Geometry of Energy Landscapes and the Optimizability of Deep Neural Networks

Simon Becker,[1] Yao Zhang,[2,1] and Alpha A. Lee[2,*]
[1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, United Kingdom
[2]Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, United Kingdom

(Received 9 August 2018; revised manuscript received 13 November 2019; accepted 6 February 2020; published 10 March 2020)

Deep neural networks are workhorse models in machine learning with multiple layers of nonlinear functions composed in series. Their loss function is highly nonconvex, yet empirically even gradient descent minimization is sufficient to arrive at accurate and predictive models. It is hitherto unknown why deep neural networks are easily optimizable. We analyze the energy landscape of a spin glass model of deep neural networks using random matrix theory and algebraic geometry. We analytically show that the multilayered structure holds the key to optimizability: Fixing the number of parameters and increasing network depth, the number of stationary points in the loss function decreases, minima become more clustered in parameter space, and the trade-off between the depth and width of minima becomes less severe. Our analytical results are numerically verified through comparison with neural networks trained on a set of classical benchmark datasets. Our model uncovers generic design principles of machine learning models.

DOI: 10.1103/PhysRevLett.124.108301

Nonlinear multiparameter fitting is ubiquitous in science, from cosmology [1] to biophysics [2]. The key challenge is nonconvexity: Typically fitting is done by finding parameters that minimize the discrepancy between model prediction and data, known as the loss function. The loss functions of nonlinear models often have many minima and minimization algorithms converge to local minima rather than the global minimum.

Nonetheless, models often used in machine learning appear to circumvent this problem. The workhorse model, deep neural networks [3], comprises multiple layers of nonlinear functions composed in series. Deep neural networks achieved near-human accuracy in tasks such as image recognition [4] and translation [5]. However, the success of the deep neural network raises two fundamental unsolved puzzles: First, industrial models have millions of parameters [6] and the loss function is highly nonconvex, yet surprisingly even a simple gradient descent algorithm is able to find accurate and predictive models. Second, it is long known that "shallow" neural networks—models that comprise a sum, rather than composition, of nonlinear functions—can approximate any smooth function [7]. However, deep neural networks empirically outperform shallower neural networks [8].

The surprising effectiveness of deep neural networks is often explained in terms of the classes of expressible functions. Seminal works show that the multilayered structure allows deep neural networks to disentangle highly curved manifolds in input space into flat manifolds [9–11]. Some argue that deep neural networks expresses "physical" functions: they can be mapped to the renormalization group [12] and implicitly imposes the physics of symmetry, locality, and compositionality [13]. However, recent numerical experiments problematize explanations based expressivity: shallower neural networks can match the accuracy of deep neural networks as long as one uses the trained deep neural network to augment the dataset by predicting labels of unlabeled data [14]. This observation suggests that deep and shallow networks are comparable in expressivity. An explanation of why deep neural networks are effective must therefore turn to whether one can actually find optimal parameters given data, i.e., optimizability.

Pioneering works show that for Gaussian random functions, critical points that take a value much larger than the global minimum are exponentially likely to be saddle points in the high dimensional limit [15–19]. Modeling a neural network as a Gaussian random function, some argue that the value that the loss function takes at most local minima is similar to the global minimum and this is why local minima are "good enough" [20–22]. However, this does not directly explain why deep neural networks, in particular, outperform shallow neural networks. Pioneering numerical studies of the energy landscape of loss functions using methods developed for molecular systems [23–26] focused on shallow neural networks.

In this Letter, we build on the spin glass model of deep neural networks introduced in Ref. [21] and derive novel analytical results describing the geometry of the loss function landscape as a function of network depth. We show that fixing the number of parameters and increasing network depth, the number of stationary points in the loss function decreases, minima become more clustered in parameter space, and the trade-off between the depth and width of minima becomes less severe. We verify our

results through comparison with neural networks trained on a set of classical benchmark datasets.

We consider a fully connected feed-forward network with $H-1$ hidden layers where layer $k-1$ has n_{k-1} nodes and each of them is connected to the n_k nodes of layer k. The networks we consider take input vectors $\mathbf{X} \in \mathbb{R}^{n_0}$ entering the 0th layer and returns scalar outputs Y from the Hth layer

$$Y(\mathbf{X}, \mathbf{w}) = q\theta\{\mathbf{W}_H^T \theta[\mathbf{W}_{H-1}^T \ldots \theta(\mathbf{W}_1^T \mathbf{X})]\}, \tag{1}$$

where the matrices $\mathbf{W}_k$ contain the weights $\mathbf{w}$ and the functions θ are the activation functions. We restrict the analysis to rectified linear units (ReLUs) $\theta(x) = \max(x, 0)$. The normalizing constant q will be specified later to compare different architectures. We label paths in the network as (i, j), where j labels any of the P paths from a given component X_i of the input vector. The quantity $w_{(i,j)}^{(k)}$ denotes the weight connecting layer $k-1$ with layer k along path (i, j).

For simplicity, we consider a classification task: Let $\zeta = \max_w |Y(\mathbf{X}, \mathbf{w})|$ be the maximum of the absolute value of the network output for admissible weight configurations. We consider a random labeling scenario where the ground truth Y_{true} takes values $\pm\zeta$ independent of input $\mathbf{X}$. Our goal is to characterize the loss function $\mathcal{L}(\mathbf{w}) = \mathbb{E}_A |Y_{\text{true}} - Y(\mathbf{X}, \mathbf{w})|$ for this randomly labeled dataset.

To make analytical progress, we map this neural network architecture onto a spin glass Hamiltonian via a series of elegant approximations introduced in Ref. [21]. We rewrite Eq. (1) by replacing the ReLUs by activation functions $A \in \{0, 1\}$,

$$Y(\mathbf{X}, \mathbf{w}) = q \sum_{i=1}^{n_0} \sum_{j=1}^{P} X_i A_{(i,j)} \prod_{k=1}^{H} w_{(i,j)}^{(k)}. \tag{2}$$

We next introduce the key approximations: First, the input of the network is assumed to consist of independent and standard normally distributed random variables. The activation functions A are independent and Bernoulli distributed with probability p of being 1. Second, the number of different weights Λ is assumed to be the Hth root of the total number of paths in the network. Moreover, among all possible weight combinations of the Λ number of weights, each configuration is assumed to appear almost equally often. Third, the weights (w_n) are assumed to satisfy, after rescaling, a spherical constraint $(1/\Lambda) \sum_{n=1}^{\Lambda} w_n^2 = 1$. This spherical constraint models regularization methods commonly used in the literature that penalizes the magnitude of the weights.

Under the three previously stated assumptions, and choosing $q = \Lambda^{-(H-1)/2}$, the loss function $\mathcal{L}(\mathbf{w})$ has the same distribution as $p\mathcal{H}_\lambda(\mathbf{w})$, where $\mathcal{H}_\lambda(\mathbf{w})$ is the H-spin spherical spin glass Hamiltonian

$$\mathcal{H}_\Lambda(\mathbf{w}) = \frac{1}{\Lambda^{(H-1)/2}} \sum_{i_1, \ldots, i_H = 1}^{\Lambda} Z_{i_1, \ldots, i_H} \prod_{k=1}^{H} w_{i_k} \tag{3}$$

and $Z_{i_1, \ldots, i_H}$ are independent, identical, and standard normally distributed.

We consider networks with different number of layers H but with the same number of parameters N_e, and the same input layer of size n_{input} and a scalar output layer. All layers aside from the scalar output layer and the input layer of size n_0 shall be assumed to be of equal size $n_1 = \cdots = n_{H-1}$. The number of network parameters is then given as $N_e = (H-2)n_1^2 + (n_0+1)n_1$ and the number of weights is

$$\Lambda = \left(\frac{\sqrt{n_0(n_0+2) + 4(H-2)N_e + 1} - (n_0+1)}{2(H-2)} \right)^{(H-1)/H} \times n_0^{1/H}. \tag{4}$$

Number of critical points.—The spin glass Hamiltonian (3) is nonconvex. We thus analyze how the number of critical points varies as a function of the number of layers. The number of critical points $\mathcal{N}_{\mathbb{C}}$ over the complex numbers is (up to complex conjugation)

$$\mathcal{N}_{\mathbb{C}} = \frac{(H-1)^\Lambda - 1}{H-2}. \tag{5}$$

The number of expected real critical points is

$$\mathcal{N}_{\mathbb{R}} = \frac{2^{\Lambda-1}(H-1)^{\Lambda/2}\Gamma(\Lambda - 1/2)}{\sqrt{\pi} H^{\Lambda-1/2}\Gamma(\Lambda)} \left[2(\Lambda-1) \times {}_2F_1\left(1, \Lambda - \frac{1}{2}; \frac{3}{2}; \frac{H-2}{H}\right) + {}_2F_1\left(1, \Lambda - \frac{1}{2}; \frac{\Lambda+1}{2}; \frac{1}{H}\right) \right], \tag{6}$$

where ${}_2F_1$ is the hypergeometric function. The expressions (5) and (6) show a similar qualitative asymptotic; cf. Fig. 1. Because of its simplicity, we sketch only a proof of Eq. (5) and refer the reader to the Supplemental Material [27] for a proof of Eq. (6), which builds on Ref. [33].

Proof of Eq. (5).—The loss function is a homogeneous symmetric random polynomial. We illustrate the link between the two for $H = 2$ when the Hamiltonian is just $\mathcal{H}_\Lambda(\mathbf{w}) = \sum_{i_1=1}^{\Lambda} (X_{i_1,i_1}/\sqrt{\Lambda}) w_{i_1}^2 + \sum_{i_1<i_2}^{\Lambda} ([X_{i_1,i_2} + X_{i_2,i_1}]/\sqrt{\Lambda}) w_{i_1} w_{i_2}$. In order to have a sum of random variables $Y_{i_1,i_2} + Y_{i_2,i_1}$ with the symmetry property $Y_{i_1,i_2} = Y_{i_2,i_1}$ to be distributed like $X_{i_1,i_2} + X_{i_2,i_1}$ one can choose $Y_{i_1,i_2} = (X_{i_1,i_2} + X_{i_2,i_1})/2 \sim \mathcal{N}(0, 1/2)$. Critical weights $\mathbf{w}$ of $\mathcal{H}_\Lambda(\mathbf{w})$ are precisely the generalized eigenvectors satisfying for $j \in \{1, \ldots, \Lambda\}$ the eigenvalue equation $\Lambda^{(1-H)/2} \sum_{i_2, \ldots, i_H = 1}^{\Lambda} Y_{\mathbf{j}, \ldots, i_H} \prod_{k=2}^{H} w_{i_k} = \lambda w_{\mathbf{j}}$, where two solutions $(\lambda, \mathbf{w}), (\lambda', \mathbf{w}')$ to the eigenproblem coincide if there is $t \neq 0$ such that $t\lambda^{H-2} = \lambda'$ and $t\mathbf{w} = \mathbf{w}'$. Substituting $\lambda = \gamma^{H-2}$ in the eigenvalue equation yields

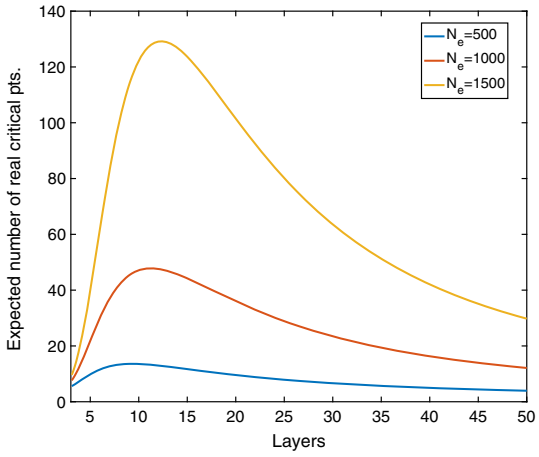

FIG. 1. The number of real critical points $\mathcal{N}_{\mathbb{R}}$, Eq. (6), in a deep neural network decreases as a function of depth for a fixed number of parameters N_e and $n_0 = 10$.

Λ-many homogeneous equations of degree $H-1$ in $\Lambda+1$ many variables $\lambda, w_1, \ldots, w_\Lambda$. The multihomogeneous Bézout's theorem [34] (Chap. 4, Sec. 2. 2) implies that such an equation has exactly $(H-1)^\Lambda$ solutions where we discard the equivalence class of the zero solution $\lambda = w_i = 0$ to end up with $(H-1)^\Lambda - 1$ solutions. Removing the $H-2$ degeneracy, due to roots of unity $e^{2\pi i/(H-2)}$, coming from the $\lambda = \gamma^{H-2}$ substitution, shows that the number of critical weights satisfies Eq. (5). This has been obtained using methods from toric geometry in Ref. [35] (Theorem 1.2) (see the Supplemental Material [27]). ∎

Figure 1 show that Eq. (5) implies that the number of critical points is a nonmonotonic function of the number of layers. Importantly, the number of critical points decreases as the number of layers increases for a deep network, thus deep networks are more optimizable because there are less critical points that traps the optimizer. Figure 1 also shows that the number of critical points increases as a function of depth for shallow networks. This agrees with the early experience with deep learning in the 1980s and 1990s—a one layer neural network is inefficient in learning compositional features, yet simply adding a few more layers to a one layer neural network causes performance to deteriorate because the number of critical points proliferates and the loss function becomes nonoptimizable [8]. The deep learning boon began when there were sufficient computational resources to train a very deep neural network.

The scaling in the number of critical points in our model deep neural network is similar to a molecular energy landscape. The number of critical points in a molecular energy landscape $\sim e^{\gamma N}$ with N the number of particles [36,37]. Equation (6) shows that

$$\mathcal{N}_{\mathbb{R}} \sim \exp\left(\frac{\log H n_0^{1/H}}{2(H-2)^{([H-1]/[2H])}} N_e^{[(H-1)/(2H)]}\right)$$

in the limit $N_e \to \infty$ and H fixed. As such, if we consider deep neural networks as a molecular system, the effective number of particles (or degrees of freedom) is $N_e^{(H-1)/(2H)}$, which depends on both the number of parameters and, intriguingly, the number of layers. Increasing depth decreases the scaling exponent, which maps to increasing the range of intermolecular potential [38].

Location of minima.—We next study where the critical points are located in weight space. Intuitively, the more clustered they are, the easier it is for an optimizer to search for minima. Let $\mathrm{Crt}(-\infty, \mathcal{E})$ denote the set of critical points for which the loss function takes values in $(-\infty, \Lambda\mathcal{E})$. For an interval $I \subset [-1,1]$ we study the number of pairs $(\mathbf{w}, \mathbf{w}')$ of critical weights in $\mathrm{Crt}(-\infty, \mathcal{E})$ with relative angle $\mathbf{w}\cdot\mathbf{w}'/\Lambda$ contained in I. This set will be denoted by $\{\mathrm{Crt}[(-\infty, \mathcal{E}), I]\}_2$. Note that the Euclidean distance $\|\mathbf{w}-\mathbf{w}'\|_2 = \sqrt{2(\Lambda - \mathbf{w}\cdot\mathbf{w}')}$. As we study large Λ asymptotics, minima occur predominantly at low energies such that we may assume that all energies are sufficiently small, i.e., $\mathcal{E}/p \in (-\infty, -\sqrt{2}/\sigma]$, where $\sigma = \sqrt{H/[2(H-1)]}$.

Our second theorem is that the upper bound to distance between minima is

$$\limsup_{\Lambda\to\infty} \frac{1}{\Lambda}\log\left(\frac{\mathbb{E}|\{\mathrm{Crt}[(-\infty,\mathcal{E}), I]\}_2|}{\mathbb{E}|\mathrm{Crt}(-\infty,\mathcal{E})|}\right) \le \sup_{r\in I}\sup_{v\in(-\infty,\mathcal{E}/p)} \Psi_H(r, v, \mathcal{E}), \tag{7}$$

where

$$\Psi_H(r,v,\mathcal{E}) = \frac{1}{2} + \frac{\mathcal{E}^2}{2p^2} + \frac{1}{2}\log\left(\frac{(H-1)(1-r^2)}{1-r^{2H-2}}\right) - \frac{1}{2}\left\langle \begin{pmatrix} v \\ v \end{pmatrix}, \Sigma_U(r)^{-1}\begin{pmatrix} v \\ v \end{pmatrix}\right\rangle + \int_{-2}^{2} \frac{\log|\sqrt{2}\sigma v - x|\sqrt{4-x^2}}{2\pi}\, dx.$$

$\Sigma(r) = -(1/H)\left(\begin{smallmatrix} b_1(r) & b_2(r) \\ b_2(r) & b_1(r)\end{smallmatrix}\right)$ is a matrix defined by

$$\alpha_H(r) = \{H - H[r^H - (H-1)(r^{H-2} - r^H)]^2\}^{-1},$$
$$b_1(r) = -H + \alpha_H(r) H^3 (r^{2H-2} - r^{2H}), \text{ and}$$
$$b_2(r) = -Hr^H - \alpha_H(r) H^3 r^{3H-4}[r^2 + H(r^2-1)^2 - 1].$$

Proof.—The full proof is in the Supplemental Material [27]. Our proof strategy combines the asymptotics for the minima of the Hamiltonian [39] (Theorem 10) with the upper bound on the angle between minima [39] (Theorem 5 and Lemma 6). ∎

Figure 2 shows that the number of minima, relative to the total number of minima, that are close to other minima [cf. Eq. (7)] increases as the number of layers increases. In other words, minima are more clustered for deeper

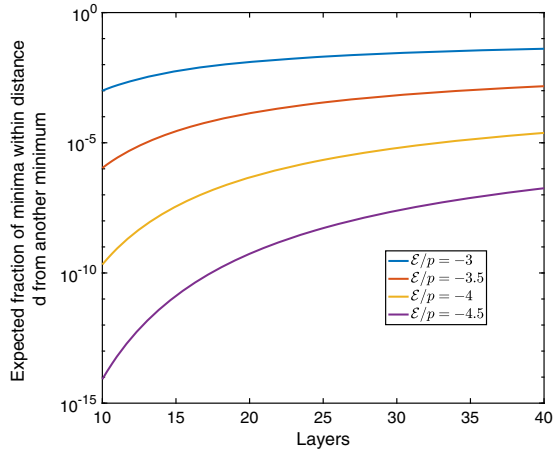

FIG. 2. Minima are more clustered for deeper networks. The figure shows the relative expected number of critical points (7) that attains a loss function value in the interval $(-\infty, \Lambda\mathcal{E})$ with $||\mathbf{w} - \mathbf{w}'||_2^2 \leq d\Lambda$ with $d = 0.02$ for fixed number of network parameters $N_e = 400$ and $n_0 = 10$.

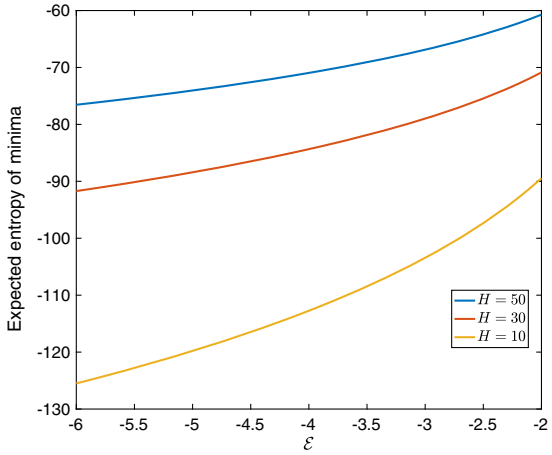

FIG. 3. Energy-entropy competition is eased by increasing network depth. The expected entropy at minima of the loss function as a function of minima depth for $N_e = 400$ network parameters, $n_0 = 10$, and $p = 0.8$.

networks, thus deep networks are more optimizable compared to shallower ones. Interestingly, minima that attain a low value of the loss function (more negative $\mathcal{E}/p$) are further apart, yet increasing network depth brings even those minima closer together in weight space.

Width of minima.—We now turn to examine how the width of minima varies with the value of loss function that it attains. To measure basin volume at minima $\mathbf{W}_q$, we consider the entropy $S(\mathbf{W}_q) = -\log\det\{\mathrm{Hess}[\mathcal{L}(\mathbf{W}_q)]\}$, with Hess being the Hessian matrix [23,24,40]. Within the harmonic approximation, larger entropy corresponds to larger basin volume. Intuitively, if wider minima are also deeper, then the function is easy to optimize, whereas functions with deep and narrow minima are difficult to optimize.

The expected entropy of the Hessian of the minima of the loss function that takes value $\Lambda\mathcal{E}$ satisfies asymptotically

$$\mathbb{E}[S(\mathrm{Hess}\mathcal{L})|\Lambda\mathcal{E}] \simeq -(\Lambda-1)\log(p) + \frac{\Lambda-1}{2}\log\left(\frac{\Lambda}{2(\Lambda-1)H(H-1)}\right) - \frac{\Lambda-1}{\pi}\int_{-\sqrt{2}}^{\sqrt{2}}\log\left|\sigma\sqrt{\frac{\Lambda}{\Lambda-1}}\frac{\mathcal{E}}{p} - t\right|\sqrt{2-t^2}\,dt. \quad (8)$$

Proof.—We start by studying a small energy interval $E = (\mathcal{E}-\varepsilon, \mathcal{E}+\varepsilon)$ around some energy $\mathcal{E}$, where we assume that the auxiliary interval $G = \sigma\sqrt{\Lambda/(\Lambda-1)}E/p$ is contained in $(-\infty, -\sqrt{2}]$, as minima of the loss function and the spin glass Hamiltonian are known to appear at low energies for large values of Λ [19]. ∎

Let $M_{\mathcal{H}_\Lambda}(\Lambda E/p)$ be the event that the Hamiltonian possesses a minimum at some energy in the interval $\Lambda E/p$. We are interested in finding the expected entropy at those points. We first rewrite this conditional expectation in terms of an auxiliary random variable $X = \sigma\mathcal{H}_\Lambda/\sqrt{\Lambda(\Lambda-1)}$ and a GOE matrix $M^{\Lambda-1}$ of size $\Lambda-1$ using the tower property and the probability distribution of the spin glass Hessian [18] (Lemma 1.1)

$$\mathbb{E}[S(\mathrm{Hess}\mathcal{H}_\Lambda)|M_{\mathcal{H}_\Lambda}(\Lambda E/p)] = \frac{\mathbb{E}(\mathbb{E}[S(\mathrm{Hess}\mathcal{H}_\Lambda)1_{M_{\mathcal{H}_\Lambda}(\Lambda E/p)}|\{\mathcal{H}_\Lambda\}])}{\mathbb{E}(\mathbb{P}[M^{\Lambda-1} \geq X, X \in G|\{X\}])}. \quad (9)$$

We now consider the asymptotic behavior of the numerator and denominator separately for large Λ. The distribution of the Hessian of $\mathcal{H}_\Lambda$ [19] (Lemma 1.1) allows us to express the numerator in terms of an auxiliary function $f_\beta(t) = \sqrt{(\Lambda-1)/(2\pi\sigma^2)}\int_G e^{-[\mathcal{E}^2(\Lambda-1)/2\sigma^2]}\log|t-x|dx$. Using the Wigner semicircle law,

$$\mathbb{E}(\mathbb{E}[S(\mathrm{Hess}\mathcal{H}_\Lambda)1_{M_{\mathcal{H}_\Lambda}(\Lambda E/p)}|\{\mathcal{H}_\Lambda\}]) \simeq -\frac{\Lambda-1}{\pi}\int_{-\sqrt{2}}^{\sqrt{2}} f_{-\sqrt{2}}(t)\sqrt{2-t^2}\,dt + \frac{\Lambda-1}{2}\log\left(\frac{\Lambda}{2(\Lambda-1)H(H-1)}\right)\mathbb{P}[M_{\mathcal{H}_\Lambda}(\Lambda E/p)]. \quad (10)$$

For the denominator in Eq. (9), we use the probability distribution of X and that the lowest eigenvalue of the random matrix $M^{\Lambda-1}$ concentrates at the lower end $-\sqrt{2}$ of the semicircle distribution for Λ large [41] (Theorem 1). Hence, it follows that $\mathbb{E}[\mathbb{P}(M^{\Lambda-1} \geq X, X \in G|\{X\})] = \sqrt{(\Lambda-1)/(2\pi\sigma^2)}\int_G e^{-[t^2(\Lambda-1)/2\sigma^2]}dt$. Having obtained asymptotic expressions for both the numerator and denominator in Eq. (9), we take the limit $\varepsilon\downarrow 0$ such that the energy

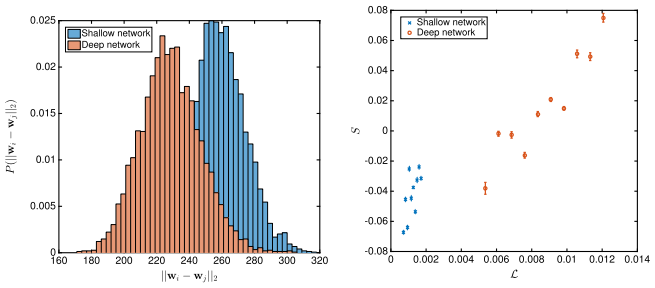

FIG. 4. Numerical experiments agree qualitatively with the analytical predictions. Left: The histogram of distances between minima. Minima in deeper networks are closer together. Right: The loss function at minima plotted against the expected entropy of minima. Lower minima are narrower but this energy-entropy trade-off is less severe for deep networks. The figures are plotted for the "Boston Housing" dataset, cf. Refs. [46,47]. To compute the expected value of entropy, we discretize the distribution of values that the loss function takes into 10 bins.

interval E shrinks down to a single energy value $\mathcal{E}$ such that Eq. (8) follows immediately.

Figure 3 shows that the lower in loss function that the minima attains, the narrower it is, thus there is an "energy-entropy" competition. The existence of energy-entropy competition is nontrivial and unlike many atomic cluster systems analyzed in the literature [42–44], where the lower minima have larger basins of attraction. However, this competition is smoothened as the number of layers increases. For very deep networks, minima that attain a very low value of loss function has almost the same width as minima that attain a high value of loss function. As such, there is less risk of minimization algorithms getting trapped in wide but very suboptimal local minima. Interestingly, the presence of energy-entropy competition is different from many molecular systems, where more stable minima appear to also have a wider basin of attraction [45].

To verify our analytical results, we consider a classical set of 10 benchmark datasets [46,47]. Figure 4 shows the results for one dataset (results for the remaining datasets, shown in the Supplemental Material [27], agree with the theory)—the distance between minima decreases as a function of depth, as shown by the shift in the distribution of pairwise distance between minima, and the tradeoff between minima depth and width is eased. Enumerating the number of critical points is numerically challenging and has only been done for particle systems with relative small number of particles [48,49], thus this is outside the scope of the present study. In the numerical experiments, the input size is 10, the shallow network comprises 2 hidden layers and 11 nodes each and the deep network comprises 6 hidden layers with 11 nodes each, such that the total number of parameters is 726. Further details are discussed in the Supplemental Material [27].

In summary, we derived a series of analytical results showing that deep networks are more optimizable than shallow networks because there are less critical points, the minima are more clustered, and the energy-entropy trade-off is eased. We verified our analytical results via a set of numerical experiments on classical benchmark datasets in machine learning. Our work sheds light on why deep learning empirically works from the perspective of optimization, as well as suggests new design principles. For example, the most optimizable machine learning architecture is one where lower minima are also wider, and we speculate that analogies between loss function and energy landscape of atomic systems [42–44] holds the key to engineering such architectures.

This work was supported by the EPSRC Grant No. EP/L016516/1 for the University of Cambridge CDT, the CCA (S. B.). A. A. L. acknowledges support from the Winton Programme for the Physics of Sustainability.

*aal44@cam.ac.uk

[1] F. Leclercq, A. Pisani, and B. D. Wandelt, in *Proceedings of the International School of Physics "Enrico Fermi"* (IOS Press, Amsterdam, 2014), p. 189.
[2] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna, PLoS Comput. Biol. **3**, e189 (2007).
[3] Y. LeCun, Y. Bengio, and G. Hinton, Nature (London) **521**, 436 (2015).
[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ImageNet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems* (2012), pp. 1097–1105.
[5] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, arXiv:1609.08144.
[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, Int. J. Comput. Vis. **115**, 211 (2015).
[7] G. Cybenko, Math. Control Signals Syst. **2**, 303 (1989).
[8] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning* (MIT Press, Cambridge, MA, 2016), Vol. 1.
[9] G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio, On the number of linear regions of deep neural networks, in *Advances in Neural Information Processing Systems* (2014), pp. 2924–2932.
[10] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, Exponential expressivity in deep neural networks through transient chaos, in *Advances in Neural Information Processing Systems* (2016), pp. 3360–3368.
[11] T. Poggio, H. Mhaskar, L. Rosasco, B. Miranda, and Q. Liao, Int. J. Automat. Comput. **14**, 503 (2017).
[12] P. Mehta and D. J. Schwab, arXiv:1410.3831.
[13] H. W. Lin, M. Tegmark, and D. Rolnick, J. Stat. Phys. **168**, 1223 (2017).
[14] J. Ba and R. Caruana, Do deep nets really need to be deep? in *Advances in Neural Information Processing Systems* (2014), pp. 2654–2662.
[15] Y. V. Fyodorov, Phys. Rev. Lett. **92**, 240601 (2004).

[16] A. J. Bray and D. S. Dean, Phys. Rev. Lett. **98**, 150201 (2007).
[17] Y. V. Fyodorov and I. