Generalisation error in learning with random features and the hidden manifold model

Federica Gerace [*1] **Bruno Loureiro** [*1] **Florent Krzakala** [2] **Marc Mézard** [2] **Lenka Zdeborová** [1]

Abstract

We study generalised linear regression and classification for a synthetically generated dataset encompassing different problems of interest, such as learning with random features, neural networks in the lazy training regime, and the hidden manifold model. We consider the high-dimensional regime and using the replica method from statistical physics, we provide a closed-form expression for the asymptotic generalisation performance in these problems, valid in both the under- and over-parametrised regimes and for a broad choice of generalised linear model loss functions. In particular, we show how to obtain analytically the so-called double descent behaviour for logistic regression with a peak at the interpolation threshold, we illustrate the superiority of orthogonal against random Gaussian projections in learning with random features, and discuss the role played by correlations in the data generated by the hidden manifold model. Beyond the interest in these particular problems, the theoretical formalism introduced in this manuscript provides a path to further extensions to more complex tasks.

1. Introduction

One of the most important goals of learning theory is to provide generalisation bounds describing the quality of learning a given task as a function of the number of samples. Existing results fall short of being directly relevant for the state-of-the-art deep learning methods (Zhang et al., 2016; Neyshabur et al., 2017). Consequently, providing tighter results on the generalisation error is currently a very active research subject. The traditional learning theory approach to generalisation follows for instance the Vapnik-Chervonenkis (Vapnik, 1998) or Rademacher (Bartlett & Mendelson, 2002) worst-case type bounds, and many of their more recent extensions (Shalev-Shwartz & Ben-David, 2014). An alternative approach, followed also in this paper, has been pursued for decades, notably in statistical physics, where the generalisation ability of neural networks was analysed for a range of "typical-case" scenario *for synthetic data arising from a probabilistic model* (Seung et al., 1992; Watkin et al., 1993; Advani et al., 2013; Advani & Saxe, 2017; Aubin et al., 2018; Candès & Sur, 2020; Hastie et al., 2019; Mei & Montanari, 2019; Goldt et al., 2019). While at this point it is not clear which approach will lead to a complete generalisation theory of deep learning, it is worth pursuing both directions.

The majority of works following the statistical physics approach study the generalisation error in the so-called teacher-student framework, where the input data are element-wise i.i.d. vectors, and the labels are generated by a teacher neural network. In contrast, in most of real scenarios the input data do not span uniformly the input space, but rather live close to a lower-dimensional manifold. The traditional focus onto i.i.d. Gaussian input vectors is an important limitation that has been recently stressed in (Mézard, 2017; Goldt et al., 2019). In (Goldt et al., 2019), the authors proposed a model of synthetic data to mimic the latent structure of real data, named the *hidden manifold model*, and analysed the learning curve of one-pass stochastic gradient descent algorithm in a two-layer neural network with a small number of hidden units also known as committee machine.

Another key limitation of the majority of existing works stemming from statistical physics is that the learning curves were only computed for neural networks with a few hidden units. In particular, the input dimension is considered large, the number of samples is a constant times the input dimension and the number of hidden units is of order one. Tight learning curves were only very recently analysed for two-layer neural networks with more hidden units. These studies addressed in particular the case of networks that have a fixed first layer with random i.i.d. Gaussian weights (Hastie et al., 2019; Mei & Montanari, 2019), or the lazy-training regime where the individual weights change only

[*]Equal contribution [1]Université Paris-Saclay, CNRS, CEA, Institut de physique théorique, 91191, Gif-sur-Yvette, France. [2]Laboratoire de Physique de École normale supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, F-75005 Paris, France. Correspondence to: Bruno Loureiro <brloureiro@gmail.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

infinitesimally during training, thus not learning any specific features (Chizat et al., 2019; Jacot et al., 2018; Geiger et al., 2019b).

In this paper we compute the generalisation error and the corresponding learning curves, i.e. the test error as a function of the number of samples for a model of high-dimensional data that encompasses at least the following cases:

- generalised linear regression and classification for data generated by the hidden manifold model (HMM) of (Goldt et al., 2019). The HMM can be seen as a single-layer generative neural network with i.i.d. inputs and a rather generic feature matrix (Louart et al., 2018; Goldt et al., 2019).
- Learning data generated by the teacher-student model with a random-features neural network (Rahimi & Recht, 2008), with a very generic feature matrix, including deterministic ones. This model is also interesting because of its connection with the lazy regime, that is equivalent to the random features model with slightly more complicated features (Chizat et al., 2019; Hastie et al., 2019; Mei & Montanari, 2019).

We give a closed-form expression for the generalisation error in the high-dimensional limit, obtained using a non-rigorous heuristic method from statistical physics known as the replica method (Mézard et al., 1987), that has already shown its remarkable efficacy in many problems of machine learning (Seung et al., 1992; Engel & Van den Broeck, 2001; Advani et al., 2013; Zdeborová & Krzakala, 2016), with many of its predictions being rigorously proven, e.g. (Talagrand, 2006; Barbier et al., 2019). While in the present model it remains an open problem to derive a rigorous proof for our results, we shall provide numerical support that the formula is indeed exact in the high-dimensional limit, and extremely accurate even for moderately small system sizes.

1.1. The model

We study high-dimensional regression and classification for a *synthetic* dataset $\mathcal{D} = \{(\boldsymbol{x}^\mu, y^\mu)\}_{\mu=1}^n$ where each sample μ is created in the following three steps: (i) First, for each sample μ we create a vector $\boldsymbol{c}^\mu \in \mathbb{R}^d$ as

$$\boldsymbol{c}^\mu \sim \mathcal{N}(0, \mathrm{I}_d)\,, \tag{1}$$

(ii) We then draw $\boldsymbol{\theta}^0 \in \mathbb{R}^d$ from a separable distribution P_θ and draw independent labels $\{y^\mu\}_{\mu=1}^n$ from a (possibly probabilistic) rule f^0:

$$y^\mu = f^0\left(\frac{1}{\sqrt{d}}\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0\right) \in \mathbb{R}\,. \tag{2}$$

(iii) The input data points $\boldsymbol{x}^\mu \in \mathbb{R}^p$ are created by a one-layer generative network with fixed and normalised weights $\mathrm{F} \in \mathbb{R}^{d\times p}$ and an activation function $\sigma : \mathbb{R} \to \mathbb{R}$, acting component-wise:

$$\boldsymbol{x}^\mu = \sigma\left(\frac{1}{\sqrt{d}}\mathrm{F}^\top \boldsymbol{c}^\mu\right)\,. \tag{3}$$

We study the problem of supervised learning for the dataset $\mathcal{D}$ aiming at achieving a low generalisation error ϵ_g on a new sample $\boldsymbol{x}^\mathrm{new}, y^\mathrm{new}$ drawn by the same rule as above, where:

$$\epsilon_\mathrm{g} = \frac{1}{4^k}\mathbb{E}_{\boldsymbol{x}^\mathrm{new}, y^\mathrm{new}}\left[(\hat{y}_{\boldsymbol{w}}(\boldsymbol{x}^\mathrm{new}) - y^\mathrm{new})^2\right]. \tag{4}$$

with $k = 0$ for regression task and $k = 1$ for classification task. Here, $\hat{y}_{\boldsymbol{w}}$ is the prediction on the new label y^new of the form:

$$\hat{y}_{\boldsymbol{w}}(\boldsymbol{x}) = \hat{f}\left(\boldsymbol{x}\cdot\hat{\boldsymbol{w}}\right). \tag{5}$$

