MACHINE LEARNING
Science and Technology

PAPER • OPEN ACCESS

The twin peaks of learning neural networks

MACHINE
LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
23 January 2024

REVISED
17 April 2024

ACCEPTED FOR PUBLICATION
30 May 2024

PUBLISHED
10 June 2024

PAPER

The twin peaks of learning neural networks

Elizaveta Demyanenko, Christoph Feinauer, Enrico M Malatesta* and Luca Saglietti

Department of Computing Sciences, Bocconi University, 20136 Milano, Italy

* Author to whom any correspondence should be addressed.

E-mail: enrico.m.malatesta@gmail.com

Keywords: mean dimension, double descent, replica method, random feature model

Supplementary material for this article is available online

Abstract

Recent works demonstrated the existence of a double-descent phenomenon for the generalization error of neural networks, where highly overparameterized models escape overfitting and achieve good test performance, at odds with the standard bias-variance trade-off described by statistical learning theory. In the present work, we explore a link between this phenomenon and the increase of complexity and sensitivity of the function represented by neural networks. In particular, we study the Boolean mean dimension (BMD), a metric developed in the context of Boolean function analysis. Focusing on a simple teacher-student setting for the random feature model, we derive a theoretical analysis based on the replica method that yields an interpretable expression for the BMD, in the high dimensional regime where the number of data points, the number of features, and the input size grow to infinity. We find that, as the degree of overparameterization of the network is increased, the BMD reaches an evident peak at the interpolation threshold, in correspondence with the generalization error peak, and then slowly approaches a low asymptotic value. The same phenomenology is then traced in numerical experiments with different model classes and training setups. Moreover, we find empirically that adversarially initialized models tend to show higher BMD values, and that models that are more robust to adversarial attacks exhibit a lower BMD.

1. Introduction

The evergrowing scale of modern neural networks often prevents a detailed understanding of how predictions relate back to the model inputs (Sejnowski 2020). While this lack of interpretability can hinder adoption in sectors with a high impact on society (Rudin 2019), the impressive performance of neural network-based models in fields like natural language processing (Vaswani *et al* 2017, OpenAI 2023, Touvron *et al* 2023), computational biology (Jumper *et al* 2021) and computer vision and image generation (Ramesh *et al* 2022, Rombach *et al* 2022) have made them the de-facto standard for many real-world applications. This tension has motivated a large interest in the field of explainable AI (XAI) (Guidotti *et al* 2018, Montavon *et al* 2018, Vilone and Longo 2020).

Deep learning models, which by now can feature hundreds of billions of parameters (Brown *et al* 2020), seemingly defy the notion that increasing model complexity should decrease generalization performance. Counter to what one would expect from statistical learning theory (Vapnik 1999), the observation has been that larger—heavily overparameterized—models often perform better (Neyshabur *et al* 2017). This has led to the question how complex the function represented by an overparameterized neural network is after training. Many lines of research suggest that neural network models are biased towards implementing simple functions, despite their large parameter count, and that this implicit bias is crucial for their good generalization performance (Valle-Perez *et al* 2018). The general problem of measuring the complexity of deep neural networks has given rise to several complexity metrics (Novak *et al* 2018) and studies on how they relate to generalization (Jiang *et al* 2019).

Connected to this, recent studies (Belkin *et al* 2019, Geiger *et al* 2019) on the effect of overparameterization in neural networks led to the rediscovery of the 'double descent' phenomenon, first observed in the statistical physics literature (Opper 1995), which is the observation that when increasing the capacity of a neural network (measured, for example, by the number of parameters) the generalization error shows a sudden peak around the interpolation point (where approximately zero training error is achieved), but then a second decrease towards a low asymptotic value is observed at higher overparameterization.

In the present work, we study the double descent phenomenon under a notion of function sensitivity based on the *mean dimension* (Hoyt and Owen 2021, Hahn *et al* 2022). The mean dimension yields a measure of the mean interaction order between input variables in a function, and can also be proved to be related to the variance of the function under local perturbations of the input features. While this notion originated in the field statistics (Liu and Owen 2006), several computational techniques have been proposed for its estimation in the context of neural networks. One of the main obstacles, however, comes from trying to characterize the sensitivity of the function over an input distribution that is strongly structured and not fully known.

In this paper, we propose to focus on the study of the *Boolean mean dimension* (O'Donnell 2014) (BMD), which involves a simple i.i.d. binary input distribution. We show how the BMD can be estimated efficiently, and provide analytical and numerical evidence of the correlation of this metric with several phenomena observed on the data used for training and testing the model.

2. Related works

2.1. Overparameterization and double descent

Several studies (Baity-Jesi *et al* 2018, Geiger *et al* 2019, Advani *et al* 2020) confirmed the robustness of the double descent phenomenology for a large variety of architectures, datasets, and learning paradigms. An analytical study of double descent in the context of the random feature model (RFM) (Rahimi and Recht 2007) was conducted rigorously for the square loss in Mei and Montanari (2019) and for generic loss by Gerace *et al* (2020) using the replica method (Mézard *et al* 1987). Double descent has then later found also in the context of one layer model learning a Gaussian mixture dataset (Mignacco *et al* 2020); similarly to the RFM, the peak in the generalization can be avoided by optimally regularizing the network. In this context in Baldassi *et al* (2020) it was also shown that choosing the optimal regularization corresponds to maximize a flatness-based measure of the loss minimizer. A range of later studies further explored this phenomenology in related settings (d'Ascoli *et al* 2020, Gerace *et al* 2022).

Different scenarios have also been shown to give rise to a similar phenomenology, such as the epoch-wise double descent and sample non-monotonicity (Nakkiran *et al* 2021) and the triple descent that can appear with noisy labels and can be regularized by the non-linearity of the activation function (d'Ascoli *et al* 2020).

In this work, we connect the usual double descent of the generalization error with the behavior of the mean dimension, which is a complexity metric that can be evaluated without requiring task-specific data.

2.2. Mean dimension and BMD

The *mean dimension* (MD), based on the analysis of variance (ANOVA) expansion (Efron and Stein 1981, Owen 2003), can be intuitively understood as a marker of the complexity of a function due to the presence of interactions between a large set of input variables.

The mean dimension has been used as a tool to analyze and compare for example neural networks (Hoyt and Owen 2021, Hahn *et al* 2022) and, with a slightly different definition, also generative models of protein sequences (Feinauer and Borgonovo 2022). The MD has the advantage that it can be calculated for a *black-box function*, without regard to the internal mechanism for calculating the input-output relation. One major drawback, however, is the intense computational cost associated with its direct estimation. This computational limitation has led to the proposal of several approximation strategies (Hoyt and Owen 2021, Hahn *et al* 2022). In some special cases, the mean dimension can be explicitly expressed as a function of the coefficients of a Fourier expansion, as seen from the relationship between the BMD and the *total influence* (O'Donnell 2014) defined in the analysis of Boolean functions (see below), and its generalization (Feinauer and Borgonovo 2022) for functions with categorical variables.

3. Mean dimension

In the next paragraphs, we first provide a general mathematical definition of the mean dimension for a square-integrable function with real-valued input distribution. We then specialize to the case of a binary input distribution and define the BMD, which will be the main quantity investigated throughout this paper.

Finally, we will discuss how to efficiently estimate the MD and the BMD through a simple Monte Carlo procedure.

3.1. Mathematical definition

To give a proper mathematical definition of the mean dimension, for a real-valued function $f(\boldsymbol{x})$ of n variables $f:\mathbb{R}^n\to\mathbb{R}$, it is convenient to introduce some notation that will be used in the rest of the paper. We will denote the set of indexes $\{1,\dots n\}$ by $[n]$. We define $\boldsymbol{x}_u$ the set of input variables x_i, with $i\in u\subseteq[n]$ and by $\boldsymbol{x}_{\backslash u}$ the set of variables for which $i\notin u$. We will also assume that $\boldsymbol{x}$ is drawn from a distribution $p(\boldsymbol{x})$. The basic idea of the mean dimension (Hahn *et al* 2022) is to derive a complexity measure for f from an expansion of the type

$$f(\boldsymbol{x})=\sum_{u\subseteq[n]}f_u\left(\boldsymbol{x}_u\right) \tag{1}$$

where the 'components' $f_u(\boldsymbol{x}_u)$ can be computed from the following recursion relation

$$f_u\left(\boldsymbol{x}_u\right)\equiv\int f(\boldsymbol{x})\,p\left(\boldsymbol{x}_{\backslash u}|\boldsymbol{x}_u\right)\,\mathrm{d}\boldsymbol{x}_{\backslash u}-\sum_{v\subset u}f_v\left(\boldsymbol{x}_v\right) \tag{2}$$

with the initial condition $f_\emptyset=\int f(\boldsymbol{x})\,p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\equiv\mathbb{E}[f]$. It can be shown that coefficients of the expansion have zero average if u is non empty

$$\int f_u\left(\boldsymbol{x}_u\right)p_u\left(\boldsymbol{x}_u\right)\,\mathrm{d}\boldsymbol{x}_u=0\qquad u\neq\emptyset \tag{3}$$

where we have denoted by $p_u(\boldsymbol{x}_u)$ the marginal probability distribution over the set u. Moreover, they satisfy orthogonality relations, namely

$$\int f_u\left(\boldsymbol{x}_u\right)f_v\left(\boldsymbol{x}_v\right)p_{u\cup v}\left(\boldsymbol{x}_{u\cup v}\right)\,\mathrm{d}\boldsymbol{x}_{u\cup v}=0\,,\qquad\text{if } u\neq v\,. \tag{4}$$

Using those relations we can write the variance of the function as a decomposition of 2^n-1 terms

$$\sigma^2=\mathbb{E}\left[f^2\right]-\mathbb{E}\left[f\right]^2=\sum_{u\subseteq[n]\backslash\emptyset}\sigma_u^2 \tag{5}$$

where

$$\sigma_u^2\equiv\int f_u^2\left(\boldsymbol{x}_u\right)\,p_u\left(\boldsymbol{x}_u\right)\,\mathrm{d}\boldsymbol{x}_u\,. \tag{6}$$

The mean dimension M_f is then defined as (Hahn *et al* 2022)

$$M_f=\sum_{u\subseteq[n]}|u|\frac{\sigma_u^2}{\sigma^2}, \tag{7}$$

i.e. a weighted sum over possible interactions, with each subset of inputs contributing based on how much they influence the variance.

3.2. Pseudo-Boolean functions and Fourier coefficients

We now derive an explicit expression for the mean dimension of n-dimensional pseudo-Boolean functions taking values on the real domain, $f:\{-1,1\}^n\to\mathbb{R}$ under the assumption of input features that are i.i.d. from $\{-1,1\}$.

Denoting by $\boldsymbol{s}\in\{-1,1\}^n$ the n-dimensional binary input of f, such a function can be uniquely written as a *Fourier expansion* (O'Donnell 2014) in terms of a finite set of *Fourier coefficients* $\hat{f}_u$, $u\subseteq[n]$ as

$$f(\boldsymbol{s})=C+\sum_i h_i s_i+\sum_{i<j}J_{ij}s_i s_j+\sum_{i<j<k}K_{ijk}s_i s_j s_k+\ldots=\sum_{u\subseteq[n]}\hat{f}_u\,\chi_u\left(\boldsymbol{s}_u\right) \tag{8}$$

where

$$\chi_u\left(\boldsymbol{s}_u\right)=\prod_{i\in u}s_i \tag{9}$$

represent the Fourier basis of the decomposition that are orthonormal $\langle \chi_u(\boldsymbol{s})\chi_v(\boldsymbol{s})\rangle = \delta_{u,v}$ with respect to the uniform distribution over $\{-1,1\}^n$, where we use the notation

$$\langle \bullet \rangle \equiv \frac{1}{2^n} \sum_{s \in \{-1,1\}^n} \bullet . \tag{10}$$

The Fourier coefficients $\hat{f}_u$ can give information about the moments of the function f with respect to the uniform distribution (10) over $\boldsymbol{s}$; for example the first moment is

$$\langle f(\boldsymbol{s}) \rangle = \hat{f}_\emptyset \tag{11}$$

whereas the variance can be obtained as

$$\sigma^2 = \langle f^2(\boldsymbol{s}) \rangle - \langle f(\boldsymbol{s}) \rangle^2 = \sum_{u \subseteq [n] \setminus \emptyset} \hat{f}_u^2 . \tag{12}$$

We can quantify the contribution c_k of interaction of order k to the variance of $f(\boldsymbol{s})$ as the ratio

$$c_k = \frac{\sum_{u \subseteq [n] \setminus \emptyset : |u| = k} \hat{f}_u^2}{\sigma^2} . \tag{13}$$

Notice that $\sum_k c_k = 1$, so that c_k can be interpreted as a (discrete) probability measure over interactions. The mean dimension of f can then be written as the mean interaction degree when weighted according to it contribution to the variance, i.e. as a weighted sum of feature influences divided by the total variance of the function, so

$$M_f \equiv \sum_{k=1}^{n} k c_k = \frac{\sum_{u \subseteq [n]} |u| \hat{f}_u^2}{\sigma^2} . \tag{14}$$

This expression is equivalent to equation (7) for pseudo-Boolean functions under the assumptions that all features are i.i.d from $\{-1,1\}$. The expression connects the notion of simplicity in terms of variance contributions to the same notion in terms of explicit expansion coefficients. Intuitively, a large mean dimension is indicating that the function fluctuates due to a large contribution of high-degree interactions.

3.3. Estimating the mean dimension through Monte Carlo

The expression of the mean dimension in (7) involves a sum over all the set of subsets of n variables, and its numerical evaluation through a brute-force approach would be intractable in high dimension. However, it can be shown that a more efficient evaluation scheme of equation (7), can be achieved through a Monte Carlo approach (Liu and Owen 2006). First, the MD can be rewritten as a sum over the n input components:

$$M_f = \frac{\sum_{i=1}^{n} \tau_i^2}{\sigma^2} \tag{15}$$

where the *influence* of the ith input component τ_i is defined as:

$$\tau_i^2 = \frac{1}{2} \int \mathrm{d}\boldsymbol{x}\, \mathrm{d}x_i'\, p(\boldsymbol{x})\, p\left(x_i' | \boldsymbol{x}_{\setminus i}\right) \left(f(\boldsymbol{x}) - f(\boldsymbol{x}^{\oplus i})\right)^2 . \tag{16}$$

And where we have denoted by $\boldsymbol{x}^{\oplus i}$ a vector $\boldsymbol{x}$ with a resampled ith coordinate. We show an original proof of this identity in appendix A.

Note that the definition of the MD for a generic input distribution in equation (16), entails a resampling procedure that presumes knowledge of the conditional distribution of a pixel given the rest of the pixel values. In the general case, this pixel is to be resampled multiple times from this conditional distribution, to compute the variance of the function under this variation of the input. This conditional distribution, however, is not a known quantity for a real dataset. For this reason, for example, some authors have proposed an 'exchange' procedure, where one randomly samples a different pixel value observed in the same dataset (Hahn *et al* 2022), however this approximation neglects the within sample correlations.

Expression (16) can be specialized to the case of binary i.i.d. inputs, where one can identify the influence functions τ_i^2 with the discrete derivatives:

$$\tau_i^2 = \left\langle \left(\mathcal{D}_i f(\boldsymbol{s})\right)^2 \right\rangle \tag{17}$$

where $\mathcal{D}_i f(\boldsymbol{s})$ denotes the i_{th} (discrete) derivative of $f(\boldsymbol{s})$, i.e.

$$\mathcal{D}_i f(\boldsymbol{s}) \equiv \frac{f(s_1,\ldots,s_i=1,\ldots,s_n) - f(s_1,\ldots,s_i=-1,\ldots,s_n)}{2} \tag{18}$$

and measures the average sensitivity of the function to a flip of the i_{th} variable. The sum of the influences $\sum_i \left\langle (\mathcal{D}_i f(\boldsymbol{s}))^2 \right\rangle$ is known in the field of the analysis of pseudo-Boolean functions as *total influence* of f (O'Donnell 2014). In terms of the Fourier expansion, we have

$$\mathcal{D}_i f(\boldsymbol{s}) = \sum_{u\subseteq[n]:i\in u} \hat{f}_u\, \chi_{u\setminus i}\left(\boldsymbol{s}_{u\setminus i}\right). \tag{19}$$

Therefore computing the mean dimension for pseudo-Boolean functions boils down to querying the function f on uniformly sampled binary sequences of length $n-1$.

3.4. BMD

In the general case, the underlying input distribution of the training dataset is not known and estimating the MD on this distribution becomes unfeasible. In the present work, we propose employing the estimation procedure presented in the last section, based on binary sequences, as an easily computable proxy of the sensitivity of the neural network function. In order to distinguish this proxy from the mean dimension over the dataset distribution, we call the resulting quantity the BMD. We show in the results below that the BMD can in some cases be computed analytically, and that it is qualitatively related to the generalization phenomenology in neural networks.

