PAPER

An analytic theory of shallow networks dynamics for hinge loss classification*

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: ML 2021

An analytic theory of shallow networks dynamics for hinge loss classification*

Franco Pellegrini** and **Giulio Biroli**

Laboratoire de Physique de l'École Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, F-75005 Paris, France
E-mail: franco.pellegrini@phys.ens.fr

Received 9 November 2021
Accepted for publication 9 November 2021
Published 29 December 2021

Online at stacks.iop.org/JSTAT/2021/124005
https://doi.org/10.1088/1742-5468/ac3a76

Abstract. Neural networks have been shown to perform incredibly well in classification tasks over structured high-dimensional datasets. However, the learning dynamics of such networks is still poorly understood. In this paper we study in detail the training dynamics of a simple type of neural network: a single hidden layer trained to perform a classification task. We show that in a suitable mean-field limit this case maps to a single-node learning problem with a time-dependent dataset determined self-consistently from the average nodes population. We specialize our theory to the prototypical case of a linearly separable data and a linear hinge loss, for which the dynamics can be explicitly solved in the infinite dataset limit. This allows us to address in a simple setting several phenomena appearing in modern networks such as slowing down of training dynamics, crossover between rich and lazy learning, and overfitting. Finally, we assess the limitations of mean-field theory by studying the case of large but finite number of nodes and of training samples.

Keywords: machine learning

Supplementary material for this article is available online

*This article is an updated version of: Pellegrini F and Biroli G 2020 An analytic theory of shallow networks dynamics for hinge loss classification *Advances in Neural Information Processing Systems* vol 33 ed H Larochelle, M Ranzato, R Hadsell, M F Balcan and H Lin (New York: Curran Associates) pp 5356–67.
**Author to whom any correspondence should be addressed.

1742-5468/21/124005+14$33.00

Contents

1. Introduction

Despite their proven ability to tackle a large class of complex problems [1], neural networks are still poorly understood from a theoretical point of view. While general theorems prove them to be universal approximators [2], their ability to obtain generalizing solutions given a finite set of examples remains largely unexplained. This behavior has been observed in multiple settings. The huge number of parameters and the optimization algorithms employed to optimize them (gradient descent and its variations) are thought to play key roles in it [3–5].

In consequence, a large research effort has been devoted in recent years to understanding the training dynamics of neural networks with a very large number of nodes [6–8]. Much theoretical insight has been gained in the training dynamics of linear [9, 10] and nonlinear networks for regression problems, often with quadratic loss and in a teacher-student setting [11–14], highlighting the evolution of correlations between data and network outputs. More generally, the input–output correlation and its effect on the landscape has been used to show the effectiveness of gradient descent [15, 16]. Other approaches have focused on infinitely wide networks to perform a mean-field analysis of the weights dynamics [17–22], or study its neural tangent kernel (NTK, or 'lazy') limit [23–26].

In this work, we investigate the learning dynamics for binary classification problems, by considering one of the most common cost functions employed in this setting: the linear hinge loss. The idea behind the hinge loss is that examples should contribute to the cost function if misclassified, but also if classified with a certainty lower than a given threshold. In our case this cost is linear in the distance from the threshold, and zero for examples classified above threshold, that we shall call *satisfied* henceforth. This specific

J. Stat. Mech. (2021) 124005

choice leads to an interesting consequence: the instantaneous gradient for each node due to *unsatisfied* examples depends on the activation of the other nodes only through their population, while that due to *satisfied* examples is just zero. Describing the learning dynamics in the mean-field limit amounts to computing the effective example distribution for a given distribution of parameters: each node then evolves 'independently' with a time-dependent dataset determined self-consistently from the average nodes population.

Contribution. We provide an analytical theory for the dynamics of a single hidden layer neural network trained for binary classification with linear hinge loss. In section 2, we obtain the mean-field theory equations for the training dynamics. Those equations are a generalizations of the ones obtained for mean-square loss in [17–22]. In section 3, we focus on linearly separable data with spherical symmetry and present an explicit analytical solution of the dynamics of the nodes parameters. In this setting we provide a detailed study of the cross-over between the lazy [23] and rich [27] learning regimes (section 3.2). Finally, we assess the limitations of mean-field theory by studying the case of large but finite number of nodes and finite number of training samples (section 3.3). The most important new effect is overfitting, which we are able to describe by analyzing corrections to mean-field theory. In section 3.4, we show that introducing a small fraction of mislabeled examples induces a slowing down of the dynamics and hastens the onset of the overfitting phase. Finally in section 4, we present numerical experiments on a realistic case, and show that the associated nodes dynamics in the first stage of training is in good agreement with our results.

The merit of the model we focused on is that, thanks to its simplicity, several effects happening in real networks can be studied analytically. Our analytical theory is derived using reasoning common in theoretical physics, which we expect can be made rigorous following the lines of [17–22]. All our results are tested throughout the paper by numerical simulations which confirm their validity.

Related works. The study of neural network dynamics with one (or few) nodes started in statistical physics [11], but was mainly focused on the online setting. More recent works on separable data [28, 29] observed the main trend of logarithmic alignment with the max margin vector under rather general settings. Mean-field analysis of the training dynamics of very wide neural networks have mainly focused on regression problems with mean-square losses [17–23], whereas fewer works [30, 31] have tackled the dynamics for classification tasks[1]. The task and architecture we focus on bears strong similarities to the one proposed in des Combes *et al* [30], but with fewer assumptions on the dataset and initialization. With respect to [30], we show the relation with mean-field treatments [17–22] and provide a full analysis of the dynamics, in particular the cross-over between rich and lazy learning. Moreover, we discuss the limitations of mean-field theory, the source of overfitting and the change in the dynamics due to mislabeling.

[1] In the NTK (or 'lazy') limit [23–25] general losses have been considered.

J. Stat. Mech. (2021) 124005

2. Mean-field equation for the density of parameters

We consider a binary classification task for N points in d dimensions $\{\mathbf{x}_n\} \subset \mathbb{R}^d$ with corresponding labels $y_n = \pm 1$. We focus on a hidden layer neural network consisting of M nodes with activation σ. The output of the network is therefore

$$f(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{M}\sum_{i=1}^{M} a_i \sigma\left(\frac{\mathbf{w}_i \cdot \mathbf{x}}{\sqrt{d}}\right), \tag{1}$$

where $\boldsymbol{\theta}_i = \{a_i, \mathbf{w}_i\}$ represents all the trainable parameters of the model: $\{\mathbf{w}_i\}$, the d-dimensional weight vectors between input and each hidden node, and $\{a_i\}$, the contributions of each node to the output. All components are initialized before training from a Gaussian distribution with zero mean and unit standard deviation. The $1/M$ in front of the sum leads to the so-called mean-field normalization [17]. In the large-M limit, this allows to do what is called a hydrodynamic treatment in physics, a procedure that have been put on a rigorous basis in this context in [17–23] (here the $\boldsymbol{\theta}_i$s play the role of particle positions). One of the main assumptions of this procedure is that in the large M-limit one can rewrite the output function in terms of the averaged nodes population (or density) $\rho(\boldsymbol{\theta})$:

$$f(\mathbf{x};\boldsymbol{\theta}) = \int \mathrm{d}\boldsymbol{\theta}\, \rho(\boldsymbol{\theta}) a \sigma\left(\frac{\mathbf{w}\cdot\mathbf{x}}{\sqrt{d}}\right). \tag{2}$$

To optimize the parameters we minimize the loss function

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N} \ell(y_n, f(\mathbf{x}_n;\boldsymbol{\theta})) \tag{3}$$

by gradient flow $\dot{\boldsymbol{\theta}} = -\beta^* \partial\mathcal{L}/\partial\boldsymbol{\theta}$ ($\ell(x,y)$ will be specified later). The dynamical equations for the parameters $\{a_i, \mathbf{w}_i\}$ read:

$$\begin{cases} \dot{a}_i &= -\dfrac{\beta}{N}\displaystyle\sum_{n=1}^{N} \frac{\partial \ell(y_n, f(\mathbf{x};\boldsymbol{\theta}))}{\partial f} \sigma\left(\frac{\mathbf{w}_i\cdot\mathbf{x}}{\sqrt{d}}\right) \\ \dot{\mathbf{w}}_i &= -\dfrac{\beta}{N}\displaystyle\sum_{n=1}^{N} \frac{\partial \ell(y_n, f(\mathbf{x};\boldsymbol{\theta}))}{\partial f} a_i \sigma'\left(\frac{\mathbf{w}_i\cdot\mathbf{x}}{\sqrt{d}}\right)\frac{\mathbf{x}}{\sqrt{d}}, \end{cases} \tag{4}$$

where we have defined the effective learning rate $\beta = \beta^*/M$. These equations show that the coupling between the different nodes has a mean-field form: it is through the function f, i.e. only through the density $\rho(\boldsymbol{\theta}, t)$. Following standard techniques one can obtain a closed hydrodynamic-like equation on $\rho(\boldsymbol{\theta}, t)$ in the large M limit:

$$\partial_t \rho(\boldsymbol{\theta}, t) = \beta \nabla_{\boldsymbol{\theta}} \left(\rho(\boldsymbol{\theta}, t) \nabla_{\boldsymbol{\theta}} \frac{\delta\mathcal{L}[\rho(\boldsymbol{\theta}, t)]}{\delta\rho(\boldsymbol{\theta}, t)}\right) \quad \rho(\boldsymbol{\theta}, 0) = \mathcal{N}(0, \mathbb{I}) \tag{5}$$

where we have made explicit that the $\mathcal{L}$ is a functional of the density ρ since it depends on $f(\mathbf{x};\boldsymbol{\theta})$, see equations (2) and (3). The convergence of the dynamical process to the hydrodynamic limit is usually assumed in the physics literature, proofs

J. Stat. Mech. (2021) 124005

(that we expect can be generalized to our case) have been worked out in [32, 33]. (See online supplementary material (https://stacks.iop.org/JSTAT/2021/124005/mmedia) for details.)

To be more concrete, in the following we consider the case of linear hinge loss, $\ell(y, f) = \mathcal{R}(h - yf)$ (h being the size of the hinge, often taken as 1), and rectified linear unit activation function: $\sigma(x) = \mathcal{R}(x) = \max(0, x)$. With this choice

$$\frac{\delta \mathcal{L}[\rho(\boldsymbol{\theta}, t)]}{\delta \rho(\boldsymbol{\theta}, t)} = -a \left\langle u(\mathbf{x}, y\,; t)\theta\left(\mathbf{w}\cdot\mathbf{x}\right) y \frac{\mathbf{w}\cdot\mathbf{x}}{\sqrt{d}} \right\rangle_{\mathbf{x},y}, \tag{6}$$

with θ being the Heaviside step function. The notation $u(\mathbf{x}, y\,; t) \equiv \mathbb{I}_{h - yf(\mathbf{x}\,;\boldsymbol{\theta}(t))>0}$ denotes the indicator function of the *unsatisfied* examples, i.e. those $(\mathbf{x}, y)$ for which the loss is positive, and $\langle\cdot\rangle_{\mathbf{x},y}$ denotes the average over examples and classes ($y = \pm 1$ for binary classification). The dynamical equations on the node parameters simplify too:

$$\begin{cases} \dot{a}_i(t) &= \dfrac{\beta}{\sqrt{d}} \mathbf{w}_i \cdot \langle u(\mathbf{x}, y\,; t)\theta\left(\mathbf{w}_i\cdot\mathbf{x}\right) y\mathbf{x}\rangle_{\mathbf{x},y} \\ \dot{\mathbf{w}}_i(t) &= \dfrac{\beta}{\sqrt{d}} a_i \langle u(\mathbf{x}, y\,; t)\theta\left(\mathbf{w}_i\cdot\mathbf{x}\right) y\mathbf{x}\rangle_{\mathbf{x},y}. \end{cases} \tag{7}$$

Remarkably, the equation on the $\mathbf{w}_i$ is very similar to the one induced by the Hebb rule in biological neural networks.

3. Analysis of a linearly separable case

We now focus on a linearly separable model, where the dynamics can be solved explicitly. We consider a reference unit vector $\hat{\mathbf{w}}^*$ in input space and examples distributed according to a spherical probability distribution $P(\mathbf{x})$. We label each example based on the sign of its scalar product with $\hat{\mathbf{w}}^*$, leading to a distribution for $y = \pm 1$: $P(\mathbf{x}, y) = P(\mathbf{x})\,\theta(y(\hat{\mathbf{w}}^*\cdot\mathbf{x}))$.

In order to be able to explore different training regimes, we adopt a rescaled loss function, similar to the one proposed in Chizat *et al* [23]:

$$\mathcal{L}^{\alpha}(\boldsymbol{\theta}) = \frac{1}{\alpha^2 N} \sum_{n=1}^{N} \mathcal{R}\left[h - \alpha y_n \left(f(\mathbf{x}_n\,;\boldsymbol{\theta}) - f(\mathbf{x}_n\,;\boldsymbol{\theta}_0)\right)\right], \tag{8}$$

where α is the rescaling parameter and $\boldsymbol{\theta}_0$ are the parameters at the beginning of training. Subtracting the initial output of the network ensures that no bias is introduced by the specific finite choice of parameters at initialization, while having no influence in the hydrodynamic limit since the output is 0 by construction.

3.1. Explicit solution for an infinite training set

We first consider the limit of infinite number of examples, and later discuss the effects induced by a finite training set.

J. Stat. Mech. (2021) 124005

The explicit solution of the training dynamics is obtained making use of the cylindrical symmetry around $\hat{\mathbf{w}}^*$, which implies that the average in the equations of motion (7) does not depend on $\mathbf{w}$, i.e.

$$\langle u(\mathbf{x}, y; t)\theta\left(\mathbf{w}\cdot\mathbf{x}\right) y\mathbf{x}\rangle_{\mathbf{x},y} = I(t)\hat{\mathbf{w}}^*, \tag{9}$$

where $I(t) \equiv \langle u(\mathbf{x}, y\,; t)\theta\left(\mathbf{w}\cdot\mathbf{x}\right) y\mathbf{x}\cdot\hat{\mathbf{w}}^*\rangle_{\mathbf{x},y}$. By plugging the identity (9) into equations (6) and (7) one finds that the hydrodynamic equation (5) can be solved by the method of the characteristic, where $\rho(\boldsymbol{\theta}, t)$ is obtained by transporting the initial condition through the equation (7). By decomposing the vector $\mathbf{w}$ in its parallel and perpendicular components with respect to $\hat{\mathbf{w}}^*$, i.e. $\mathbf{w} = w^{\parallel}\hat{\mathbf{w}}^* + \mathbf{w}_{\perp}$, and using the solution $\rho(\boldsymbol{\theta}, t)$, one finds that the parameters $\boldsymbol{\theta}$ at time t are distributed in law as:

$$\begin{cases} a(t) & \overset{d}{\sim} & a(0)\cosh(\gamma(t)) + w^{\parallel}(0)\sinh(\gamma(t)) \\ w^{\parallel}(t) & \overset{d}{\sim} & w^{\parallel}(0)\cosh(\gamma(t)) + a(0)\sinh(\gamma(t)); \\ \mathbf{w}_{\perp}(t) & \overset{d}{\sim} & \mathbf{w}_{\perp}(0) \end{cases} \qquad \gamma(t) = \frac{\beta}{\alpha\sqrt{d}}\int_0^t I(t)\mathrm{d}t, \tag{10}$$

where $a(0), w^{\parallel}(0), \mathbf{w}_{\perp}(0)$ are given by the initial condition distributions: since all initial components of $\mathbf{w}$ were taken as i.i.d. Gaussian, so is $w^{\parallel}(0)$ and every component of $\mathbf{w}_{\perp}(0)$ for any choice of basis. Using the distribution of $\boldsymbol{\theta}$ at time t, one can then compute $\langle u(\mathbf{x}, y\,; t)\theta\left(\mathbf{w}\cdot\mathbf{x}\right) y\mathbf{x}\cdot\hat{\mathbf{w}}^*\rangle_{\mathbf{x},y}$ and hence obtain a self-consistent equation on $I(t)$, which completes the mean-field solution. Similarly, one can obtain explicitly the output function and the indicator function which acquire a simple form:

$$f(\mathbf{x}\,;\boldsymbol{\theta}) = \frac{\sinh(2\gamma(t))}{2\sqrt{d}}\hat{\mathbf{w}}^*\cdot\mathbf{x}, \tag{11}$$

$$u(\mathbf{x}, y\,; t) = \theta\left(\frac{2h\sqrt{d}}{\alpha\,\sinh(2\gamma(t))} - y\hat{\mathbf{w}}^*\cdot\mathbf{x}\right) \tag{12}$$

where we have used that $f(\mathbf{x}\,;\boldsymbol{\theta}) = 0$ at $t = 0$. As expected, both functions have cylindrical symmetry around $\hat{\mathbf{w}}^*$. The analytical derivation of these results and the following ones is presented in the SM. Since by definition $I(t) \geqslant 0$ the function $\gamma(t)$ is monotonously increasing and starts from zero at $t = 0$. To be more specific, we consider two cases: normally distributed data with unit variance in each dimension, and uniform data on the d-dimensional unit sphere. The corresponding self-consistent equations on $\gamma(t)$ read respectively:

$$\dot{\gamma}(t) = \frac{\beta I^N(0)}{\alpha\sqrt{d}}\left(1 - \exp\left[-\frac{2h^2 d}{\alpha^2\sinh^2(2\gamma(t))}\right]\right), \tag{13}$$

$$\dot{\gamma}(t) = \frac{\beta I^S(0)}{\alpha\sqrt{d}}\left(1 - \max\left(0, 1 - 4h^2\, d/(\alpha^2\sinh^2(2\gamma(t)))\right)^{\frac{d-1}{2}}\right), \tag{14}$$

where $I^N(0) = 1/\sqrt{2\pi}$ and $I^S(0) = \Gamma\left(\frac{d+2}{2}\right)/\left(\Gamma\left(\frac{d+1}{2}\right) d\sqrt{\pi}\right)$. Both equations imply that $\gamma(t) \sim t$ for small t and $\gamma(t) \sim \ln t$ for large t.

J. Stat. Mech. (2021) 124005

We have now gained a full analytical description of the training dynamics: the node parameters evolve in time following equation (10). Note that their trajectory is independent of the training parameters and the initial distribution, which only affect the time dependence, i.e. the 'clock' $\gamma(t)$. The change of the output function is given by equation (11), where one sees that only the amplitude of $f(x, \boldsymbol{\theta})$ varies with time and is governed by $\gamma(t)$. The amplitude increases monotonically so that more examples can be classified above the margin h at later times; the more examples are classified the slower becomes the increase of $\gamma(t)$ and hence the dynamics.

Our theoretical prediction can be directly compared with a simple numerical experiment. Figure 1 shows the training of a network with $M = 400$ on Gaussian input data. The top panels (a) and (b) compare the analytical evolution of the network parameters a_i and $w_i^{\parallel}$ obtained from equation (10) to the numerical one. In (c) we plot $\gamma(t)$ (computed numerically) showing that it grows linearly in the beginning and logarithmically at longer times, as expected from theory. In (d) we show a scatter plot illustrating that the time when an example is satisfied is proportional to its projection on the reference vector, following on average our estimate based on equation (12). Overall, the agreement with the analytical solution is very good. The spread around the analytical solution in (d) is a finite-M effect, that we will analyze in section 3.3. The departure from the analytical result (10) happens at large time when the finiteness of the training set starts to matter (the larger is the training set the larger is this time). In fact, for any finite number of examples the empirical average over unsatisfied examples deviates from its population average and the dynamics is modified eventually, and ultimately stops when the whole training set is classified beyond margin. We study this regime in section 3.3.

3.2. Lazy learning and rich learning regimes

The presence of the factor α in the loss function (8) allows us to explore explicitly the crossover between different learning regimes, in particular the '*lazy learning*' regime corresponding to $\alpha \to \infty$ [23]. The dynamical equations can be studied in this limit by introducing $\overline{\gamma}(t) = \alpha\gamma(t)$. For concreteness, let us focus on the case of normally distributed data. Taking the $\alpha \to \infty$ limit of equation (13) one finds the equation for $\overline{\gamma}(t)$:

$$\dot{\overline{\gamma}}(t) = \frac{\beta I^N(0)}{\sqrt{d}} \left(1 - \exp\left[-\frac{2h^2\, d}{4\overline{\gamma}(t)^2}\right]\right). \tag{15}$$

As for the evolution of the parameters and the output function, we obtain:

$$\begin{cases} a_i(t) - a_i(0) & = w_i^{\parallel}(0)\dfrac{\overline{\gamma}(t)}{\alpha} + O(\alpha^{-2}) \\ w_i^{\parallel}(t) - w_i^{\parallel}(0) & = a_i(0)\dfrac{\overline{\gamma}(t)}{\alpha} + O(\alpha^{-2}) \end{cases}; \qquad \alpha f(\mathbf{x}\,;\boldsymbol{\theta}) = \frac{\overline{\gamma}(t)}{\sqrt{d}}\hat{\mathbf{w}}^* \cdot \mathbf{x}. \tag{16}$$

The equations above provide an explicit solution of lazy learning dynamics and illustrate its main features: the $\boldsymbol{\theta}_i$ evolves very little and along a fixed direction, in this case given by $(w_i^{\parallel}(0), a_i(0), 0)$. Despite the small changes in the nodes parameters, of the order of

J. Stat. Mech. (2021) 124005

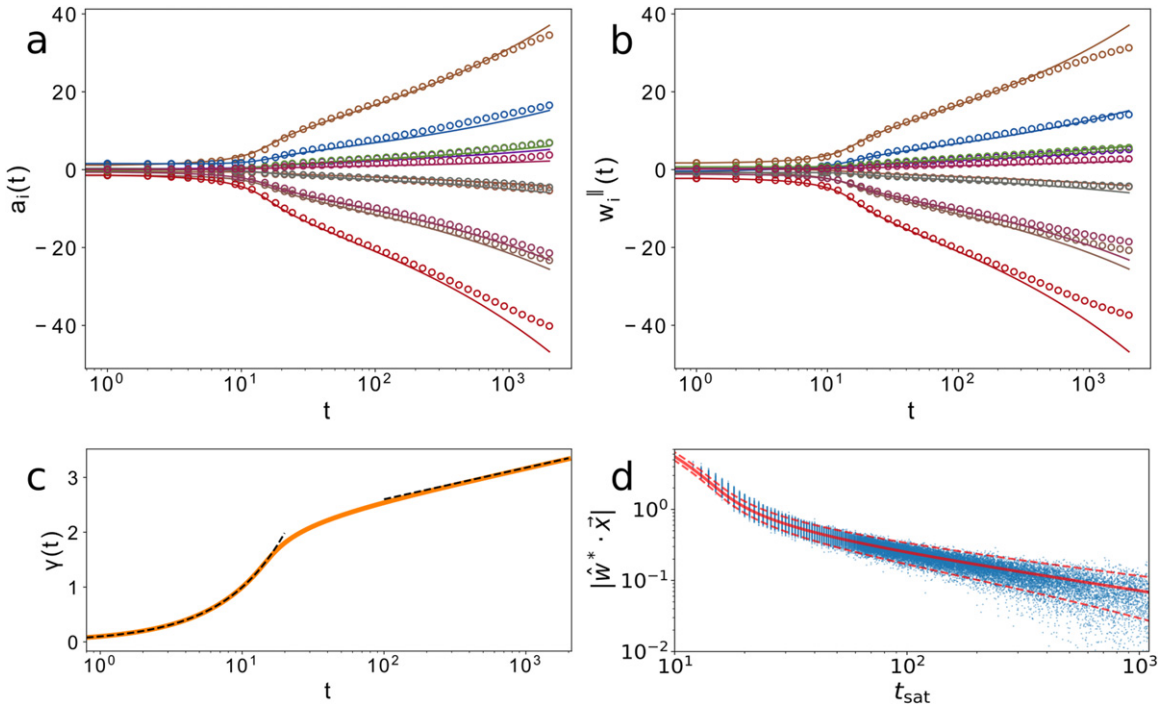

Figure 1. Training of a network with $M = 400$, $N = 10^5$, $d = 100$, $\alpha = 1.0$, $h = 1$, $\beta^* = 10^3$, for $t_{\max} = 2 \times 10^3$ timesteps (until all examples are classified) with final generalization error ~ 0.01 evaluated on 10^5 examples. Data and initial parameters are taken from a normal distribution of zero mean and width 1 per dimension. (a) and (b) Evolution of ten of the $a_i(t)$s in (a) and of the $w_i^{\parallel}(t)$s in (b) during training (circles) compared to our theoretical prediction (lines) for the same initial values. (c) Evolution of $\gamma(t)$ obtained through numerical integration of equation (13) for the parameters of this example. The dashed lines represent the linear approximation near $t = 0$ and the logarithmic slope $\log(t)/4$ for large γ (shifted with a fitted constant). (d) Projection of examples on the vector $\hat{\mathbf{w}}^*$ as a function of the time t_{sat} when they are first satisfied. The red line is the estimate of our theory, the dashed lines represent our estimate for a standard deviation due to the finite number of nodes M (see section 3.3).

$1/\alpha$, the network does learn since classification is performed through $\alpha f(\mathbf{x}; \boldsymbol{\theta})$ which has an order one change even for $\alpha \to \infty$. In this regime, the correlation between a and $w^{\parallel}$ only increases slightly, but this is enough for classification, since an infinite amount of displacements in the right direction is sufficient to solve the problem.

On the contrary, when α is of order one or smaller, the dynamics is in the so-called '*rich learning*' regime [27]. At the beginning of learning, the initial evolution of the $\boldsymbol{\theta}_i$s follows the same linear trajectories of the lazy-learning regime. However, at later stages, the trajectories are no more linear and the norm of the weights increases exponentially in $\gamma(t)$, stopping only at very large values of γ when all nodes are almost aligned with $\hat{\mathbf{w}}^*$ (for small α). Note that, as observed in Geiger *et al* [34], with the

J. Stat. Mech. (2021) 124005

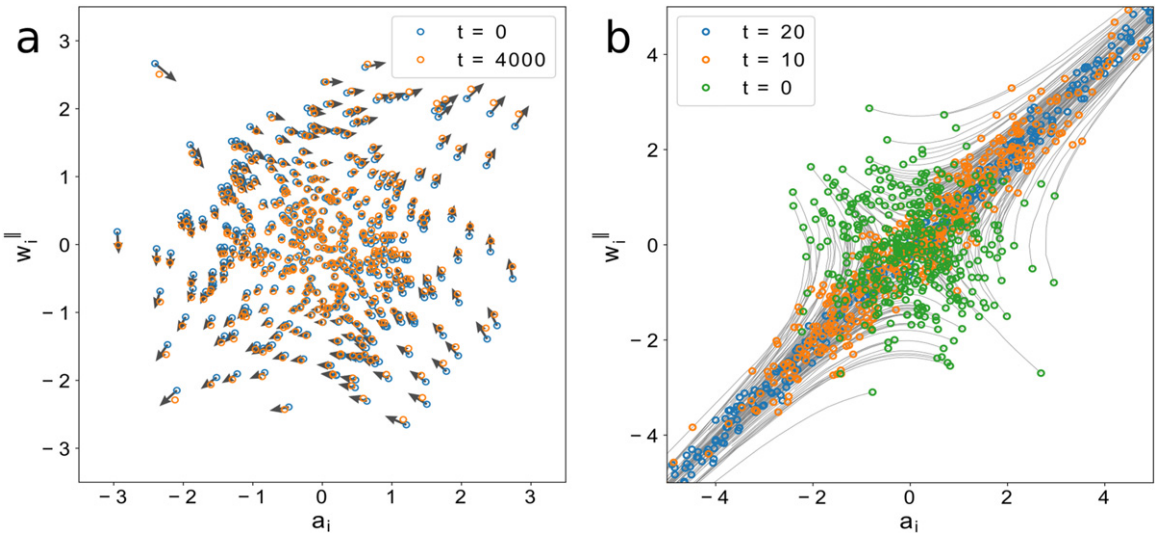

Figure 2. Evolution of a_i and $w_i^{\parallel}$ for a network with $M = 400$, $N = 10^4$, $d = 100$, $h = 1$ in two different regimes. Data and initial parameters are taken from a normal distribution of zero mean and width 1 per dimension. (a) First and last step of a case with $\alpha = 10^3$ (learning rate $\beta^* = 10^4$, training set is fitted by $t = 3000$, final generalization error ~ 0.04). The arrows indicate the analytical derivative at $t = 0$, showing that the evolution is approximately linear. (b) Initial steps (time indicated in legend) of a case with $\alpha = 10^{-3}$ (learning rate $\beta^* = 1$, training set is fitted by $t = 300$, final generalization error ~ 0.02). The gray lines follow the evolution of each node.

standard normalization $1/\sqrt{M}$ it would be the parameter $\alpha\sqrt{M}$ governing the crossover between the two regimes.

We compare the two dynamical evolutions in figure 2. The left panel (a) shows the displacement of parameters between initialization and full classification (zero training loss) for a network with $\alpha = 10^3$. As expected, the displacement is small and linear. A very different evolution takes place for $\alpha = 10^{-3}$ in the right (b) panel. The trajectories are non-linear, and all nodes approach large values close to the $a = w^{\parallel}$ line at the end of the training. Correspondingly, the initially isotropic Gaussian distribution evolves toward one with covariance matrix $\cosh(2\gamma)$ on the diagonal and $\sinh(2\gamma)$ off diagonal.

Note that for all values of α, even very large ones, the trajectories of the $\boldsymbol{\theta}_i$s are identical and given by equation (10). What differs is the 'clock' $\gamma(t)$, in particular for large α the system remains for a much longer time in the lazy regime. This is true as long as the number of training samples is infinite. Instead, if the number of data is finite, the dynamics stops once the whole training set is fitted: for large α this happens before the system is able to leave the lazy regime, whereas for small α a full non-linear (rich) evolution takes place. Hence, the finiteness of the training set leads to very distinct

J. Stat. Mech. (2021) 124005

dynamics and profoundly different 'trained' models (having both fitted the training dataset) with possibly different generalization properties [25, 34, 35].

3.3. Beyond mean-field theory

The solution we presented in the previous sections holds in the limit of an infinite number of nodes and of training data. Here we study the corrections to this asymptotic limit, and discuss the new phenomena that they bring about.

Finite number of nodes. In the large M limit the a_i and $\mathbf{w}_i$ are Gaussian i.i.d. variables. By the central limit theorem, the function (2) concentrates around its average, and has negligible fluctuations of the order of $1/\sqrt{M}$ when $M \to \infty$. If M is large but finite (keeping an infinite training set), these fluctuations of $f(x, \boldsymbol{\theta})$ are responsible for the leading corrections to mean-field theory. In the SM we compute explicitly the variance of the output function, $\lim_{M\to\infty} M \operatorname{Var}[f(x, \boldsymbol{\theta})] = \sigma_f^2(t)$, with

$$\sigma_f^2(t) \equiv ((5\cosh^2(2\gamma(t)) - 2\cosh(2\gamma(t)) - 3)(\hat{\mathbf{w}}^* \cdot \mathbf{x})^2 + 2\cosh(2\gamma(t))|\mathbf{x}|^2)/(4d). \tag{17}$$

The main effect of this correction is to induce a spread in the dynamics, e.g. of the data with same satisfaction time. This phenomenon is shown in figure 1(d) for $M = 400$, where we compare the numerical spread to an estimate of the values of $\hat{\mathbf{w}}^* \cdot \mathbf{x}$ such that the hinge is equal to the average plus or minus one standard deviation (details on this estimate in the SM).

Finite number of data. We now consider a finite but large number of examples N (keeping infinite the number of nodes). In the large N limit the empirical average over the data in $\langle u(\mathbf{x}, y\,; t)\theta\,(\mathbf{w}\cdot\mathbf{x})\,y\,\mathbf{x}\rangle_{\mathbf{x},y}$ converges to its mean $I(t)\hat{\mathbf{w}}^*$. The main effect of considering a finite N is that the empirical average fluctuates around this value. Using the central limit theorem we show in the SM that the leading correction to the asymptotic result reads:

$$\langle u(\mathbf{x}, y\,; t)\theta\,(\mathbf{w}\cdot\mathbf{x})\,y\mathbf{x}\rangle_{\mathbf{x},y} = I(t)\hat{\mathbf{w}}^* + \frac{J(t)}{\sqrt{N}}\boldsymbol{\delta}\mathbf{w}_\perp + O(1/N), \tag{18}$$

where $\boldsymbol{\delta}\mathbf{w}_\perp$ is a unitary random vector perpendicular to $\hat{\mathbf{w}}^*$ and $J(t) \equiv \sqrt{(d-1)f^U(t)/2}$. The term $f^U(t) \equiv \langle u(\mathbf{x}, y\,; t)\rangle_{\mathbf{x},y}$, the fraction of unsatisfied examples at time t, controls the strength of the correction, as expected since only unsatisfied data contribute to the empirical average $\langle\cdot\rangle_{\mathbf{x},y}$. The vector on the rhs of (18) is the one toward which all the $\mathbf{w}_i$ align, see equation (10). Therefore, the main effect of the correction (18) is for the nodes parameters to align along a direction which is slightly different from $\hat{\mathbf{w}}^*$ and dependent on the training set. This naturally induces different accuracies between the training and the test sets, i.e. it leads to *overfitting*[2]. Note that the strength of the signal, $I(t)$, is roughly of the order of the fraction of unsatisfied data $f^U(t)$, whereas the noise due to

[2] The two accuracies instead coincide for $N \to \infty$, since all possible data are seen during the training and no overfitting is present in the asymptotic limit.

J. Stat. Mech. (2021) 124005

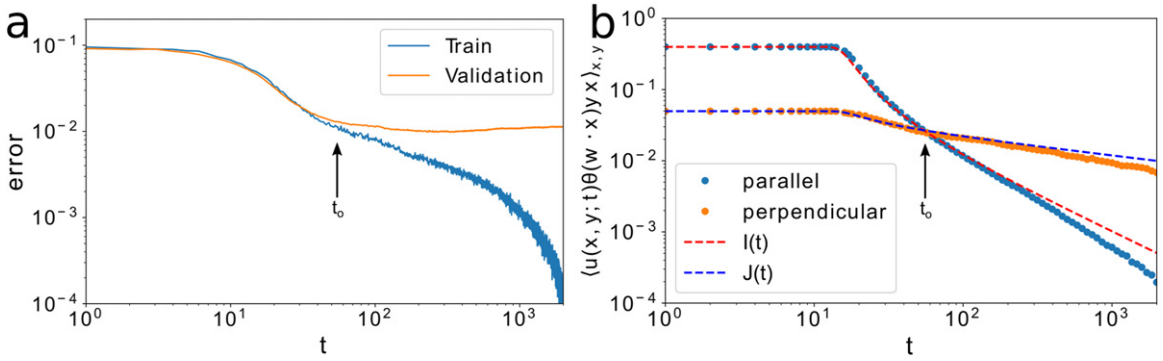

Figure 3. (a) Training (blue) and generalization (orange) error (fraction of misclassified examples), during training with same parameters as figure 1. (b) Components of $\langle u(\mathbf{x}, y; t)\theta\left(\mathbf{w}\cdot\mathbf{x}\right) y\mathbf{x}\rangle_{\mathbf{x},y}$ along $\hat{\mathbf{w}}^*$ (parallel) and perpendicular to it, during training. The dots are numerical results for the same training show in (a). The lines represent our analytical predictions $I(t)$ and $J(t)/\sqrt{N}$ for the same parameters.

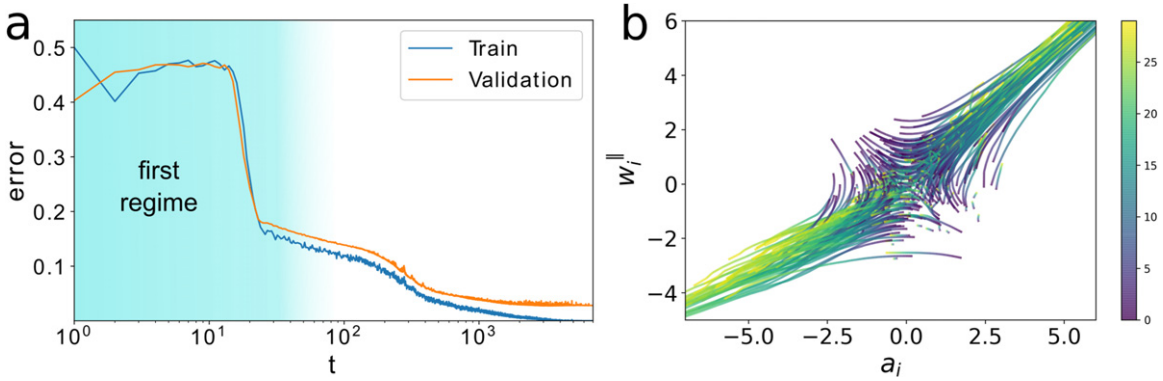

Figure 4. (a) Training (blue) and generalization (orange) error for a network with $M = 400$, trained on $N = 10^4$ MNIST data ($d = 784$), with parity labels. Inputs are only rescaled by a factor $1/255$, no further processing is done. The training is performed with $\beta^* = 1000$, $\alpha = 1$, $h = 1$ and the validation error on 10^4 examples is $\sim$0.03 after 2000 evolution steps. The shaded area represents the area where our theory applies. (b) Evolution of a_i and $w_i^{\parallel}$ in the first 30 steps of training. The color (see color bar) represents the step of evolution.

the finite training set is proportional to the square root of it. The larger the time, the smaller $f^U(t)$ is, hence the stronger are the fluctuations with respect to the signal. In figure 3(b), we compute numerically the components of $\langle u(\mathbf{x}, y\,; t)\theta\left(\mathbf{w}\cdot\mathbf{x}\right) y\mathbf{x}\rangle_{\mathbf{x},y}$ parallel and perpendicular to $\hat{\mathbf{w}}^*$, and compare them to $I(t)$ and $J(t)/\sqrt{N}$. Remarkably, we find a very good agreement even for times when $J(t)/\sqrt{N}$ is no longer a small correction. This suggest that an estimate of the time t_o when overfitting takes place is given by $I(t_o) = J(t_o)/\sqrt{N}$. We test this conjecture in (a): indeed the two contributions are of the same order of magnitude for $t_o \sim 50$, which is around the time when training and validation errors diverge.

J. Stat. Mech. (2021) 124005

3.4. Mislabeling

We now briefly address the effects due to noise in the labels, see the SM for detailed results and numerical experiments. Mislabeling is introduced by flipping the label of a small fraction δ of the examples. The main effect is to decrease the strength of the signal, $I(t)$, since the mislabeled data lead to an opposite contribution in (9) with respect to the correctly labeled ones. In the asymptotic limit of infinite N and M, the reduction of the signal slows down the dynamics, which stops when the number of unsatisfied correct examples equals the one of mislabeled ones. For large but finite N, the noise $J(t)/\sqrt{N}$ is enhanced with respect to the signal because its strength is related to the fraction of *all* unsatisfied examples, and not just the correctly labeled ones. Hence, overfitting is stronger and takes place earlier with respect to the case analyzed before.

4. Discussion and experiment

We have provided an analytical theory for the dynamics of a single hidden layer neural network trained for binary classification with linear hinge loss. We have found two dynamical regimes: a first one, correctly accounted for by mean-field theory, in which every node has its own dynamics with a time-dependent dataset determined self-consistently from the average nodes population. During this evolution the nodes parameters align with the direction of the reference classification vector. In the second regime, which is not accounted for by mean-field theory, the noise due to the finite training set becomes important and overfitting takes place. The merit of the model we focused on is that, thanks to its simplicity, several effects happening in real networks can be studied in detail analytically. Several works have shown distinct dynamical regimes in the training dynamics: first the network learns coarse grained properties, later on it captures the finer structure, and eventually it overfits [8, 13, 36, 37]. Given the simplicity of the dataset considered, we expect our model to describe the first regime but not the second one, which would need a more complex model of data.

In particular, the effective one-dimensional nature of the $\mathbf{w}$ evolution is due to the cylindrical symmetry of the data, resulting in a direction-independent expression for the integral in equation (9). In a more general setting, we can still expect to recover a similar behavior at the beginning of training, where the difference between the two classes averages dominates most of the dynamics. After that, the integral will depend more and more on the direction of $\mathbf{w}$, leading to specialization and a departure from our simple model. To test this conjecture, we train our network to classify the parity of MNIST handwritten digits [38]. To establish a relationship with our case, we define $\hat{\mathbf{w}}^*$ as the direction of the difference between the averages of the two parity sets. We can now define $w^{\parallel}$ for each node, and study the dynamics of $a_i, w_i^{\parallel}$. We report in figure 4 the evolution of these parameters in the early steps of training, in which the training loss decreases of 65% of its initial value (figure 4(a)). The evolution of the parameters (figure 4(b)) bears a strong resemblance with our findings, see the remarkable similarity with figure 2(b). A similar experiment for even richer datasets (CIFAR10 and ImageNet) is presented in the SM.

J. Stat. Mech. (2021) 124005

Broader impact

Given the purely theoretical scope of this paper, it does not seem to present any foreseeable societal consequence.

Acknowledgments

We thank S d'Ascoli and L Sagun for discussions, and M Wyart for exchanges about his work on a similar model [39]. We acknowledge funding from the French government under management of Agence Nationale de la Recherche as part of the 'Investissements d'avenir' program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and from the Simons Foundation collaboration 'Cracking the Glass Problem' (No. 454935 to G Biroli).

References

[1] LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436–44
[2] Barron A R 1993 Universal approximation bounds for superpositions of a sigmoidal function *IEEE Trans. Inf. Theory* **39** 930–45
[3] Poggio T, Kawaguchi K, Liao Q, Miranda B, Rosasco L, Boix X, Hidary J and Mhaskar H 2017 Theory of deep learning: III. Explaining the non-overfitting puzzle (arXiv:1801.00173)
[4] Suggala A, Prasad A and Ravikumar P K 2018 Connecting optimization and regularization paths *Advances in Neural Information Processing Systems* vol 31 ed S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi and R Garnett (New York: Curran Associates) pp 10608–19
[5] Gidel G, Bach F and Lacoste-Julien S 2019 Implicit regularization of discrete gradient dynamics in linear neural networks *Advances in Neural Information Processing Systems* vol 32 ed H Wallach, H Larochelle, A Beygelzimer, F d'Alché Buc, E Fox and R Garnett (New York: Curran Associates) pp 3202–11
[6] Dauphin Y N, Pascanu R, Gulcehre C, Cho K, Ganguli S and Bengio Y 2014 Identifying and attacking the saddle point problem in high-dimensional non-convex optimization *Advances in Neural Information Processing Systems* pp 2933–41
[7] Sagun L, Bottou L and LeCun Y 2016 Singularity of the hessian in deep learning (arXiv:1611.07476)
[8] Baity-Jesi M, Sagun L, Geiger M, Spigler S, Ben Arous G, Cammarota C, LeCun Y, Wyart M and Biroli G 2019 Comparing dynamics: deep neural networks versus glassy systems *J. Stat. Mech.* 124013
[9] Saxe A, Mcclelland J and Ganguli S 2014 Exact solutions to the nonlinear dynamics of learning in deep linear neural networks *Int. Conf. Learning Representations* pp 1–22
[10] Lampinen A K and Ganguli S 2019 An analytic theory of generalization dynamics and transfer learning in deep linear networks *Int. Conf. Learning Representations*
[11] Saad D and Solla S A 1995 Exact solution for on-line learning in multilayer neural networks *Phys. Rev. Lett.* **74** 4337–40
[12] Advani M S and Saxe A M 2017 High-dimensional dynamics of generalization error in neural networks (arXiv:1710.03667)
[13] Goldt S, Advani M S, Saxe A M, Krzakala F and Zdeborová L 2019 Generalisation dynamics of online learning in over-parameterised neural networks (arXiv:1901.09085)
[14] Yoshida Y, Karakida R, Okada M and Amari S-I 2019 Statistical mechanical analysis of learning dynamics of two-layer perceptron with multiple output units *J. Phys. A: Math. Theor.* **52** 184002
[15] Du S S, Zhai X, Poczos B and Singh A 2019 Gradient descent provably optimizes over-parameterized neural networks *Int. Conf. Learning Representations*
[16] Arora S, Du S S, Hu W, Li Z and Wang R 2019 Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks *36th Int. Conf. Machine Learning (ICML)* (9–15 June 2019) pp 477–502
[17] Mei S, Montanari A and Nguyen P-M 2018 A mean field view of the landscape of two-layer neural networks *Proc. Natl Acad. Sci. USA* **115** E7665–71

J. Stat. Mech. (2021) 124005

[18] Mei S, Misiakiewicz T and Montanari A 2019 Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit *Proc. 32nd Conf. Learning Theory (Proc. Machine Learning Research (PMLR))* vol 99 (Phoenix, USA, 25–28 June 2019) ed A Beygelzimer and D Hsu pp 2388–464
[19] Kadmon J and Sompolinsky H 2016 Optimal architectures in a solvable model of deep networks *Advances in Neural Information Processing Systems* vol 29 ed D D Lee, M Sugiyama, U V Luxburg, I Guyon and R Garnett (New York: Curran Associates) pp 4781–9
[20] Rotskoff G M and Vanden-Eijnden E 2018 Trainability and accuracy of neural networks: an interacting particle system approach (arXiv:1805.00915)
[21] Araújo D, Oliveira R I and Yukimura D 2019 A mean-field limit for certain deep neural networks (arXiv:1906.00193)
[22] Nguyen P-M 2019 Mean field limit of the learning dynamics of multilayer neural networks (arXiv:1902.02880)
[23] Chizat L, Oyallon E and Bach F 2019 On lazy training in differentiable programming *NeurIPS 2019: 33rd Conf. Neural Information Processing Systems* (Vancouver, Canada)
[24] Jacot A, Gabriel F and Hongler C 2018 Neural tangent kernel: convergence and generalization in neural networks *Advances in Neural Information Processing Systems* vol 31 ed S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi and R Garnett (New York: Curran Associates) pp 8571–80
[25] Lee J, Xiao L, Schoenholz S, Bahri Y, Novak R, Sohl-Dickstein J and Pennington J 2019 Wide neural networks of any depth evolve as linear models under gradient descent *Advances in Neural Information Processing Systems* vol 31 ed H Wallach, H Larochelle, A Beygelzimer, F d'Alché Buc, E Fox and R Garnett (New York: Curran Associates) pp 8572–83
[26] Geiger M, Jacot A, Spigler S, Gabriel F, Sagun L, d'Ascoli S, Biroli G, Hongler C and Wyart M 2020 Scaling description of generalization with number of parameters in deep learning *J. Stat. Mech.* 023401
[27] Woodworth B, Gunasekar S, Lee J D, Moroshko E, Savarese P, Golan I, Soudry D and Srebro N 2020 Kernel and rich regimes in overparametrized models (arXiv:2002.09277)
[28] Soudry D, Hoffer E, Nacson M S, Gunasekar S and Srebro N 2018 The implicit bias of gradient descent on separable data *J. Mach. Learn. Res.* **19** 2822–78
[29] Liao Z and Couillet R 2018 The dynamics of learning: a random matrix approach *Proc. Machine Learning Research (PMLR)* (Stockholm, Sweden, 10–15 July 2018) vol 80 ed J Dy and A Krause (Stockholmsmässan) pp 3072–81
[30] des Combes R T, Pezeshki M, Shabanian S, Courville A and Bengio Y 2018 On the learning dynamics of deep neural networks (arXiv:1809.06848)
[31] Nacson M S, Lee J, Gunasekar S, Savarese P H P, Srebro N and Soudry D 2019 Convergence of gradient descent on separable data *Proc. Machine Learning Research (PMLR)* (16–18 April 2019) vol 89 ed K Chaudhuri and M Sugiyama pp 3420–8
[32] Kipnis C and Landim C 1998 *Scaling Limits of Interacting Particle Systems* vol 320 (Berlin: Springer)
[33] Serfaty S 2014 Coulomb gases and Ginzburg–Landau vortices (arXiv:1403.6860)
[34] Geiger M, Spigler S, Jacot A and Wyart M 2019 Disentangling feature and lazy training in deep neural networks (arXiv:1906.08034)
[35] Arora S, Du S S, Hu W, Li Z, Salakhutdinov R R and Wang R 2019 On exact computation with an infinitely wide neural net *Advances in Neural Information Processing Systems* vol 32 ed H Wallach, H Larochelle, A Beygelzimer, F d'Alché Buc, E Fox and R Garnett (New York: Curran Associates) pp 8141–50
[36] Saad D and Solla S A 1996 Dynamics of on-line gradient descent learning for multilayer neural networks *Advances in Neural Information Processing Systems* pp 302–8
[37] Bordelon B, Canatar A and Pehlevan C 2020 Spectrum dependent learning curves in kernel regression and wide neural networks (arXiv:2002.02561)
[38] LeCun Y, Cortes C and Burges C J 2010 MNIST handwritten digit database (AT & T Labs) (http://yann.lecun.com/exdb/mnist)
[39] Paccolat J, Petrini L, Geiger M, Kevin T and Wyart M 2020 Geometric compression of invariant manifolds in neural nets (arXiv:2007.11471)

J. Stat. Mech. (2021) 124005