The weights $\hat{\boldsymbol{w}} \in \mathbb{R}^p$ are learned by minimising a loss function with a ridge regularisation term (for $\lambda \geq 0$) and defined as

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\mathrm{argmin}}\left[\sum_{\mu=1}^n \ell(y^\mu, \boldsymbol{x}^\mu\cdot\boldsymbol{w}) + \frac{\lambda}{2}||\boldsymbol{w}||_2^2\right], \tag{6}$$

where $\ell(\cdot,\cdot)$ can be, for instance, a logistic, hinge, or square loss. Note that although our formula is valid for any f^0 and $\hat{f}$, we take $f^0 = \hat{f} = \mathrm{sign}$, for the classification tasks and $f^0 = \hat{f} = \mathrm{id}$ for the regression tasks studied here. We now describe in more detail the above-discussed reasons why this model is of interest for machine learning.

Hidden manifold model: The dataset $\mathcal{D}$ can be seen as generated from the *hidden manifold model* introduced in (Goldt et al., 2019). From this perspective, although $\boldsymbol{x}^\mu$ lives in a p dimensional space, it is parametrised by a latent d-dimensional subspace spanned by the rows of the matrix F which are "hidden" by the application of a scalar non-linear function σ. The labels y^μ are drawn from a generalised linear rule defined on the latent d-dimensional subspace via eq. (2). In modern machine learning parlance, this can be seen as data generated by a one-layer generative neural network, such as those trained by generative adversarial networks or variational auto-encoders with random Gaussian inputs $\mathbf{c}^\mu$ and a rather generic weight matrix F (Goodfellow et al., 2014; Kingma & Welling, 2013; Louart et al., 2018; Seddik et al., 2020).

Random features: The model considered in this paper is also an instance of the random features learning discussed in (Rahimi & Recht, 2008) as a way to speed up kernel-ridge-regression. From this perspective, the $\boldsymbol{c}^\mu$s $\in \mathbb{R}^d$ are regarded as a set of d-dimensional i.i.d. Gaussian data points, which are projected by a feature matrix $\mathrm{F} = (\boldsymbol{f}_\rho)_{\rho=1}^p \in \mathbb{R}^{d\times p}$ into a higher dimensional space, followed by a non-linearity σ. In the $p \to \infty$ limit of infinite

number of features, performing regression on $\mathcal{D}$ is equivalent to kernel regression on the $\mathbf{c}^\mu$s with a deterministic kernel $K(\mathbf{c}^{\mu_1}, \mathbf{c}^{\mu_2}) = \mathbb{E}_{\mathbf{f}}\left[\sigma(\mathbf{f}\cdot\mathbf{c}^{\mu_1}/\sqrt{d})\cdot\sigma(\mathbf{f}\cdot\mathbf{c}^{\mu_2}/\sqrt{d})\right]$ where $\mathbf{f}\in\mathbb{R}^d$ is sampled in the same way as the rows of F. Random features are also intimately linked with the lazy training regime, where the weights of a neural network stay close to their initial value during training. The training is lazy as opposed to a "rich" one where the weights change enough to learn useful features. In this regime, neural networks become equivalent to a random feature model with correlated features (Chizat et al., 2019; Du et al., 2018; Allen-Zhu et al., 2019; Woodworth et al., 2019; Jacot et al., 2018; Geiger et al., 2019b).

1.2. Contributions and related work

The main contribution of this work is a closed-form expression for the generalisation error $\epsilon_{\rm g}$, eq. (8), that is valid in the high-dimensional limit where the number of samples n, and the two dimensions p and d are large, but their respective ratios are of order one, and for generic sequence of matrices F satisfying the following *balance conditions*:

$$\frac{1}{\sqrt{p}}\sum_{i=1}^{p} w_i^{a_1}w_i^{a_2}\cdots w_i^{a_s}\mathrm{F}_{i\rho_1}\mathrm{F}_{i\rho_2}\cdots\mathrm{F}_{i\rho_q} = O(1), \quad (7)$$

where $\{\boldsymbol{w}^a\}_{a=1}^r$ are r independent samples from the Gibbs measure (14), and $\rho_1,\rho_2,\cdots,\rho_q \in \{1,\cdots,d\}$, $a_1,a_2,\cdots,a_s\in\{1,\cdots,r\}$ are an arbitrary choice of subset of indices, with $s,q\in\mathbb{Z}_+$. The non-linearities $f^0,\hat{f},\sigma$ and the loss function ℓ can be arbitrary. Our result for the generalisation error stems from the replica method and we conjecture it to be exact for convex loss functions ℓ. It can also be useful for non-convex loss functions but in those cases it is possible that the so-called replica symmetry breaking (Mézard et al., 1987) needs to be taken into account to obtain an exact expression. In the present paper we hence focus on convex loss functions ℓ and leave the more general case for future work. The final formulas are simpler for non-linearities σ that give zero when integrated over a centred Gaussian variable, and we hence focus on those cases.

An interesting application of our setting is ridge regression, i.e. taking $\hat{f}(x)=x$ with square loss, and random i.i.d. Gaussian feature matrices. For this particular case (Mei & Montanari, 2019) proved an equivalent expression. Indeed, in this case there is an explicit solution of eq. (6) that can be rigorously studied with random matrix theory. In a subsequent work (Montanari et al., 2019) derived heuristically a formula for the special case of random i.i.d. Gaussian feature matrices for the maximum margin classification, corresponding to the hinge loss function in our setting, with the difference, however, that the labels y^μ are generated from the $\mathbf{x}^\mu$ instead of the variable $\mathbf{c}^\mu$ as in our case.

Our main technical contribution is thus to provide a generic formula for the model described in Section 1.1 for any loss function and for fairly generic features F, including for instance deterministic ones.

The authors of (Goldt et al., 2019) analysed the learning dynamics of a neural network containing several hidden units using a one-pass stochastic gradient descent (SGD) for exactly the same model of data as here. In this online setting, the algorithm is never exposed to a sample twice, greatly simplifying the analysis as what has been learned at a given epoch can be considered independent of the randomness of a new sample. Another motivation of the present work is thus to study the sample complexity for this model (in our case only a bounded number of samples is available, and the one-pass SGD would be highly suboptimal).

An additional technical contribution of our work is to derive an extension of the equivalence between the considered data model and a model with Gaussian covariate, that has been observed and conjectured to hold rather generically in both (Goldt et al., 2019; Montanari et al., 2019). While we do not provide a rigorous proof for this equivalence, we show that it arises naturally using the replica method, giving further evidence for its validity.