4. Analytical results

We now derive an analytic expression for the mean dimension in the special case of the RFM (Rahimi and Recht 2007, Goldt *et al* 2019, Loureiro *et al* 2021, Baldassi *et al* 2022), focusing on the same high dimensional regime where the double descent phenomenon can be detected. In the next sections, we will define the model, the learning task and the high dimensional limit precisely, and we will sketch the analytical derivation of the expression for the BMD.

4.1. Model definition and learning task

The RFM is a two-layer neural network with random and fixed first-layer weights (also called features) and trainable second-layer weights. Given a D-dimensional input, $\boldsymbol{x}\in\mathbb{R}^D$, and denoting by $F\in\mathbb{R}^{D\times N}$ the $D\times N$ frozen feature matrix, the pre-activation of the RFM is given by:

$$\hat{y}(\boldsymbol{w};\boldsymbol{x}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} w_i\, \sigma\left(\frac{1}{\sqrt{D}}\sum_{k=1}^{D} F_{ki} x_k\right) \tag{20}$$

where $\boldsymbol{w}$ is an N-dimensional weight vector and σ is a (usually non-linear) function. The parameter N indicates the number of features in the RFM and can be varied to change the degree of over-parametrization of the model. As in Baldassi *et al* (2022), we will hereafter focus on the case of i.i.d. standard normal distributed feature components $F_{ki}\sim\mathcal{N}(0,1)$, although the formalism allows for a simple extension to a generic fixed feature map, under a simple weak correlation requirement (see Gerace *et al* 2020, Loureiro *et al* 2021 for additional details).

We consider a classification task defined by a training dataset of size P, denoted as $\mathcal{D}=\{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^{P}$. The inputs are assumed to be i.i.d. with first and second moments fixed respectively to $\mathbb{E}x_i = 0$ and $\mathbb{E}x_i^2 = 1$. Note that, for example, both binary input components $x_i\in\{-1,1\}$ and Gaussian components $x_i\sim\mathcal{N}(0,1)$ satisfy the above assumption. The binary labels $y^\mu\in\{-1,1\}$ are assumed to be produced by a 'teacher' linear model $\boldsymbol{w}^T\in\mathbb{R}^D$, with normalized weights on the D-sphere $\|\boldsymbol{w}^T\|_2^2 = D$, according to:

$$y^\mu = \mathrm{sign}\left(\frac{1}{\sqrt{D}}\sum_{k=1}^{D} w_k^T x_k^\mu\right), \qquad \mu\in[P]. \tag{21}$$

The learning task is then framed as an optimization problem with generic loss function ℓ and ridge regularization

$$\boldsymbol{w}_\star = \underset{\boldsymbol{w}\in\mathbb{R}^N}{\arg\min}\left[\sum_{\mu=1}^{P}\ell\left(y^\mu, \hat{y}^\mu(\boldsymbol{w};\boldsymbol{x}^\mu)\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2\right], \tag{22}$$

where λ is a positive external parameter controlling the regularization strength. In the following we will consider the two most common convex loss functions, namely the mean squared error (MSE) and the cross-entropy (CE) losses, defined as

$$\ell_{\mathrm{mse}}\left(y,\hat{y}\right) = \frac{1}{2}\left(y^{\mu}-\hat{y}^{\mu}\right)^{2} \tag{23a}$$

$$\ell_{\mathrm{ce}}\left(y,\hat{y}\right) = \log\left(1+\mathrm{e}^{-y\hat{y}}\right) . \tag{23b}$$

We analyze the learning problem in the high-dimensional limit where the number of features, input components and training-set size diverge $N, D, P \to \infty$ at constant rates $\alpha \equiv P/N = \mathcal{O}(1)$ and $\alpha_D \equiv D/N = \mathcal{O}(1)$. In this limit, strong concentration properties allow for a deterministic characterization of the above-defined learning problem in terms of a finite set of scalar quantities called order parameters. In the next sections, and in detail in the appendices, we will sketch the derivation of this reduced description.

4.2. Rephrasing the problem in terms of the Boltzmann measure

The learning task in (22) can be characterized within a statistical physics framework. One can introduce a probability measure over the weights $\boldsymbol{w}$ in terms of the Boltzmann distribution

$$\boldsymbol{w} \sim p_{\beta}\left(\boldsymbol{w};\mathcal{D}\right) = \frac{\mathrm{e}^{-\beta\sum_{\mu=1}^{P}\ell\left(y^{\mu},\hat{y}^{\mu}\left(\boldsymbol{w};\boldsymbol{x}^{\mu}\right)\right)-\frac{\beta\lambda}{2}\sum_{i=1}^{N}w_i^2}}{Z_{\beta}} \tag{24}$$

where β is the inverse temperature, the loss function in (22) plays the role of an energy, and the partition function Z_{β} is a normalization factor that reads

$$Z_{\beta} = \int \mathrm{d}\boldsymbol{w}\,\mathrm{e}^{-\beta\sum_{\mu=1}^{P}\ell\left(y^{\mu},\hat{y}^{\mu}\left(\boldsymbol{w};\boldsymbol{x}^{\mu}\right)\right)-\frac{\beta\lambda}{2}\sum_{i=1}^{N}w_i^2} . \tag{25}$$

The distribution $p_{\beta}(\boldsymbol{w};\mathcal{D})$ can be interpreted in a Bayesian setting as the posterior distribution over the weights $\boldsymbol{w}$ given a dataset $\mathcal{D}$, and (24) corresponds to Bayes theorem, where the term $e^{-\beta\sum_{\mu}\ell\left(y^{\mu},\hat{y}^{\mu}\left(\boldsymbol{w};\boldsymbol{x}^{\mu}\right)\right)}$, corresponds to the likelihood and $e^{-\frac{\beta\lambda}{2}\|\boldsymbol{w}\|_2^2}$ is the prior distribution over the weights.

In the zero-temperature limit, when $\beta \to \infty$, the probability measure $p_{\beta}(\boldsymbol{w};\mathcal{D})$ concentrates on the solutions to the optimization problem in (22). To characterize the typical (i.e. the most probable) properties of these solutions, one needs to perform an average over the possible realizations of the training set $\mathcal{D}$ and of the features F, computing the free-energy of the system

$$f = -\lim_{\beta\to\infty}\lim_{N\to\infty}\frac{1}{\beta N}\mathbb{E}_{\mathcal{D},F}\ln Z_{\beta} . \tag{26}$$

The computation of this 'quenched' average can be achieved via the replica method (Mézard *et al* 1987) from spin-glass theory, which reduces the characterization of the solutions of (22) to the determination of a finite set of scalar quantities called order parameters (Engel and Van den Broeck 2001, Malatesta 2023).

In appendix B.1, we sketch the replica calculation for the free energy, first presented in Gerace *et al* (2020), in the simplifying case of an odd non-linear activation σ.

4.3. Analytical determination of the BMD in the RFM

We now derive an analytic expression for the BMD which can be efficiently evaluated for a trained RFM. The definition (15) reads

$$M_f(\boldsymbol{w}) \equiv \frac{\frac{1}{2}\sum_{k=1}^{D}\left\langle\left(\hat{y}\left(\boldsymbol{w};\boldsymbol{x}\right)-\hat{y}\left(\boldsymbol{w};\boldsymbol{x}^{\oplus k}\right)\right)^2\right\rangle}{\left\langle\hat{y}^2\left(\boldsymbol{w};\boldsymbol{x}\right)\right\rangle-\left\langle\hat{y}\left(\boldsymbol{w};\boldsymbol{x}\right)\right\rangle^2} . \tag{27}$$

where $\langle\bullet\rangle$ and $x^{\oplus k}$, defined in (10) and (16), entail an expectation over i.i.d. uniform binary inputs. In appendix B.2, we perform the annealed averages appearing in the numerator and the denominator separately, obtaining the expression:

$$M_f(\boldsymbol{w}) = \frac{\frac{1}{N}\sum_{ij}\bar{\Psi}_{ij}w_iw_j}{\frac{1}{N}\sum_{ij}\Psi_{ij}w_iw_j} \tag{28}$$

where we defined

$$\Omega_{ij} \equiv \frac{1}{D}\sum_{k=1}^{D}F_{ki}F_{kj} . \tag{29a}$$

$$\bar{\Psi}_{ij} \equiv \bar{\kappa}_\star^2 \Omega_{ii} \mathbb{I}_{ij} + \bar{\kappa}_0^2 \Omega_{ij} + \bar{\kappa}_1^2 \Omega_{ij}^2, \tag{29b}$$

$$\Psi_{ij} \equiv \kappa_\star^2 \mathbb{I}_{ij} + \kappa_1^2 \Omega_{ij}. \tag{29c}$$

