The jamming transition as a paradigm to understand the loss landscape of deep neural networks

Mario Geiger,[1, *] Stefano Spigler,[1, *] Stéphane d'Ascoli,[2, 3] Levent
Sagun,[2, 1] Marco Baity-Jesi,[4] Giulio Biroli,[2, 3] and Matthieu Wyart[1]

[1]*Institute of Physics, EPFL, CH-1015 Lausanne, Switzerland*
[2]*Institut de Physique Théorique, Université Paris-Saclay, CEA, CNRS, F-91191 Gif-sur-Yvette, France*
[3]*Laboratoire de Physique Statistique, École Normale Supérieure,*
PSL Research University, F-75005 Paris, France
[4]*Department of Chemistry, Columbia University, 10027 New York, USA*
(Dated: June 18, 2019)

Deep learning has been immensely successful at a variety of tasks, ranging from classification to artificial intelligence. Learning corresponds to fitting training data, which is implemented by descending a very high-dimensional loss function. Understanding under which conditions neural networks do not get stuck in poor minima of the loss, and how the landscape of that loss evolves as depth is increased remains a challenge. Here we predict, and test empirically, an analogy between this landscape and the energy landscape of repulsive ellipses. We argue that in fully-connected deep networks a phase transition delimits the over- and under-parametrized regimes where fitting can or cannot be achieved. In the vicinity of this transition, properties of the curvature of the minima of the loss (the spectrum of the hessian) are critical. This transition shares direct similarities with the jamming transition by which particles form a disordered solid as the density is increased, which also occurs in certain classes of computational optimization and learning problems such as the perceptron. Our analysis gives a simple explanation as to why poor minima of the loss cannot be encountered in the overparametrized regime. Interestingly, we observe that the ability of fully-connected networks to fit random data is independent of their depth, an independence that appears to also hold for real data. We also study a quantity Δ which characterizes how well ($\Delta < 0$) or badly ($\Delta > 0$) a datum is learned. At the critical point it is power-law distributed on several decades, $P_+(\Delta) \sim \Delta^\theta$ for $\Delta > 0$ and $P_-(\Delta) \sim (-\Delta)^{-\gamma}$ for $\Delta < 0$, with exponents that depend on the choice of activation function. This observation suggests that near the transition the loss landscape has a hierarchical structure and that the learning dynamics is prone to avalanche-like dynamics, with abrupt changes in the set of patterns that are learned.

PACS numbers: 64.70.Pf,65.20.+w,77.22.-d

I. INTRODUCTION

Deep neural networks are now central tools for a variety of tasks including image classification [1, 2], speech recognition [3] and the development of artificial intelligence that can for example master the game of Go beyond human level [4, 5]. A neural network represents a (very high-dimensional) function f that depends on a large number of parameters N [2]. These parameters are learned so as to correctly classify P training data by minimizing some loss function $\mathcal{L}$, generally via stochastic gradient descent (a kind of noisy version of gradient descent). There is great flexibility in the network architecture, loss function and minimization protocol one can use. These features are ultimately selected to optimize the classification of previously unseen data, or *generalization*. Although the current progress in designing [6, 7] and training [8] networks that generalize well is undeniable, it remains mostly empirical. A general theory explaining and fostering this success is lacking, and central questions remain to be clarified. First, since the loss function is generally not convex, why doesn't the learning dynamics get stuck in poorly performing minima with high loss? In other words, under which conditions can one guarantee that training data are well fitted? Second, what are the benefits of deeper networks? On the one hand it is often argued, and proved in some cases, that the advantage of deep networks stems from their enhanced expressive power, i.e. their ability to build complex functions with a much smaller number of parameters than needed for shallow networks [9–13]. Indeed if deep networks are able to fit data with less parameters, then they are likely to generalize better. On the other hand, one can handcraft neural networks that fit even structure-less, random data with a rather small number of parameters $N \sim P$ [14–17]. These results for the static capacity of networks appear to be independent of depth [16, 17]. Yet, it is unclear whether such parsimonious solutions can be found dynamically in practice simply by descending the loss function, and whether depth can help finding them. More generally, how is the loss landscape affected by depth?

Complex physical systems with non-convex energy landscapes featuring an exponentially large number of local minima are called glassy [18]. Does the landscape of deep learning fall into a known class of glassy systems?

---
* These two authors contributed equally.

Along this line, an analogy between deep networks and mean-field glasses (p-spins) has been proposed [19], in which the learning dynamics is expected to get stuck in the highest minima of the loss, which are the most abundant. Yet, several numerical and rigorous works [20–23] (the latter focusing on shallow and very overparametrized networks) suggest a different landscape geometry where the loss function is characterized by a connected level set. Furthermore, studies of the Hessian of the loss function [24–26] and of the learning dynamics [27, 28] support that the landscape is characterized by an abundance of flat directions, even near its bottom, at odds with traditional glassy systems.

In the last decade several works have unveiled an analogy between the physical phenomenon of jamming [29, 30] and phase transitions taking place in certain classes of computational optimization and learning problems [31–33], in particular the perceptron [33, 34] — the simplest neural network performing linear classification. In this work we push this analogy further and show that the geometry of the training loss landscape and the training dynamics of fully connected deep neural networks is affected by a jamming transition similar to that of repulsive ellipses [30]. As illustrated in Fig. 2, jamming occurs in packings of particles interacting through a finite-range potential $\mathcal{U}$, when the particle density ϕ reaches some critical value ϕ_c. At that point, particles can no longer be accommodated without touching each other and the system becomes a solid with singular landscape properties, embodied for example in the spectrum of the Hessian of $\mathcal{U}$ [35, 36], that at the transition displays many (almost) flat directions. Particles of different shapes, such as spheres and ellipses, can lead to different jamming scenarios [37–40].
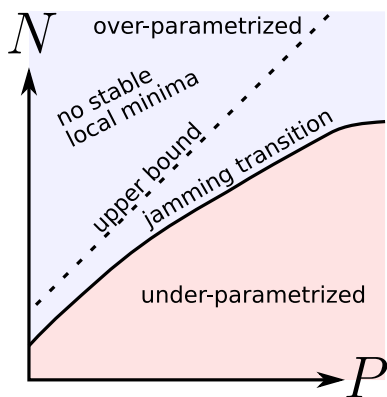

FIG. 1. N: degrees of freedom, P: training examples.

Here we show that for two commonly used loss functions (cross-entropy and quadratic hinge), fully-connected deep networks undergo a jamming transition too, below which all data are correctly fitted and above which they are not, both for real data (images) and random data [41]. In both cases the transition appears to be solely controlled by the number of parameters of the

network N, independently of depth. For random data, the transition takes place as the quantity P/N increases toward some critical value P/N^*. For the hinge loss, using results from the jamming literature we argue that $P/N^* \geq C_0$ where C_0 is a constant that we can measure *a posteriori* once learning took place. To hold, this result requires the network output to remain sensitive to all its weights during training, as we observe empirically in the examples we study. This view supports that the dynamics cannot get stuck in poor minima in the overparametrized regime where networks tend to operate, because there are not enough constraints to form minima in that regime. We also find that the jamming transition is sharp and the landscape appears to fall in the same universality class independently of depth (as long as at least one hidden layer is present). Differently from the (non-convex) perceptron, that was proven to lie in the same universality class as spherical particles [33, 34], we show that deep networks instead jam in a manner similar to ellipses. From our analysis we deduce the singular properties of the spectrum of the Hessian of the loss, which indeed must display many flat directions. We find empirically that other key quantities (the fraction of data which are almost correctly or almost incorrectly classified) display power-law behaviours on several decades, with new exponents. In glassy systems, such power-laws reveal properties that cannot be reached by studying the Hessian, in particular the fact that the dynamics occurs via broadly distributed avalanches [42–44], indicative of a hierarchical organization of the landscape [45]. This observation thus suggests that these properties also characterize deep networks near the transition. Note that in this work we focus on training and the ability of deep neural networks to fit a dataset. The implications and the relations with generalization between jamming and generalization are investigated in [46, 47].

II. ANALOGY BETWEEN JAMMING AND DEEP LEARNING

A. Jamming

Understanding the energy landscape — in particular the properties of the Hessian, referred to as vibrational properties in this context — in disordered systems of interacting particles is a long-standing and practically important problem [48]. It was realized that for purely repulsive, finite-range particles, such properties are singular near the jamming transition where the system becomes a solid [35, 36], allowing one to develop and test theories for the vibrations of glasses, that turn out to apply in a broader class of systems where the interactions do not necessarily have finite range [29]. Here we shall follow the same strategy for deep networks, where the role of the "interaction potential" is played by the choice of loss function. Finite-range interactions are mimicked by the *hinge loss*, for which we predict a sharp transition when
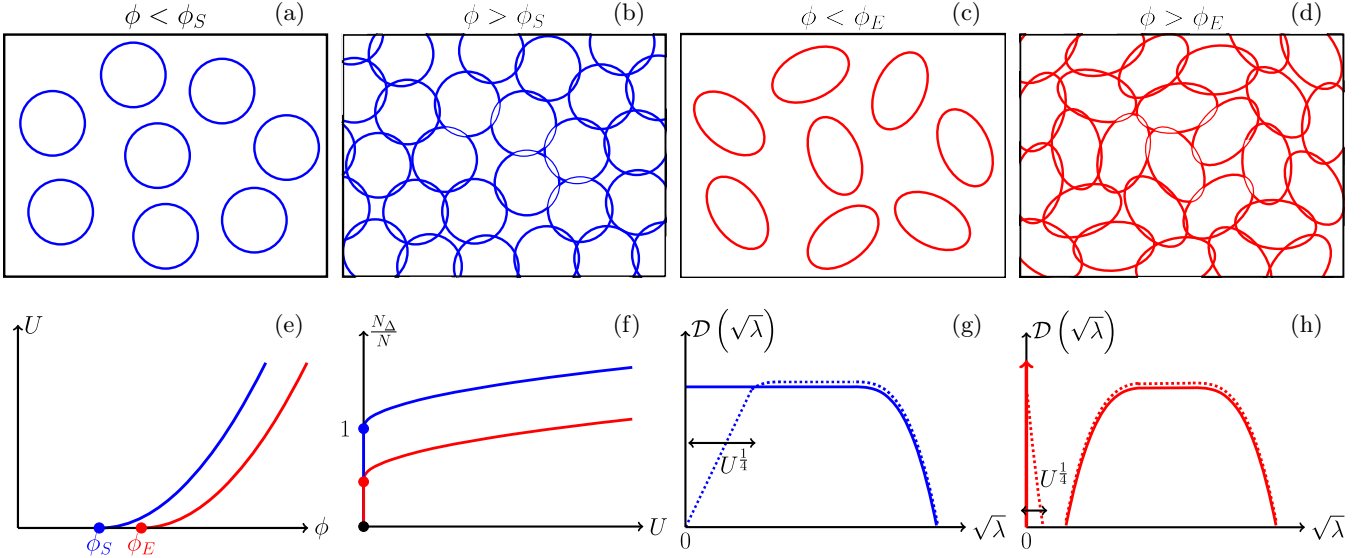
FIG. 2. Sketch of the jamming transition for repulsive spheres and ellipses. (a,b,c,d) Both systems transition from a fluid to a solid as the density passes some threshold, noted ϕ_S for spheres and ϕ_E for ellipses. (e) For denser packings, the potential energy $\mathcal{U}$ becomes finite. (f) The ratio N_Δ/N between the number of particles in contact N_Δ (corresponding to unsatisfied constraints) and the number of degrees of freedom N jumps discontinuously to a finite value, which is unity for spheres but smaller for ellipses. (g,h) This difference has dramatic consequence on the energy landscape, in particular on the spectrum of the Hessian. In both cases, the spectrum becomes non-zero at jamming, but it displays a delta function with finite weight for ellipses (indicating strictly flat directions), followed by a gap with no eigenvalues, followed by a continuous spectrum (h, full line). For spheres, there is no delta function nor gap (g, full line). As one enters the jammed phase, in both cases a characteristic scale $\lambda \sim \sqrt{\mathcal{U}}$ appears in the spectrum (g and h, dotted lines).

going from an overparametrized to an underparametrized regime. At the transition, the Hessian is singular and displays an abundance of low-energy modes. For other types of losses — such as for the commonly used cross-entropy loss defined below — the transition exists but its effects on the spectrum are expected to be less sharp (see discussion below).

We start by recalling some results on the jamming transition. We will first discuss the case of spherical particles, since it has been studied thoroughly and is easier to formalize. The behavior of elliptical particles will be discussed later on. Consider spheres of radius R at positions $\{\mathbf{r}_i\}$, corresponding to a total number of degrees of freedom $\tilde{N}$. We denote by $r_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\|$ the distance between particles i and j, and define their overlap $\Delta_{ij} = 2R - r_{ij}$. Two particles are said to be in contact if $\Delta_{ij} > 0$, and N_Δ denotes the number of such contacts. We label by μ all the possible pairs of particles (ij) and by m the sets of contacts. We consider the following potential energy:

$$\mathcal{U} = \sum_{\mu \in m} \frac{1}{2}\Delta_\mu^2. \tag{1}$$

We denote by N the effective number of degrees of freedom which affect the variables Δ_μ. It is in general smaller than $\tilde{N}$ because of (i) global translations or rotations of the system and (ii) "rattlers", i.e. particles which make no contact with the others, whose degrees of freedom are irrelevant as far as the solid phase is concerned.

As the jamming transition is approached from above (large density ϕ), $\mathcal{U} \to 0$ as sketched in Fig. 2, implying that $\Delta_\mu \to 0 \ \forall \mu \in m$. As argued in [49], for each $\mu \in m$ the constraint $\Delta_\mu = 0$ defines a manifold of dimension $N - 1$. Satisfying N_Δ such equations thus generically leads to a manifold of solutions of dimension $N - N_\Delta$. Imposing that solutions exist thus implies that, at jamming, one has

$$N_\Delta \leq N. \tag{2}$$

Note that this argument implicitly assumes that the N_Δ constraints are independent. In disordered systems this assumption is generally correct in practice, but it may break down if symmetries are present, which is the case e.g. in crystals where Eq. (2) can be violated.

An opposite bound can be obtained for spheres by considerations of stability, by imposing that in a stable minimum the Hessian must be positive definite [35]. The Hessian is an $N \times N$ matrix which can be written as [50]

$$\mathcal{H}_U = \sum_{\mu \in m} \nabla\Delta_\mu \otimes \nabla\Delta_\mu + \sum_{\mu \in m} \Delta_\mu \nabla \otimes \nabla\Delta_\mu \equiv \mathcal{H}_0 + \mathcal{H}_p, \tag{3}$$

where $\mathcal{H}_0$ and $\mathcal{H}_p$ correspond to the first and second sum, respectively. $\mathcal{H}_0$ is positive semi-definite, since it is the sum of N_Δ matrices of rank unity; thus $\mathrm{rk}(\mathcal{H}_0) \leq N_\Delta$, implying that the kernel of $\mathcal{H}_0$ is at least of dimension $N - N_\Delta$. On the other hand for *spheres* — but not for ellipses, and this will have major consequences

— $\mathcal{H}_p$ is negative definite, which simply stems from the fact that the second-order contribution of the displacement to the distance between two points is always positive - a straightforward application of the Pythagoras theorem. It is easy to show [35] that any non-zero vector $|u\rangle$ belonging to the kernel of $\mathcal{H}_0$ must satisfy $\langle u|\mathcal{H}_U|u\rangle = \langle u|\mathcal{H}_p|u\rangle < 0$ [51]. Thus stability requires that $\mathrm{rk}(\mathcal{H}_0) = N$, implying that $N_\Delta \geq N$. Together with Eq. (2) that leads to $N_\Delta = N$: as spheres jam the number of degrees of freedom and the number of constraints (stemming from contacts) are equal, as empirically observed [52]. This property is often called *isostaticity*: when it holds, mean-field arguments [53–55] predict that the density of vibrational modes $D(\sqrt{\lambda})$ displays a plateau up to vanishingly small λ, as observed numerically [35, 36] and sketched in Fig. 2G.

However, for *ellipses* [37] (and as we shall see, for deep networks), this argument breaks down because $\mathcal{H}_p$ is not negative definite. Whether such a matrix has positive eigenvalues or not plays a role of utmost importance in the selection of the universality class of the jamming transition, and it has major consequences on the singularity of the landscape, as it can be evinced from the spectrum of the Hessian matrix. Indeed, for ellipses stability and jamming can, and generically do, occur at:

$$N_\Delta/N < 1, \tag{4}$$

a situation that is referred to as *hypostatic*. The density of vibrational modes at jamming must then display a delta function in zero of magnitude $1 - N_\Delta/N$, corresponding to the kernel of $\mathcal{H}_0$ ($\mathcal{H}_p$ vanishes at jamming since $\Delta_\mu \to 0 \; \forall \mu \in m$). Mean-field arguments applied to hypostatic materials [40, 56] predict that at larger λ, the spectrum presents a gap before becoming continuous again, as sketched in Fig. 2H. Away from jamming the effects of $\mathcal{H}_p$ kick in and broaden the delta function by an amount proportional to the typical value of the overlap $\Delta \sim \sqrt{\mathcal{U}}$, as sketched in Fig. 2H.

We now show that even in the hypostatic case, stability can be constraining. Let us denote by E_- the vector space spanned by the negative eigenvalues of $\mathcal{H}_p$, whose dimension very close to jamming is denoted N_-. Stability then imposes that the intersection of the kernel of $\mathcal{H}_0$ and E_- is zero, which is possible only if

$$N_\Delta \geq N_-. \tag{5}$$

Finally, another key structural property of the jamming transition is contained in the distribution $P_+(\Delta)$ of *positive overlaps*, sometimes referred to as *forces* (the force between two particles is Δ when $\Delta > 0$), and the distribution $P_-(\Delta)$ of *gaps* ($\Delta < 0$) between particles. It was shown that even if a packing of spherical particles is linearly stable, paths in the phase space that lower the energy are easily found unless both distributions are critical, with $P_+(\Delta) \sim \Delta^\theta$ and $P_-(\Delta) \sim (-\Delta)^{-\gamma}$, with $\gamma \geq (1-\theta)/2$ [42, 57], as numerically confirmed in [58, 59]. For a broad class of dynamics, this bound must be saturated [43], a scenario referred to as *marginal stability*

which implies that the dynamics proceeds via power-law distributed events (called avalanches) in which the set of constraints change. Calculations in infinite dimensions [45, 60] showed that marginal stability is associated with a hierarchical organization of minima of the energy (a phenomenon referred to as *replica symmetry breaking* [61]), and exponents were found to follow $\gamma = 0.41269\ldots$ and $\theta = 0.42311\ldots$ which appear accurate even in finite dimensions [57, 62].

B. Deep Learning

Set-up: We consider a binary classification problem, with a set of P distinct training data denoted as $\{\mathbf{x}_\mu, y_\mu\}_{\mu=1,\ldots,P}$. The vector x_μ is the datum itself, which lives in dimension d (e.g. it could be an image), and $y_\mu = \pm 1$ is its label. A network architecture corresponds to a function $f(\mathbf{x}; \mathbf{W})$, where $\mathbf{W}$ denotes the vector of parameters and $f(\mathbf{x}; \mathbf{W})$ corresponds to the output of the network shown in Fig. 3. In this scheme, each neuron sums the activity of all the neurons in the previous layer with some weights, sketched as connections in Fig. 3 (each connection thus corresponds to one parameter $W_{\alpha,\beta}^{(i)}$). Next, a bias $B_\alpha^{(i)}$ is added to this sum (one additional parameter per neuron) to obtain the so-called pre-activation ($a_\alpha^{(i)}$ in the picture and in the equations). The neuron activity is then a non-linear function ρ of that pre-activation: in what follows we will deal mainly with $\rho(a) = a\theta(a)$ — the so-called rectified linear unit — but we will also present some results with $\rho(a) = \tanh(a)$. The computation is done iteratively from the first layer (close to the input $\mathbf{x}$) to the last one (the output $f(\mathbf{x}; \mathbf{W})$):

$$f(\mathbf{x}; \mathbf{W}) \equiv a^{(L+1)}, \tag{6}$$

$$a_\beta^{(i)} = \sum_\alpha W_{\alpha,\beta}^{(i)} \rho\left(a_\alpha^{(i-1)}\right) - B_\beta^{(i)}, \tag{7}$$

$$a_\beta^{(1)} = \sum_\alpha W_{\alpha,\beta}^{(1)} x_\alpha - B_\beta^{(1)}. \tag{8}$$

In our notation the vector $\mathbf{W}$ contains all the parameters, including the biases. $\mathbf{W}$ is learned by minimizing a cost function, which can generically be written $\mathcal{L}(\mathbf{W}) = \frac{1}{P} \sum_{\mu=1}^{P} \ell\left(y_\mu, f(\mathbf{x}_\mu; \mathbf{W})\right)$. A widely chosen kind of loss is the cross entropy, $\ell(y, f) = \log\left(1 + e^{-yf}\right)$. Another common choice is the hinge loss, defined as $\ell(y, f) = \frac{1}{2}\Delta(y, f)^2 \theta(\Delta(y, f)) = \frac{1}{2}\max(0, \Delta(y, f))^2$, where we have introduced the data overlap

$$\Delta(y, f) \equiv \epsilon - yf, \tag{9}$$

with $\epsilon > 0$ being a constant. In what follows we choose $\epsilon = 1/2$ without loss of generality [63]. These loss functions take such a simple form only for a binary classification task, with labels $y = \pm 1$, where $\ell(y, f) \equiv \ell(y \cdot f)$; the two loss functions are compared in Fig. 4. In the
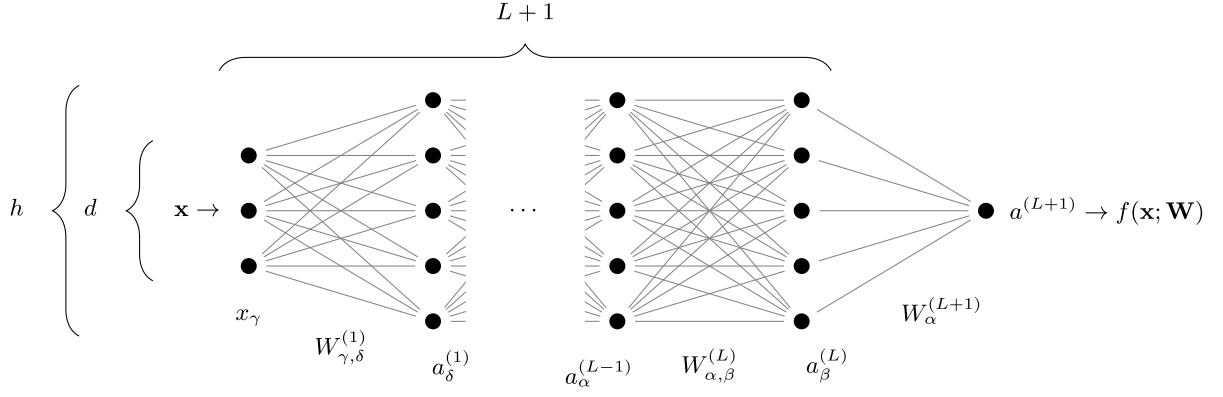
FIG. 3. Architecture of a fully-connected network with L hidden layers of constant size h. Points indicate neurons, connections between them are characterized by a weight. Biases are not represented here.

Particles	vs	Neural networks
positions of particles (N degrees of freedom) $\leftrightarrow$		parameters of the network (N degrees of freedom)
pairs of particles (ij)	$\leftrightarrow$	patterns μ
energy $\mathcal{U}$	$\leftrightarrow$	loss $\mathcal{L}$
long range interaction	$\leftrightarrow$	(for instance) cross-entropy
finite range interaction	$\leftrightarrow$	hinge loss
particle density ϕ	$\leftrightarrow$	number of data divided by the number of parameters P/N
separate two particles	$\leftrightarrow$	fit a datum
force distribution	$\leftrightarrow$	density of unsatisfied patterns $P_+(\Delta)$
gap distribution	$\leftrightarrow$	density of satisfied patterns $P_-(\Delta)$

TABLE I. Correspondence between the jargon of particle systems and that of neural networks.

hinge loss, the condition $\Delta_\mu = \Delta\left(y_\mu, f(\mathbf{x}_\mu; \mathbf{W})\right) < 0$ ensures that the datum μ is *satisfied* — that is, correctly classified by a margin ϵ. The data which do not respect this margin will be referred to as *unsatisfied* (not to be confused with *misclassified* data, for which $y_\mu f(\mathbf{x}_\mu) < 0$) — the number of such data will be denoted as N_Δ. With this definition, $\mathcal{L}$ is formally identical to $\mathcal{U}$ in Eq. (1) as already noted for the perceptron [55], and it can be written as $\mathcal{L}(\mathbf{W}) = \frac{1}{P} \sum_{\mu \in m} \frac{1}{2}\Delta_\mu^2$, where m is the set of unsatisfied patterns. The correspondence between interacting particles and neural networks is summarized in Table I.

Performance of the hinge loss and its extension to multi-class problems: This section can be skipped at a first reading. We tested in the context of image classification that the hinge loss performs as well as the cross entropy on a state-of-the-art architecture [64]: we ran the implementation [65] for CIFAR-10 and we retrained it by replacing the cross entropy by the hinge loss. To compare the two losses in a standard setting we adapted the hinge loss for multiple classes, although in what follows we only study binary classification. To predict the label of an input $\mathbf{x}_\mu$ among 10 possible labels $c = 0, \ldots, 9$, the network's last layer returns as output a list of 10 values $f_{\mu,c}$: each $f_{\mu,c}$ can be interpreted as the probability that c is the predicted label. Let $t_{\mu,c}$ be the true target labels:

for each μ, $t_{\mu,c}$ is equal to 1 if c is equal to the label of $\mathbf{x}_\mu$ and -1 otherwise. Multiclass hinge loss can then be written as

$$\mathcal{L} = \frac{1}{10P} \sum_{\mu,c} (\epsilon - t_{\mu,c} f_{\mu,c})^2 \theta(\epsilon - t_{\mu,c} f_{\mu,c}). \qquad (10)$$

We obtained an error of 3.72% by running their original code (they report on github an error of 3.68%) and 3.61%, 3.65%, 3.82% in three runs with the hinge loss.

Effective number of parameters: Following the argument developed after Eq. (1), we expect that at the transition point where the loss becomes non-zero, Eq. (2) will hold true and $N_\Delta \leq N$. (Related arguments were recently made for a quadratic loss [23]. In this case, we expect that the landscape will be related to that of floppy spring networks, whose spectra were predicted in [56]). Just as is the case for the jamming of particles, here we must pay attention to the effective number of degrees of freedom that do affect the output, $N_{\text{eff}}(\mathbf{W})$. in the space of functions going from the neighborhoods of the training set to real numbers, we consider the manifold of functions $f(\mathbf{x}; \mathbf{W})$ obtained by varying $\mathbf{W}$. We denote by $N_{\text{eff}}(\mathbf{W})$ the dimension of the tangent space of this manifold at $\mathbf{W}$. We discuss in Appendix A how $N_{\text{eff}}(\mathbf{W})$ can be measured. In general we have $N_{\text{eff}}(\mathbf{W}) \leq N$. Sev-
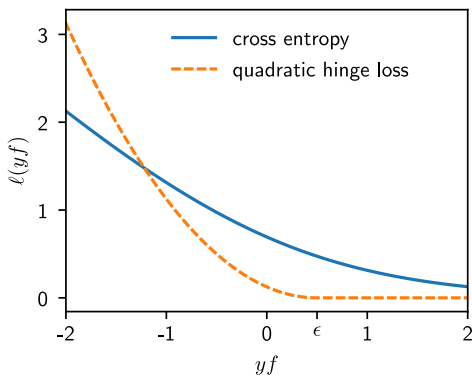
FIG. 4. Cross entropy and hinge loss functions. If the network classifies two classes with labels $y = \pm 1$ then the loss can be written as $\ell(y, f) = \ell(yf)$. The plot shows the two cases studied in this work, namely the cross-entropy and the hinge loss; for the latter, a parameter $\epsilon = \frac{1}{2}$ has been used.

eral reasons can make $N_{\text{eff}}(\mathbf{W})$ strictly smaller than N, including:

- The signal does not propagate in the network, i.e. $f(\mathbf{x}; \mathbf{W}) = C_1$ where C_1 is a constant for all $\mathbf{x}$ in the neighborhood of the training points $\mathbf{x}_\mu$. In that case, the manifold is of dimension unity and $N_{\text{eff}}(\mathbf{W}) = 1$. This situation will occur for a poor initialization of the weights as discussed in [66], or for example if all biases are too negative on the neurons of one layer for ReLU activation function. It can also occur if the data $\mathbf{x}_\mu$ are chosen in an adversarial manner for a given choice of initial weights. For example, one can choose input patterns so as to not activate the first layer of neurons (which is possible if the number of such neurons is not too large). Poor transmission will be enhanced (and adversarial choices of data will be made simpler) if the architecture presents some bottlenecks. In the situation where $N_{\text{eff}}(\mathbf{W}) = 1$, it is very simple to obtain local minima of the loss at finite loss values, even when the model has many parameters.

- The activation function is linear, then the output function is an affine function of the input, leading to $N_{\text{eff}} \leq d + 1$. Dimension-dependent bounds will also exist if the activation function is polynomial (because the output function then is also restricted to be polynomial).

- Symmetries are present in the network, e.g. the scale symmetry in ReLU networks. It will reduce one degrees of freedom per node.

- Some neurons are never active e.g. in the ReLU case, their associated weights do not contribute to N_{eff}.

Thus there are $N - N_{\text{eff}}$ directions in parameters space that do not affect the function. These directions will lead to zero modes in the Hessian at any minima of the loss. In what follows we consider stability with respect to the N_{eff} directions orthogonal to those, which thus affect the output function. Our results on the impossibility to get stuck in bad minima are expressed in terms of N_{eff}. However, as reported in Appendix A, we find empirically that for a proper initialization of the weights and rectangular fully connected networks, $N_{\text{eff}} \approx N$ (the difference is small and equal to the number of hidden neurons, and only results from the symmetry associated with each ReLU neuron). Henceforth to simplify notations we will use the symbol N to represent the number of effective parameters. In the following sections, the Hessians are computed with respect to all the N parameters.

Constraints on the stability of minima: Let us suppose (and justify later) that for a fixed number of data P, if N is sufficiently large then gradient descent with proper weights initialization leads to $\mathcal{L} = 0$, whereas if N is very small after training $\mathcal{L} > 0$. Consider that N is increased from a small value. At some value N^* the loss obtained after training will approach zero [67], i.e. $\lim_{N \to N^*} \mathcal{L} = 0$. In analogy with the behavior of packings of particles, we refer to this point as the jamming transition. At the transition the stability constraint developed in Eq. (5) above also applies if the derivative of $f(\mathbf{x}; \mathbf{W})$ is continuous, which holds true if the non-linear function ρ is smooth. Thus we have:

$$P \geq N_\Delta \geq N_- , \qquad (11)$$

since $P \geq N_\Delta$ (the number of unsatisfied patterns is obviously smaller than the total number of patterns).

We shall assume that the fraction $N_-/N \equiv C_0$ of negative eigenvalues of $\mathcal{H}_p$ does not vanish in the large N limit. In Appendix B we provide an argument supporting this result in the case of a specific non-linear function (ReLU) and random data, that yields $C_0 = 1/2$ independently of depth. It implies that unlike for spheres, but just like ellipses, $\mathcal{H}_p$ is not negative definite: we are therefore in the hypostatic scenario where one expects $N_\Delta < N^*$ at jamming, a point at which the spectrum must display a fraction of flat directions, as well as stiff ones, as described in Fig. 2H.

Moreover from this assumption and Eq.11, we obtain that stability cannot be obtained for $N \geq P/C_0$. For larger N, the dynamics cannot get stuck in a bad minimum, because in this over-parametrized regime there are not enough constraints to form them. It implies for the jamming transition that:

$$N^* \leq P/C_0. \qquad (12)$$

Notice that this bound is expected to be valid for any monotonic cost function, as for instance the cross entropy (the Hessian can always be decomposed as in Eq. (3)). However, the spectrum of the Hessian would be different [68].
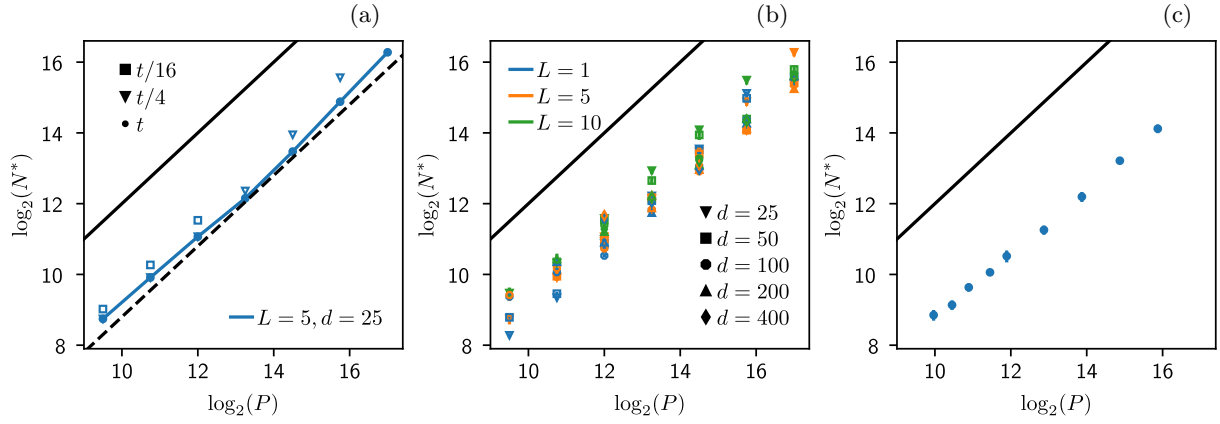
FIG. 5. Jamming transition with random data. (a) N^* vs number of data P for different learning times as indicated in legend, where $t = 10^6$ steps and a cross-entropy loss function is used. The curves at small times (orange and green) are shown as diverging to indicate the absence of the transition. The dotted black line toward which the dynamics appear to converge has slope 1, supporting $N^* \sim P$ at long times. Here $L = 5$ and $d = 25$. (b) N^* vs number of data P after $t = 10^6$ for various depths L and input dimensions d as indicated in legend, using the same loss function. The transition shows little dependency on L and d. (c) Same plot as (b) for a network with hinge loss, with $d = h$ and $t = 2 \cdot 10^6$. In the three plots (a,b,c), the black line indicates the theoretical upper bound: $P/N^* = 1/2 - N_c/N$ derived for the hinge loss.

Smooth *vs* non-smooth output function: In our numerical study below, we consider the most common choice for the non-linear function ρ, namely the rectified linear unit (ReLU): $\rho(a) = a\,\Theta(a) = \max(0, a)$. In that case, as stated above we expect for random data the spectrum of $\mathcal{H}_p$ to be symmetric (a fact that appears to also hold true for the image dataset we use, see below), thus $N_-/N = 1/2$. Yet, with the ReLU, $f(\mathbf{x}; \mathbf{W})$ is not continuous and presents cusps, so that the Hessian needs not be positive definite for stability and Eq. (11) needs to be modified. Introducing the number of directions N_c presenting cusps, stability implies $N_\Delta \geq N_- - N_c$ leading to $C_0 N^* \leq P + N_c$. Empirically we find that $N_c/N^* \in [0.21, 0.25]$ both for random data and images as reported in Appendix C, implying that:

$$N^* \leq 4P. \tag{13}$$

For comparison, below we also present results for networks with tanh activation functions. In that case the landscape is smooth and the system ends up in minima without negative eigenvalues. For such networks the spectrum of $\mathcal{H}_p$ is not exactly symmetric, and we observe $C_0 = N_-/N \approx 0.43$.

Main results: Overall, our analysis supports that

1. In the case of hinge loss there is a sharp transition for $N^* \leq P/C_0$ ($N^* < 4P$ with the ReLU), below which the loss converges to some non-zero value (under-parametrized phase) and above which it becomes null (over-parametrized phase).

2. At that point the fraction N_Δ/N of unsatisfied constraints per degree of freedom jumps to a finite value, see Fig. 6 (a,b).

3. Unlike for spheres or the perceptron, isostaticity

$N_\Delta/N = 1$ cannot be guaranteed. Instead one expects generically $N_\Delta/N < 1$ as for ellipses.

4. We are thus in the hypostatic universality class, where the scaling properties of the spectrum of the Hessian near jamming are prescribed in Fig.2.

In the next sections, we confirm these predictions in numerical experiments and observe the generalization properties at and beyond the transition point.

III. FOR RANDOM DATA THE TRANSITION OCCURS FOR $N \sim P$

We begin the numerical study of the transition between the overparametrized and underparametrized regime in the case of random data, taken to lie on the d-dimensional hyper-sphere of radius $\sqrt{d}$, $\mathbf{x}_\mu \in \mathcal{S}^d$ with random label $y_\mu = \pm 1$. The source code used to generate the simulations described in this section and the following ones is available at https://github.com/mariogeiger/nn_jamming. We proceed as follows: we build a network with a number of weights N large enough for it to be able to fit the whole dataset without errors. Next, we reduce the number of weights by decreasing the width h while keeping the depth L fixed, until the network cannot correctly classify all the data anymore within the chosen learning time. We denote this transition point N^*.

We have noticed (data not shown) that the precise location of the transition point P/N^* has a mild dependence on the dynamics (ADAM versus regular SGD, choice of batch size, learning rate schedule, etc...): the same holds true for the jamming of repulsive particles, where the choice of the dynamics affects the precise value of the critical density ϕ_c, but not the critical behaviour close to this point.

