriga-Alonso, Rasmussen, and Aitchison 2018; Yang 2019; Lee et al. 2020) and others.

Problem setup We consider a teacher-student setting where $\mathbf{x} \sim \mathbb{P}_d$, $\tau \sim \mathcal{N}(0, {\sigma_\tau}^2)$ and $y = f_d(\mathbf{x}) + \tau$, with d the input space dimension, $\mathbf{x} \in \mathbb{R}^d$ the input feature vector, τ the noise, y the output value, $(\mathbb{P}_d)_{d\in\mathbb{N}}$ a family of probability distributions over $\mathbb{R}^d$ such that $\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_d}(\mathbf{x}^T\mathbf{x})$ is bounded as $d \to \infty$; and $f_d : \mathbb{R}^d \longrightarrow \mathbb{R}$ a family of functions verifying that $\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_d}\left(f_d(\mathbf{x})^2\right)$ is bounded as $d \to \infty$. These assumptions, also taken in (Bordelon, Canatar, and Pehlevan 2020; Canatar, Bordelon, and Pehlevan 2021), are quite mild as they only exclude pathological behaviors where the variances of the input or the output explode at infinity.

We sample a number n of examples, i.i.d. from the teacher. The training set then consists of:

$$\begin{aligned} \mathbf{X} &= (\mathbf{x}_1, ..., \mathbf{x}_n)^T \in \mathbb{R}^{n\times d} \\ \mathbf{y} &= f_d(\mathbf{X}) + \mathbf{t} = (y_1, ..., y_n)^T \in \mathbb{R}^n \end{aligned} \tag{5}$$

where $f_d(\mathbf{X}) = (f_d(\mathbf{x}_1), ..., f_d(\mathbf{x}_n))^T \in \mathbb{R}^n$ and $\mathbf{t} = (\tau_1, ..., \tau_n)^T \in \mathbb{R}^n$ is a noise sample. We study the generalisation error of particular kernel regressions and Gaussian process regressions trained on this data.

Gaussian process and kernel regression Consider an NNGP $z \sim \mathcal{GP}(0, K)$, where K is the NNGP kernel obtained with the infinite-width limit of equation (2), and taking the width of the output layer to be 1, thus yielding an output $z(\mathbf{x}) \in \mathbb{R}$. Standard Bayesian inference in Gaussian process regression (Rasmussen and Williams 2006) gives us that the predicted $\bar{z}$, conditional on $(\mathbf{X}, \mathbf{y}, \mathbf{x})$ follows a Gaussian distribution $\bar{z}|\mathbf{X}, \mathbf{y}, \mathbf{x} \sim \mathcal{N}(\bar{\mu}, \bar{K})$ where:

$$\bar{\mu}_{\mathbf{X},\mathbf{y},\mathbf{x}} = \mathbf{k}_{\mathbf{x},\mathbf{X}}^T(\mathbf{K}_{\mathbf{X},\mathbf{X}} + \sigma_\epsilon^2\mathbf{I}_n)^{-1}\mathbf{y} \tag{6}$$

$$\bar{K}_{\mathbf{X},\mathbf{y},\mathbf{x}} = K(\mathbf{x},\mathbf{x}) - \mathbf{k}_{\mathbf{x},\mathbf{X}}^T(\mathbf{K}_{\mathbf{X},\mathbf{X}} + \sigma_\epsilon^2\mathbf{I}_n)^{-1}\mathbf{k}_{\mathbf{x},\mathbf{X}} \tag{7}$$

if we assume a noise model $z|\mathbf{y} \sim \mathcal{N}(z, \sigma_\epsilon^2)$, and $\mathbf{K}_{\mathbf{X},\mathbf{X}} \in \mathbb{R}^{n\times n}$ with $(\mathbf{K}_{\mathbf{X},\mathbf{X}})_{i,j} := K(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{k}_{\mathbf{x},\mathbf{X}} \in \mathbb{R}^n$ with $(\mathbf{k}_{\mathbf{x},\mathbf{X}})_i := K(\mathbf{x}, \mathbf{x}_i)$. Furthermore, the mean prediction $\bar{\mu}_{\mathbf{X},\mathbf{y},\mathbf{x}}$ of GP regression is also the solution of kernel ridge regression with the same kernel K and ridge parameter σ_ϵ^2 (Rasmussen and Williams 2006).

Double Descent from the Perspective of Gaussian Processes

Previous studies of double descent in neural networks using random matrix theory (Mei and Montanari 2019; Hastie et al. 2019; Advani, Saxe, and Sompolinsky 2020; Liu, Liao, and Suykens 2020) took the network width N to infinity, along with the dimension of the input space, and the number of training examples. This motivated us to study double-descent with NNGPs.

We first characterise a counterpart of the NNGP kernel with finite width since a well-defined kernel is a prerequisite to leverage the theories of GP regression and kernel regression. We were able to identify the random kernel underpinning the empirical covariance matrix by assuming that the width of the last layer is finite; this alone will allow us to derive insights on the double-descent phenomenon.

Then, we study the limiting spectral distribution of the empirical covariance matrix of features. We derive a non-trivial relationship between the spectral distributions of the empirical NNGP kernel random matrix and the actual NNGP kernel random matrix.

Finally, we isolate the dependency to these spectral distributions in the generalisation errors of GP regression and kernel regression. This allows us to interpret the double-descent phenomenon concerning γ, attributing it to the spectrum of the related random kernel being a perturbation of the NNGP kernel limit's spectrum. The degree of this perturbation varies depending on γ and this variation is mirrored in the spectral distributions.

A Width-dependent Random Kernel

Consider $z \sim \mathcal{GP}(0, K)$ obtained with the infinite width limit of (2). Finding a dependence with the width N is not straightforward in the case of NNGPs because, at this point, the network width has already been taken to infinity. Our idea is therefore to study the behavior of a counterpart of the Gaussian process z *before* the width is taken to infinity.

Following (Lee et al. 2018), we consider the output $h^{L,N}(\mathbf{x}_i) = \sum_{k=1}^N \mathbf{W}_k^L \phi\left(\mathbf{h}_k^{L-1}(\mathbf{x}_i)\right)$ of a random network 2 with $L \geq 2$, $N_1, ..., N_{L-2} = \infty$, $N_{L-1} = N$ and $N_L = 1$, i.e. where all widths have been taken to infinity with the exception of the last one.

Proposition 1. *The covariance matrix of the evaluations of $h^{L,N}$, conditional on the pre-activations, satisfies, for all pairs of training data points $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d$ (rows of $\mathbf{X}$), that:*

$$\begin{aligned} \left(\mathbf{K}^N_{\mathbf{X},\phi,\mathbf{h}^{L-1}}\right)_{i,j} &= \mathbb{E}_{\mathbf{W}^L}\left(h^{L,N}(\mathbf{x}_i)h^{L,N}(\mathbf{x}_j)|\{\mathbf{h}_k^{L-1}\}_{k=1}^N\right) \\ &= \frac{1}{N}\sum_{k=1}^N \phi\left(\mathbf{h}_k^{L-1}(\mathbf{x}_i)\right)\phi\left(\mathbf{h}_k^{L-1}(\mathbf{x}_j)\right) \end{aligned}$$

where the expectation is thus taken over the last-layer weights $\mathbf{W}^L$, and it is an unbiased estimator of $K(\mathbf{x}_i, \mathbf{x}_j)$ with variance $\mathrm{Var}\left(\left(\mathbf{K}^N_{\mathbf{X},\phi,\mathbf{h}^{L-1}}\right)_{i,j}\right) = \mathcal{O}_{N\to\infty}(\frac{1}{N})$.

The random matrix $\mathbf{K}^N_{\mathbf{X},\phi,\mathbf{h}^{L-1}}$ is the empirical covariance matrix of the features created by the NNGP (2) before the last width is taken to infinity.

Conditionally on $\mathbf{X}$, the values $K(\mathbf{x}_i, \mathbf{x}_j)$ are constant and the $\left(\mathbf{K}^N_{\mathbf{X},\phi,\mathbf{h}^{L-1}}\right)_{i,j}$ are random variables whose randomness stems from $\mathbf{h}^{L-1}$. In turn, $\mathbf{K}^N_{\mathbf{X},\phi,\mathbf{h}^{L-1}}$ satisfies the

kernel property (Rasmussen and Williams 2006):

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \left(\mathbf{K}^N_{\mathbf{X},\phi,\mathbf{h}^{L-1}}\right)_{i,j} = \frac{1}{N}\mathbf{a}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{a} \geq 0 \quad (8)$$

for all $\mathbf{a} = (a_1, ..., a_n)^T \in \mathbb{R}^n$, where $\mathbf{\Phi} \in \mathbb{R}^{N\times n}$ with $\mathbf{\Phi}_{jk} = \phi\left(\mathbf{h}_j^{L-1}(\mathbf{x}_k)\right)$, which holds for any realisation of the random matrix $\mathbf{\Phi}$. Note, however, that $\mathbf{K}^N_{\mathbf{X},\phi,\mathbf{h}^{L-1}}$ does not define a kernel as it is not a well-defined function of $(\mathbf{x}, \mathbf{x}')$ but merely a countable family of random variables that can be indexed on $(\mathbf{x}, \mathbf{x}')$. This is problematic in our setting since the covariance function in a Gaussian process must be a kernel with respect to the full, continuous, space.

We next propose a way of converting the aforementioned family of random variables into a random kernel, i.e. a kernel-valued random variable.

Theorem 1. *For $N \in (1, \infty)$, and $L \geq 2$ there exists a probability space $(\Omega_N, \mathcal{A}_N, \mathbb{P}_N)$ and a random variable $K^{\phi,L,N} : \Omega_N \longrightarrow \mathbb{R}^{(\mathbb{R}^d)^2}$ with image in the functional space $\mathbb{R}^{(\mathbb{R}^d)^2}$ such that :*

1. *$K^{\phi,L,N}(\omega)$ is a kernel for all $\omega \in \Omega_N$,*
2. *for all sets of points $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T \in \mathbb{R}^{n\times d}$, the random matrix $\mathbf{K}^N_{\mathbf{X},\phi,\mathbf{h}^{L-1}}$, and the random matrix $\mathbf{K}^{\phi,L,N}_{\mathbf{X},\mathbf{X}}$, defined as,*

$$\Omega_N \longrightarrow \mathbb{R}^{n\times n} \qquad \omega \mapsto \left(K^{\phi,L,N}(\omega)(\mathbf{x}_i, \mathbf{x}_j)\right)_{ij\in[n]}$$

follow the same distribution: for all $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d$, $\mathbb{E}_{K^{\phi,L,N}}\left(K^{\phi,L,N}(\mathbf{x}_i, \mathbf{x}_j)\right) = K^{\phi,L}(\mathbf{x}_i, \mathbf{x}_j)$, where the expectation is taken over the random kernel $K^{\phi,L,N}$.

We have thus defined a random variable $K^{\phi,L,N}$ over a functional space, whose realisations are kernel functions interpolating the random matrices of interest. When there is no ambiguity, we use K^N to denote $K^{\phi,L,N}$. We can now study the random matrices $\mathbf{K}^N_{\mathbf{X},\mathbf{X}}$, whose randomness stems from the random kernel function K^N and the random matrix $\mathbf{X}$, using the more convenient definition of $\mathbf{K}^N_{\mathbf{X},\phi,\mathbf{h}^{L-1}}$, whose randomness stems from the random variables $\mathbf{h}^{L-1}$ and the random matrix $\mathbf{X}$. Conditionally on K^N, the corresponding Gaussian process $z_{K^N} \sim \mathcal{GP}(0, K^N)$ is well-defined and Bayesian inference can be performed with equations (6-7).

Limiting Spectral Distributions

The following theorem relates the limiting spectral distribution of the *actual* NNGP kernel random matrix $\mathbf{K}_{\mathbf{X},\mathbf{X}}$ and the *empirical* NNGP kernel random matrix $\mathbf{K}^N_{\mathbf{X},\mathbf{X}}$.

Theorem 2. *Consider an NNGP obtained with the infinite-width limit of (2) with $L \geq 2$, $N_L = 1$ and the non-linear activation ϕ, a measurable, Lipschitz function. Consider the associated NNGP kernel denoted K, the associated random kernel function K^N and the random matrix $\mathbf{K}^N_{\mathbf{X},\mathbf{X}}$ defined by Theorem 1 for kernel K. Then, the random matrix $\mathbf{K}_{\mathbf{X},\mathbf{X}}$ admits, in the limit $n, d \to \infty$ with fixed ratio $\frac{n}{d} = \psi \in (0, \infty)$, a limiting nonrandom spectral measure μ^ϕ_ψ. Furthermore, in the limit $N, n, d \to \infty$ with fixed ratio $\frac{n}{N} = \gamma \in (0, \infty)$, $\frac{n}{d} = \psi \in (0, \infty)$, the empirical spectral distribution of $\mathbf{K}^N_{\mathbf{X},\mathbf{X}}$ converges in distribution to the non-random measure $\rho^\gamma_{MP} \boxtimes \mu^\phi_\psi$.*

The proof of Theorem 2, given in the Appendix, relies on a recent result in random matrix theory (Theorem 1 in (Banna, Merlevede, and Peligrad 2015)). As a corollary, for deep linear networks, if the data covariance matrix $\mathbf{X}\mathbf{X}^T$ admits a limiting spectral distribution μ_ψ, then the limiting spectral distribution of $\mathbf{K}^N_{\mathbf{X},\mathbf{X}}$ is $\rho^\gamma_{MP} \boxtimes \mu_\psi$. In particular, if the covariance matrix is isotropic, then $\mu_\psi = \rho^\psi_{MP}$ and the limiting spectral distribution is the Marchenko-Pastur map of a Marchenko-Pastur distribution $\rho^\gamma_{MP} \boxtimes \rho^\psi_{MP}$.

Here, we have made an important distinction between the random matrices $\mathbf{K}^N_{\mathbf{X},\mathbf{X}}$ and $\mathbf{K}_{\mathbf{X},\mathbf{X}}$, which was not made in previous works (Fan and Wang 2020). Indeed, it is not the same thing to consider the NNGP kernel K, which appears *after* the width of a neural network is taken to infinity, and its counterpart K^N *before* the width is taken to infinity (which should be called the *empirical* NNGP kernel).

Theorem 2 tells us how the spectral distribution of the empirical covariance matrix of the features created by the neural network (2) depends on the *actual* conjugate kernel of its associated NNGP.

The important fact to notice for the interpretation of the double-descent curve in neural networks is that, in the extremely overparametrised regime $\gamma \to 0$, the spectral distribution becomes that of the NNGP kernel itself. Indeed, the fixed-point equation (1), which characterises the Marchenko-Pastur map of μ^ϕ_ψ, becomes:

$$S(z) = \int \frac{1}{x - z} d\mu^\phi_\psi(x) = S_{\mu^\phi_\psi}(z). \quad (9)$$

To put it differently from a spectral perspective, the neural network exhibits behavior akin to its corresponding NNGP in the highly overparametrised regime. Subsequently, we explore how the generalization error of the corresponding GP and kernel regressions hinges on this spectral distribution and mirrors the double-descent pattern.

Double Descent in NNGPs and Kernel Regression

We are now in position to study the generalisation error of the corresponding Gaussian process and kernel regressions. We will calculate the generalisation error of kernel regression with kernel K^N:

$$E_{\mathcal{K}}(K^N) := \mathbb{E}_{\mathbf{x},y,\mathbf{X},\mathbf{y}}\left((\bar{\mu}^{K^N}_{\mathbf{X},\mathbf{y},\mathbf{x}} - y)^2\right) \quad (10)$$

and the generalisation error of Gaussian process regression with GP z_{K^N}:

$$E_{\mathcal{GP}}(K^N) := \mathbb{E}_{\mathbf{x},\mathbf{X},\mathbf{y}}\left(\mathbb{E}_{y,\bar{z}_{K^N}}\left((\bar{z}_{K^N} - y)^2 | \mathbf{X}, \mathbf{y}, \mathbf{x}\right)\right) \quad (11)$$

where $\bar{\mu}^{K^N}_{\mathbf{X},\mathbf{y},\mathbf{x}}$ is the prediction mean of the Gaussian process regression with prior z_{K^N}, $\bar{z}_{K^N}|\mathbf{X}, \mathbf{y}, \mathbf{x}$ is the posterior distribution of Gaussian process regression with prior z_{K^N}, and the expectations are taken over the out-of-sample data and

the training samples. These predictions depend on the realisation of the kernel function K^N. We study these generalisation errors when all quantities go to infinity and averaging over the random kernel using $n = \gamma N$, and $d = \frac{\gamma N}{\psi}$:

$$\begin{aligned} E_{\mathcal{K}}(\gamma, \psi) &:= \lim_{N\to\infty} \mathbb{E}_{K^N}\left(E_{\mathcal{K}}(K^N)\right) \\ E_{\mathcal{GP}}(\gamma, \psi) &:= \lim_{N\to\infty} \mathbb{E}_{K^N}\left(E_{\mathcal{GP}}(K^N)\right) \end{aligned} \tag{12}$$

The following theorem highlights the dependence of the generalisation errors with some terms of interest that solely depend on the spectral measure that we studied in the previous section. The limits of these spectral measures will give us the double-descent behavior.