And the coefficients κ are defined as expectations of derivatives of the activation function over a standard Gaussian measure $Dz = \frac{e^{-z^2/2}}{\sqrt{2\pi}} \mathrm{d}z$:

$$\kappa_0 = \int Dz \sigma(z), \qquad \kappa_1 = \int Dz \sigma'(z), \qquad \kappa_2 = \int Dz \sigma^2(z), \tag{30a}$$

$$\bar{\kappa}_0 = \kappa_1, \qquad \bar{\kappa}_1 = \int Dz \sigma''(z), \qquad \bar{\kappa}_2 = \int Dz \left(\sigma'(z)\right)^2, \tag{30b}$$

$$\kappa_\star^2 = \kappa_2 - \kappa_1^2 - \kappa_0^2, \qquad \bar{\kappa}_\star^2 = \bar{\kappa}_2 - \bar{\kappa}_1^2 - \bar{\kappa}_0^2. \tag{30c}$$

As we show in the appendix, the above expression (28) is universal: evaluating the MD with respect to a different i.i.d. input distributions with matching first and second moments would give exactly the same result.

Moreover, note that the evaluation of expression (28) no longer involves a Monte-Carlo over the input distribution, with a major gain in computational cost. In appendix B.2, we show the agreement of this compact formula with the computationally more expensive Monte Carlo estimation of the BMD.

The mean dimension therefore explicitly depends on the model parameters $\boldsymbol{w}$. The evaluation of the typical BMD of a trained RFM can thus be computed by taking an expectation over the zero-temperature Boltzmann measure for the weights derived in the replica computation, $M_f = \mathbb{E}_{\mathcal{D},F} \langle M_f(\boldsymbol{w}) \rangle_{\boldsymbol{w}}$. The notation $\langle \bullet \rangle_{\boldsymbol{w}} \equiv \int \mathrm{d}\boldsymbol{w} \bullet p_\infty(\boldsymbol{w}; \mathcal{D})$ is thus used to indicate an average over the posterior distribution in equation (24), in the large β limit.

In the case of the replica computation for an odd activation function, that we reported in appendix B.1, one can simplify further expression (28) by recognizing that $\bar{\kappa}_1 = 0$ and that $\Omega_{ii} = 1$ when the feature components have second moment equal to 1. In this case, the numerator and the denominator can be directly expressed in terms of the order parameters of the model:

$$M_f = 1 + (\bar{\kappa}_2 - \kappa_2) \frac{q_d}{Q_d} \tag{31}$$

where

$$q_d \equiv \mathbb{E}_{\mathcal{D},F} \left\langle \frac{1}{N} \sum_{i=1}^{N} w_i^2 \right\rangle_{\boldsymbol{w}}, \tag{32a}$$

$$p_d \equiv \mathbb{E}_{\mathcal{D},F} \left\langle \frac{1}{N} \sum_{i,j=1}^{N} \Omega_{ij} w_i w_j \right\rangle_{\boldsymbol{w}}, \tag{32b}$$

$$Q_d \equiv \kappa_\star^2 q_d + \kappa_1^2 p_d. \tag{32c}$$

The order parameters q_d, p_d can be computed by solving saddle point equations as shown in appendix B.1.

Notice that in the case of a linear activation function the BMD is always 1 since a flip in the inputs will induce always the same response.

In figure 1 we show the plot of the generalization error and the corresponding BMD of the RFM at a fixed α_T, as a function of $1/\alpha$ for the MSE (left panels) and CE loss (right panels). As shown in Gerace *et al* (2020), for small regularization λ the generalization error develops a peak approximately where the model starts to fit all training data. In the case of the MSE loss, this threshold is often called interpolation threshold and it is located at $N = P$. When using the CE loss, this happens when the projected data become linearly separable and the exact location of the threshold strongly depends on the input statistics and features. Exactly in the correspondence of the generalization error peak the BMD displays its own peak, meaning that the function implemented by the network is more sensitive to perturbation of the inputs.

An interesting insight can be deduced from the behavior of the BMD at the optimal value of regularization for the RFM (dashed red curves in figure 1). While the generalization error becomes monotonic as the over-parametrization is increased, the BMD still reaches a peak at first and then descends to 1 only in the kernel limit $N/P \to \infty$. This might be surprising since the ground-truth linear model, the teacher, has BMD equal to 1 and one would expect the best generalizing RFM to achieve the best possible approximation of this function and therefore to match its BMD. However, blind minimization of the BMD is not compatible with good generalization, as seen from the performance of the RFM with very large

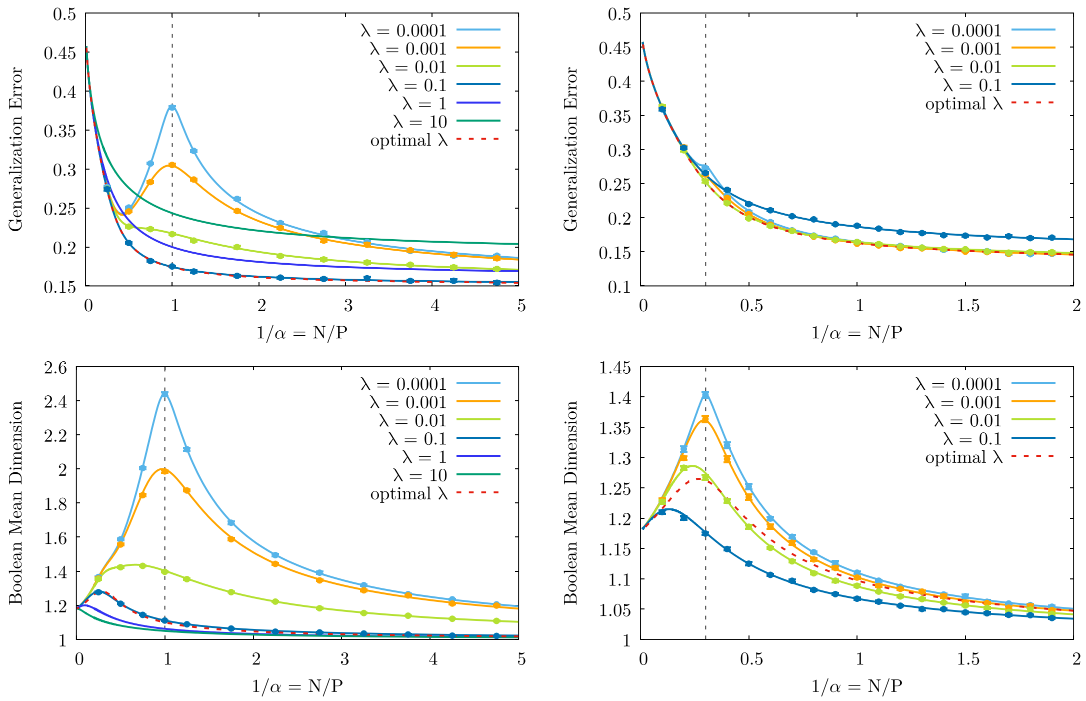

Figure 1. Generalization error (top panels) and BMD (bottom panels) as a function of the overparameterization degree $1/\alpha = N/P$, for fixed $\alpha_T = P/D = 3$ and with $\sigma = \tanh$. The left and right panels represents respectively the case of the MSE and of the CE loss. Several value of the regularization λ are displayed, together with the optimal one (which was found by minimizing the generalization error for each value of α, see red dashed line). As it can be seen in both plots, for small regularization λ, the location of the peak in the generalization error exactly coincides with the one in the BMD (vertical dashed lines). As one increases the regularization the peak in the both the generalization and the BMD is milded.

regularization λ. The explanation of this comes from the architectural mismatch between the linear teacher and the RFM: according to the GET the RFM learning problem is equivalent to a linear problem with an additional noise with an intensity regulated by the degree of non-linearity of the activation function (d'Ascoli *et al* 2020). This noise initially forces the under-parameterized RFM to overstretch its parameters to fit the data, causing an increased sensitivity to input perturbations. As the over-parameterization is increased, the RFM becomes equivalent to an optimally regularized linear model (Gerace *et al* 2020) and the BMD slowly drops to 1 in this limit.

