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: October 4, 2018)

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, and puts forward the surprising result that the ability of fully connected networks to fit random data is independent of their depth. Our observations suggests that this independence also holds 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, $P_+(\Delta) \sim \Delta^\theta$ for $\Delta > 0$ and $P_-(\Delta) \sim (-\Delta)^{-\gamma}$ for $\Delta < 0$, with $\theta \approx 0.3$ and $\gamma \approx 0.2$. 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? 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] and phase transitions taking place in certain classes of computational optimization and learning problems [30–32], in particular the perceptron [32, 33] — 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 [29]. As illustrated in Fig. 1, 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 $\mathcal{U}$ [34, 35], that at the transition displays many (almost) flat directions. Particles of different shapes, such as spheres and ellipses, can lead to different jamming scenarios [36–39]. 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. 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 $r = P/N$ increases toward some critical value r_c. For the hinge loss, using results from the jamming literature we argue that $r_c \geq C_0$ where C_0 is a constant independent of depth. 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 over-parametrized 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 [32, 33], 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, 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 [40–42], indicative of a hierarchical organization of the landscape [43]. 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 will be investigated in a future publication.

II. ANALOGY BETWEEN JAMMING AND DEEP LEARNING

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 [44]. It was realized that for purely repulsive, finite-range particles, such properties are singular near the jamming transition where the system becomes a solid [34, 35], 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 [45]. 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 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, and will be investigated in another study.

We start by recalling some results on the jamming transition. Consider spherical particles 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:

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

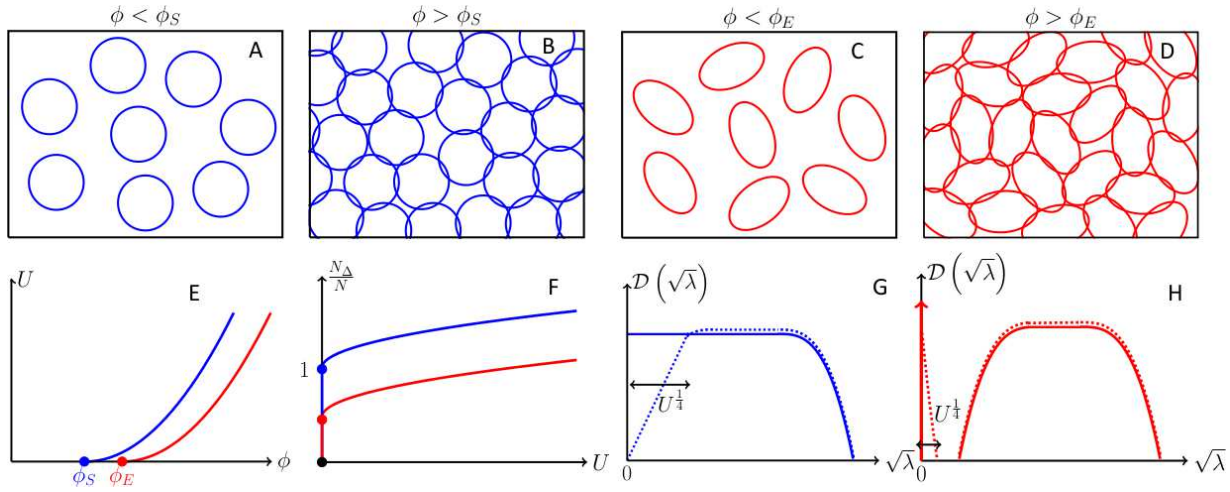

FIG. 1: Sketch of the jamming transition for repulsive spheres (blue) and ellipses (red). (A,B,C,D) Both systems transition from a fluid to a solid as the density passes some threshold, noted ϕ_S for spheres and ϕ_E for ellipses. (E) For denser packings, the potential energy $\mathcal{U}$ becomes finite. (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).

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. 1, implying that $\Delta_\mu \to 0\ \forall \mu \in m$. As argued in [46], 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 [34]. The Hessian is an $N \times N$ matrix which can be written as

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

where $\mathcal{H}_0$ and $\mathcal{H}_p$ correspond to the first and second sum, respectively. $\mathcal{H}_0$ is positive semi-definite. 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* $\mathcal{H}_p$ 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 [34] that any non-zero vector $|u\rangle$ belonging to the kernel of $\mathcal{H}_0$ must satisfy $\langle u|\mathcal{H}_U|u\rangle = \langle u|\mathcal{H}_p|u\rangle < 0$ [64]. 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 [47]. This property is often called *isostaticity*: when it holds, mean-field arguments [48–50] predict that the density of vibrational modes $D(\sqrt{\lambda})$ displays a plateau up to vanishingly small λ, as observed numerically [34, 35] and sketched in Fig. 1G.

However, for *ellipses* [36] (and as we shall see, for deep networks), this argument breaks down because $\mathcal{H}_p$ is not negative definite, and stability and jamming can occur at $N_\Delta/N < 1$: such a situation is referred to as *hypostatic*.

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 $r = P/N$
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.

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 [39, 51] predict that at larger λ, the spectrum presents a gap before becoming continuous again, as sketched in Fig. 1H. 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. 1H.