Finally, the analysis of our formula for particular machine learning tasks of interest allows for an analytical investigation of a rich phenomenology that is also observed empirically in real-life scenarios. In particular

- The double descent behaviour, as termed in (Belkin et al., 2019) and exemplified in (Spigler et al., 2019), is exhibited for the non-regularized logistic regression loss. The peak of worst generalisation does not corresponds to $p=n$ as for the square loss (Mei & Montanari, 2019), but rather corresponds to the threshold of linear separability of the dataset. We also characterise the location of this threshold, generalising the results of (Candès & Sur, 2020) to our model.
- When using projections to approximate kernels, it has been observed that orthogonal features F perform better than random i.i.d. (Choromanski et al., 2017). We show that this behaviour arises from our analytical formula, illustrating the "unreasonable effectiveness of structured random orthogonal embeddings"(Choromanski et al., 2017).
- We compute the phase diagram for the generalisation error for the hidden manifold model and discuss the dependence on the various parameters, in particular the ratio between the ambient and latent dimensions.

2. Main analytical results

We now state our two main analytical results. The replica computation used here is in spirit similar to the one per-

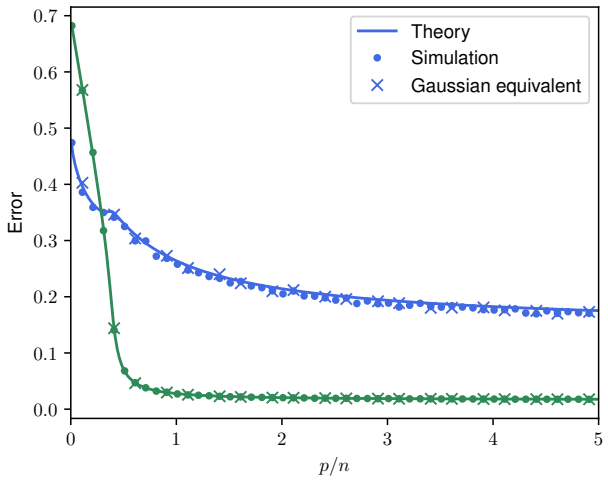

Figure 1. Comparison between theory (full line), and simulations with dimension $d = 200$ on the original model (dots), eq. (3), with $\sigma = \text{sign}$, and the Gaussian equivalent model (crosses), eq. (17), for logistic loss, regularisation $\lambda = 10^{-3}$, $n/d = 3$. Labels are generated as $y^\mu = \text{sign}\left(\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0\right)$ and $\hat{f} = \text{sign}$. Both the training loss (green) and generalisation error (blue) are depicted. The theory and the equivalence with the Gaussian model are observed to be very accurate even at dimensions as small as $d = 200$.

formed in a number of tasks for linear and generalised linear models (Gardner & Derrida, 1989; Seung et al., 1992; Kabashima et al., 2009; Krzakala et al., 2012), but requires a significant extension to account for the structure of the data. We refer the reader to the supplementary material Sec. 3 for the detailed and lengthy derivation of the final formula. The resulting expression is conjectured to be exact and, as we shall see, observed to be accurate even for relatively small dimensions in simulations. Additionally, these formulas reproduce the rigorous results of (Mei & Montanari, 2019), in the simplest particular case of a Gaussian projection matrix and ridge regression task. It remains a challenge to prove them rigorously in broader generality.

2.1. Generalisation error from replica method

Let F be a feature matrix satisfying the balance condition stated in eq. (7). Then, in the high-dimensional limit where $p, d, n \to \infty$ with $\alpha = n/p$, $\gamma = d/p$ fixed, the generalisation error, eq. (4), of the model introduced in Sec. (4) for σ such that its integral over a centered Gaussian variable is zero (so that $\kappa_0 = 0$ in eq. (17)) is given by the following easy-to-evaluate integral:

$$\lim_{n\to\infty} \epsilon_g = \mathbb{E}_{\lambda,\nu}\left[\left(f^0(\nu) - \hat{f}(\lambda)\right)^2\right], \tag{8}$$

where $f^0(.)$ is defined in (2), $\hat{f}(.)$ in (5) and (ν, λ) are jointly Gaussian random variables with zero mean and covariance matrix:

$$\Sigma = \begin{pmatrix} \rho & M^\star \\ M^\star & Q^\star \end{pmatrix} \in \mathbb{R}^2 \tag{9}$$

with $M^\star = \kappa_1 m_s^\star$, $Q^\star = \kappa_1^2 q_s^\star + \kappa_\star^2 q_w^\star$. The constants $\kappa_\star, \kappa_1$ depend on the nonlinearity σ via eq. (17), and $q_s^\star, q_w^\star, m_s^\star$, defined as:

$$\begin{aligned} \rho &= \frac{1}{d}||\boldsymbol{\theta}^0||^2 & q_s^\star &= \frac{1}{d}\mathbb{E}||\mathrm{F}\hat{\boldsymbol{w}}||^2 \\ q_w^\star &= \frac{1}{p}\mathbb{E}||\hat{\boldsymbol{w}}||^2 & m_s^\star &= \frac{1}{d}\mathbb{E}\left[(\mathrm{F}\hat{\boldsymbol{w}})\cdot\boldsymbol{\theta}^0\right] \end{aligned} \tag{10}$$

The values of these parameters correspond to the solution of the optimisation problem in eq. (6), and can be obtained as the fixed point solutions of the following set of self-consistent *saddle-point equations*:

$$\begin{cases} \hat{V}_s = \frac{\alpha\kappa_1^2}{\gamma V}\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\ \mathcal{Z}(y,\omega_0)\left(1 - \partial_\omega \eta(y,\omega_1)\right)\right], \\ \hat{q}_s = \frac{\alpha\kappa_1^2}{\gamma V^2}\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\ \mathcal{Z}(y,\omega_0)\left(\eta(y,\omega_1) - \omega_1\right)^2\right], \\ \hat{m}_s = \frac{\alpha\kappa_1}{\gamma V}\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\ \partial_\omega \mathcal{Z}(y,\omega_0)\left(\eta(y,\omega_1) - \omega_1\right)\right], \\ \\ \hat{V}_w = \frac{\alpha\kappa_\star^2}{V}\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\ \mathcal{Z}(y,\omega_0)\left(1 - \partial_\omega \eta(y,\omega_1)\right)\right], \\ \hat{q}_w = \frac{\alpha\kappa_\star^2}{V^2}\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\ \mathcal{Z}(y,\omega_0)\left(\eta(y,\omega_1) - \omega_1\right)^2\right], \end{cases}$$

$$\begin{cases} V_s = \frac{1}{\hat{V}_s}\left(1 - z\, g_\mu(-z)\right), \\ q_s = \frac{\hat{m}_s^2 + \hat{q}_s}{\hat{V}_s^2}\left[1 - 2z g_\mu(-z) + z^2 g'_\mu(-z)\right] \\ \quad\quad - \frac{\hat{q}_w}{(\lambda+\hat{V}_w)\hat{V}_s}\left[-z g_\mu(-z) + z^2 g'_\mu(-z)\right], \\ m_s = \frac{\hat{m}_s}{\hat{V}_s}\left(1 - z\, g_\mu(-z)\right), \\ \\ V_w = \frac{\gamma}{\lambda+\hat{V}_w}\left[\frac{1}{\gamma} - 1 + z g_\mu(-z)\right], \\ q_w = \gamma\frac{\hat{q}_w}{(\lambda+\hat{V}_w)^2}\left[\frac{1}{\gamma} - 1 + z^2 g'_\mu(-z)\right], \\ \quad\quad - \gamma\frac{\hat{m}_s^2 + \hat{q}_s}{(\lambda+\hat{V}_w)\hat{V}_s}\left[-z g_\mu(-z) + z^2 g'_\mu(-z)\right], \end{cases} \tag{11}$$

written in terms of the following auxiliary variables $\xi \sim \mathcal{N}(0,1)$, $z = \frac{\lambda+\hat{V}_w}{\hat{V}_s}$ and functions:

$$\begin{aligned} \eta(y,\omega) &= \underset{x\in\mathbb{R}}{\text{argmin}}\left[\frac{(x-\omega)^2}{2V} + \ell(y,x)\right], \\ \mathcal{Z}(y,\omega) &= \int \frac{\mathrm{d}x}{\sqrt{2\pi V^0}} e^{-\frac{1}{2V^0}(x-\omega)^2}\delta\left(y - f^0(x)\right) \end{aligned} \tag{12}$$

where $V = \kappa_1^2 V_s + \kappa_\star^2 V_w$, $V^0 = \rho - \frac{M^2}{Q}$, $Q = \kappa_1^2 q_s + \kappa_\star^2 q_w$, $M = \kappa_1 m_s$, $\omega_0 = \left(M/\sqrt{Q}\right)\xi$ and $\omega_1 = \sqrt{Q}\xi$. In the above, we assume that the matrix $\mathrm{F}\mathrm{F}^\top \in \mathbb{R}^{d\times d}$ associated to the feature map F has a well behaved spectral density, and denote g_μ its Stieltjes transform.

The training loss on the dataset $\mathcal{D} = \{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^n$ can also be obtained from the solution of the above equations as

$$\lim_{n\to\infty} \epsilon_t = \frac{\lambda}{2\alpha} q_w^\star + \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y, \omega_0^\star\right) \ell\left(y, \eta(y, \omega_1^\star)\right)\right] \quad (13)$$

where as before $\xi \sim \mathcal{N}(0,1)$, $y \sim \mathrm{Uni}(\mathbb{R})$ and $\mathcal{Z}, \eta$ are the same as in eq. (12), evaluated at the solution of the above saddle-point equations $\omega_0^\star = \left(M^\star/\sqrt{Q^\star}\right)\xi$, $\omega_1^\star = \sqrt{Q^\star}\xi$.

Sketch of derivation — We now sketch the derivation of the above result. A complete and detailed account can be found in Sec. 3 of the supplementary material. The derivation is based on the key observation that in the high-dimensional limit the asymptotic generalisation error only depends on the solution $\hat{\boldsymbol{w}} \in \mathbb{R}^p$ of eq. (5) through the scalar parameters $(q_s^\star, q_w^\star, m_s^\star)$ defined in eq. (10). The idea is therefore to rewrite the high-dimensional optimisation problem in terms of only these scalar parameters.

The first step is to note that the solution of eq. (6) can be written as the average of the following Gibbs measure

$$\pi_\beta(\boldsymbol{w}|\{\boldsymbol{x}^\mu, y^\mu\}) = \frac{1}{\mathcal{Z}_\beta} e^{-\beta\left[\sum_{\mu=1}^n \ell(y^\mu, \boldsymbol{x}^\mu\cdot\boldsymbol{w}) + \frac{\lambda}{2}||\boldsymbol{w}||_2^2\right]}, \quad (14)$$

in the limit $\beta \to \infty$. Of course, we have not gained much, since an exact calculation of π_β is intractable for large values of n, p and d. This is where the replica method comes in. It states that the distribution of the free energy density $f = -\log \mathcal{Z}_\beta$ (when seen as a random variable over different realisations of dataset $\mathcal{D}$) associated with the measure μ_β concentrates, in the high-dimensional limit, around a value f_β that depends only on the averaged *replicated partition function* $\mathcal{Z}_\beta^r$ obtained by taking $r > 0$ copies of $\mathcal{Z}_\beta$:

$$f_\beta = \lim_{r\to 0^+} \frac{\mathrm{d}}{\mathrm{d}r} \lim_{p\to\infty} \left[-\frac{1}{p}\left(\mathbb{E}_{\{\boldsymbol{x}^\mu, y^\mu\}} \mathcal{Z}_\beta^r\right)\right]. \quad (15)$$

Interestingly, $\mathbb{E}_{\{\boldsymbol{x}^\mu, y^\mu\}} \mathcal{Z}_\beta^r$ can be computed explicitly for $r \in \mathbb{N}$, and the limit $r \to 0^+$ is taken by analytically continuing to $r > 0$ (see Sec. 3 of the supplementary material). The upshot is that $\mathcal{Z}^r$ can be written as

$$\mathbb{E}_{\{\boldsymbol{x}^\mu, y^\mu\}} \mathcal{Z}_\beta^r \propto \int \mathrm{d}q_s \mathrm{d}q_w \mathrm{d}m_s \; e^{p\Phi_\beta^{(r)}(m_s, q_s, q_w)} \quad (16)$$

where Φ_β - known as the replica symmetric potential - is a concave function depending only on the following scalar parameters:

$$q_s = \frac{1}{d}||\mathrm{F}\boldsymbol{w}||^2, \quad q_w = \frac{1}{p}||\boldsymbol{w}||^2, \quad m_s = \frac{1}{d}(\mathrm{F}\boldsymbol{w})\cdot\boldsymbol{\theta}^0$$

for $\boldsymbol{w} \sim \pi_\beta$. In the limit of $p \to \infty$, this integral concentrates around the extremum of the potential $\Phi_\beta^{(0)}$ for any β. Since the optimisation problem in eq. (5) is convex, by construction as $\beta \to \infty$ the overlap parameters $(q_s^\star, q_w^\star, m_s^\star)$ satisfying this optimisation problem are precisely the ones of eq. (10) corresponding to the solution $\hat{\boldsymbol{w}} \in \mathbb{R}^p$ of eq. (5).

In summary, the replica method allows to circumvent the hard-to-solve high-dimensional optimisation problem eq. (5) by directly computing the generalisation error in eq. (4) in terms of a simpler scalar optimisation. Doing gradient descent in $\Phi_\beta^{(0)}$ and taking $\beta \to \infty$ lead to the saddle-point eqs. (11).

2.2. Replicated Gaussian Equivalence

The backbone of the replica derivation sketched above and detailed in Sec. 3 of the supplementary material is a central limit theorem type result coined as the *Gaussian equivalence theorem* (GET) from (Goldt et al., 2019) used in the context of the "replicated" Gibbs measure obtained by taking r copies of (14). In this approach, we need to assume that the "balance condition" (7) applies with probability one when the weights w are sampled from the replicated measure. We shall use this assumption in the following, checking its self-consistency via agreement with simulations.

It is interesting to observe that, when applying the GET in the context of our replica calculation, the resulting asymptotic generalisation error stated in Sec. 2.1 is equivalent to the asymptotic generalisation error of the following linear model:

$$\boldsymbol{x}^\mu = \kappa_0 \mathbf{1} + \kappa_1 \frac{1}{\sqrt{d}} \mathrm{F}^\top \boldsymbol{c}^\mu + \kappa_\star \, \boldsymbol{z}^\mu, \quad (17)$$

with $\kappa_0 = \mathbb{E}\left[\sigma(z)\right]$, $\kappa_1 \equiv \mathbb{E}\left[z\sigma(z)\right]$, $\kappa_\star^2 \equiv \mathbb{E}\left[\sigma(z)^2\right] - \kappa_0^2 - \kappa_1^2$, and $\boldsymbol{z}^\mu \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_p)$. We have for instance, $(\kappa_0, \kappa_1, \kappa_\star) \approx \left(0, \frac{2}{\sqrt{3\pi}}, 0.2003\right)$ for $\sigma = \mathrm{erf}$ and $(\kappa_0, \kappa_1, \kappa_\star) = \left(0, \sqrt{\frac{2}{\pi}}, \sqrt{1-\frac{2}{\pi}}\right)$ for $\sigma = \mathrm{sign}$, two cases explored in the next section. This equivalence constitutes a result with an interest in its own, with applicability beyond the scope of the generalised linear task eq. (6) studied here.

Equation (17) is precisely the mapping obtained by (Mei & Montanari, 2019), who proved its validity rigorously in the particular case of the square loss and Gaussian random matrix F using random matrix theory. The same equivalence arises in the analysis of kernel random matrices (Cheng & Singer, 2013; Pennington & Worah, 2017) and in the study of online learning (Goldt et al., 2019). The replica method thus suggests that the equivalence actually holds in a much larger class of learning problem, as conjectured as well in (Montanari et al., 2019), and numerically confirmed in all our numerical tests. It also potentially allows generalisation of the analysis in this paper for data coming from a learned generative adversarial network, along the lines of (Seddik et al., 2019; 2020).

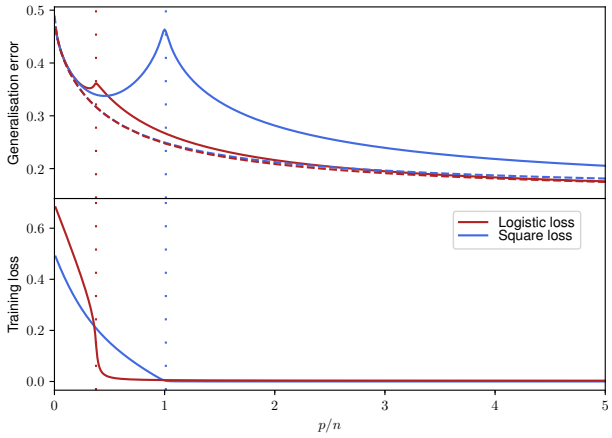

Figure 2. Upper panel: Generalisation error evaluated from eq. (8) plotted against the number of random Gaussian features per sample $p/n = 1/\alpha$ and fixed ratio between the number of samples and dimension $n/d = \alpha/\gamma = 3$ for logistic loss (red), square loss (blue). Labels are generated as $y^\mu = \text{sign}\left(\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0\right)$, data as $\boldsymbol{x}^\mu = \text{sign}\left(\text{F}^\top \boldsymbol{c}^\mu\right)$ and $\hat{f} = \text{sign}$ for two different values of regularisation λ, a small penalty $\lambda = 10^{-4}$ (full line) and a value of lambda optimised for every p/n (dashed line). Lower panel: The training loss corresponding to $\lambda = 10^{-4}$ is depicted.

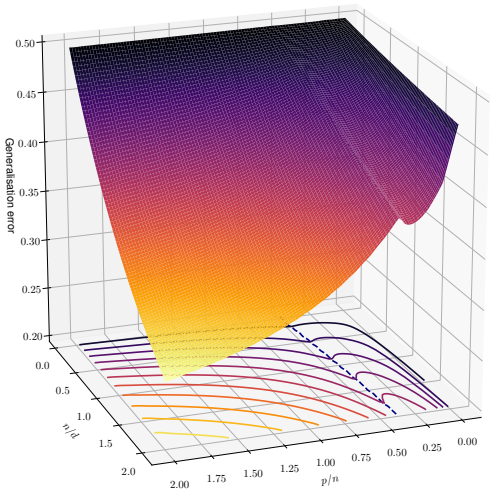

Figure 3. Generalisation error of the logistic loss at fixed very small regularisation $\lambda = 10^{-4}$, as a function of $n/d = \alpha/\gamma$ and $p/n = 1/\alpha$, for random Gaussian features. Labels are generated with $y^\mu = \text{sign}\left(\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0\right)$, the data $\boldsymbol{x}^\mu = \text{sign}\left(\text{F}^\top \boldsymbol{c}^\mu\right)$ and $\hat{f} = \text{sign}$. The interpolation peak happening where data become linearly separable is clearly visible here.

Fig. 1 illustrates the remarkable agreement between the result of the generalisation formula, eq. (8) and simulations both on the data eq. (3) with $\sigma(x) = \text{sign}(x)$ non-linearity, and on the Gaussian equivalent data eq. (17) where the non-linearity is replaced by rescaling by a constant plus noise. The agreement is flawless as implied by the theory in the high-dimensional limit, testifying that the used system size $d = 200$ is sufficiently large for the asymptotic theory to be relevant. We observed similar good agreement between the theory and simulation in all the cases we tested, in particular in all those presented in the following.

3. Applications of the generalisation formula

3.1. Double descent for classification with logistic loss

Among the surprising observations in modern machine learning is the fact that one can use learning methods that achieve zero training error, yet their generalisation error does not deteriorate as more and more parameters are added into the neural network. The study of such “interpolators” have attracted a growing attention over the last few years (Advani & Saxe, 2017; Spigler et al., 2019; Belkin et al., 2019; Neal et al., 2018; Hastie et al., 2019; Mei & Montanari, 2019; Geiger et al., 2019a; Nakkiran et al., 2019), as it violates basic intuition on the bias-variance trade-off (Geman et al., 1992). Indeed classical learning theory suggests that generalisation should first improve then worsen when increasing model complexity, following a U-shape curve. Many methods, including neural networks, instead follow a so-called "double descent curve" (Belkin et al., 2019) that displays two regimes: the "classical" U-curve found at low number of parameters is followed at high number of parameters by an interpolation regime where the generalisation error decreases monotonically. Consequently neural networks do not drastically overfit even when using much more parameters than data samples (Breiman, 1995), as actually observed already in the classical work (Geman et al., 1992). Between the two regimes, a "peak" occurs at the interpolation threshold (Opper & Kinzel, 1996; Engel & Van den Broeck, 2001; Advani & Saxe, 2017; Spigler et al., 2019). It should, however, be noted that existence of this "interpolation" peak is an independent phenomenon from the lack of overfitting in highly over-parametrized networks, and indeed in a number of the related works these two phenomena were observed separately (Opper & Kinzel, 1996; Engel & Van den Broeck, 2001; Advani & Saxe, 2017; Geman et al., 1992). Scaling properties of the peak and its relation to the jamming phenomena in physics are in particular studied in (Geiger et al., 2019a).

Among the simple models that allow to observe this behaviour, random projections —that are related to lazy training and kernel methods— are arguably the most natural one. The double descent has been analysed in detail in the present model in the specific case of a square loss on a regression task with random Gaussian features (Mei & Montanari, 2019). Our analysis allows to show the generality and the robustness of the phenomenon to other tasks,

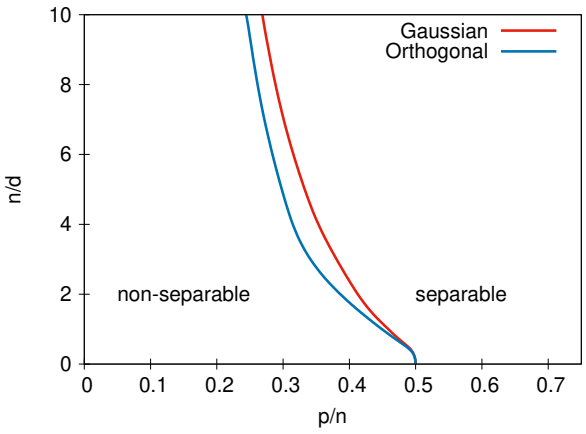

Figure 4. The position of the interpolation peak in logistic regression with $\lambda = 10^{-4}$, where data become linearly separable, as a function of the ratio between the number of samples n and the dimension d. Labels are generated with $y^\mu = \text{sign}\left(\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0\right)$, the data $\boldsymbol{x}^\mu = \text{sign}\left(\text{F}^\top \boldsymbol{c}^\mu\right)$ and $\hat{f} = \text{sign}$. The red line is with Gaussian random features, the blue line with orthogonal features. We see that for linear separability we need smaller number of projections p with orthogonal random features than with Gaussian.

matrices and losses. In Fig. 2 we compare the double descent as present in the square loss (blue line) with the one of logistic loss (red line) for random Gaussian features. We plot the value of the generalisation error at small values of the regularisation λ (full line), and for optimal value of λ (dashed line) for a fixed ratio between the number of samples and the dimension n/d as a function of the number of random features per sample p/n. We also plot the value of the training error (lower panel) for a small regularisation value, showing that the peaks indeed occur when the training loss goes to zero. For the square loss the peak appears at $1/\alpha = p/n = 1$ when the system of n linear equations with p parameters becomes solvable. For the logistic loss the peak instead appears at a value $1/\alpha^*$ where the data $\mathcal{D}$ become linearly separable and hence the logistic loss can be optimised down to zero. These values $1/\alpha^*$ depends on the value n/d, and this dependence is plotted in Fig. 4. For very large dimension d, i.e. $n/d \to 0$ the data matrix X is close to iid random matrix and hence the $\alpha^*(n/d = 0) = 2$ as famously derived in classical work by Cover (Cover, 1965). For $n/d > 0$ the α^* is growing ($1/\alpha^*$ decreasing) as correlations make data easier to linearly separate, similarly as in (Candès & Sur, 2020).

Fig. 2 also shows that better error can be achieved with the logistic loss compared to the square loss, both for small and optimal regularisations, except in a small region around the logistic interpolation peak. In the Kernel limit, i.e. $p/n \to \infty$, the generalization error at optimal regularisation saturates at $\epsilon_g(p/n \to \infty) \simeq 0.17$ for square loss and at $\epsilon_g(p/n \to \infty) \simeq 0.16$ for logistic loss. Fig. 3 then depicts a 3D plot of the generalisation error also illustrating the position of the interpolation peak.

3.2. Random features: Gaussian versus orthogonal

Kernel methods are a very popular class of machine learning techniques, achieving state-of-the-art performance on a variety of tasks with theoretical guarantees (Schölkopf et al., 2002; Rudi et al., 2017; Caponnetto & De Vito, 2007). In the context of neural network, they are the subject of a renewal of interest in the context of the Neural Tangent Kernel (Jacot et al., 2018). Applying kernel methods to large-scale "big data" problems, however, poses many computational challenges, and this has motivated a variety of contributions to develop them at scale, see, e.g., (Rudi et al., 2017; Zhang et al., 2015; Saade et al., 2016; Ohana et al., 2019). Random features (Rahimi & Recht, 2008) are among the most popular techniques to do so.

Here, we want to compare the performance of random projection with respect to structured ones, and in particular orthogonal random projections (Choromanski et al., 2017) or deterministic matrices such as real Fourier (DCT) and Hadamard matrices used in fast projection methods (Le et al., 2013; Andoni et al., 2015; Bojarski et al., 2016). Up to normalisation, these matrices have the same spectral density. Since the asymptotic generalisation error only depends on the spectrum of $\text{FF}^\top$, all these matrices share the same theoretical prediction when properly normalised, see Fig. 5. In our computation, left- and right-orthogonal invariance is parametrised by letting $\text{F} = \text{U}^\top \text{DV}$ for $\text{U} \in \mathbb{R}^{d\times d}$, $\text{V} \in \mathbb{R}^{p\times p}$ two orthogonal matrices drawn from the Haar measure, and $\text{D} \in \mathbb{R}^{d\times p}$ a diagonal matrix of rank $\min(d, p)$. In order to compare the results with the Gaussian case, we fix the diagonal entries $d_k = \max(\sqrt{\gamma}, 1)$ of D such that an arbitrary projected vector has the same norm, on average, to the Gaussian case.

Fig. 5 shows that random orthogonal embeddings always outperform Gaussian random projections, in line with empirical observations, and that they allow to reach the kernel limit with fewer number of projections. Their behaviour is, however, qualitatively similar to the one of random i.i.d. projections. We also show in Fig. 4 that orthogonal projections allow to separate the data more easily than the Gaussian ones, as the phase transition curve delimiting the linear separability of the logistic loss get shifted to the left.

3.3. The hidden manifold model phase diagram

In this subsection we consider the hidden manifold model where p-dimensional x data lie on a d-dimensional manifold, we have mainly in mind $d < p$. The labels y are generated using the coordinates on the manifold, eq. (2).

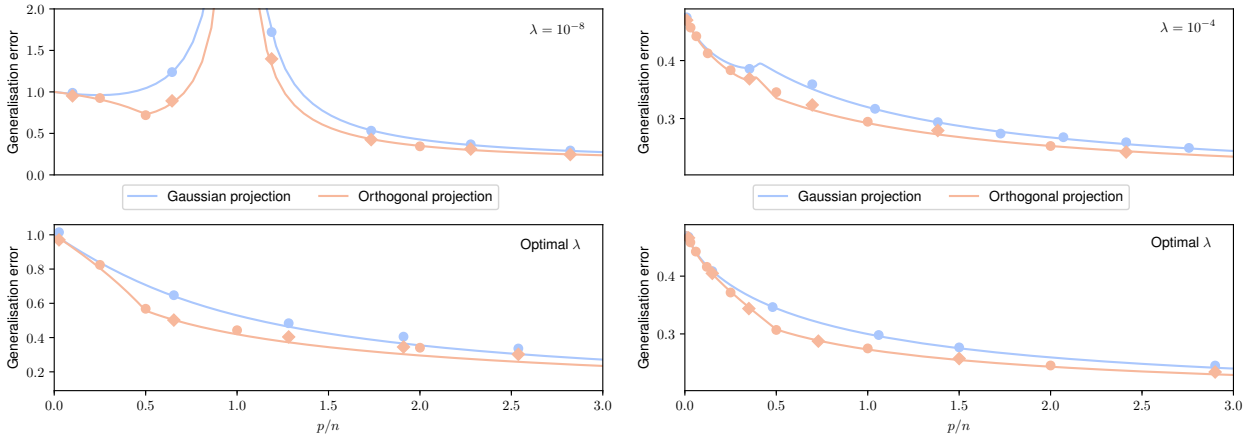

Figure 5. Generalisation error against the number of features per sample p/n, for a regression problem (left) and a classification one (right). **Left (ridge regression):** We used $n/d = 2$ and generated labels as $y^\mu = \boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0$, data as $\boldsymbol{x}^\mu = \text{sign}\left(\text{F}^\top \boldsymbol{c}^\mu\right)$ and $\hat{f}(x) = x$. The two curves correspond to ridge regression with Gaussian (blue) versus orthogonal (red) projection matrix F for both $\lambda = 10^{-8}$ (top) and optimal regularisation λ (bottom). **Right (logistic classification)**: We used $n/d = 2$ and generated labels as $y^\mu = \text{sign}\left(\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0\right)$, data as $\boldsymbol{x}^\mu = \text{sign}\left(\text{F}^\top \boldsymbol{c}^\mu\right)$ and $\hat{f} = \text{sign}$. The two curves correspond to a logistic classification with again Gaussian (blue) versus orthogonal (red) projection matrix F for both $\lambda = 10^{-4}$ and optimal regularisation λ. In all cases, full lines is the theoretical prediction, and points correspond to gradient-descent simulations with $d = 256$. For the simulations of orthogonally invariant matrices, we results for Hadamard matrices (dots) and DCT Fourier matrices (diamonds).

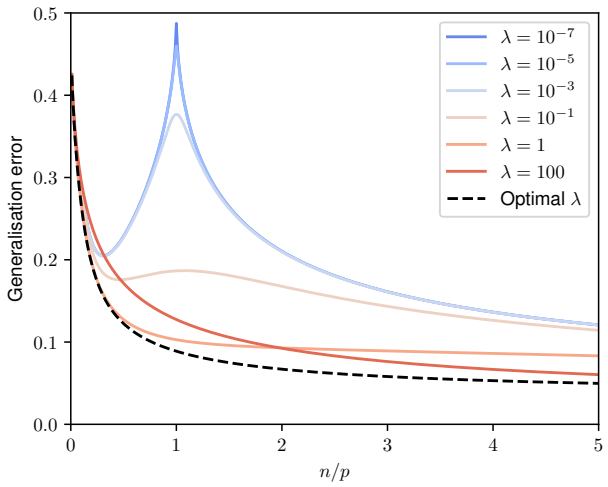

Figure 6. Generalisation error against the number of samples per dimension, $\alpha = n/p$, and fixed ratio between the latent and data dimension, $d/p = 0.1$, for a classification task with square loss on labels generated as $y^\mu = \text{sign}\left(\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0\right)$, data $\boldsymbol{x}^\mu = \text{erf}\left(\text{F}^\top \boldsymbol{c}^\mu\right)$ and $\hat{f} = \text{sign}$, for different values of the regularisation λ (full lines), including the optimal regularisation value (dashed).

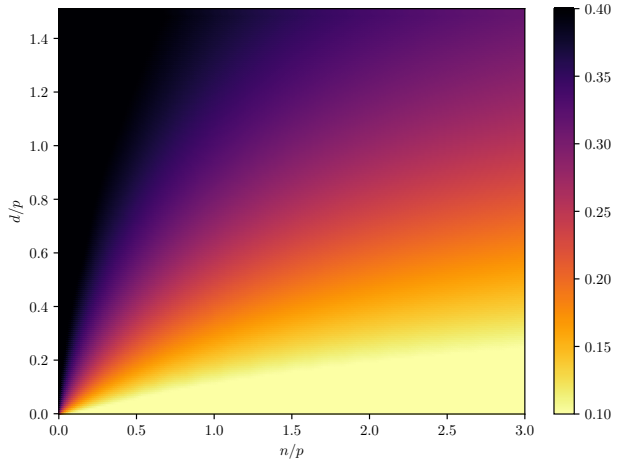

Figure 7. Heat-map of the generalisation errors as a function of the number of samples per data dimension n/p against the ratio of the latent and data dimension d/p, for a classification task with square loss on labels $y^\mu = \text{sign}\left(\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0\right)$ and data $\boldsymbol{x}^\mu = \text{erf}\left(\text{F}^\top \boldsymbol{c}^\mu\right)$ for the optimal values of the regularisation λ.

In Fig. 6 we plot the generalisation error of classification with the square loss for various values of the regularisation λ. We fix the ratio between the dimension of the sub-manifold and the dimensionality of the input data to $d/p = 0.1$ and plot the learning curve, i.e. the error as a function of the number of samples per dimension. Depending on the value of the regularisation, we observe that the interpolation peak, which is at $\alpha = 1$ at very small regularisation (here the over-parametrised regime is on the left hand side), decreases for larger regularisation λ. A similar behaviour has been observed for other models in the past, see e.g. (Opper & Kinzel, 1996). Finally Fig. 6 depicts the error for optimised regularisation parameter in the black dashed line. For large number of samples we observe the generalisa-

tion error at optimal regularisation to saturate in this case at $\epsilon_g(\alpha \to \infty) \to 0.0325$. A challenge for future work is to see whether better performance can be achieved on this model by including hidden variables into the neural network.

Fig. 7 then shows the generalisation error for the optimised regularisation λ with square loss as a function of the ratio between the latent and the data dimensions d/p. In the limit $d/p \gg 1$ the data matrix becomes close to a random iid matrix and the labels are effectively random, thus only bad generalisation can be reached. Interestingly, as d/p decreases to small values even the simple classification with regularised square loss is able to "disentangle" the hidden manifold structure in the data and to reach a rather low generalisation error. The figure quantifies how the error deteriorates when the ratio between the two dimensions d/p increases. Rather remarkably, for a low d/p a good generalisation error is achieved even in the over-parametrised regime, where the dimension is larger than the number of samples, $p > n$. In a sense, the square loss linear classification is able to locate the low-dimensional subspace and find good generalisation even in the over-parametrised regime as long as roughly $d \lesssim n$. The observed results are in qualitative agreement with the results of learning with stochastic gradient descent in (Goldt et al., 2019) where for very low d/p good generalisation error was observed in the hidden manifold model, but a rather bad one for $d/p = 0.5$.

Acknowledgements

This work is supported by the ERC under the European Union's Horizon 2020 Research and Innovation Program 714608-SMiLe, as well as by the French Agence Nationale de la Recherche under grant ANR-17-CE23-0023-01 PAIL and ANR-19-P3IA-0001 PRAIRIE. We also acknowledge support from the chaire CFM-ENS "Science des données". We thank Google Cloud for providing us access to their platform through the Research Credits Application program. BL was partially financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

Advani, M., Lahiri, S., and Ganguli, S. Statistical mechanics of complex neural systems and high dimensional data. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(03):P03014, 2013.

Advani, M. S. and Saxe, A. M. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*, 2017.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252, 2019.

Andoni, A., Indyk, P., Laarhoven, T., Razenshteyn, I., and Schmidt, L. Practical and optimal lsh for angular distance. In *Advances in neural information processing systems*, pp. 1225–1233, 2015.

Aubin, B., Maillard, A., Krzakala, F., Macris, N., Zdeborová, L., et al. The committee machine: computational to statistical gaps in learning a two-layers neural network. In *Advances in Neural Information Processing Systems*, pp. 3223–3234, 2018.

Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborová, L. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

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

Bojarski, M., Choromanska, A., Choromanski, K., Fagan, F., Gouy-Pailler, C., Morvan, A., Sakr, N., Sarlos, T., and Atif, J. Structured adaptive and random spinners for fast machine learning computations. *arXiv preprint arXiv:1610.06209*, 2016.

Breiman, L. Reflections after refereeing papers for nips. *The Mathematics of Generalization*, pp. 11–15, 1995.

Candès, E. J. and Sur, P. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *Ann. Statist.*, 48(1): 27–42, 02 2020. doi: 10.1214/18-AOS1789. URL https://doi.org/10.1214/18-AOS1789.

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Cheng, X. and Singer, A. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 02(04):1350010, 2013.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems 32*, pp. 2933–2943. 2019.

Choromanski, K. M., Rowland, M., and Weller, A. The unreasonable effectiveness of structured random orthogonal embeddings. In *Advances in Neural Information Processing Systems*, pp. 219–228, 2017.

Cover, T. M. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, EC-14(3):326–334, 1965.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. *ICLR 2019, arXiv preprint arXiv:1810.02054*, 2018.

Engel, A. and Van den Broeck, C. *Statistical mechanics of learning*. Cambridge University Press, 2001.

Gardner, E. and Derrida, B. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983, 1989.

Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., d'Ascoli, S., Biroli, G., Hongler, C., and Wyart, M. Scaling description of generalization with number of parameters in deep learning. *arXiv preprint arXiv:1901.01608*, 2019a.

Geiger, M., Spigler, S., Jacot, A., and Wyart, M. Disentangling feature and lazy learning in deep neural networks: an empirical study. *arXiv preprint arXiv:1906.08034*, 2019b.

Geman, S., Bienenstock, E., and Doursat, R. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Goldt, S., Mézard, M., Krzakala, F., and Zdeborová, L. Modelling the influence of data structure on learning in neural networks. *arXiv preprint arXiv:1909.11500*, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

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

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Kabashima, Y., Wadayama, T., and Tanaka, T. A typical reconstruction limit for compressed sensing based on lp-norm minimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(09):L09003, 2009.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Krzakala, F., Mézard, M., Sausset, F., Sun, Y., and Zdeborová, L. Statistical-physics-based reconstruction in compressed sensing. *Physical Review X*, 2(2):021005, 2012.

Le, Q., Sarlós, T., and Smola, A. Fastfood-approximating kernel expansions in loglinear time. In *Proceedings of the international conference on machine learning*, volume 85, 2013.

Louart, C., Liao, Z., Couillet, R., et al. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

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

Mézard, M. Mean-field message-passing equations in the hopfield model and its generalizations. *Physical Review E*, 95(2):022117, 2017.

Mézard, M., Parisi, G., and Virasoro, M. *Spin glass theory and beyond: an introduction to the Replica Method and its applications*, volume 9. World Scientific Publishing Company, 1987.

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. The generalization error of max-margin linear classifiers: high-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: where bigger models and more data hurt. *ICLR 2020, arXiv preprint arXiv:1912.02292*, 2019.

Neal, B., Mittal, S., Baratin, A., Tantia, V., Scicluna, M., Lacoste-Julien, S., and Mitliagkas, I. A modern take on the bias-variance tradeoff in neural networks. *arXiv preprint arXiv:1810.08591*, 2018.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 5947–5956, 2017.

Ohana, R., Wacker, J., Dong, J., Marmin, S., Krzakala, F., Filippone, M., and Daudet, L. Kernel computations from large-scale random features obtained by optical processing units. *arXiv preprint arXiv:1910.09880*, 2019.

Opper, M. and Kinzel, W. Statistical mechanics of generalization. In *Models of neural networks III*, pp. 151–209. Springer, 1996.

Pennington, J. and Worah, P. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems 30*, pp. 2637–2646. 2017.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pp. 1177–1184. 2008.

Rudi, A., Carratino, L., and Rosasco, L. Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems*, pp. 3888–3898, 2017.

Saade, A., Caltagirone, F., Carron, I., Daudet, L., Drémeau, A., Gigan, S., and Krzakala, F. Random projections through multiple optical scattering: approximating kernels at the speed of light. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6215–6219. IEEE, 2016.

Schölkopf, B., Smola, A. J., Bach, F., et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Seddik, M. E. A., Tamaazousti, M., and Couillet, R. Kernel random matrices of large concentrated data: the example of gan-generated images. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7480–7484. IEEE, 2019.

Seddik, M. E. A., Louart, C., Tamaazousti, M., and Couillet, R. Random matrix theory proves that deep learning representations of gan-data behave as gaussian mixtures. *arXiv preprint arXiv:2001.08370*, 2020.

Seung, H. S., Sompolinsky, H., and Tishby, N. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.

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

Spigler, S., Geiger, M., d'Ascoli, S., Sagun, L., Biroli, G., and Wyart, M. A jamming transition from under-to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52 (47):474001, 2019.

Talagrand, M. The Parisi formula. *Annals of mathematics*, 163:221–263, 2006.

Vapnik, V. N. *The nature of statistical learning theory*. Wiley, New York, 1st edition, September 1998. ISBN 978-0-471-03003-4.

Watkin, T. L., Rau, A., and Biehl, M. The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65(2):499, 1993.

Woodworth, B., Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Kernel and deep regimes in overparametrized models. *arXiv preprint arXiv:1906.05827*, 2019.

Zdeborová, L. and Krzakala, F. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65 (5):453–552, 2016.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *ICLR 2017, arXiv preprint arXiv:1611.03530*, 2016.

Zhang, Y., Duchi, J., and Wainwright, M. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research*, 16(1):3299–3340, 2015.