Note that in the large dataset limit, when $\alpha, \alpha_T \to \infty$ with $\alpha_D = \mathcal{O}(1)$, a secondary peak for the BMD of the RFM emerges around $\alpha_D = 1$, i.e. when the number of parameters of the RFM is the same as the number of input features. This peak is caused by the insurgence of singular values in the spectrum of the covariance matrix Ω and is more accentuated at lower values of the regularization. Since modern deep networks operate in a completely different regime from the large dataset limit specified above, we expect this secondary peak not to be visible in realistic settings. For example, in the above plots in the low regularization regime, this peak is overshadowed by the main BMD peak. We analyze this phenomenology in detail in appendix C.

5. Numerical results

In the following subsections, we explore numerically the robustness of the BMD phenomenology analyzed in the RFM, considering different types of data distribution, model architecture and learning task.

Furthermore, we show that adversarially initialized models also display higher BMD, and that the increased sensitivity associated with a large BMD can hinder the robustness of the model against random perturbations of the training inputs.

Finally, we show that the location of the BMD peak is robust to the choice of input statistics used for its measurement, even in non-i.i.d. settings.

5.1. Experimental setup

In the following subsections, each panel displays the performance of a large number of different model architectures with varying degree of over-parameterization, trained on different datasets. Except where specified otherwise, all model are initialized with the common *Xavier* method (Glorot and Bengio 2010) and use the Adam optimizer (Kingma and Ba 2014), with batch size 128 and learning rate 10^{-4}. No specific early stopping criterion is implemented. As in other works analysing the double descent, we experiment with different levels of uniformly random label noise during training (which is introduced by corrupting a random fraction of labels), which tends to make the double descent peak more pronounced (Nakkiran *et al* 2021). We discuss the effect of label noise below.

5.1.1. Model architectures

We consider different types of model architectures:

- RFM, described above, where the number of hidden neurons in the first (fixed) layer controls the degree of over-parameterization.
- Two-layer fully-connected network (MLP) with tanh activation, where the number of hidden neurons in the first layer controls the degree of over-parameterization.
- ResNet-18: a family of minimal ResNet (He *et al* 2016) architectures based on the implementation of (Nakkiran *et al* 2021). The structure is finalized with fully connected and softmax layers. As in Nakkiran *et al* (2021), we control the over-parameterization of the model by changing the number of channels in the convolutional layers. Namely, the 4 ResNet blocks contain convolutional layers of widths $[k, 2k, 4k, 8k]$, with k varying from 1 to 20.

Both RFM and two-layer fully connected networks in our experiments use hyperbolic tangent activation functions and have weights initialized from a Gaussian distribution and bias terms initialized with zeros. The loss function optimized during training is the cross-entropy loss with L_2 regularization (the intensity of the regularization is set to zero if not specified otherwise).

5.1.2. Data preprocessing

In the following experiments, we use continuous inputs during the training of the models, normalizing the input features to lie within the $[-1,1]$ interval. While such normalizations are common in preprocessing pipelines, here this procedure has also the benefit of matching the range of variability of the training inputs with that of the randomly i.i.d. sampled binary sequences used to estimate the BMD. We explore the effect of different normalization ranges in appendix section E.

5.2. MD and generalization peaks as a function of overparametrization

In figure 2 we show train and test error, and the BMD for an RFM trained with and without label noise on binary MNIST (even vs odd digits) as a function of the hidden layer width. In figure 3, we instead consider a two-layer MLP trained on 10-digits MNIST (varying width) and a ResNet-18 trained on CIFAR10 (varying number of channels), both with label noise. In the multi-label case, we are defining the BMD of the network as the average of the BMDs over the classes, where the output of the network is a vector of predicted log-probabilities for each class (i.e. there is a log-softmax activation in the last layer).

5.2.1. Position of the BMD peak

The BMD displays a peak around the point where the number of parameters of the model allows it to reach zero training error, in close correspondence with the generalization error peak. We find this phenomenology to be robust with respect to the model class, the dataset, and the over-parameterization procedure. Notice however, that standard optimizers based on SGD are able to implicitly regularize the trained models and can strongly reduce the peaking behavior, as already observed in the context of double descent. In the presented figures we introduced label-noise, which ensures the presence of over-fitting and is thus able to restore both peaks.

An important observation is that, in order to see this phenomenology, it is not necessary to account for the training input distribution for the evaluation of the MD, which would not be possible in the case of real data. In fact, in the over-fitting regime, it is possible to detect an increased sensitivity of the neural network function for multiple input distributions, including the i.i.d. binary inputs entailed in the BMD evaluation. This is explored further in section 5.6.

5.2.2. Asymptotic behavior of the BMD

When the degree of parametrization of the model is further increased, the BMD decreases and settles on an asymptotic value. The decrease of the BMD in the number of parameters is faster with lower label noise, see figure 2(left panel vs. right panel). The asymptotic value, reached in the limit of an infinite number of parameters, is task- and model-dependent. For example, in figure 2, the functions learned by the RFMs no longer approximate a linear model (BMD equal to 1), and are instead bound to higher values of the BMD.

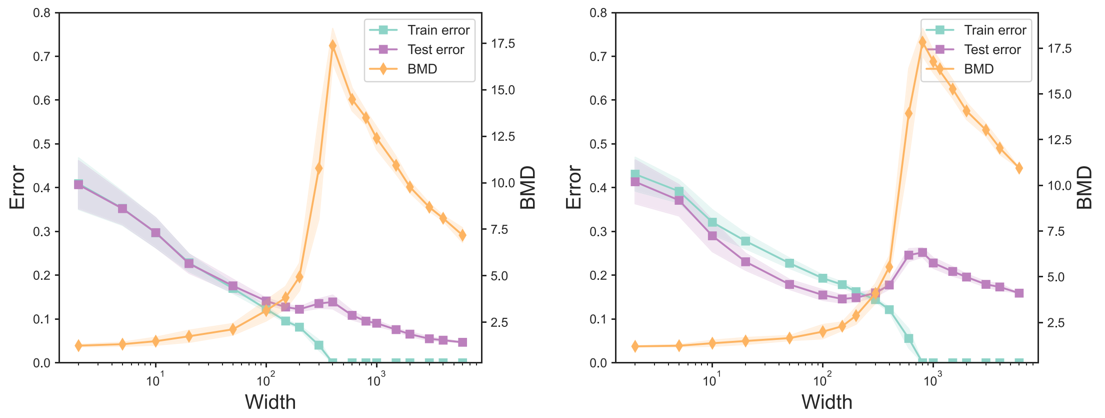

Figure 2. Train error (turquoise), test error (violet) and BMD (orange) curves of the random feature model trained on the MNIST dataset with binary labels on 5 K train samples with 0% label noise (left) and 10% label noise (right), tested on 5 K samples. The resulting plots represent an average (and standard deviations) obtained repeating 20 different times the experiment.

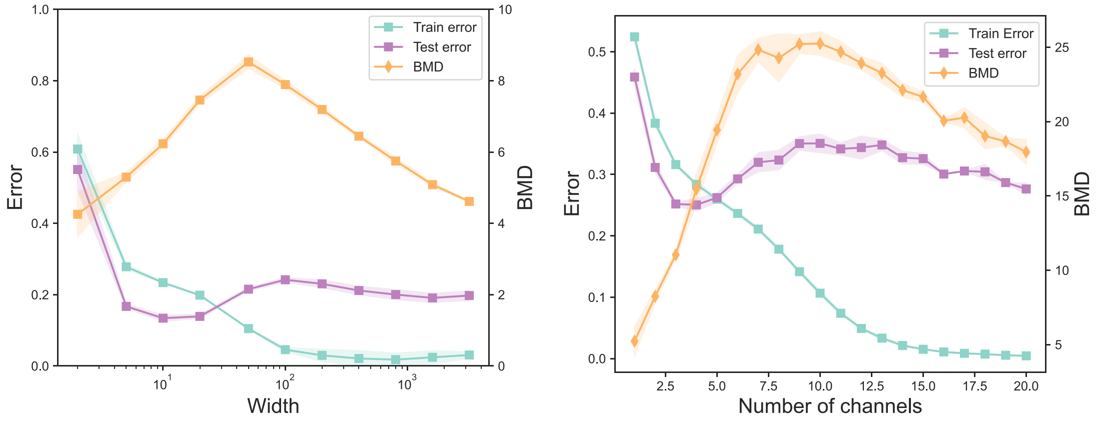

Figure 3. (Left) Train error (turquoise), test error (violet) and BMD (orange) of the two-layer fully-connected network trained on the MNIST dataset with 10 labels on 20 K train samples with 20% label noise, tested on the 5 K samples. The resulting plot represents an average (and standard deviations) obtained repeating 20 different times the experiment. (Right) Train error, test error and BMD of ResNet-18 trained on the CIFAR-10 with 15% label noise in the train set. The resulting plot represents an average (and standard deviations) obtained repeating 5 different times the experiment.