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

$$N_\Delta \geq N_-. \tag{4}$$

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 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 $\gamma \geq (1-\theta)/2$ [40, 52], as numerically confirmed in [53, 54]. For a broad class of dynamics, this bound must be saturated [41], 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 [43, 55] showed that marginal stability is associated with a hierarchical organization of minima of the energy (a phenomenon referred to as *replica symmetry breaking* [56]), and exponents were found to follow $\gamma = 0.41269\ldots$ and $\theta = 0.42311\ldots$ which appear accurate even in finite dimensions [52, 57].

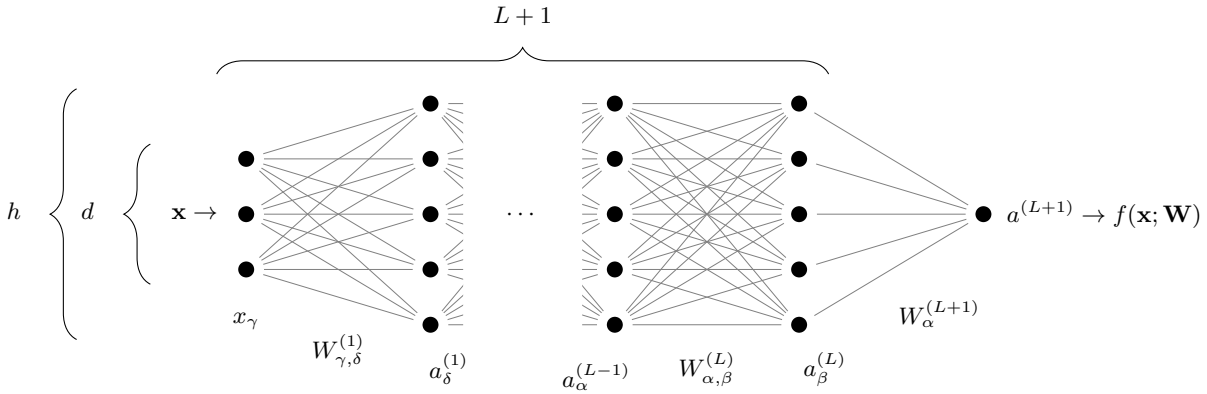

FIG. 2: Architecture of a fully-connected network with L hidden layers of constant size h. Points indicate neurons, connections between them are characterized by a weight. Biases are not represented here.

Deep Learning: 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. 2. In this scheme, each neuron sums the activity of all the neurons in the previous layer with some weights, sketched as connections in Fig. 2 (each connection thus corresponds to one parameter $W^{(i)}_{\alpha,\beta}$). Next, a bias $B^{(i)}_\alpha$ is added to this sum (one additional parameter per neuron) to obtain the so-called pre-activation ($a^{(i)}_\alpha$ in the picture and in the equations). The neuron activity is then a non-linear function ρ of that pre-activation. The computation is done iteratively from the first layer (close to the input $\mathbf{x}$) to the last one (the output $f(\mathbf{x};\mathbf{W})$):

$$f(\mathbf{x};\mathbf{W}) \equiv a^{(L+1)}, \tag{5}$$

$$a^{(i)}_\beta = \sum_\alpha W^{(i)}_{\alpha,\beta}\,\rho\left(a^{(i-1)}_\alpha\right) - B^{(i)}_\beta, \tag{6}$$

$$a^{(1)}_\beta = \sum_\alpha W^{(1)}_{\alpha,\beta}\,x_\alpha - B^{(1)}_\beta. \tag{7}$$

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

$$\Delta(y,f) \equiv \epsilon - yf, \tag{8}$$

with $\epsilon > 0$ being a constant. In what follows we choose $\epsilon = 1/2$ without loss of generality [64]. 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 [50], 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. We tested [65] in the context of image classification that the hinge loss performs as well as the cross entropy on a state-of-the-art architecture [58]. 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. 3.

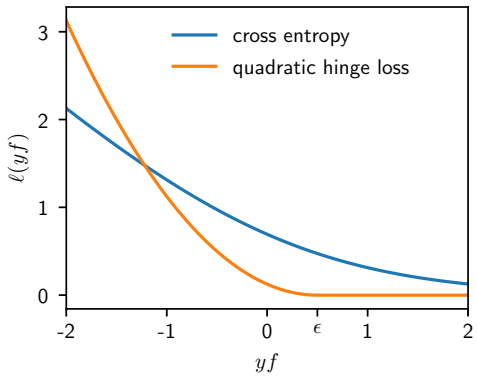

FIG. 3: Cross entropy and hinge loss functions. If the network classifies two classes with labels $y = \pm 1$ then the loss can be written as $\ell(y,f) = \ell(yf)$. The plot shows the two cases studied in this work, namely the cross-entropy and the hinge loss; for the latter, a parameter $\epsilon = \frac{1}{2}$ has been used.

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 [51]). Just as is the case for the jamming of particles, here and in what follows N needs to be defined as the effective number of degrees of freedom that do affect the output (which turns out to be essentially the number of parameters in our empirical studies below). As discussed in Appendix A, it can be defined as the dimension of the vector space spanned by the gradients $\nabla_{\mathbf{W}} f(\mathbf{x};\mathbf{W})$ as $\mathbf{x}$ takes values in the neighborhood of the $\mathbf{x}_\mu$. In practice,

several effects could make N smaller than the number of parameters $\tilde{N}$, including: (i) hidden neurons that are always inactive (for any input pattern), which are equivalent to rattlers in the jamming of particles — the weights connecting them should not be counted as degrees of freedom; (ii) internal transformations that leave the output invariant — for the network we use below, multiplying the incoming weights and bias of a given neuron by a factor and dividing the outgoing ones by the same factor does not change the output: this removes effectively one degree of freedom per hidden neuron; (iii) for deep *linear* networks — i.e. without a non-linear activation function ρ — the network boils down to a product of the weight matrices of the different layers, and its effective dimension cannot exceed the dimension d of such resulting matrix. (iv) If the network does not transmit the signal; i.e. if $f(\mathbf{x};\mathbf{W})$ is independent of $\mathbf{x}$, then $N = 1$. In our empirical study below, we find that for proper initial conditions only the effect (ii) is present, and we take it into account when defining N (this effect is anyway a very small portion of the total number of parameters in wide networks). In other words, the network remains sensitive to all its weights during learning.

The stability constraint developed in Eq. (4) 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. It implies that at jamming the number of parameters N^* satisfies

$$r_c \equiv \frac{P}{N^*} \geq \frac{N_-}{N^*}\,, \tag{9}$$

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 (see Appendix B for an argument supporting this result in the case of a specific non-linear function and random data, where we find $C_0 = 1/2$ independently of depth). Eq. (9) implies that (i) unlike for spheres, but just like ellipses, $\mathcal{H}_p$ is not negative definite: we are 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. 1H. (ii) The dynamics cannot get stuck in a bad minimum if $N > P/C_0$, because in this over-parametrized regime there are not enough constraints to form a minimum.

In our numerical study below, we follow 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, it can be argued (see Appendix B) that for random data the spectrum of $\mathcal{H}_p$ is symmetric (a fact that appears to also hold true for the image dataset we use, see below), thus $N_- = N/2$. Yet, with the ReLU, $f(\mathbf{x};\mathbf{W})$ is not continuous and presents cusps, so that Eq. (9) needs to be modified. Introducing the number of directions N_c presenting cusps, stability implies $N_\Delta > N_- - N_c$ and

$$r_c \geq 1/2 - N_c/N^*\,. \tag{10}$$

Empirically we find that $N_c/N^* \in [0.21, 0.25]$ both for random data and images as reported in Appendix C, implying $r_c \geq \frac{1}{4}$.

Overall, our analysis supports that (i) in the case of hinge loss there is a sharp transition for $N^* \leq C_0 P$, below which the loss converges to some non-zero value and above which it becomes null; (ii) at that point the fraction of constraints per degree of freedom N_Δ/N jumps to a finite value; (iii) the spectrum of the Hessian must present a delta function, a gap and a continuous spectrum at the transition. In the two next sections we confirm these predictions.

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 first consider the cross-entropy loss. As initial condition for the dynamics we use the default initialization of `pytorch` [66]. 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 $\min(P/2, 1024)$, batch normalization is also used. In Fig. 4A 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. 4B 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 $P/N^* = 1/2 - N_c/N$,

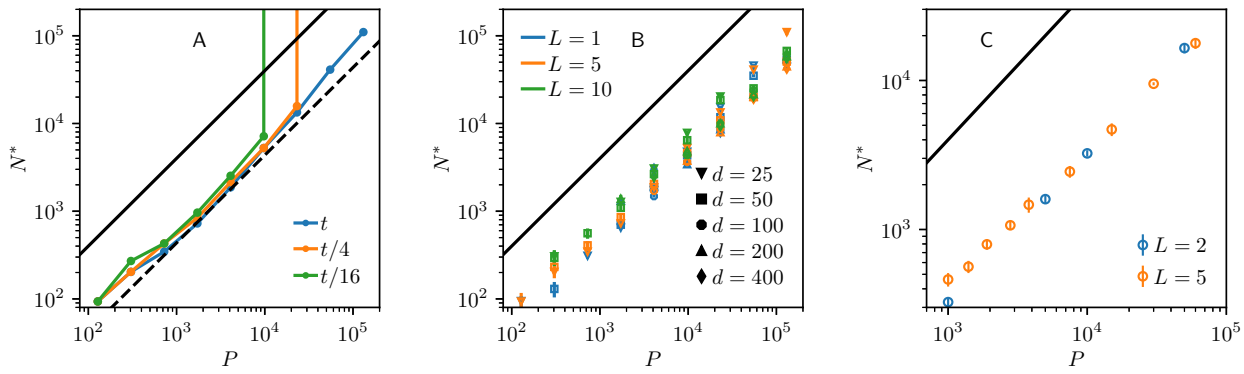

FIG. 4: Jamming transition with random data. (A) N^* vs number of data P for different learning times as indicated in legend, where $t = 10^6$ steps and a cross-entropy loss function is used. The curves at small times (orange and green) are shown as diverging to indicate the absence of the transition. 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.

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.

We have noticed (data not shown) that the precise location of the transition point $r_c = P/N^*$ 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.

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 [59], no batch normalization and $t = 2 \cdot 10^6$ steps of ADAM [60] 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. 4C: results are very similar to that of the cross-entropy loss.

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 two 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. 5A reports the ratio N_Δ/N as a function of $r = P/N$ 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 $r < 2.8$ and $r > 2.9$), a jump has occurred from 0 to $N_\Delta/N \approx 0.75$, a result consistent with the bound of Eq. (4) implying $N_\Delta/N \geq (N_- - N_c)/N \gtrapprox 0.25$. For $r \in [2.8, 2.9]$, 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. 5B 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 $r = P/N$) but very small when quantities are expressed in terms of potential energy $\mathcal{U}$ (analogous to $\mathcal{L}$) [47].

For the sake of completeness we also show the number of misclassified data as a function of the loss in Fig. 5C. 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.

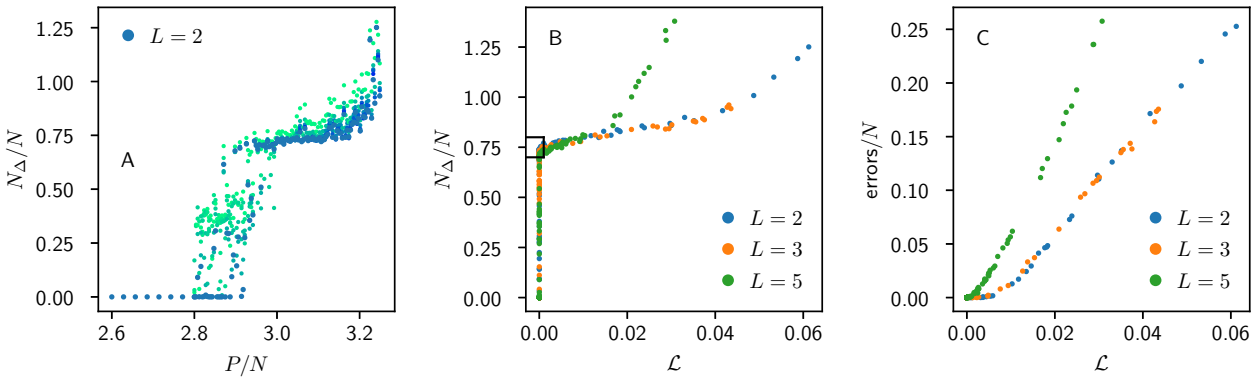

FIG. 5: Behaviour near the transition for random data. (A) Number of unsatisfied constraints N_Δ per parameter N as a function of $r = P/N$. Collections of vertical points correspond to the same run, but with different learning times from green (short time, starting at $3 \cdot 10^5$ steps) to blue (10^7 steps). The data support a discontinuous jump in this quantity at some $r_c \in [2.8, 2.9]$ 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 $r > 2.9$. In the interval $r \in [2.8, 2.9]$, 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 [29]. 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.

V. SPECTRUM OF THE HESSIAN OF THE LOSS NEAR JAMMING

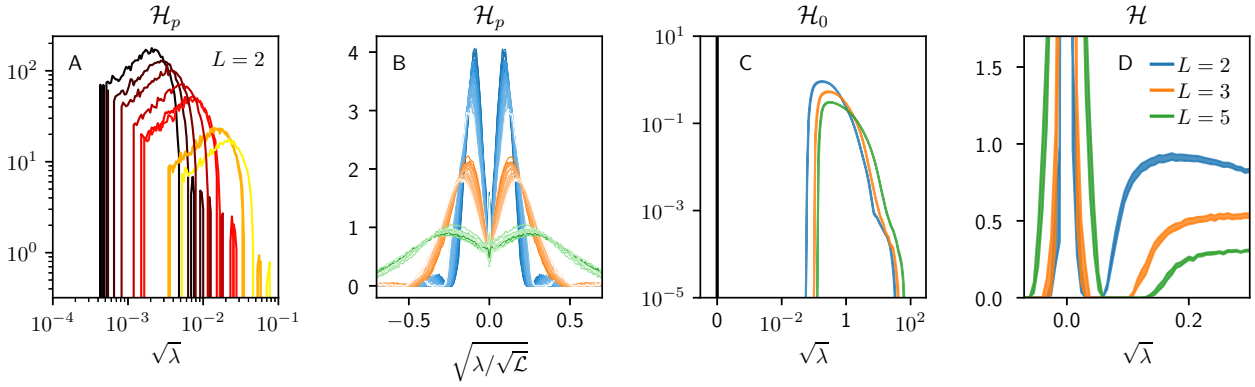

FIG. 6: The data shown in this figure concern the underparametrized points close to the transition for random data, which in Fig. 5B are enclosed in a black rectangle. (A) Positive part of the spectrum of $\mathcal{H}_p$ for ten distinct runs in the underparametrized phase close to the transition. The 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.

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. 5B. 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. 6A 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. 6B 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. This figure also shows that the spectrum of $\mathcal{H}_p$ is essentially symmetric, as argued in Appendix B. (The apparent zero modes comes from the symmetries induced by the homogeneity of the ReLU function discussed above: there is one such mode per neuron, and our definition of N takes this effect into account). In Fig. 6C the spectrum of $\mathcal{H}_0$ is shown, which depends very mildly on the distance from the transition. As expected for hypostatic situations [51], it displays a delta function, a gap, and a continuous spectrum. Finally in Fig. 6D, the spectrum of $\mathcal{H}$ is shown. As expected from Section II, the delta function of H_0 becomes blurred on a scale $\lambda \sim \sqrt{\mathcal{L}}$.

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 overparametrized regime, where they signal how much room is left around a minimum of the loss to fit additional patterns. In Fig. 7 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 $\gamma \approx 0.41269\ldots$).

In the case of spheres, the two exponents are related by an inequality that happens to be saturated [40, 52]. 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 [41]) 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 [41], a property associated with a hierarchical organization of the loss landscape [42, 43]. 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.

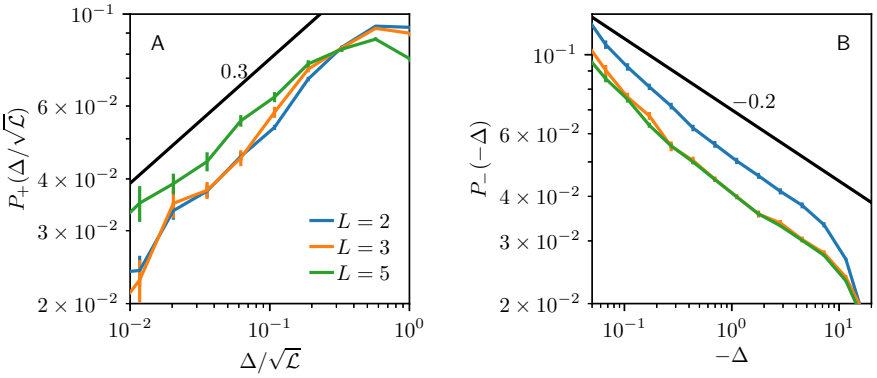

FIG. 7: (A) Distribution of re-scaled overlaps $z \equiv \Delta/\sqrt{\mathcal{L}} > 0$ near threshold, supporting that $P_+(z) \sim z^\theta$ with an exponent $\theta \approx 0.3$ that does not vary with L in the range probed. (B) The distribution of gaps $P_-(\Delta) \sim |\Delta|^{-\gamma}$ for $\Delta < 0$, with $\gamma \approx 0.2$, which again does not vary with L.

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. 9 for $P \geq 10^4$).

In Fig. 8 we show that a jamming transition is also found for real data with a discontinuous behavior of N_Δ/N. Fig. 8A shows the number of unsatisfied patterns per parameter N_Δ/N increasing P at fixed N, and in Fig. 8B the same quantity is plotted against the loss. As for random data, the latter is less noisy. In Fig. 8C 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. 5A-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. 8E-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. 8D, 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. 9 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

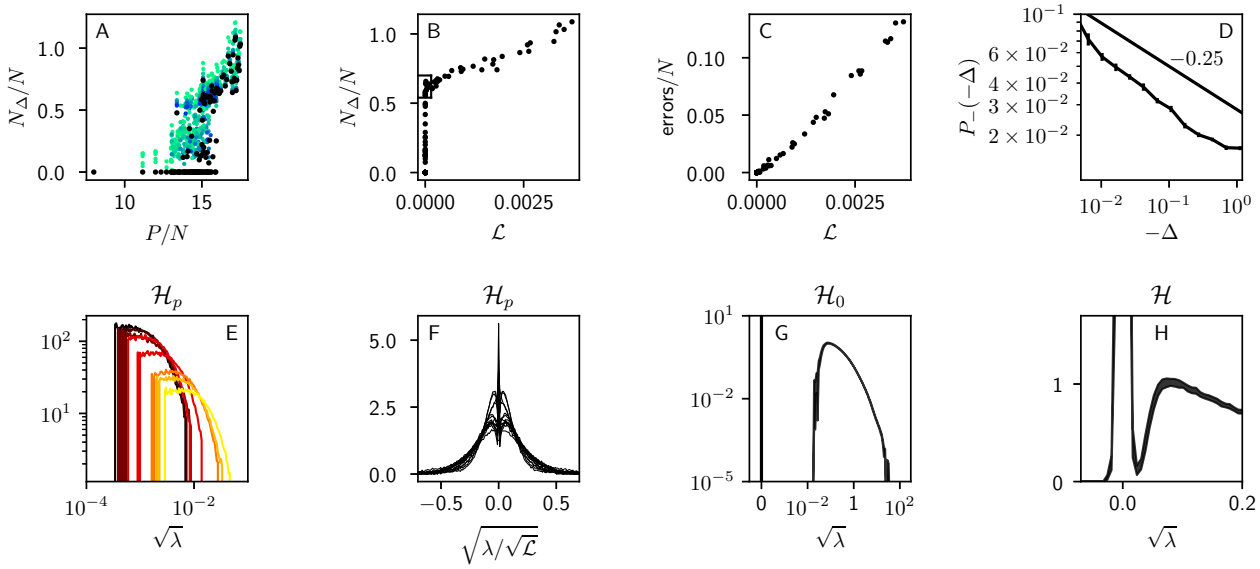

FIG. 8: Results with the MNIST dataset, keeping the first 10 PCA components. $d = 10$, $h = 30$ and $L = 5$ ($N = 3900$), varying $P = 1, \ldots, 70k$. (A) The number of unsatisfied patterns N_Δ/N jumps discontinuously when $r = P/N$ is increased. (B) The same quantity is less noisy when plotted against 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.

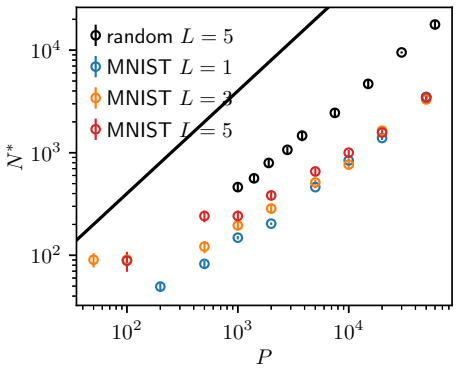

FIG. 9: Results with the MNIST dataset, keeping the first 10 PCA components (see main text), with $d = 10$ and varying P and h. The plot shows the number of parameters N^* at the jamming transition. For comparison, we also show the theoretical upper bound (solid curve) and the results found with random data (black points). The maximum number of steps is $2 \cdot 10^6$.

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 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. 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. 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 $P/N < r_c$ where r_c is $O(1)$. 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.

Role of depth: We showed that depth is not helpful 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 showed 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 [61].

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.

Acknowledgments

We thank C. Brito, C. Cammarota, T.S. Cohen, S. Franz, Y. LeCun, F. Krzakala, R. Ravasio, P. Urbani and L. Zdeborova for helpful discussions. This work was partially supported by the grant from the Simons Foundation (#454935 Giulio Biroli, #454953 Matthieu Wyart). M.W. thanks the Swiss National Science Foundation for support under Grant No. 200021-165509.

The manuscript [62], which appeared at the same time than ours, shows that the critical properties of the jamming transition found for the non-convex perceptron [33] hold more generally in some shallow networks. This universality is an intriguing result. Understanding the connection with our findings, which show instead a jamming transition similar to that of ellipses, is certainly worth future studies.

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, in *Advances in neural information processing systems* (2012), pp. 1097–1105.
[2] Y. LeCun, Y. Bengio, and G. Hinton, Nature **521**, 436 (2015).
[3] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al., IEEE Signal processing magazine **29**, 82 (2012).
[4] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Nature **529**, 484 (2016).
[5] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al., Nature **550**, 354 (2017).
[6] Y. LeCun, Y. Bengio, et al., The handbook of brain theory and neural networks **3361**, 1995 (1995).
[7] K. He, X. Zhang, S. Ren, and J. Sun, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.
[8] S. Ioffe and C. Szegedy, in *International conference on machine learning* (2015), pp. 448–456.
[9] G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio, in *Advances in neural information processing systems* (2014), pp. 2924–2932.
[10] M. Bianchini and F. Scarselli, IEEE transactions on neural networks and learning systems **25**, 1553 (2014).
[11] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, arXiv preprint arXiv:1606.05336 (2016).
[12] R. Eldan and O. Shamir, in *Conference on Learning Theory* (2016), pp. 907–940.
[13] H. Lee, R. Ge, T. Ma, A. Risteski, and S. Arora, arXiv preprint arXiv:1702.07028 (2017).
[14] E. Gardner, Journal of physics A: Mathematical and general **21**, 257 (1988).
[15] R. Monasson and R. Zecchina, Physical review letters **75**, 2432 (1995).
[16] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, arXiv preprint arXiv:1611.03530 (2016).
[17] E. B. Baum, Journal of complexity **4**, 193 (1988).
[18] L. Berthier and G. Biroli, Reviews of Modern Physics **83**, 587 (2011).
[19] A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, and Y. LeCun, in *Artificial Intelligence and Statistics* (2015), pp. 192–204.
[20] C. D. Freeman and J. Bruna, arXiv preprint arXiv:1611.01540 (2016).
[21] E. Hoffer, I. Hubara, and D. Soudry, in *Advances in Neural Information Processing Systems* (2017), pp. 1729–1739.
[22] D. Soudry and Y. Carmon, arXiv preprint arXiv:1605.08361 (2016).
[23] Y. Cooper, arXiv preprint arXiv:1804.10200 (2018).
[24] L. Sagun, L. Bottou, and Y. LeCun, arXiv preprint arXiv:1611.07476 (2016).

[25] L. Sagun, U. Evci, V. U. Güney, Y. Dauphin, and L. Bottou, ICLR 2018 Workshop Contribution, arXiv:1706.04454 (2017).
[26] A. J. Ballard, R. Das, S. Martiniani, D. Mehta, L. Sagun, J. D. Stevenson, and D. J. Wales, Physical Chemistry Chemical Physics (2017).
[27] Z. C. Lipton, arXiv preprint arXiv:1602.07320 (2016).
[28] M. Baity-Jesi, L. Sagun, M. Geiger, S. Spigler, G. B. Arous, C. Cammarota, Y. LeCun, M. Wyart, and G. Biroli, Proceedings of Machine Learning Research **80**, 324 (2018), arXiv preprint arXiv:1803.06969.
[29] A. J Liu, S. R Nagel, W. Saarloos, and M. Wyart (2010).
[30] F. Krzakala and J. Kurchan, Physical Review E **76**, 021122 (2007).
[31] L. Zdeborová and F. Krzakala, Physical Review E **76**, 031131 (2007).
[32] S. Franz, G. Parisi, M. Sevelev, P. Urbani, and F. Zamponi, SciPost Physics **2**, 019 (2017).
[33] S. Franz and G. Parisi, Journal of Physics A: Mathematical and Theoretical **49**, 145001 (2016).
[34] M. Wyart, L. E. Silbert, S. R. Nagel, and T. A. Witten, Physical Review E **72**, 051306 (2005).
[35] L. E. Silbert, A. J. Liu, and S. R. Nagel, Phys. Rev. Lett. **95**, 098301 (2005).
[36] A. Donev, I. Cisse, D. Sachs, E. A. Variano, F. H. Stillinger, R. Connelly, S. Torquato, and P. M. Chaikin, Science **303**, 990 (2004).
[37] M. Mailman, C. F. Schreck, C. S. O'Hern, and B. Chakraborty, Phys. Rev. Lett. **102**, 255501 (2009).
[38] Z. Zeravcic, N. Xu, A. J. Liu, S. R. Nagel, and W. van Saarloos, Europhys. Lett. **87**, 26001 (2009).
[39] C. Brito, H. Ikeda, P. Urbani, M. Wyart, and F. Zamponi, arXiv preprint arXiv:1807.01975 (2018).
[40] M. Wyart, Phys. Rev. Lett. **109**, 125502 (2012).
[41] M. Müller and M. Wyart, Annual Review of Condensed Matter Physics **6**, 177 (2015).
[42] S. Franz and S. Spigler, Physical Review E **95**, 022139 (2017).
[43] P. Charbonneau, J. Kurchan, G. Parisi, P. Urbani, and F. Zamponi, Nature Communications **5** (2014).
[44] A. Anderson, *Amorphous Solids: Low Temperature Properties*, vol. 24 of *Topics in Current Physics* (Springer, Berlin, 1981).
[45] M. Wyart, Annales de Phys **30**, 1 (2005).
[46] A. V. Tkachenko and T. A. Witten, Phys. Rev. E **60**, 687 (1999).
[47] C. S. O'Hern, L. E. Silbert, A. J. Liu, and S. R. Nagel, Phys. Rev. E **68**, 011306 (2003).
[48] E. DeGiuli, A. Laversanne-Finot, G. A. Düring, E. Lerner, and M. Wyart, Soft Matter **10**, 5628 (2014).
[49] L. Yan, E. DeGiuli, and M. Wyart, EPL (Europhysics Letters) **114**, 26003 (2016).
[50] S. Franz, G. Parisi, P. Urbani, and F. Zamponi, Proceedings of the National Academy of Sciences **112**, 14539 (2015).
[51] G. Düring, E. Lerner, and M. Wyart, Soft Matter **9**, 146 (2013).
[52] E. Lerner, G. During, and M. Wyart, Soft Matter **9**, 8252 (2013).
[53] E. Lerner, G. Düring, and M. Wyart, EPL (Europhysics Letters) **99**, 58003 (2012).
[54] P. Charbonneau, E. I. Corwin, G. Parisi, and F. Zamponi, Physical Review Letters **109**, 205501 (2012), URL `http://link.aps.org/doi/10.1103/PhysRevLett.109.205501`.
[55] P. Charbonneau, J. Kurchan, G. Parisi, P. Urbani, and F. Zamponi, Journal of Statistical Mechanics: Theory and Experiment **2014**, 10009 (2014).
[56] M. Mézard, G. Parisi, and M. Virasoro, *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, vol. 9 (World Scientific Publishing Company, 1987).
[57] P. Charbonneau, E. I. Corwin, G. Parisi, and F. Zamponi, Physical Review Letters **114**, 125504 (2015).
[58] X. Gastaldi, arXiv preprint arXiv:1705.07485 (2017).
[59] A. M. Saxe, J. L. McClelland, and S. Ganguli, arXiv preprint arXiv:1312.6120 (2013).
[60] D. P. Kingma and J. Ba, arXiv preprint arXiv:1412.6980 (2014).
[61] R. Shwartz-Ziv and N. Tishby, arXiv preprint arXiv:1703.00810 (2017).
[62] P. U. S. Franz, S. Hwang, arXiv preprint arXiv:1809.09945 (2018).
[63] 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.
[64] 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.
[65] We ran the implementation `https://github.com/mariogeiger/pytorch_shake_shake` for CIFAR-10 and we retrained it by replacing the cross entropy by the hinge loss (adapted for multiple classes) and no other modification of their code. 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% when replacing the loss function by the hinge loss. The results are very similar.
[66] 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.

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. 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:

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

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

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

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

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

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

where the elements of the matrix G are defined as

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

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

In Fig. 10 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.

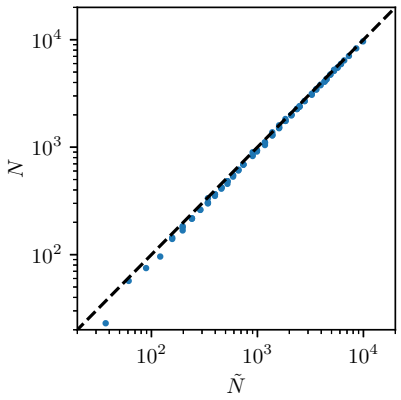

FIG. 10: Results with the MNIST dataset, keeping the first 10 PCA components. $d = 10$ and $L = 3$, varying P and h. Effective N vs total number of parameters $\tilde{N}$. N is always smaller than $\tilde{N}$ because there is a symmetry per each ReLU-neuron in the network.

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.

First, we argue that it must be so for $\hat{\mathcal{H}}_\mu$. It is equivalent to show that $\mathrm{tr}(\hat{\mathcal{H}}_\mu^n) = 0$ for any *odd* n.

$$\mathrm{tr}(\hat{\mathcal{H}}_\mu^n) = \sum_{i_1, i_2, \ldots, i_n} \hat{\mathcal{H}}^\mu_{i_1, i_2} \hat{\mathcal{H}}^\mu_{i_2, i_3} \cdots \hat{\mathcal{H}}^\mu_{i_n, i_1}, \tag{B1}$$

where the indices $i_1, \dots, i_n$ stand for synapses connecting a pair of neurons (i.e. each index is associated with a synaptic weight $W^{(j)}_{\alpha,\beta}$: 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 $W^{(j)}_{\alpha,\beta}$ and $W^{(k)}_{\gamma,\delta}$ reads

$$\hat{\mathcal{H}}^{\mu;(jk)}_{\alpha\beta;\gamma\delta} = \sum_{\pi_0,\dots,\pi_L} \theta(a^\mu_{L,\pi_L}) \cdots \theta(a^\mu_{1,\pi_1}) x^\mu_{\pi_0} \cdot \partial_{W^{(j)}_{\alpha,\beta}} \partial_{W^{(k)}_{\gamma,\delta}} \left[W^{(L+1)}_{\pi_L} W^{(L)}_{\pi_L,\pi_{L-1}} \cdots W^{(1)}_{\pi_1\pi_0} \right]. \tag{B2}$$

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 $\mathrm{tr}(\hat{\mathcal{H}}^n_\mu)$ averaged over the random labels is zero for odd n. We find that the rank of $\hat{\mathcal{H}}_\mu$ is proportional to the number of active nodes. Therefore the spectrum of $\hat{\mathcal{H}}_\mu$ only contains $O(\sqrt{N})$ non zero eigenvalues. Since this number diverges we expect the spectrum to be self-averaging, even-though with two different scalings for the delta peak and the set of non-zero eigenvalues. In consequence, for any realisation of the data, odd moments are expected to be zero and the spectrum must be symmetric.

We thus have to show that $\mathrm{tr}(\hat{\mathcal{H}}^n_\mu)$ changes sign under the symmetry mentioned above for odd n. Note that the sum in Eq.B2 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. (B1) 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. (B2)). For each index equal to $L+1$ (the last layer), there are exactly two terms: $\hat{\mathcal{H}}^\mu_{j,L+1}\hat{\mathcal{H}}^\mu_{L+1,k}$ (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. (B1) 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, concluding our argument showing that $\hat{\mathcal{H}}_\mu$ has a symmetric spectrum. 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. These are the terms that contribute to $\mathrm{tr}(\mathcal{H}^n_p)$, which therefore it is also expected to vanish in the large N limit for all odd n.

Note that the sets of arguments presented above are not at a level of a rigorous proof, for which a careful analysis of sub-leading corrections would be needed.

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

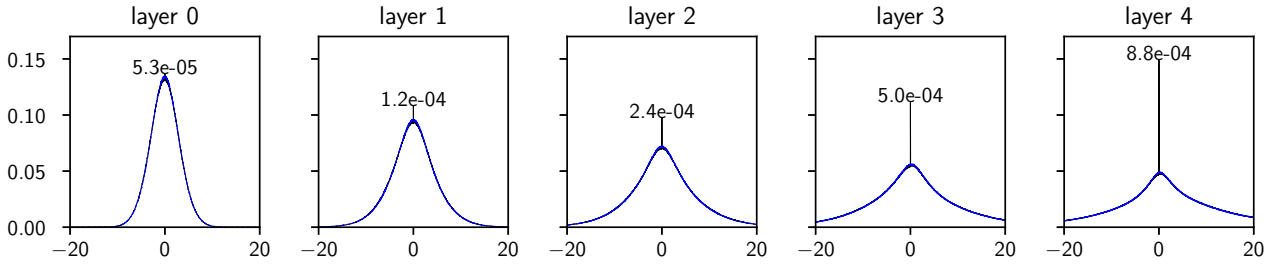

FIG. 11: Density of the pre-activations for each layers with $L = 5$ and random data, averaged over all the runs just above the jamming transition with that architecture. Black: distribution obtained over the training set. Blue: previously unseen random data (the two curves are on top of each other except 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.

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