Cross-entropy loss: We first consider the cross-entropy loss — the results are qualitatively similar to those with the hinge loss. As initial condition for the dynamics we use the default initialization of `pytorch` [69]. The system then evolves according to a stochastic gradient descent (SGD) with a learning rate of 10^{-2} for $5 \cdot 10^5$ steps and 10^{-3} for $5 \cdot 10^5$ steps; the batch size is set to $\min(P/2, 1024)$; only in this case, with the cross-entropy loss, batch normalization is also used. In Fig. 5 (a) we show how N^* depends on the total learning time: the larger is the learning time the more the asymptotic relationship N^* vs P is consistent with an asymptotic linear behaviour. Note that for large P and small times, errors are always present and the transition cannot be found.

In Fig. 5 (b) we show N^* versus the number of data P after $t = 10^6$ steps for several depths L and input dimensions d (we checked that $t = 10^6$ is enough to get convergence to the conjectured asymptotic linear behaviour for all depths investigated). It is noteworthy that (i) the points always lie below the theoretical upper bound $P/N^* = 1/2 - N_c/N$, and (ii) the transition does not appear to depend on L and d. Surprisingly, this result indicates that in the present setup the ability of fully connected networks to fit random data is independent of the depth. As we shall see, we observe the same independence on depth for the image data studied below.

Hinge loss: In order to test the dependence of our results on the specific choice of the loss function, we performed the same experiment using the hinge loss. In this case we used an orthogonal initialization [70], no batch normalization and $t = 2 \cdot 10^6$ steps of ADAM [71] with batch size $= P$ and a learning rate starting at 10^{-4}, progressively divided by 10 every 250k steps. The location of the transition is shown in Fig. 5 (c): results are very similar to that of the cross-entropy loss.

Hinge _v.s._ cross-entropy loss from a conceptual perspective: As shown above, both losses appears to lead to similar performances. As shown in this section, both of them also displays a transition where all data are fitted. Yet, the nature of this transition is harder to investigate for the cross-entropy. Indeed in that case the total loss is never zero, except if the output and therefore the weights diverge. Thus in the overparametrized phase, the learning dynamic never settles, and the weights slowly drift to infinity. In practice, users stop learning at finite times (which is not needed for the hinge loss where the dynamics really stops in the overparametrized regime when the loss vanishes). Working at finite time however blurs true critical behavior near jamming, as discussed in [47].

IV. THE TRANSITION IS HYPOSTATIC

From the analysis of Section II, the number of constraints per parameter N_Δ/N is expected to jump discontinuously at the transition. To test this prediction we consider several architectures, both with $N \approx 8000$ and

$d = h$ but with different depths $L = 2$, $L = 3$ and $L = 5$. The vicinity of the transition is studied by varying P around the transition value. We used the hinge loss with the same gradient descent dynamics as described above, for a duration of 10^7 steps. Fig. 6 (a) reports the ratio N_Δ/N as a function of the ratio P/N and of the learning time, as detailed in caption. It is clear that in the range where N_Δ/N has reached a stationary value (i.e. for $P/N < 2.8$ and $P/N > 2.9$), a jump has occurred from 0 to $N_\Delta/N \approx 0.75$, a result consistent with the bound of Eq. (5) implying $N_\Delta/N \geq (N_- - N_c)/N \gtrsim 0.25$. For $P/N \in [2.8, 2.9]$, the dynamics has not yet converged and the data are somewhat scattered. This observation is presumably the signature of the usual slowing down that occurs near critical points.

Fig. 6 (b) shows the same quantity N_Δ/N, now plotted as a function of the loss $\mathcal{L}$. Strikingly, all the scatter is gone, and one observes a clear discontinuous behaviour for $\mathcal{L} \to 0$. Interestingly, this state of affairs is very similar to the jamming transition of particles, for which the noise in the data due to finite size effects is quite strong when quantities are expressed in terms of the density ϕ (analogous to P/N) but very small when quantities are expressed in terms of potential energy $\mathcal{U}$ (analogous to $\mathcal{L}$) [52].

For the sake of completeness we also show the number of misclassified data as a function of the loss in Fig. 6 (c). The number of misclassified data increases monotonically — and initially very slowly — with the loss. Indeed, close to the jamming threshold in the underparametrized phase, if $0 < \Delta_\mu < \epsilon$ the pattern μ is well classified but the corresponding gap Δ_μ is positive: unsatisfied constraints do not lead to misclassification right away.

In Fig.6 (d) we show that N_Δ/N vs the loss $\mathcal{L}$ exhibits a sharp transition also for networks with tanh activation functions.

V. SPECTRUM OF THE HESSIAN OF THE LOSS NEAR JAMMING

The Hessian is a key feature of landscapes, as it characterizes its curvature, and it is also a central aspect of the theoretical description above. In this section we systematically analyze the spectra of $\mathcal{H}$, $\mathcal{H}_0$ and $\mathcal{H}_p$. To test the predictions on the singularity of the Hessian matrix, we need to focus on the underparametrized data points near the transition. These points are contained in the black rectangle on the left side of Fig. 6 (b). The networks that we use are relatively small, but, for reference, it would be possible to compute the spectrum of the Hessian also for large networks, as discussed e.g. in [72, 73]. The setting is as above: the network uses the hinge loss and is trained with ADAM with full batches (batch size $= P$), orthogonal initialization and no batch normalization.

Relu networks: At the end of each run, we compute the hessian $\mathcal{H}$ of the loss $\mathcal{L}$, as well as the two terms $\mathcal{H}_0$ and $\mathcal{H}_p$ contributing to it, as defined in Eq. (3). Fig. 7
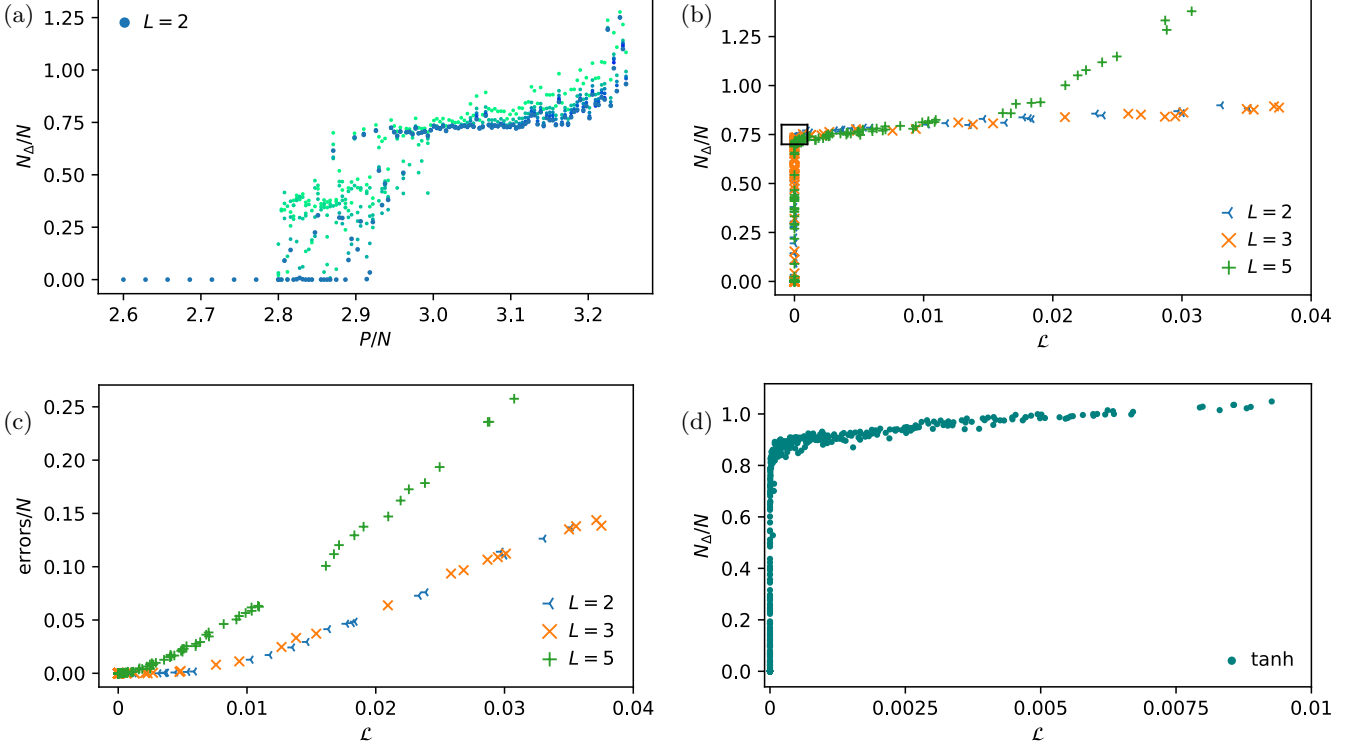
FIG. 6. Behaviour near the transition for random data. (a) Number of unsatisfied constraints N_Δ per parameter N as a function of P/N. Collections of vertical points correspond to the same run, but with different learning times from green (short time, starting at $3 \cdot 10^5$ steps) to blue (10^7 steps). The data support a discontinuous jump in this quantity at some $P/N \in [2.8, 2.9]$ at asymptotically long times. Indeed, outside that range the learning dynamics appear to have converged to zero for $r < 2.8$, and to some value > 0.7 for $P/N > 2.9$. In the interval $P/N \in [2.8, 2.9]$, data are still evolving in time. (b) N_Δ/N vs $\mathcal{L}$ follows a curve with almost no scatter for all $L = 2, 3, 5$. This is similar to the jamming transition where finite size noise is eliminated when quantities are plotted against the potential energy, rather than the packing fraction [30]. The black rectangle on the left side of the plot (small loss $\mathcal{L}$ and finite ratio N_Δ/N) marks the points in the underparametrized phase that are close to the transition. (c) Relationship between the number of misclassified data (data points with negative $y_\mu f(\mathbf{x}_\mu)$) and $\mathcal{L}$, displaying a smooth behavior. (d) N_Δ/N vs $\mathcal{L}$ for a network with $L = 3$, but with tanh activation functions rather than ReLUs.

(a) shows the positive part of the spectrum of $\mathcal{H}_p$ for different values of the loss, illustrating that the dependence on the latter is very significant. In Fig. 7 (b) we confirm that the spectrum of $\mathcal{H}_p$ collapses when the eigenvalues are re-scaled by $\mathcal{L}^{1/2}$, as expected from Section II. The key observation is that these spectra are symmetric, as argued in Appendix B. We also don't observe any accumulation of eigenvalues at $\lambda = 0$, except for the trivial zero modes stemming from the scaling symmetry of ReLU neurons (whose number is the total number of hidden neurons, much smaller than the number of weights). Fig. 7 (c) shows the spectrum of $\mathcal{H}_0$ at the end of training for runs close to the jamming transition. As expected it is semi-positive definite, with a delta peak at $\lambda = 0$ corresponding to $N - N_\Delta$ modes. It is followed by a gap and a continuous spectrum, as predicted near the jamming transition of particles if $N_\Delta < N$ [56] (which occurs for elliptic particles [40]). As the loss increases, N_Δ increases and the gap is reduced. Finally in Fig. 7 (d), the spectrum of $\mathcal{H}$ is shown. Interestingly the spectrum of

the Hessian is not positive definite, but present some unstable modes. This phenomenon stems from our choice of ReLU activation function, which leads to cusps in the landscape as quantified in the C. Such cusps can stabilize directions that would be unstable according to the Hessian.

Tanh networks: On the contrary, networks with tanh activation functions exhibit a smooth landscape, and in principle the loss is able to reach minima without any negative eigenvalues, since there are no cusps that could possibly stabilize them. Indeed, when minimizing a tanh-network with $P = 11000$ random patterns and $N = 2232$ parameters, we observe that after 10 million ADAM steps there remain only 11 negative eigenvalues (between $-5 \cdot 10^{-5}$ and $-2 \cdot 10^{-8}$), and after 100 million ADAM steps only 6 were left (between $-4 \cdot 10^{-6}$ and $-2 \cdot 10^{-8}$). For comparison, in ReLU networks the number of negative eigenvalues at the end of training is about $10\%N$.

In Fig. 7 (e-g) we show the spectrum of the matrices $\mathcal{H}_0, \mathcal{H}_p$ and $\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_p$, for a tanh-network at jam-
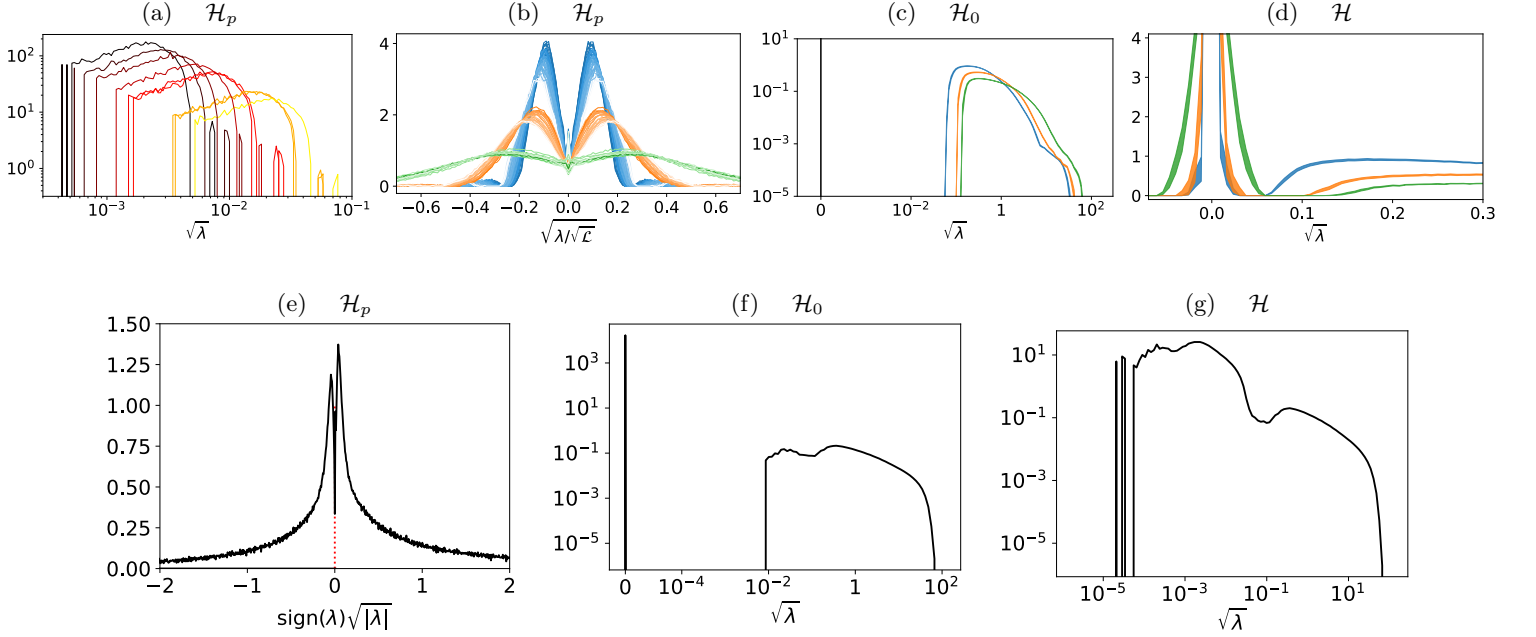
FIG. 7. The data shown in this figure concern the underparametrized points close to the transition for random data, which in Fig. 6 (b) are enclosed in a black rectangle. (a) Positive part of the spectrum of $\mathcal{H}_p$ for ten distinct runs in the underparametrized phase close to the transition. The associated loss value grows from black (low) to yellow (high). (b) These spectra collapse when plotted in terms of $\lambda/\sqrt{\mathcal{L}}$ as expected. Lighter colors correspond to higher losses. Note that they appear symmetric, in agreement with our hypothesis estimating the number of negative modes (an argument that explains this fact can be found in Appendix B). Colors are as in (d): $L = 2$ (blue), $L = 3$ (red) and $L = 5$ (green). (c) The spectrum of $\mathcal{H}_0$ contains a delta function in zero of weight $N - N_\Delta$, followed by a gap, followed by a continuous spectrum, as expected for hypostatic systems. (d) The spectrum of the total Hessian $\mathcal{H}$ has a similar shape, excepted that the delta function is blurred. Note that $\mathcal{H}$ has negative eigenvalues. These directions may in fact be stabilized by the N_c cusps of the linear rectifier, or alternatively may indicate that the learning dynamics did not converge to a local minimum yet. The thickness of each line correspond to the standard deviation. (e-g) Spectrum of the matrices $\mathcal{H}_p, \mathcal{H}_0$ and $\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_p$ for tanh networks, respectively. Notice that the spectrum of $\mathcal{H}_p$ is no longer symmetric, compared to the ReLU case: the faction of negative eigenvalues is $C_0 \approx 0.43$.

ming. The matrix $\mathcal{H}_0$ is qualitatively similar to what was observed in ReLU-networks: it presents a delta peak in 0 and a gapped bulk of positive eigenvalues. The matrix $\mathcal{H}_p$ appears quite different, since it is no longer symmetric: the number N_- of negative eigenvalues is approximately $0.43N$ — therefore $C_0 = 0.43$ instead of $C_0 = 0.5$. The total Hessian $\mathcal{H}$ is not gapped in the present case, even though it displays two peaks. In order to clearly have a gap we would have to sample points closer to jamming (with a smaller loss, since $\mathcal{H}_p$ is proportional to $\sqrt{\mathcal{L}}$ close to jamming).

Overall, as we move from the under-parametrized phase to the over-parametrized one the situation is as follows:

1. N below N^*: There are many constraints with respect to the number of variable N, $\mathcal{H}_0$ is almost full rank and can easily compensate the negative eigenvalues of $\mathcal{H}_p$. The spectrum of $\mathcal{H}_p$ is symmetric.

2. N approaching N^* from below: The rank of $\mathcal{H}_0$ decreases but it does not go below $C_0 N$ since it has to compensate the vanishingly small negative eigenvalues of $\mathcal{H}_p$.

3. As N is large enough, the dynamics finds a global

minimum at $\mathcal{L} = 0$ and $\mathcal{H}_p$ vanishes.

VI. DISTRIBUTION OF GAPS REVEALS NEW SINGULAR BEHAVIOUR

We now study the distribution of *gaps* $\Delta < 0$ and *overlaps* $\Delta > 0$, which play an important role near jamming. Positive Δ's are associated with unsatisfied patterns — which increase the loss of the system — whereas negative Δ's correspond to satisfied patterns — which are correctly classified with a margin ϵ and do not contribute to the loss. The latter offer an important measure not only at the jamming transition, but also in the over-parametrized regime, where they signal how much room is left around a minimum of the loss to fit additional patterns. In Fig. 8 (a,b) we show the two distributions for different depths $L = 2, 3, 5$ (positive Δ's have been rescaled by $\mathcal{L}^{1/2}$). Remarkably, they behave as power laws for about two decades, $P_+(\Delta/\sqrt{\mathcal{L}}) \sim (\Delta/\sqrt{\mathcal{L}})^\theta$ and $P_-(\Delta) \sim |\Delta|^{-\gamma}$, with novel exponents $\theta \approx 0.3$ and $\gamma \approx 0.2$ that appear to differ from those found for the jamming of particles (which are $\theta \approx 0.42311\ldots$ and $\gamma \approx 0.41269\ldots$). For comparison, in Fig. 8 (c,d) we show

the distribution of the same variables for tanh-networks, that also display power-law behaviors but with different exponents $\theta \approx 0.2$ and $\gamma \approx 0.16$.

In the case of spheres, the two exponents are related by an inequality that happens to be saturated [42, 57]. The inequality comes from arguments on the stability of jammed packings, and the fact that it is saturated (which can be proven for certain dynamics [43]) implies that such systems are *marginally stable*: they display an abundance of low-energy excitations and are prone to avalanche dynamics and crackling response when perturbed [43], a property associated with a hierarchical organization of the loss landscape [44, 45]. The presence of such power laws for deep networks thus suggests they are marginally stable as well, and that the learning dynamics may occur by avalanches where the unsatisfied constraints change by bursts. This will be subject of detailed studies in a future paper.

VII. IMAGE DATA: MNIST

We now consider a dataset called MNIST, which consists of a collection of black and white pictures of 28×28 pixels depicting handwritten digits from 0 to 9. The labels y_μ in principle would be the digits themselves ($y_\mu \in \{0, \ldots, 9\}$), but to compare more directly with our previous experiments we gathered all the digits into two groups (even and odd numbers) with labels $y_\mu = \pm 1$. The architecture of the network is as in the previous sections: the d inputs are fed to a cascade of L fully-connected layers with h neurons, that in the end result in a single scalar output. The loss function used is the hinge loss.

If we kept the original input size of $28 \times 28 = 784$ then the majority of the network's weights would be necessarily concentrated in the first layer (the width h cannot be too large in order to be able to compute the Hessian). To avoid this issue, we opted for a reduction of the input size. We performed a principal component analysis (PCA) on the whole dataset and we identified the 10 dimensions that carry the most variance; then we used the components of each image along these directions as a new input of dimension $d = 10$. This projection hardly diminishes the performance of the network (we find the generalization accuracy to be larger than 90% at the jamming transition in Fig. 10 for $P \geq 10^4$).

In Fig. 9 we show that a jamming transition is also found for real data with a discontinuous behavior of N_Δ/N. Fig. 9 (a) shows the number of unsatisfied patterns per parameter N_Δ/N increasing P at fixed N, and in Fig. 9 (b) the same quantity is plotted against the loss. As for random data, the latter is less noisy. In Fig. 9 (c) we show that the number of misclassified data (i.e. the number of patterns with $y_\mu f(\mathbf{x}_\mu) < 0$) grows smoothly with the loss. These plots depict the same scenario as we found for random data, namely the one presented in Fig. 6 (a-c), except for the magnitude of the density of constraints at the transition with $N_\Delta/N \approx 0.5$ rather

than $N_\Delta/N \approx 0.7$ as observed before. Hence, the number of unsatisfied patterns at the transition is not universal.

Also the spectrum of the Hessian matrix is similar to that of random data. In Fig. 9 (e-h) we show the positive part of the spectrum of $\mathcal{H}_p$, the total spectrum of $\mathcal{H}_p$, the spectrum of $\mathcal{H}_0$ and the spectrum of the total Hessian $\mathcal{H}$, respectively. As with random data: the matrix $\mathcal{H}_p$ has a symmetric spectrum and the matrix $\mathcal{H}_0$ has a finite number of zero modes and a gapped continuous distribution of modes at high energy. The spectrum of the total Hessian is again similar to that of $\mathcal{H}_0$, where the delta function in zero has been smeared.

The distribution of gaps (negative Δ's) is plotted in Fig. 9 (d), suggesting a power law with an exponent $\gamma = 0.25$ that is slightly larger than the value found for random data, $\gamma \approx 0.2$. It is unclear whether this difference is significant. We observed that the distribution of overlaps (positive Δ's) has large sample to sample variations (not shown), and the acquisition of enough statistics to measure it extensively will be done elsewhere.

A key difference between random and structured data however is the location N^* of the transition, shown in Fig. 10 versus the number P of patterns. For a fixed number P of MNIST pictures we ran several simulations with networks of different sizes, and found in this way the lowest value N^* for which all patterns could still be classified correctly. In the figure we present the results for two network architectures of different depths $L = 1, 3, 5$ (the width h was varied in order to control the network size). Key results are that (i) N^* is essentially independent of depth, especially at larger P and (ii) the minimum number of parameters N^* to fit the data is significantly smaller than for random data, a difference that seems to increase with P. The behavior of N^* in the (hypothetical) limit $P \to \infty$ could be indeed different from the linear scaling of random data: a sub-linear scaling or even a finite asymptotic value are possible alternatives. More generally, how the data structure affects the location of the transition $N^*(P)$ is an important question for the future.

VIII. CONCLUSION

By slightly changing the loss function — i.e. by considering the hinge loss rather than the commonly used cross entropy, a change that does not degrade performance — we could recast the problem of minimizing the loss function of deep networks into a constraint satisfaction problem with continuous degrees of freedom. This kind of problem has been abundantly studied in physics, in particular in the context of the jamming of particles, and some theoretical tools developed in that field readily apply to deep networks. In particular from this analogy one predicts a sharp transition as the number of parameters is reduced, separating a region where all constraints can be satisfied (that is, all the data are perfectly fitted) and the loss is zero after learning, and a region where the ratio
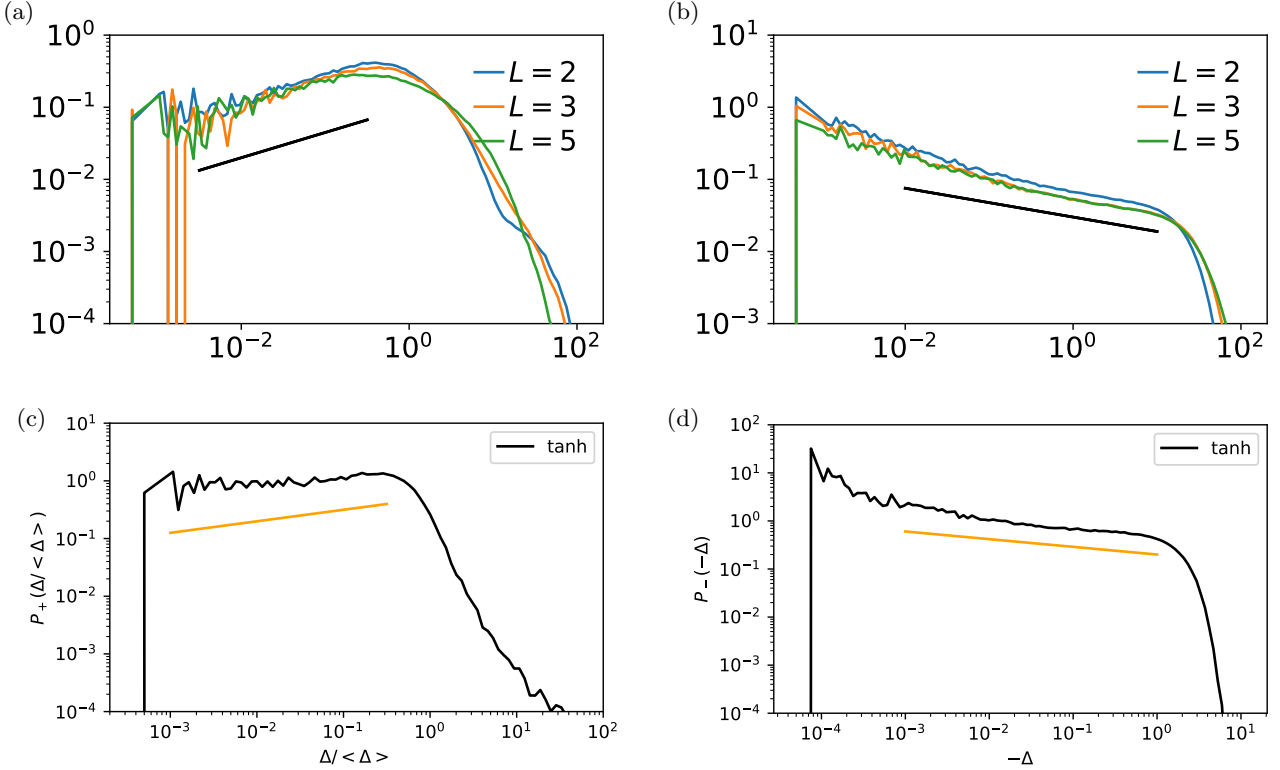
FIG. 8. (a) Distribution of re-scaled overlaps $z \equiv \Delta/\sqrt{\mathcal{L}} > 0$ near threshold, supporting that $P_+(z) \sim z^\theta$ with an exponent $\theta \approx 0.3$ that does not vary with L in the range probed. (b) The distribution of gaps $P_-(\Delta) \sim |\Delta|^{-\gamma}$ for $\Delta < 0$, with $\gamma \approx 0.2$, which again does not vary with L. (c-d) Distribution of overlaps and gaps for tanh-networks. The exponents in this case are different: $\theta \approx 0.2$, $\gamma \approx 0.16$.

of the number of unsatisfied constraints to the number of parameters is of order one. This ratio jumps discontinuously at the transition, where it attains a value smaller than one. Near that point, the spectrum of the Hessian is singular, reminiscent of a critical behavior. One key finding is that deep learning falls into the hypostatic universality class, similar to that of ellispes. We also observe a scaling behavior and new exponents characterizing how well constraints are satisfied or not (through the distributions $P_-(\Delta)$ and $P_+(\Delta)$, respectively). This bears comparison with the known behavior of packings of particles — where such singularities signal marginality and avalanche-type response — and of the perceptron (the simplest, shallow, neural network), that lie in the same universality class. Yet there is no theory so far to explain these exponents for deep networks. These results also shed light on some aspects of deep learning:

Not getting stuck in poor minima of the loss: Our analysis supports that in the overparametrized regime, the dynamics does not get stuck in poor minima because the number of constraints to satisfy (data to fit) P is too small to hamper minimization: the system is in an easy satisfiable phase. In particular assuming that a certain operator (namely the matrix $\mathcal{H}_p$) has a fraction of negative eigenvalues (which we could show in the case

of the ReLU activation function and random data, and confirm numerically) implies that no poor minima exist if $P/N < P/N^* = \mathcal{O}(1)$. Here N is the number of effective degrees of freedom of the network, which in all the cases we studied is essentially equal to the number of parameters. This argument does not rule out the possibility that, with a very poor choice of initial condition, a poor minimum of the loss can be found. This is the case in particular if the network does not propagate the signal (then $N = 1$ in our formalism, independently of the number of parameters). Presumably usual tricks used to train deep networks (batch normalization, residual links, proper weight initialization, ...) ensure that the sensitivity of the network to its parameters is preserved during training so that N is indeed similar to the number of parameters, a hypothesis that would be useful to test in a broader setting.

In the under-parametrized phase the network gets stuck at a positive loss, either because the ground state is no longer at zero loss or because the system is trapped in an excited local minimum. The fact that the jamming transition itself depends on the dynamics (as is the case for the jamming of particles) suggests that in the under-parametrized case the network is in a local minimum.
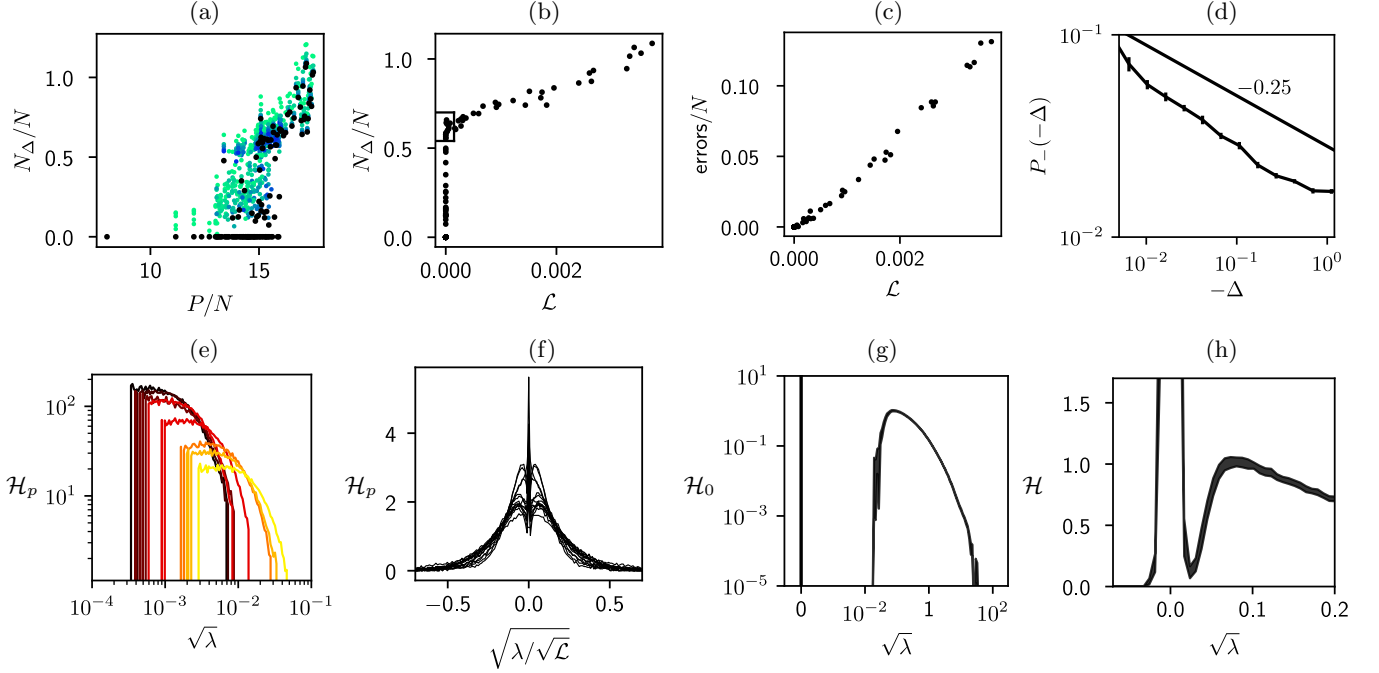
Role of depth: We observed that depth is not helpful

FIG. 9. Results with the MNIST dataset, keeping the first 10 PCA components. $d = 10$, $h = 30$ and $L = 5$ ($N = 3900$), varying $P = 1, \ldots, 70k$. (a) The number of unsatisfied patterns N_Δ/N jumps discontinuously when $r = P/N$ is increased. (b) The same quantity is less noisy when plotted against the loss. (c) The number of misclassified data is a smooth function of the loss. (d) Distribution of the negative gaps ($\Delta < 0$), with a tentative exponent $\gamma = 0.25$. In the second row (e-h), the Hessian of the runs contained in the rectangle of plot (b) are shown: (e) positive part of the spectrum of $\mathcal{H}_p$, in logarithmic scale; (f) the total spectrum of $\mathcal{H}_p$ appears to be symmetric; (g) the spectrum of $\mathcal{H}_0$ presents a delta function in zero and a gapped continuous spectrum at high frequencies; (h) the spectrum of the total Hessian $\mathcal{H}$ resembles that for random data: the delta function in the spectrum of $\mathcal{H}_0$ is smeared.

to fit random data in fully connected networks: increasing depth and reducing width so that the total number of weights is fixed does not allow to fit the data with less parameters. We have also observed that this finding continues to hold in a realistic case based on MNIST. This may seem to clash with mathematical results, such as [9–11, 13], which establish that depth enhances expressivity. However, we tackle the question of expressivity for *realistic* data and learning protocols, which is quite different. Our results, that need confirmation by further studies on a broader range of data, point toward a negative answer for fully connected networks. It may be that the added expressive power of deep networks is only useful for architectures exploiting the symmetry and hierarchy in the data (e.g. as in convolutional networks). Alternatively, depth may play a role in accelerating the learning dynamics [74].

Reference point for network architectures: key properties of deep networks, including the learning dynamics and the generalization power, are believed to be affected by the landscape geometry. We have argued that there exists a critical line $N^*(P)$ where the landscape is singular (with both flat and stiff directions), suggesting that it will be a useful reference point to study dynamics and generalization. Concerning the former, our observations suggest that learning near threshold may

occur by avalanches, that is, by abrupt changes in the set of data that are correctly classified. In practice, networks are generally trained in the overparametrized regime $N \gg N^*$. It would be interesting to investigate whether the learning dynamics, at intermediate times where many data are not fitted yet, resembles the dynamics near threshold and displays bursts of changes in the constraints. Concerning the latter, we have studied the effect of jamming on generalization since this article was first written, as appears in [46, 47].

ACKNOWLEDGMENTS

The manuscript [75], which appeared at the same time than ours, shows that the critical properties of the jamming transition found for the non-convex perceptron [34] hold more generally in some shallow networks. This universality is an intriguing result. Understanding the con-
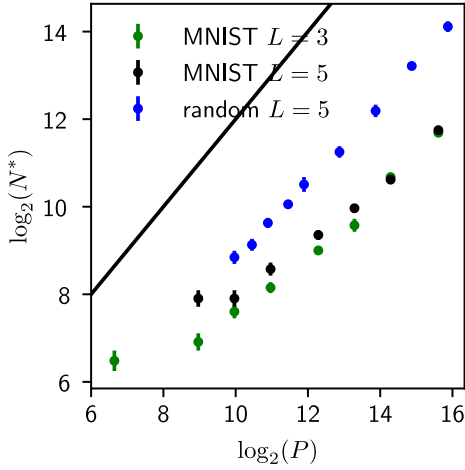
FIG. 10. Results with the MNIST dataset, keeping the first 10 PCA components (see main text), with $d = 10$ and varying P and h. The plot shows the number of parameters N^* at the jamming transition. For comparison, we also show the theoretical upper bound (solid curve) and the results found with random data (black points). The maximum number of steps is $2 \cdot 10^6$.

nection with our findings, which show instead a jamming transition similar to that of ellipses, is certainly worth future studies.

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems* (2012) pp. 1097–1105.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," Nature **521**, 436 (2015).

[3] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal processing magazine **29**, 82–97 (2012).

[4] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, *et al.*, "Mastering the game of go with deep neural networks and tree search," Nature **529**, 484 (2016).

[5] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, *et al.*, "Mastering the game of go without human knowledge," Nature **550**, 354 (2017).

[6] Yann LeCun, Yoshua Bengio, *et al.*, "Convolutional networks for images, speech, and time series," The handbook of brain theory and neural networks **3361**, 1995 (1995).

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016) pp. 770–778.

[8] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning* (2015) pp. 448–456.

[9] Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio, "On the number of linear regions of deep neural networks," in *Advances in neural information processing systems* (2014) pp. 2924–2932.

[10] Monica Bianchini and Franco Scarselli, "On the complexity of neural network classifiers: A comparison between shallow and deep architectures," IEEE transactions on neural networks and learning systems **25**, 1553–1565 (2014).