Theorem 3. *Under the same assumptions as in our Theorem 2, the limiting generalisation errors $E_{\mathcal{K}}(\gamma, \psi)$ and $E_{\mathcal{GP}}(\gamma, \psi)$ can be expressed:*

$$\begin{aligned} E_{\mathcal{K}}(\gamma, \psi) = D(\gamma,\psi) + C(\gamma, \psi)g(\gamma, \psi)^2 \\ + B(\gamma, \psi)g_2(\gamma, \psi) + A(\gamma, \psi)g(\gamma, \psi) \end{aligned} \tag{13}$$

$$\begin{aligned} E_{\mathcal{GP}}(\gamma, \psi) = \bar{D}(\gamma,\psi) + C(\gamma, \psi)g(\gamma, \psi)^2 \\ + B(\gamma, \psi)g_2(\gamma, \psi) + \bar{A}(\gamma, \psi)g(\gamma, \psi) \end{aligned} \tag{14}$$

where:

$$\begin{aligned} g(\gamma, \psi) &:= \int_0^\infty \frac{1}{\lambda + \sigma_\epsilon^2} \mathrm{d}(\rho_{MP}^\gamma \boxtimes \mu_\psi^\phi)(\lambda) \\ g_2(\gamma, \psi) &:= \int_0^\infty \frac{1}{(\lambda + \sigma_\epsilon^2)^2} \mathrm{d}(\rho_{MP}^\gamma \boxtimes \mu_\psi^\phi)(\lambda) \end{aligned} \tag{15}$$

and $A(\gamma, \psi), \bar{A}(\gamma, \psi), B(\gamma, \psi), C(\gamma, \psi), D(\gamma, \psi), \bar{D}(\gamma, \psi)$ are bounded with respect to $n, N, d \to \infty$, and $B(\gamma, \psi)$ is non-zero.

The proof of Theorem 3, which is provided in the Appendix, relies on the diagonalisation of the kernel random matrix $\mathbf{K}^N_{\mathbf{X},\mathbf{X}}$ and exploits Theorem 2 to compute the expectation of the inverse of the eigenvalues in the limit of infinite quantities. Note that we do not need to compute the coefficients A, B, C and D analytically. Indeed, our objective is not to provide a fine-grained analysis of the generalisation error, for which other expressions are already available in the literature ((Jacot et al. 2020; Canatar, Bordelon, and Pehlevan 2021; Simon et al. 2022) and others), but rather to isolate the terms that contribute to the coarse-grained double-descent behavior, in order to provide a neat interpretation. Indeed, the boundedness of A, B, C and D and the boundedness away from zero of B are sufficient to capture the double descent behaviour as per the following corollary.

Corollary 1. *Suppose that the assumptions of Theorem 2 hold true. Then, in the limit of $\sigma_\epsilon \to 0$ (noise-free), the generalisation error $E_{\mathcal{K}}(\gamma, \psi)$ exhibits a double descent with respect to γ. More precisely, the asymptote for the under-parametrised regime is given by:*

$$\lim_{\gamma\to\infty} E_{\mathcal{K}}(\gamma, \psi) = D(\gamma, \psi). \tag{16}$$

The asymptote for the interpolation threshold is given by:

$$\lim_{\gamma\to 1^-} E_{\mathcal{K}}(\gamma, \psi) = \infty \qquad \lim_{\gamma\to 1^+} E_{\mathcal{K}}(\gamma, \psi) = \infty. \tag{17}$$

Finally the asymptote $\lim_{\gamma\to 0} E_{\mathcal{K}}(\gamma, \psi)$ for the over-parametrised regime is finite and given by

$$\begin{aligned} D(\gamma,\psi) + C(\gamma, \psi)\left(\int_0^\infty \frac{1}{\lambda}\mathrm{d}(\mu_\psi^\phi)(\lambda)\right)^2 \\ + B(\gamma, \psi)\int_0^\infty \frac{1}{\lambda^2}\mathrm{d}(\mu_\psi^\phi)(\lambda) \\ + A(\gamma, \psi)\int_0^\infty \frac{1}{\lambda}\mathrm{d}(\mu_\psi^\phi)(\lambda) \end{aligned}$$

The result also holds for $E_{\mathcal{GP}}(\gamma, \psi)$, replacing $A(\gamma, \psi), D(\gamma, \psi)$ by $\bar{A}(\gamma, \psi), \bar{D}(\gamma, \psi)$.

The possibility of convergence to a finite value in the over-parametrised regime is enabled by the behavior of the Marchenko-Pastur map, as explained by equation (9). Indeed, the empirical spectral distribution converges to that of the *actual* NNGP kernel matrix. The divergence at the interpolation threshold is due to eigenvalues becoming arbitrarily close to zero, due to a structural property independent of the input data distribution: the strictly positive support of the nonrandom measure $\rho_{MP}^\gamma \boxtimes \mu_\psi^\phi$ becomes arbitrarily close to zero when $\gamma \to 1$. In practice, the divergence is reduced by the effects of regularisation (in the case of NNGP regression, the noise model). More details are given in the Appendix.

Experiments

In this section, we provide empirical evidence demonstrating the accuracy of our predictions of the spectral distribution of NNGP kernel random feature matrices, as well as the manifestation of the double-descent phenomenon in the generalisation error of NNGP kernel regression. These experiments were carried out using GPU resources on Google Colab.

We have simulated the empirical spectral distribution of the kernel random matrix $\mathbf{K}^{\hat{N}}_{\mathbf{X},\mathbf{X}}$ for high values of N, n, d for ReLU and tanh with both a synthetic dataset with data drawn from an isotropic multivariate Gaussian distribution $\mathbb{P}_d = \mathcal{N}(0, \frac{1}{d} I_d)$, and the MNIST dataset (LeCun 2012).

As illustrated in Figure 1, we find excellent agreement with the theoretical prediction of the limiting spectral distributions. We used the Marchenko-Pastur fixed point equation (1) to compute the limiting spectral distribution $\rho_{MP}^\gamma \boxtimes \mu_\psi^\phi$, by iterating over the recursive sequence it defines in the Stieltjes transform space and then inverting the Stieljes transform using the inversion formula. In the case of synthetic data drawn from $\mathbb{P}_d = \mathcal{N}(0, \frac{1}{d} I_d)$ and with no nonlinearity, the actual NNGP kernel matrix can be characterised exactly by $\mu_\psi^\phi = \rho_{MP}^\psi$. In general, the actual NNGP kernel is not known, hence we estimated the actual NNGP kernel matrix by sampling $\mathbf{K}^{\hat{N}}_{\mathbf{X},\mathbf{X}}$ with a very large value of $\hat{N} \gg n$,relying on the fact that $\rho_{MP}^0 \boxtimes \mu_\psi^\phi = \mu_\psi^\phi$. We focused on a subset of MNIST restricted to digits "0" and "1" in order to simplify the structure of the covariance matrices and their spectral distributions.

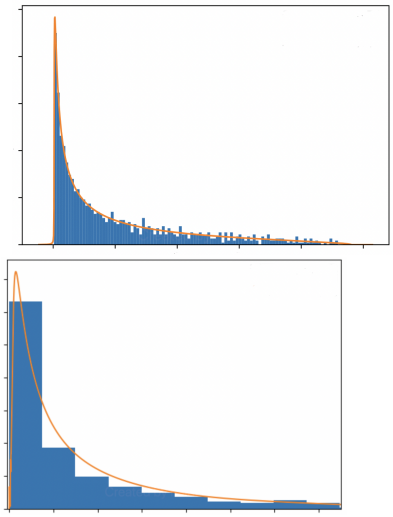

Figure 1: Simulated empirical spectral distribution (blue histograms) versus theoretical limiting spectral distribution (orange curves) of the empirical covariance matrix $\mathbf{K}^N_{\mathbf{X},\mathbf{X}}$. The X-axis are indexed by the eigenvalues of the kernel random matrices by increments of 1 and 2 for the top and bottom figure respectively, and the Y-axis are indexed by the probability density by increments of 0.5 and 0.05 for the top and bottom figures, respectively. Top, we use a two-layer NNGP without non-linearities with teacher distribution $\mathcal{N}(0, \frac{1}{d} I_d)$ using $N = 300$, $n = 200$, and $d = 400$. Bottom, we use a two-layer ReLU NNGP on a subset of MNIST taking $N = 600$, $n = 300$, and $d = 784$ (pixels on MNIST images). The simulated distribution is obtained by sampling from the random matrix, and the theoretical distibution is obtained by solving the Marchenko-Pastur equation.

We have simulated the generalisation errors of NNGP kernel regression on the same datasets. To calculate the generalisation errors, we relied on the spectral universality assumption (SUA) (Sollich and Halees 2002) to estimate eigenfunctions (and hence coefficients $A(\gamma, \psi), B(\gamma, \psi), C(\gamma, \psi), D(\gamma, \psi)$), which states that in high dimension eigenfunctions become unstructured and can be approximated by independent Gaussian entries. Determining in which cases the SUA is valid is still an active area of research (Karoui 2010; Cheng and Singer 2013; Fan and Montanari 2015; Liu, Liao, and Suykens 2020; Lu and Yau 2023). In the case of isotropic data and no nonlinearity, the SUA is exact (Karoui 2010), which helps explain the very good agreement shown in Figure 2.

In the case of MNIST with ReLU, there is evidence that the SUA does apply to some extent, see e.g. (Simon et al. 2022) who also use SUA to estimate generalisation errors of kernel regression in high dimensions on MNIST. Note that, while we rely on the SUA to provide numerical values for the coefficients $A(\gamma, \psi), \bar{A}(\gamma, \psi), B(\gamma, \psi), C(\gamma, \psi), D(\gamma, \psi), \bar{D}(\gamma, \psi)$ our general theoretical results only require boundedness of these coefficients. As can be seen in Figure 2, we find reasonable agreement. The main sources of discrepancies stem from the fact that we only use a small subset (300 examples) to estimate the empirical spectral distribution and that the SUA may not be completely accurate in this particular setting. Although our predictions for the generalisation error are not perfectly accurate, we emphasise that they correctly predict the double-descent phenomenon, and thus support our claim that the double-descent phenomenon is only driven by the spectral distribution and its dependence with the width.

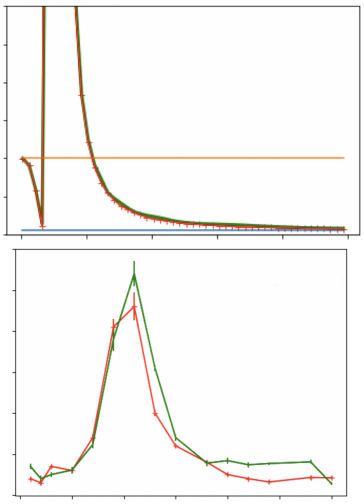

Figure 2: Simulated generalisation error (red curves) versus theoretical limiting generalisation error (green curves), as a function of $\frac{1}{\gamma}$. The X-axis are indexed by $\frac{1}{\gamma}$ by increments of 2 and 0.5 for the top and bottom figure respectively, and the Y-axis are indexed by the generalisation error by increments of 0.5 and 1 for the top and bottom figure respectively. Top, the simulated error is obtained by sampling from the kernel random matrix of a single-hidden-layer NNGP with no non-linearity under a teacher distribution $\mathcal{N}(0, \frac{1}{d} I_d)$. The theoretical distibution and asymptotes (underparameterized asymptote in orange and overparameterized asymptote in blue) are obtained by integrating $\frac{1}{\lambda}$ and $\frac{1}{\lambda^2}$ over the map of the distribution ρ^{ψ}_{MP}. Bottom, the simulated error is obtained by sampling from the kernel random matrix of a two-hidden-layer ReLU NNGP on a subset of MNIST, and the theoretical distribution is obtained by sampling from the Marchenko-Pastur map of the empirical NNGP kernel matrix $\mathbf{K}^{\hat{N}}_{\mathbf{X},\mathbf{X}}$, and from independent Gaussian distributions in lieu of eigenfunctions.

Related Work

The properties of stochastic gradient descent (SGD) have been proposed as an explanation for the favourable generalisation power of DNNs in the over-parametrised regime; for instance, the tendency to escape saddle points (Criscitiello and Boumal 2019) could explain how solutions that generalise well are selected over all others. The neural tangent ker-

nel describes the dynamics of SGD in the functional space and its relationship to the generalisation power of DNNs is well documented (Jacot, Gabriel, and Hongler 2018; Adlam and Pennington 2020a; Bordelon, Canatar, and Pehlevan 2020; Cao et al. 2020; Geiger et al. 2020).

The favourable generalisation properties of DNNs in the over-parametrised regime may also be explained using Bayesian methods and other kernel machines in a way that is unrelated to SGD training; indeed, the good performance of NNGP regression (Lee et al. 2018, 2020) provides compelling evidence in this direction. A related argument is that the parameter-function map is exponentially biased towards Kolmogorov simple functions (Dingle, Camargo, and Louis 2018; Valle-Pérez, Camargo, and Louis 2018; Mingard et al. 2019); since the data on which DNNs are trained has structure, this inductive bias leads to good generalisation in the over-parametrised regime. Due to large differences in the sizes of the basins of attraction (Schaper and Louis 2014), SGD converges to functions with a probability that is remarkably close to the Bayesian posterior probability that a DNN expresses upon random sampling of parameters (Mingard et al. 2020). These ideas are still being debated (Ghorbani et al. 2020; Wilson and Izmailov 2020; Belkin 2021).

The seminal work of Vallet, Cailton, and Refregier (1989); Seung and Sompolinsky (1992) on the double-descent phenomenon and the subsequent developments in (Mei and Montanari 2019; Hastie et al. 2019; Advani, Saxe, and Sompolinsky 2020; Liu, Liao, and Suykens 2020) suggest that the favourable generalisation power of DNNs is an intrinsic characteristic of the set of functions that these models can learn, as generalisation errors are computed analytically and independently from the learning algorithm. Methods of statistical physics have traditionally been the tool of choice for obtaining closed-form formulae in this setting (Engel, von Guericke, and den Broeck 2012). Recent works have provided analytical expressions for the generalisation error of high-dimensional kernel regressions (Canatar, Bordelon, and Pehlevan 2021; Bordelon, Canatar, and Pehlevan 2020; Jacot et al. 2020; Simon et al. 2022; Cui et al. 2022). In particular, (Jacot et al. 2020) and (Simon et al. 2022) rely on the spectral universality assumption, just as we do to estimate the coefficients in our formula. As pointed out by (Simon et al. 2022), other works take the spectral universality assumption implicitly via, for instance, the replica method (Bordelon, Canatar, and Pehlevan 2020; Canatar, Bordelon, and Pehlevan 2021). Our computation of the generalization error is thus similar to the works of (Jacot et al. 2020; Simon et al. 2022). Their results however hold for frozen kernels and the dependence with the width is not studied.

The limiting spectral distributions of the kernel random matrices that we study in this paper were first investigated in (Fan and Wang 2020). Our results are, however, stronger since they require less restrictive assumptions on the data generating process and the non-linear activations; for instance, we do not assume the non-linear activation to be twice differentiable nor the columns of the input data matrix to be "(ϵ, B)-orthonormal". This was made possible by deriving the analytical expression as a function of an implicit quantity: the spectral measure of the actual NNGP kernel (the assumptions taken in (Fan and Wang 2020) are precisely needed to derive an explicit formula for the spectrum of the actual NNGP kernel). Furthermore, we emphasise that the link between the "CK" kernel random matrix in (Fan and Wang 2020) and the actual conjugate (NNGP) kernel is not straightforward. This subtle distinction enables a more transparent interpretation of the the double-descent phenomenon. This distinction was already noticed and exploited by (Louart, Liao, and Couillet 2018), who found an expression of the limiting spectrum depending on the actual NNGP kernel matrix (not only its spectrum) under very general assumptions. In this paper, we introduce a novel proof technique enabled by the recent result of (Banna, Merlevede, and Peligrad 2015) which allows us to isolate the dependency in the spectrum under mild assumptions such as boundedness of the second moment of the data distribution.

The double-descent behaviour in the learning curves of high-dimensional kernel regression (including the NNGP and neural tangent kernels as particular cases) has been described in (Canatar, Bordelon, and Pehlevan 2021; Bordelon, Canatar, and Pehlevan 2020). Our work improves on this line of research by introducing the idea of width-dependent kernels, which is especially well-suited to the context of DNNs where double descent manifests as the network width tends to infinity. Recent studies of the double-descent phenomenon have focused on random features regressions in the case of shallow networks (Gerbelot, Abbara, and Krzakala 2020; Liao, Couillet, and Mahoney 2020; Emami et al. 2020; Gerace et al. 2020; Li, Zhou, and Gretton 2021; Adlam and Pennington 2020b; Belkin, Hsu, and Xu 2020; Chen and Schaeffer 2021; Bosch et al. 2022; D'Ascoli et al. 2020), or kernel regression with no dependence on the width (Liu, Liao, and Suykens 2020; Mallinar et al. 2022).

Conclusions

In this paper, we have exploited results from random matrix theory to offer a new perspective on the double descent phenomenon in FCNs through the lens of Gaussian process kernels. We have derived analytical expressions for the generalisation error under teacher-student scenarios, which are applicable to networks of arbitrary depth and a large family of nonlinearities. This analysis allows us to predict the double descent behaviour as the width of the last layer changes relatively to the number of examples, and understand it as simply arising from the discrepancy between the spectrum of width-dependent random kernel (corresponding the empirical covariance matrix of the features), and that of the width-independent NNGP kernel. Finally, we hope that the tools we have developed will motivate further research on the properties of the generalisation error of neural networks.

Acknowledgements

Work supported by the SIRIUS Centre (Res. Council of Norway, project 237889), and the EPSRC projects ConCur (EP/V050869/1) and UK FIRES (EP/S019111/1). For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

References

Adlam, B.; and Pennington, J. 2020a. The Neural Tangent Kernel in High Dimensions: Triple Descent and a Multi-Scale Theory of Generalization. In *Proc. of PMLR*, 74–84.

Adlam, B.; and Pennington, J. 2020b. Understanding Double Descent Requires A Fine-Grained Bias-Variance Decomposition. In *Proc. of NeurIPS*.

Advani, M.; Saxe, A.; and Sompolinsky, H. 2020. High-dimensional dynamics of generalization error in neural networks. *Neural networks : the official journal of the International Neural Network Society*, 132: 428–446.

Banna, M.; Merlevede, F.; and Peligrad, M. 2015. On the limiting spectral distribution for a large class of symmetric random matrices with correlated entries. *Stochastic Processes and their Applications*, 125.

Belkin, M. 2021. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *arXiv preprint arXiv:2105.14368*.

Belkin, M.; Hsu, D.; Ma, S.; and Mandal, S. 2019. Reconciling modern machine-learning practice and the classical bias-variance tradeoff. *Proc. Nat. Academy of Sciences*, 32.

Belkin, M.; Hsu, D.; and Xu, J. 2020. Two Models of Double Descent for Weak Features. *SIAM Journal on Mathematics of Data Science*, 2(4): 1167–1180.

Berthier, R.; Bach, F.; and Gaillard, P. 2020. Tight Nonparametric Convergence Rates for Stochastic Gradient Descent under the Noiseless Linear Model. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *NeurIPS*, volume 33, 2576–2586. Curran Associates, Inc.

Bordelon, B.; Canatar, A.; and Pehlevan, C. 2020. Spectrum Dependent Learning Curves in Kernel Regression and Wide Neural Networks. In *Proc. PMLR*, 1024–1034.

Bosch, D.; Panahi, A.; Özcelikkale, A.; and Dubhash, D. 2022. Double Descent in Random Feature Models: Precise Asymptotic Analysis for General Convex Regularization. *arXiv*.

Canatar, A.; Bordelon, B.; and Pehlevan, C. 2021. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1).

Cao, Y.; Fang, Z.; Wu, Y.; Zhou, D. X.; and Gu, Q. 2020. Towards Understanding the Spectral Bias of Deep Learning. *arXiv preprint: arXiv:1912.01198*.

Chen, Z.; and Schaeffer, H. 2021. Conditioning of Random Feature Matrices: Double Descent and Generalization Error. *arXiv:2110.11477*.

Cheng, X.; and Singer, A. 2013. The Spectrum of Random Inner-Product Kernel Matrices. *Random Matrices: Theory and Applications*, 02(04): 1350010.

Criscitiello, C.; and Boumal, N. 2019. Efficiently escaping saddle points on manifolds. *NeurIPS*, 32.

Cui, H.; Loureiro, B.; Krzakala, F.; and Zdeborová , L. 2022. Generalization error rates in kernel regression: the crossover from the noiseless to noisy regime. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11): 114004.

Daniely, A.; Frostig, R.; and Singer, Y. 2016. Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity. *NeurIPS*, 29.

D'Ascoli, S.; Refinetti, M.; Biroli, G.; and Krzakala, F. 2020. Double Trouble in Double Descent: Bias and Variance(s) in the Lazy Regime. In *ICML*, volume 119, 2280–2290.

Dingle, K.; Camargo, C. Q.; and Louis, A. A. 2018. Input–output maps are strongly biased towards simple outputs. *Nature Communications*, 9(1): 1–7.

Emami, M.; Sahraee-Ardakan, M.; Pandit, P.; Rangan, S.; and Fletcher, A. K. 2020. Generalization Error of Generalized Linear Models in High Dimensions. In *PMLR*.

Engel, A.; von Guericke, O.; and den Broeck, C. V. 2012. *Statistical mechanics of learning*. Cambridge Univ. Press.

Fan, Z.; and Montanari, A. 2015. The spectral norm of random inner-product kernel matrices. *Probability Theory and Related Fields*, 173: 27–85.

Fan, Z.; and Wang, Z. 2020. Spectra of the Conjugate Kernel and Neural Tangent Kernel for Linear-Width Neural Networks. *NeurIPS*, 33.

Garriga-Alonso, A.; Rasmussen, C. E.; and Aitchison, L. 2018. Deep convolutional networks as shallow gaussian processes. *arXiv preprint arXiv:1808.05587*.

Geiger, M.; Jacot, A.; Spigler, S.; Gabriel, F.; Sagun, L.; d'Ascoli, S.; Biroli, G.; Hongler, C.; and Wyart, M. 2020. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*.

Geiger, M.; Petrini, L.; and Wyart, M. 2020. Perspective: A Phase Diagram for Deep Learning unifying Jamming, Feature Learning and Lazy Training. *arXiv preprint arXiv:2012.15110*.

Gerace, F.; Loureiro, B.; Krzakala, F.; Mezard, M.; and Zdeborova, L. 2020. Generalisation error in learning with random features and the hidden manifold model. In *ICML*.

Gerbelot, C.; Abbara, A.; and Krzakala, F. 2020. Asymptotic errors for convex penalized linear regression beyond Gaussian matrices. In *ICML*.

Geronimo, J.; and Hill, T. 2002. Necessary and Sufficient Condition that the Limit of Stieltjes Transforms is a Stieltjes Transform. *Journal of Approximation Theory*.

Ghorbani, B.; Mei, S.; Misiakiewicz, T.; and Montanari, A. 2020. When Do Neural Networks Outperform Kernel Methods? *NeurIPS*, 33.

Györfi, L.; Kohler, M.; Krzyzak, A.; and Walk, H. 2002. Springer.

Hannun, A.; Case, C.; Casper, J.; B. Catanzaro, G. D.; Elsen, E.; Prenger, R.; Satheesh, S.; Sengupta, S.; Coates, A.; and Ng, A. Y. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Hastie, T.; Montanari, A.; Rosset, S.; and Tibshirani, R. 2019. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint: arXiv:1903.08560*.

Jacot, A.; Gabriel, F.; and Hongler, C. 2018. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *NeurIPS*, 31.

Jacot, A.; Simsek, B.; Spadaro, F.; Hongler, C.; and Gabriel, F. 2020. Kernel Alignment Risk Estimator: Risk Prediction from Training Data. In *NeurIPS*, volume 33.

Karoui, N. E. 2010. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1).

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deepconvolutional neural networks. *Advances in Neural Information Processing Systems*, 25.

LeCun, Y. 2012. The MNIST database.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature*, 521(7553): 436–444.

Lee, J.; Bahri, Y.; Novak, R.; Schoenholz, S. S.; Pennington, J.; and Sohl-Dickstein, J. 2018. Deep Neural Networks as Gaussian Processes. *ICLR*.

Lee, J.; Schoenholz, S.; Pennington, J.; Adlam, B.; Xiao, L.; Novak, R.; and Sohl-Dickstein, J. 2020. Finite versus infinite neural networks: an empirical study. *NeurIPS*, 33.

Li, Z.; Zhou, Z.-H.; and Gretton, A. 2021. Towards an Understanding of Benign Overfitting in Neural Networks. *arXiv:2103.14723*.

Liao, Z.; Couillet, R.; and Mahoney, M. W. 2020. A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent. *NeurIPS*, 33.

Liu, F.; Liao, Z.; and Suykens, J. 2020. Kernel regression in high dimension: Refined analysis beyond double descent. *arXiv preprint: arXiv:2010.02681*.

Livan, G.; Novaes, M.; and Vivo, P. 2018. *Introduction to Random Matrices*. Springer International Publishing.

Louart, C.; Liao, Z.; and Couillet, R. 2018. A Random Matrix Approach to Neural Networks. *The Annals of Applied Probability*, 28(2): 1190–1248.

Lu, Y. M.; and Yau, H.-T. 2023. An Equivalence Principle for the Spectrum of Random Inner-Product Kernel Matrices with Polynomial Scalings. arXiv:2205.06308.

Mallinar, N.; Simon, J. B.; Abedsoltan, A.; Pandit, P.; Belkin, M.; and Nakkiran, P. 2022. Benign, Tempered, or Catastrophic: A Taxonomy of Overfitting. *arXiv*.

Marchenko, V.; and Pastur, L. 1967. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 72.

Matthews, A. G. d. G.; Rowland, M.; Hron, J.; Turner, R. E.; and Ghahramani, Z. 2018. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*.

Mei, S.; and Montanari, A. 2019. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*.

Mingard, C.; Skalse, J.; Valle-Pérez, G.; Martínez-Rubio, D.; Mikulik, V.; and Louis, A. A. 2019. Neural networks are a priori biased towards Boolean functions with low entropy. *arXiv preprint arXiv:1909.11522*.

Mingard, C.; Valle-Pérez, G.; Skalse, J.; and Louis, A. A. 2020. Is SGD a Bayesian sampler? Well, almost. *arXiv preprint arXiv:2006.15191*.

Mohri, M.; Rostamizadeh, A.; and A.Talwalkar. 2012. *Foundations of Machine Learning*. MIT Press.

Nakkiran, P.; Kaplun, G.; Bansal, Y.; Yang, T.; Barak, B.; and Sutskever, I. 2021. Deep double descent: where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12): 124003.

Neal, R. 1994. *Bayesian Learning for Neural Networks*. Ph.D. thesis, University of Toronto.

Novak, R.; Xiao, L.; Lee, J.; Bahri, Y.; Yang, G.; Hron, J.; Abolafia, D. A.; Pennington, J.; and Sohl-Dickstein, J. 2018. Bayesian deep convolutional networks with many channels are gaussian processes. *arXiv preprint arXiv:1810.05148*.

Poole, B.; S.Lahiri; Raghu, M.; Sohl-Dickstein, J.; and Ganguli, S. 2016. Exponential expressivity in deep neural networks through transient chaos. *NeurIPS*, 29.

Rasmussen, C.; and Williams, C. 2006. *Gaussian Processes for Machine Learning*. MIT Press.

Schaper, S.; and Louis, A. A. 2014. The arrival of the frequent: how bias in genotype-phenotype maps can steer populations to local optima. *PloS one*, 9(2): e86635.

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

Seung, H. S.; and Sompolinsky, H. 1992. Statistical mechanics of learning from examples. *Physical Review A.*, 45.

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

Simon, J. B.; Dickens, M.; Karkada, D.; and DeWeese, M. R. 2022. The Eigenlearning Framework: A Conservation Law Perspective on Kernel Regression and Wide Neural Networks. arXiv:2110.03922.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sollich, P.; and Halees, A. 2002. Learning Curves for Gaussian Process Regression: Approximations and Bounds. *Neural Computation*, 14(6): 1393–1428.

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

Vallet, F.; Cailton, J.-G.; and Refregier, P. 1989. Linear and Nonlinear Extension of the Pseudo-Inverse Solution for Learning Boolean Functions. *EPL*, 9(4): 315–320.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. Springer.

Wigner, E. 1955. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62.

Wilson, A. G.; and Izmailov, P. 2020. Bayesian Deep Learning and a Probabilistic Perspective of Generalization. *Advances in Neural Information Processing Systems*, 33.

Yang, G. 2019. Wide feedforward or recurrent neural networks of any architecture are gaussian processes. In *NeurIPS*, 9951–9960.