5.2.3. Visibility of the BMD peak and Label Noise

The double-descent generalization peak can be a very subtle phenomenon when the learning task is too coherent and the noise level in the data is too weak. With this type of data, the phenomenon can be made more evident (Nakkiran *et al* 2021) by adding label noise to the training data. This strategy naturally reduces the signal-to-noise ratio and increases the over-fitting potential during training. The BMD peak, however, seems to be easily identifiable even with zero label noise, (see left panel of figure 2) where the generalization peak is less pronounced. Note that the BMD does not require any data (neither training nor test) in order to be estimated, so it can be used as a black-box test for assessing the proximity to the separability threshold and therefore as a signal of over-fitting.

5.2.4. Impact of regularization

It has been shown that regularizing the model weakens the double-descent peak and that, at the optimal value of the regularization intensity, the generalization error smoothly decreases with the degree of over-parameterization. Similarly, the BMD peak can be dampened by adding stronger regularization, as shown in figure 4.

5.2.5. BMD and training set size

In this section, we investigate the effect of varying the number of training samples for a fixed model capacity and training procedure. By increasing the number of training samples, starting from a low number, the same model can switch from being over- to under-parameterized. Therefore increasing the number of training

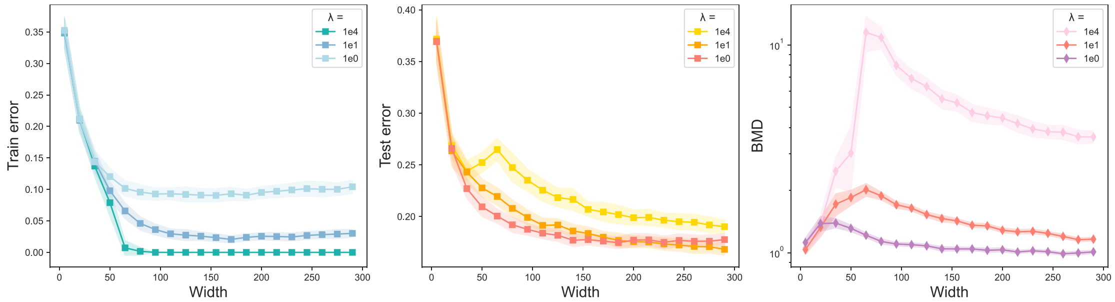

Figure 4. Impact of training with L2 regularization of the random feature model using the MNIST dataset with 10 labels, 200 train samples with no label noise and evaluating on 5 K test samples. The loss used is cross-entropy. In all plots, the curves are colored by the strength of the regularization weight λ. (Left) Regularization effect on the train error. (Center) Regularization effect on the test error. (Right) Regularization effect on the BMD. The generalization error smoothly decreases with the degree of over-parameterization. Similarly, the BMD peak can be dampened by adding stronger regularization.

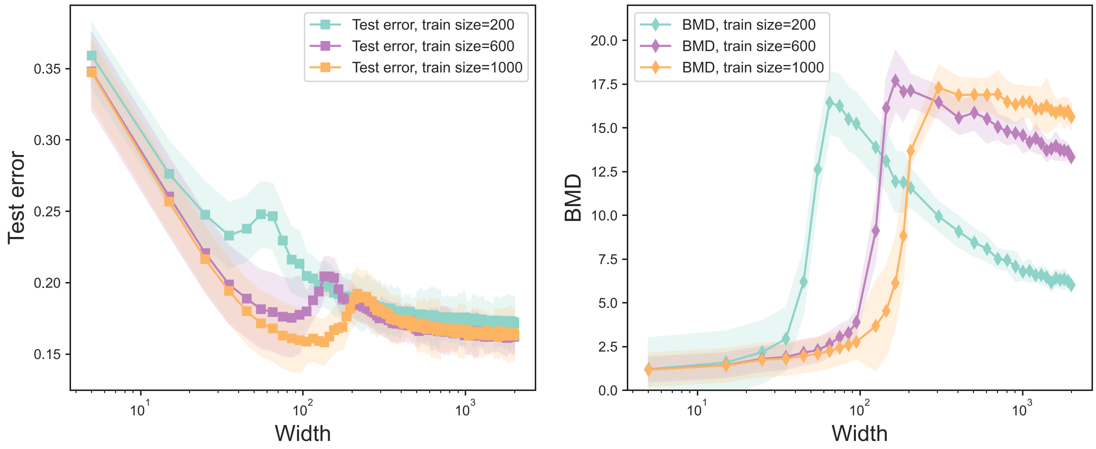

Figure 5. Effect of changing the training set size on the test error (left panel) and the BMD (right panel) of the random feature model using the MNIST dataset with 10 labels, with no label noise and evaluating on 200 test samples. The resulting plot represents an average (and standard deviations) obtained repeating 15 different times the experiment.

samples has two effects on the test error curve: on the one hand, increasing the number of training samples decreases the test error, shifting the test error curve mostly downwards. On the other hand, increasing the number of training samples increases the capacity at which the double descent peak occurs since a higher capacity is needed until the training set is effectively memorized. This shifts the test error curve (and the BMD curve) to the right. This effect can be seen in figure 5.

5.3. BMD and adversarial initialization

In this section, we analyze the BMD of two-layer fully connected networks under adversarial initialization (Liu *et al* 2020) on the MNIST dataset. This initialization scheme can be used to artificially hinder the generalization performance of the model, forcing it to converge on a bad minimum of the loss. We here aim to show that the initialization has also an effect on the BMD of the model, increasing the sensitivity of the network.

The adversarial initialization protocol works as follows. We train a two-layer fully connected network in two different phases: in the first phase, we push the network towards an adversarial initialization by pretraining the model with 100% label noise for a fixed amount of epochs; in the second phase, we train the model on the original dataset, with no label noise, for 200 epochs. The resulting plot, in figure 6(left panel), represents an average over 15 different realizations of the experiment and shows the effect of the length of the pretraining phase on both generalization performance and BMD of the network. In agreement with our analysis, we observe a simultaneous increase of the two metrics when the adversarial initialization phase is longer and the network is driven towards worse generalization.

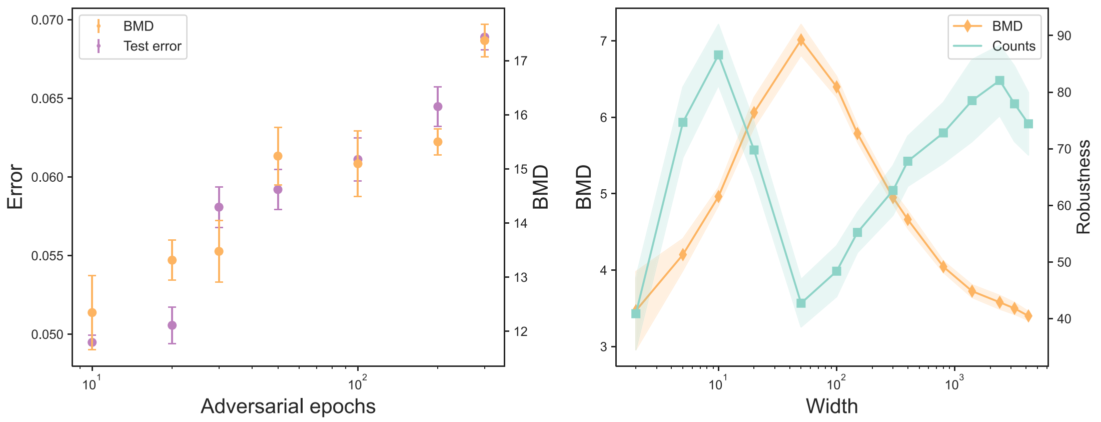

Figure 6. (Left): BMD (orange points), and test error (violet points) estimated for a two-layer fully connected network of width 10^3, trained according to the adversarial initialization protocol described in section 5.3 on 20 K samples of the MNIST dataset. On the horizontal axis we vary the number of pretraining epochs, and plot the corresponding increase in the generalization error and the BMD of the model after the second learning stage. The points represent an average of 15 different realizations of the experiment. (Right): BMD (orange line) and counts (turquoise line) estimated for a two-layer fully connected network trained on the MNIST dataset using 20 K train samples with 20% label noise and tested on 5 K samples. Counts represent the average amount of sign flips of random pixels of a correctly predicted test image that are necessary to fool the model to a wrong class label. The amount is averaged over all the correctly predicted test data samples. The resulting plot represents an average of 40 different realizations of the experiment. We observe that higher values of the BMD correspond to lower robustness of the model and vice versa.

5.4. BMD and robustness against adversarial attacks

In this section, we analyze the connection between BMD of a model and its robustness to adversarial attacks. We consider a two-layer fully-connected network trained on MNIST with 10 classes. We define as our robustness measure the average count of sign flips of randomly chosen pixels, needed to change the model prediction on a test sample that was previously classified correctly. The lower the counts, the lower the robustness of the model. Varying the capacity of the model by varying the width of the hidden layer, we plot this robustness measure against the BMD of the model in figure 6(right panel). We observe that BMD and robustness strongly anti-correlate, with the peak in BMD coinciding with a minimum of robustness.

5.5. Pixel-wise contributions to BMD

The MD as expressed in equation (15) is proportional to a sum of contributions τ_i^2 of single features indexed by i. Similar to (Hahn *et al* 2022), we plot these contributions in figure 7 as a heatmap, where the bright spots indicate features that contribute strongly to the MD. We show four heatmaps, corresponding to different capacities and at different distances from the BMD peak, for a two-layer fully connected network trained on MNIST.

Note that the colors are normalized to the $[0,1]$ range, so that very bright spots correspond to pixels that contribute to the BMD the most. It can be seen that for under-parametrized networks few pixels give the largest contribution to the BMD. Near the BMD peak, a large fraction of the pixels in the center of the image dominate the BMD, and for even larger capacities we again have fewer pixels with maximal values. This can be interpreted as the classifier losing 'focus' at the interpolation point and paying attention to fewer patterns in the over-parametrized regime.

5.6. Different distributions for estimating BMD

In BMD estimates for the previous experiments, equation (16), we focused on the case of i.i.d. binary input features. In the RFM, however, we have shown analytically that there exists a universality for the MD when one considers separable input distributions with the same first and second moments. In the numerical experiments, we have also shown evidence that the BMD peak can still provide insights into the behavior of the neural network function on the training and test data, which follow very different input statistics. To explore in detail the role of the input statistics, and of the presence of correlations in the input features, we measure the MD by resampling the inputs from different distributions: in figure 8 we plot the normalized MD curves for features sampled from:

- a uniform binary distribution (BMD).
- a standard normal (Gaussian) distribution $\mathcal{N}(0,1)$.

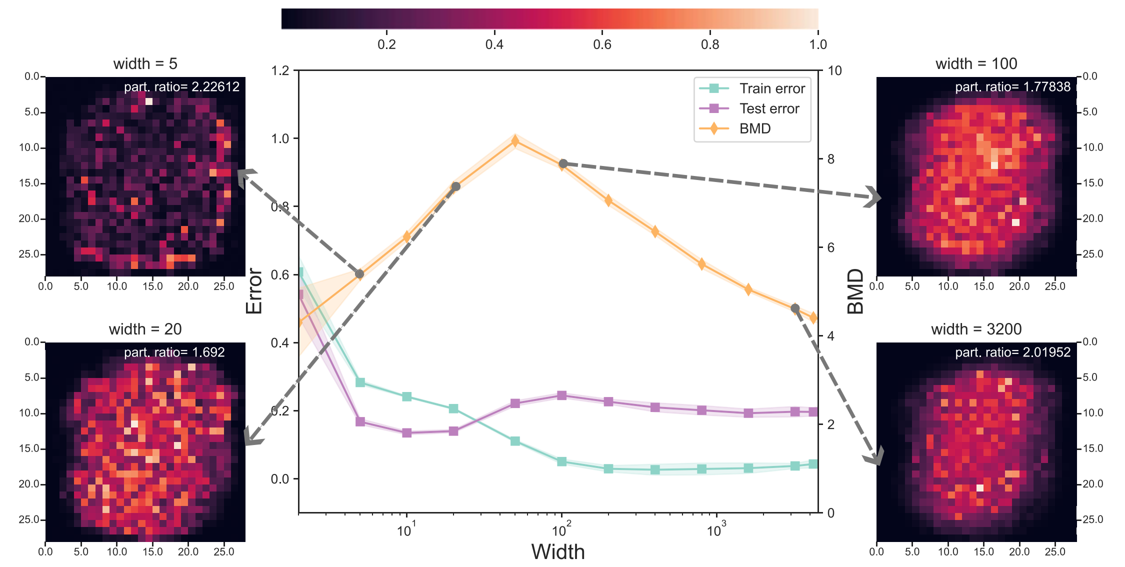

Figure 7. Heatmaps of the pixel contributions (τ_i^2 for $1 \leqslant i \leqslant 784$) estimated on the two-layer fully connected network trained on 20 K samples of the MNIST dataset with 10 classes, 20% label noise and normalized to lie within $[0,1]$ interval. The (rescaled) participation ratio is defined as $n \times \frac{\sum_{i=1}^n \tau_i^2}{(\sum_{i=1}^n \tau_i)^2}$. After the rescaling, a participation ratio of 1 indicates a uniform distribution of pixel contributions, while a value of n indicates a distribution concentrated over a single pixel. The heatmaps correspond to the contributions estimated with respect to label 0 for the models of different capacities (hidden layer dimensions) and represent only one seed, while the resulting curves on the plot represent an average over 20 different runs of the experiment.

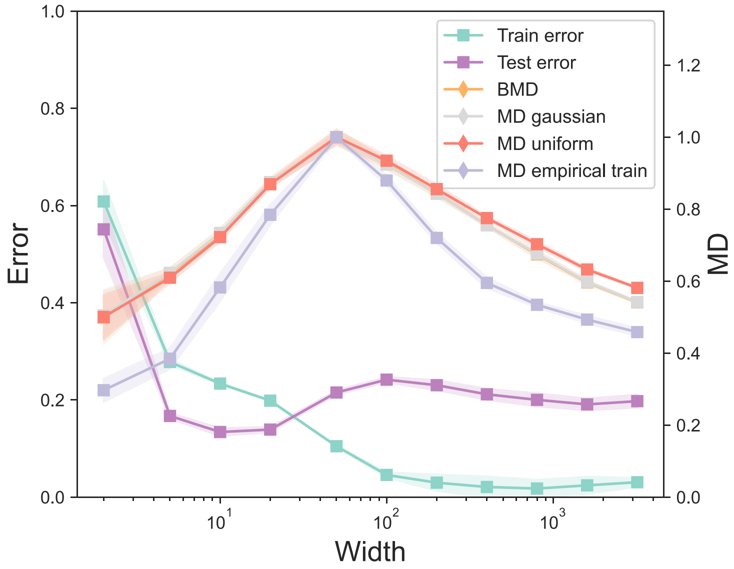

Figure 8. Mean dimensions estimated using Monte Carlo (equation (16)) w.r.t. different distributions of the two-layer fully connected network trained on the 20 K of MNIST samples with 20% label noise and tested on 5 K samples. The MD values are normalized to lie within [0,1] interval. The choice of the distribution does not affect the location of the peak. Moreover, distributions, which first two moments coincide (e.g. binary uniform Unif$\{-1,1\}$ and Gaussian $\mathcal{N}(0,1)$) yield the same MD pattern. The resulting plot represents an average over 20 different runs of the experiment.

- a uniform distribution in the range $[-1,1]$.
- empirical distribution of the training data with random uniform resampling in the range $[-1,1]$.

As one can see in figure 8, the MD curves estimated with binary and Gaussian i.i.d. inputs, with matching moments, are identical. With the uniform distribution, the second moment is $1/3$ and this results in a slightly rescaled MD curve. Introducing correlations in the inputs, in the MD estimated over the training data distribution, the curve still shows a similar behavior, and importantly the peak is found at the same value.

6. Discussion

In this work, we analyzed the BMD as a tool for assessing the sensitivity of neural network functions. In the treatable setting of the RFM, we derived an exact characterization of the behavior of this metric as a function of the degree of overparameterization of the model. Notably, we found a strong correlation between the sharp increase of the BMD and the increase of the generalization error around the interpolation threshold. This finding indicates that as the neural network starts to overfit the noise in the data, the learned function becomes more sensitive to small perturbations of the input features. Importantly, while the double descent curve requires test data to be observed, the BMD can signal this type of failure mode by using information from the neural network alone. The same phenomenology appears in more realistic scenarios with different architectures and datasets, where factors influencing double descent, like regularization and label noise, are also found to affect the BMD in similar fashion. Furthermore, we demonstrated that the BMD is informative about the vulnerability of trained models to adversarial attacks, despite assuming an input distribution that is very different from that of the training dataset.

Our study raises intriguing questions regarding the potential applications of BMD for regularization purposes. Another interesting future direction could be to investigate how comparing the BMDs achieved by a highly parametrized neural network trained on different datasets can help assess the effective dimensionality of the training data and the complexity of the discriminative tasks. Finally, it could be interesting to extend the study of the BMD in the RFM in the framework of a polynomial teacher model, recently analyzed in Aguirre-López *et al* (2024).

Data availability statement

No new data were created or analysed in this study.

Acknowledgments

We thank Emanuele Borgonovo for the insightful discussions. Enrico Malatesta and Luca Saglietti aknowledge the MUR-Prin 2022 funding, Prot. 20229T9EAT and 2022E3WYTY, financed by the EU (Next Generation EU).

ORCID iDs

Elizaveta Demyanenko https://orcid.org/0009-0002-4366-6825
Enrico M Malatesta https://orcid.org/0000-0001-8558-6175

References

Advani M S, Saxe A M and Sompolinsky H 2020 High-dimensional dynamics of generalization error in neural networks *Neural Netw.* **132** 428–46

Aguirre-López F, Franz S and Pastore M 2024 Random features and polynomial rules (arXiv:2402.10164)

Baity-Jesi M, Sagun L, Geiger M, Spigler S, Arous G B, Cammarota C, LeCun Y, Wyart M and Biroli G 2018 Comparing dynamics: deep neural networks versus glassy systems *Int. Conf. on Machine Learning* (PMLR) pp 314–23

Baldassi C, Lauditi C, Malatesta E M, Pacelli R, Perugini G and Zecchina R 2022 Learning through atypical phase transitions in overparameterized neural networks *Phys. Rev.* E **106** 014116

Baldassi C, Malatesta E M, Negri M and Zecchina R 2020 Wide flat minima and optimal generalization in classifying high-dimensional gaussian mixtures *J. Stat. Mech.* 124012

Belkin M, Hsu D, Ma S and Mandal S 2019 Reconciling modern machine-learning practice and the classical bias–variance trade-off *Proc. Natl Acad. Sci.* **116** 15849–54

Brown T *et al* 2020 Language models are few-shot learners *Advances in Neural Information Processing Systems* vol 33 pp 1877–901

d'Ascoli S, Refinetti M, Biroli G and Krzakala F 2020 Double trouble in double descent: bias and variance (s) in the lazy regime *Int. Conf. on Machine Learning* (PMLR) pp 2280–90

d'Ascoli S, Sagun L and Biroli G 2020 Triple descent and the two kinds of overfitting: Where & why do they appear? *Advances in Neural Information Processing Systems* vol 33 pp 3058–69

Efron B and Stein C 1981 The jackknife estimate of variance *Ann. Stat.* 586–96

Engel A and Van den Broeck C 2001 *Statistical Mechanics of Learning* (Cambridge University Press)

Feinauer C and Borgonovo E 2022 Mean dimension of generative models for protein sequences *bioRxiv Preprint* https://doi.org/10.1101/2022.12.12.520028 (posted online 17 December 2022)

Geiger M, Spigler S, d'Ascoli S, Sagun L, Baity-Jesi M, Biroli G and Wyart M 2019 Jamming transition as a paradigm to understand the loss landscape of deep neural networks *Phys. Rev.* E **100** 012115

Gerace F, Loureiro B, Krzakala F, Mezard M and Zdeborova L 2020 Generalisation error in learning with random features and the hidden manifold model *Proc. 37th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research* vol 119) ed H III Daumé and A Singh (PMLR) pp 3452–62

Gerace F, Saglietti L, Mannelli S S, Saxe A and Zdeborová L 2022 Probing transfer learning with a model of synthetic correlated datasets *Mach. Learn.: Sci. Technol.* **3** 015030

Glorot X and Bengio Y 2010 Xavier initialization *J. Mach. Learn. Res.*

Goldt S, Mézard M, Krzakala F and Zdeborová L 2019 Modelling the influence of data structure on learning in neural networks: the hidden manifold model *Phys. Rev.* X **10** 041044

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F and Pedreschi D 2018 A survey of methods for explaining black box models *ACM Comput. Surv. (CSUR)* **51** 1–42

Hahn R, Feinauer C and Borgonovo E 2022 The mean dimension of neural networks–what causes the interaction effects? (arXiv:2207.04890)

He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 770–8

Hoyt C and Owen A B 2021 Efficient estimation of the anova mean dimension, with an application to neural net classification *SIAM/ASA J. Uncertain. Quantification* **9** 708–30

Jiang Y, Neyshabur B, Mobahi H, Krishnan D and Bengio S 2019 Fantastic generalization measures and where to find them (arXiv:1912.02178)

Jumper J *et al* 2021 Highly accurate protein structure prediction with alphafold *Nature* **596** 583–9

Kingma D P and Ba J 2014 Adam: a method for stochastic optimization (arXiv:1412.6980)

Liu R and Owen A B 2006 Estimating mean dimensionality of analysis of variance decompositions *J. Am. Stat. Assoc.* **101** 712–21

Liu S, Papailiopoulos D and Achlioptas D 2020 Bad global minima exist and sgd can reach them *Advances in Neural Information Processing Systems* vol 33 pp 8543–52

Loureiro B, Gerbelot C, Cui H, Goldt S, Krzakala F, Mézard M and Zdeborová L 2021 Capturing the learning curves of generic features maps for realistic data sets with a teacher-student model (arXiv:2102.08127)

Malatesta E M 2023 High-dimensional manifold of solutions in neural networks: insights from statistical physics (arXiv:2309.09240)

Mei S and Montanari A 2019 The generalization error of random features regression: precise asymptotics and double descent curve (arXiv:1908.05355)

Mézard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications* vol 9 (World Scientific Publishing Company)

Mignacco F, Krzakala F, Lu Y, Urbani P and Zdeborova L 2020 The role of regularization in classification of high-dimensional noisy gaussian mixture *Int. Conf. on Machine Learning* (PMLR) pp 6874–83

Montavon G, Samek W and Müller K-R 2018 Methods for interpreting and understanding deep neural networks *Digit. Signal Process.* **73** 1–15

Nakkiran P, Kaplun G, Bansal Y, Yang T, Barak B and Sutskever I 2021 Deep double descent: where bigger models and more data hurt *J. Stat. Mech.* 124003

Neyshabur B, Bhojanapalli S, McAllester D and Srebro N 2017 Exploring generalization in deep learning *Advances in Neural Information Processing Systems* p 30

Novak R, Bahri Y, Abolafia D A, Pennington J and Sohl-Dickstein J 2018 Sensitivity and generalization in neural networks: an empirical study *Int. Conf. on Learning Representations*

O'Donnell R 2014 *Analysis of Boolean Functions* (Cambridge University Press)

OpenAI 2023 *Gpt-4 Technical Report* (available at: https://arxiv.org/abs/2303.08774)

Opper M 1995 Statistical mechanics of learning: generalization *The Handbook of Brain Theory and Neural Networks* pp 922–5

Owen A B 2003 The dimension distribution and quadrature test functions *Stat. Sin.* 1–17

Rahimi A and Recht B 2007 Random features for large-scale kernel machines *Neural Information Processing Systems* vol 3 (Citeseer) p 5

Ramesh A, Dhariwal P, Nichol A, Chu C and Chen M 2022 Hierarchical text-conditional image generation with clip latents (arXiv:2204.06125)

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B 2022 High-resolution image synthesis with latent diffusion models *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 10684–95

Rudin C 2019 Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead *Nat. Mach. Intell.* **1** 206–15

Sejnowski T J 2020 The unreasonable effectiveness of deep learning in artificial intelligence *Proc. Natl Acad. Sci.* **117** 30033–8

Touvron H *et al* 2023 Llama 2: open foundation and fine-tuned chat models (available at: https://arxiv.org/abs/2307.09288)

Valle-Perez G, Camargo C Q and Louis A A 2018 Deep learning generalizes because the parameter-function map is biased towards simple functions (arXiv:1805.08522)

Vapnik V N 1999 An overview of statistical learning theory *IEEE Trans. Neural Netw.* **10** 988–99

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I 2017 Attention is all you need *Advances in Neural Information Processing Systems* p 30

Vilone G and Longo L 2020 Explainable artificial intelligence: a systematic review (arXiv:2006.00093)