[11] Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein, "On the expressive power of deep neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 70, edited by Doina Precup and Yee Whye Teh (PMLR, International Convention Centre, Sydney, Australia, 2017) pp. 2847–2854.

[12] Ronen Eldan and Ohad Shamir, "The power of depth for feedforward neural networks," in *Conference on Learning Theory* (2016) pp. 907–940.

[13] Holden Lee, Rong Ge, Tengyu Ma, Andrej Risteski, and Sanjeev Arora, "On the ability of neural nets to express distributions," in *Proceedings of the 2017 Conference on Learning Theory*, Proceedings of Machine Learning Research, Vol. 65, edited by Satyen Kale and Ohad Shamir (PMLR, Amsterdam, Netherlands, 2017) pp. 1271–1296.

[14] Elizabeth Gardner, "The space of interactions in neural network models," Journal of physics A: Mathematical and general **21**, 257 (1988).

[15] Rémi Monasson and Riccardo Zecchina, "Weight space structure and internal representations: a direct approach to learning and generalization in multilayer neural networks," Physical review letters **75**, 2432 (1995).

[16] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, "Understanding deep learning requires rethinking generalization," International Conference on Learning Representations (2017).

[17] Eric B Baum, "On the capabilities of multilayer perceptrons," Journal of complexity **4**, 193–215 (1988).

[18] Ludovic Berthier and Giulio Biroli, "Theoretical perspective on the glass transition and amorphous materials," Reviews of Modern Physics **83**, 587 (2011).

[19] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun, "The loss surfaces of multilayer networks," in *Artificial Intelligence and Statistics* (2015) pp. 192–204.

[20] C Daniel Freeman and Joan Bruna, "Topology and geometry of deep rectified network optimization landscapes," International Conference on Learning Representations (2017).

[21] Elad Hoffer, Itay Hubara, and Daniel Soudry, "Train longer, generalize better: closing the generalization gap in large batch training of neural networks," in *Advances in Neural Information Processing Systems* (2017) pp. 1729–1739.

[22] Daniel Soudry and Yair Carmon, "No bad local minima: Data independent training error guarantees for multilayer neural networks," arXiv preprint arXiv:1605.08361 (2016).

[23] Yaim Cooper, "The loss landscape of overparameterized neural networks," arXiv preprint arXiv:1804.10200 (2018).

[24] Levent Sagun, Léon Bottou, and Yann LeCun, "Singularity of the hessian in deep learning," International Conference on Learning Representations (2017).

[25] Levent Sagun, Utku Evci, V. Uğur Güney, Yann Dauphin, and Léon Bottou, "Empirical analysis of the hessian of over-parametrized neural networks," ICLR 2018 Workshop Contribution, arXiv:1706.04454 (2017).

[26] Andrew J Ballard, Ritankar Das, Stefano Martiniani, Dhagash Mehta, Levent Sagun, Jacob D Stevenson, and David J Wales, "Energy landscapes for machine learning," Physical Chemistry Chemical Physics (2017).

[27] Zachary C Lipton, "Stuck in a what? adventures in weight space," International Conference on Learning Representations (2016).

[28] Marco Baity-Jesi, Levent Sagun, Mario Geiger, Stefano Spigler, Gerard Ben Arous, Chiara Cammarota, Yann LeCun, Matthieu Wyart, and Giulio Biroli, "Comparing dynamics: Deep neural networks versus glassy systems," in *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 80, edited by Jennifer Dy and Andreas Krause (PMLR, Stockholmsmässan, Stockholm Sweden, 2018) pp. 314–323.

[29] M. Wyart, "On the rigidity of amorphous solids," Annales de Phys **30**, 1–113 (2005).

[30] Andrea J. Liu, Sidney R. Nagel, W Saarloos, and Matthieu Wyart, *Dynamical Heterogeneities in Glasses, Colloids, and Granular Media* (2010).

[31] Florent Krzakala and Jorge Kurchan, "Landscape analysis of constraint satisfaction problems," Physical Review E **76**, 021122 (2007).

[32] Lenka Zdeborová and Florent Krzakala, "Phase transitions in the coloring of random graphs," Physical Review E **76**, 031131 (2007).

[33] Silvio Franz, Giorgio Parisi, Maxime Sevelev, Pierfrancesco Urbani, and Francesco Zamponi, "Universality of the sat-unsat (jamming) threshold in non-convex continuous constraint satisfaction problems," SciPost Physics **2**, 019 (2017).

[34] Silvio Franz and Giorgio Parisi, "The simplest model of jamming," Journal of Physics A: Mathematical and Theoretical **49**, 145001 (2016).

[35] Matthieu Wyart, Leonardo E Silbert, Sidney R Nagel, and Thomas A Witten, "Effects of compression on the vibrational modes of marginally jammed solids," Physical Review E **72**, 051306 (2005).

[36] L. E. Silbert, A. J. Liu, and S. R. Nagel, "Vibrations and diverging length scales near the unjamming transition," Phys. Rev. Lett. **95**, 098301 (2005).

[37] Aleksandar Donev, Ibrahim Cisse, David Sachs, Evan A. Variano, Frank H. Stillinger, Robert Connelly, Salvatore Torquato, and P. M. Chaikin, "Improving the density of jammed disordered packings using ellipsoids," Science **303**, 990–993 (2004).

[38] Mitch Mailman, Carl F. Schreck, Corey S. O'Hern, and Bulbul Chakraborty, "Jamming in systems composed of frictionless ellipse-shaped particles," Phys. Rev. Lett. **102**, 255501 (2009).

[39] Z. Zeravcic, N. Xu, A. J. Liu, S. R. Nagel, and W. van Saarloos, "Excitations of ellipsoid packings near jamming," Europhys. Lett. **87**, 26001 (2009).

[40] Carolina Brito, Harukuni Ikeda, Pierfrancesco Urbani, Matthieu Wyart, and Francesco Zamponi, "Universality of jamming of non-spherical particles," arXiv preprint arXiv:1807.01975 (2018).

[41] This transition influences the generalization properties of deep networks, too. This has been observed, for instance, in [46, 76], and studied by the authors in [47] (preprint).

[42] Matthieu Wyart, "Marginal stability constrains force and pair distributions at random close packing," Phys. Rev. Lett. **109**, 125502 (2012).

[43] Markus Müller and Matthieu Wyart, "Marginal stability in structural, spin, and electron glasses," Annual Review of Condensed Matter Physics **6**, 177–200 (2015).

[44] Silvio Franz and Stefano Spigler, "Mean-field avalanches in jammed spheres," Physical Review E **95**, 022139 (2017).

[45] Patrick Charbonneau, Jorge Kurchan, Giorgio Parisi, Pierfrancesco Urbani, and Francesco Zamponi, "Fractal free energy landscapes in structural glasses," Nature Communications **5** (2014).

[46] Stefano Spigler, Mario Geiger, Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Matthieu Wyart, "A jamming transition from under-to over-parametrization affects loss landscape and generalization," NIPS Workshop "Integration of Deep Learning Theories" contribution, arXiv preprint arXiv:1810.09665 (2018).

[47] Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart, "Scaling description of generalization with number of parameters in deep learning," arXiv preprint arXiv:1901.01608 (2019).

[48] A.C. Anderson, *Amorphous Solids: Low Temperature Properties*, edited by W. A. Phillips, Topics in Current Physics, Vol. 24 (Springer, Berlin, 1981).

[49] Alexei V. Tkachenko and Thomas A. Witten, "Stress propagation through frictionless granular material," Phys. Rev. E **60**, 687–696 (1999).

[50] A similar decomposition has been used in [25, 77, 78].

[51] Once again, this statement is true except for global translation or rotation of the systems, whose number however is fixed in the large N limit and disappears when the ratio N_Δ/N is considered.

[52] Corey S. O'Hern, Leonardo E. Silbert, Andrea J. Liu, and Sidney R. Nagel, "Jamming at zero temperature and zero applied stress: The epitome of disorder," Phys. Rev. E **68**, 011306–011324 (2003).

[53] Eric DeGiuli, Adrien Laversanne-Finot, Gustavo Alberto Düring, Edan Lerner, and Matthieu Wyart, "Effects of coordination and pressure on sound attenuation, boson peak and elasticity in amorphous solids," Soft Matter **10**, 5628–5644 (2014).

[54] Le Yan, Eric DeGiuli, and Matthieu Wyart, "On variational arguments for vibrational modes near jamming," EPL (Europhysics Letters) **114**, 26003 (2016).

[55] Silvio Franz, Giorgio Parisi, Pierfrancesco Urbani, and Francesco Zamponi, "Universal spectrum of normal modes in low-temperature glasses," Proceedings of the National Academy of Sciences **112**, 14539–14544 (2015).

[56] Gustavo Düring, Edan Lerner, and Matthieu Wyart, "Phonon gap and localization lengths in floppy materials," Soft Matter **9**, 146–154 (2013).

[57] Edan Lerner, Gustavo During, and Matthieu Wyart, "Low-energy non-linear excitations in sphere packings," Soft Matter **9**, 8252–8263 (2013).

[58] E. Lerner, G. Düring, and M. Wyart, "Toward a microscopic description of flow near the jamming threshold," EPL (Europhysics Letters) **99**, 58003 (2012).

[59] Patrick Charbonneau, Eric I. Corwin, Giorgio Parisi, and Francesco Zamponi, "Universal microstructure and mechanical stability of jammed packings," Physical Review Letters **109**, 205501– (2012).

[60] Patrick Charbonneau, Jorge Kurchan, Giorgio Parisi, Pierfrancesco Urbani, and Francesco Zamponi, "Exact theory of dense amorphous hard spheres in high dimension. iii. the full replica symmetry breaking solution," Journal of Statistical Mechanics: Theory and Experiment **2014**, 10009 (2014).

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

[62] Patrick Charbonneau, Eric I Corwin, Giorgio Parisi, and Francesco Zamponi, "Jamming criticality revealed by removing localized buckling excitations," Physical Review Letters **114**, 125504 (2015).

[63] The parameter ϵ fixes the scale of the loss, in the sense that, if one rescales both ϵ and the weights of the last layer by the same factor α, then all the observables are equivariant with respect to this transformation ($\Delta \to \alpha\Delta$, $\mathcal{L} \to \alpha^2\mathcal{L}$, $N_\Delta \to N_\Delta$, ...). Consequently, since any positive value of ϵ leads to the same behavior of the system, we have arbitrarily fixed its value to $\epsilon = \frac{1}{2}$. If ϵ were 0, the network would try to enforce $f(\mathbf{x}; \mathbf{W}) \equiv 0$ regardless of the specific pattern.

[64] Xavier Gastaldi, "Shake-shake regularization of 3-branch residual networks," International Conference on Learning Representations (2017).

[65] https://github.com/mariogeiger/pytorch_shake_shake.

[66] Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein, "Deep information propagation," arXiv preprint arXiv:1611.01232 (2016).

[67] For finite P, N^* will present fluctuations induced by differences of initial conditions. The fluctuations of P/N^* are however expected to vanish in the limit where P and N^* become large. This phenomenon is well-known for the jamming of particles, and is referred to as finite size effects.

[68] For the cross entropy, when the data become separable true minima exists only for diverging weights, a complication that does not occur with the hinge loss.

[69] Weights and biases are initialized with a uniform distribution on $[-\sigma, \sigma]$, where $\sigma^2 = 1/f_{in}$ and f_{in} is the number of incoming connections.

[70] Andrew M Saxe, James L McClelland, and Surya Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," International Conference on Learning Representations (2014).

[71] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," International Conference on Learning Representations (2015).

[72] Ryan P Adams, Jeffrey Pennington, Matthew J Johnson, Jamie Smith, Yaniv Ovadia, Brian Patton, and James Saunderson, "Estimating the spectral density of large implicit matrices," arXiv preprint arXiv:1802.03451 (2018).

[73] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao, "An investigation into neural net optimization via hessian eigenvalue density," arXiv preprint arXiv:1901.10159 (2019).

[74] Ravid Shwartz-Ziv and Naftali Tishby, "Opening the black box of deep neural networks via information," arXiv preprint arXiv:1703.00810 (2017).

[75] P. Urbani S. Franz, S. Hwang, "Jamming in multilayer supervised learning models," arXiv preprint arXiv:1809.09945 (2018).

[76] Madhu S Advani and Andrew M Saxe, "High-dimensional dynamics of generalization error in neural networks," arXiv preprint arXiv:1710.03667 (2017).

[77] Jeffrey Pennington and Yasaman Bahri, "Geometry of neural network loss surfaces via random matrix theory," in *International Conference on Machine Learning* (2017) pp. 2798–2806.

[78] James Martens, "New insights and perspectives on the natural gradient method," arXiv preprint arXiv:1412.1193 (2014).

Appendix A: Effective number of degrees of freedom

Due to several effects discussed in the main text, the function $f(\mathbf{x}; \mathbf{W})$ can effectively depend on less variables that the number of parameters, and thus reduce the dimension of the space spanned by the gradients $\nabla_\mathbf{W} f(\mathbf{x}; \mathbf{W})$ that enters in the theory. For instance, there could be symmetries that reduce the number of effective degrees of freedom (e.g. each ReLU activation function has one of such symmetries, since one can rescale inputs and outputs in such a way that the post-activation is left invariant); another reason could be that a neuron might never activate for all the training data, thus effectively reducing the number of neurons in the network; furthermore, we expect that the network's true dimension would also be reduced if its architecture presents some bottlenecks, is poorly designed or poorly initialized. For

example if all biases are too negative on the neurons of one layer in the Relu case, the network does not transmit any signals, leading to $N = 1$ and to the possible absence of unstable directions even if the number of parameters is very large.

It is tempting to define the effective dimension by considering the dimension of the space spanned by $\nabla_{\mathbf{W}} f(\mathbf{x}_\mu; \mathbf{W})$ as μ varies. This definition is not practical for small number of samples P however, because this dimension would be bounded by P. We can overcome such a problem by considering a neighborhood of each point $\mathbf{x}_\mu$, where the network's function and its gradient can be expanded in the pattern space:

$$f(\mathbf{x}) \approx f(\mathbf{x}_\mu) + (\mathbf{x} - \mathbf{x}_\mu) \cdot \nabla_{\mathbf{x}} f(\mathbf{x}_\mu), \quad \text{(A1)}$$

$$\nabla_{\mathbf{W}} f(\mathbf{x}) \approx \nabla_{\mathbf{W}} f(\mathbf{x}_\mu) + (\mathbf{x} - \mathbf{x}_\mu) \cdot \nabla_{\mathbf{x}} \nabla_{\mathbf{W}} f(\mathbf{x}_\mu). \quad \text{(A2)}$$

Varying the pattern μ and the point $\mathbf{x}$ in the neighborhood of $\mathbf{x}_\mu$, we can build a family M of vectors:

$$M = \{ \nabla_{\mathbf{W}} f(\mathbf{x}_\mu) + (\mathbf{x} - \mathbf{x}_\mu) \cdot \nabla_{\mathbf{x}} \nabla_{\mathbf{W}} f(\mathbf{x}_\mu) \}_{\mu, \mathbf{x}}. \quad \text{(A3)}$$

We then define the effective dimension N as the dimension of M. Because of the linear structure of M, it is sufficient to consider, for each μ, only $d + 1$ values for x, e.g. $x - x_\mu = 0, \hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_d$, where $\hat{\mathbf{e}}_n$ is the unit vector along the direction n. The effective dimension is therefore

$$N = \mathrm{rk}(G), \quad \text{(A4)}$$

where the elements of the matrix G are defined as

$$G_{i,\alpha} \equiv \partial_{W_i} f(\mathbf{x}_\mu) + \hat{\mathbf{e}}_n \cdot \nabla_{\hat{\mathbf{e}}_n} \partial_{W_i} f(\mathbf{x}_\mu), \quad \text{(A5)}$$

with $\alpha \equiv (\mu, n)$. The index n ranges from 0 to d, and $\hat{\mathbf{e}}_0 \equiv 0$.

In Fig. 11 we show the effective number of parameters N versus the total number of parameters $\tilde{N}$, in the case of a network with $L = 3$ layers trained on the first 10 PCA components of the MNIST dataset. There is no noticeable difference between the two quantities: the only reduction is due to the symmetries induced by the ReLU functions (there is one such symmetry per neuron. Indeed the ReLU function $\rho(z) = z\Theta(z)$ satisfies $\Lambda \rho(z/\Lambda) \equiv \rho(z)$.) We observed the same results for random data.

Appendix B: $\mathrm{sp}(H_p)$ is symmetric for ReLu activation function and random data

We consider $\mathcal{H}_p = -\sum_\mu y_\mu \rho(\Delta_\mu) \hat{\mathcal{H}}_\mu$, where $\hat{\mathcal{H}}_\mu$ is the Hessian of the network function $f(\mathbf{x}_\mu; \mathbf{W})$ and ρ is the Relu function. We want to argue that the spectrum of $\mathcal{H}_p$ is symmetric in the limit of large N.
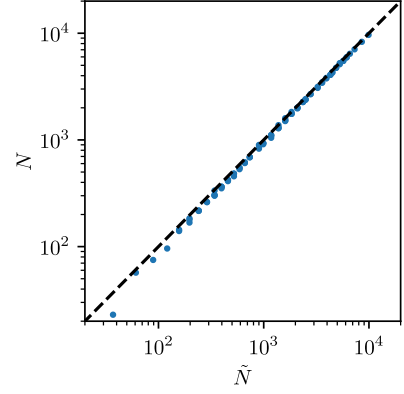

FIG. 11. Results with the MNIST dataset, keeping the first 10 PCA components. $d = 10$ and $L = 3$, varying P and h. Effective N vs total number of parameters $\tilde{N}$. N is always smaller than $\tilde{N}$ because there is a symmetry per each ReLU-neuron in the network.

We do two main hypothesis: First, the trace of any finite power of $\mathcal{H}_p$ is self-averaging (concentrates) with respect to the average over the random data:

$$\frac{1}{N} \mathrm{tr}(\hat{\mathcal{H}}_p{}^n) = \frac{1}{N} \overline{\mathrm{tr}(\hat{\mathcal{H}}_p{}^n)}.$$

Second,

$$\frac{1}{N} \sum_{\mu_1, \cdots, \mu_n} \overline{y_{\mu_1} \rho(\Delta_{\mu_1}) \cdots y_{\mu_n} \rho(\Delta_{\mu_n}) \mathrm{tr}(\hat{\mathcal{H}}_{\mu_1} \cdots \hat{\mathcal{H}}_{\mu_n})} =$$

$$\frac{1}{N} \sum_{\mu_1, \cdots, \mu_n} \overline{y_{\mu_1} \rho(\Delta_{\mu_1}) \cdots y_{\mu_n} \rho(\Delta_{\mu_n})} \, \overline{\mathrm{tr}(\hat{\mathcal{H}}_{\mu_1} \cdots \hat{\mathcal{H}}_{\mu_n})}$$

The first hypothesis is natural since $\hat{\mathcal{H}}_p$ is a very large random matrix, for which the density of eigenvalues is expected to become a non-fluctuating quantity. The second hypothesis is more tricky: it is natural to assume that the trace concentrates, however one also need to show that the sub-leading corrections to the self-averaging of the trace can be neglected.

Using these two hypothesis and the result, showed below, that

$$\overline{\mathrm{tr}(\hat{\mathcal{H}}_{\mu_1} \cdots \hat{\mathcal{H}}_{\mu_n})} = 0 \quad \text{(B1)}$$

for all n odds, one can conclude that all odds traces of $\hat{\mathcal{H}}_p$ are zero. This implies that the spectrum of $\hat{\mathcal{H}}_p$ is symmetric, more precisely that the fractions of negative and positive eigenvalues are equal.

In order to show that the statement (B1) above holds, let us argue first that $\overline{\mathrm{tr}(\hat{\mathcal{H}}_\mu^n)} = 0$ for any odd n.

$$\mathrm{tr}(\hat{\mathcal{H}}_\mu^n) = \sum_{i_1, i_2, \ldots, i_n} \hat{\mathcal{H}}_{i_1, i_2}^\mu \hat{\mathcal{H}}_{i_2, i_3}^\mu \cdots \hat{\mathcal{H}}_{i_n, i_1}^\mu, \quad \text{(B2)}$$
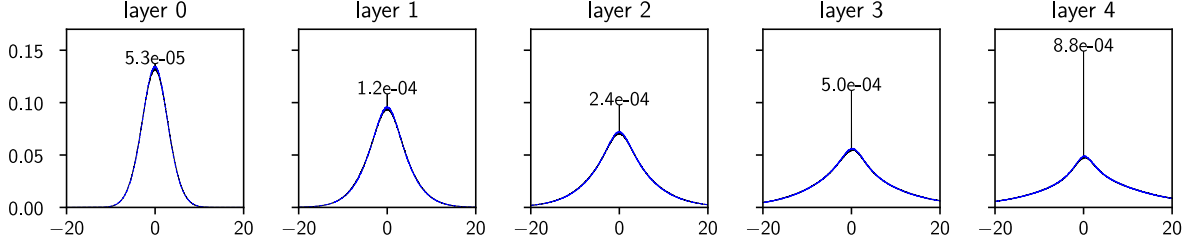
FIG. 12. Density of the pre-activations for each layers with $L = 5$ and random data, averaged over all the runs just above the jamming transition with that architecture. Black: distribution obtained over the training set. Blue: previously unseen random data (the two curves are on top of each other except for the delta in zero). The values indicate the mass of the peak in zero, which is only present when the training set is considered.

where the indices $i_1, \ldots, i_n$ stand for synapses connecting a pair of neurons (i.e. each index is associated with a synaptic weight $W_{\alpha,\beta}^{(j)}$: we are not writing all the explicit indexes for the sake of clarity). The term of the hessian obtained when differentiating with respect to weights $W_{\alpha,\beta}^{(j)}$ and $W_{\gamma,\delta}^{(k)}$ reads

$$
\hat{\mathcal{H}}_{\alpha\beta;\gamma\delta}^{\mu;(jk)} = \sum_{\pi_0, \ldots, \pi_L} \theta(a_{L,\pi_L}^\mu) \cdots \theta(a_{1,\pi_1}^\mu) x_{\pi_0}^\mu \cdot
$$
$$
\cdot \, \partial_{W_{\alpha,\beta}^{(j)}} \partial_{W_{\gamma,\delta}^{(k)}} \left[W_{\pi_L}^{(L+1)} W_{\pi_L,\pi_{L-1}}^{(L)} \cdots W_{\pi_1 \pi_0}^{(1)} \right]. \quad \text{(B3)}
$$

where we denoted with a the inputs in the nodes of the network. Our argument is based on a symmetry of the problem with random data: changing the sign of the weight of the last layer $W^{(L+1)} \longrightarrow -W^{(L+1)}$ and changing the labels $y_\mu \longrightarrow -y_\mu$ leaves the loss unchanged. We will show that this symmetry implies that $\text{tr}(\hat{\mathcal{H}}_\mu^n)$ averaged over the random labels is zero for odd n.

In fact, note that the sum in Eq.B3 contains a weight per each layer in the network, with the exception of the two layers j, k with respect to which we are deriving. This implies that any element of the hessian matrix where we have not differentiated with respect to the last layer ($j, k < L+1$) is an odd function of the last layer $W^{(L+1)}$, meaning that if $W^{(L+1)} \longrightarrow -W^{(L+1)}$, then the sign of all these Hessian elements is inverted as well.

If in the argument of the sum in Eq. (B2) there is no index belonging to the last layer, then the whole term changes sign under the transformation $W^{(L+1)} \longrightarrow -W^{(L+1)}$. Suppose now that, on the contrary, there are m terms with one index belonging to the last layer (we need not consider the case of two indices both belonging to the last layer because the corresponding term in the Hessian would be 0, as one can see in Eq. (B3)). For each index equal to $L + 1$ (the last layer), there are exactly two terms: $\hat{\mathcal{H}}_{j,L+1}^\mu \hat{\mathcal{H}}_{L+1,k}^\mu$ (for some indexes j, k). Since j, k cannot be $L + 1$ too, this implies that the number m of terms with an index belonging to the last layer is always even. Consequently, when the sign of $W^{(L+1)}$ is reversed, the argument of the sum in Eq. (B2) is multiplied by $(-1)^{n-m}$ (once for each term *without* an index

belonging to the last layer), which is equal to -1 if n is odd. The same symmetry can be used to show that a matrix made of an odd product of matrices $\hat{\mathcal{H}}_\mu$, such as $\hat{\mathcal{H}}_\mu \hat{\mathcal{H}}_{\mu'} \hat{\mathcal{H}}_{\mu''}$, must also have a symmetric spectrum, concluding our argument.

Appendix C: Density of pre-activations for ReLU activation functions

The densities of pre-activation (i.e. the value of the neurons before applying the activation function) is shown in Fig. 12 for random data. It contains a delta distribution in zero. The number N_c of pre-activations equal to zero when feeding a network $L = 5$ all its random dataset is $N_c \approx 0.21N$, corresponding to the number of directions in phase space where cusps are present in the loss function. For MNIST data we find $N_c \approx 0.19N$. By taking $L = 2$ and random data we find $N_c \approx 0.25N$. In these directions, stability can be achieved even if the hessian would indicate an instability. For this reason, instead of N_- in Equation 11 one should use $N/2 - N_c \approx 0.25N$.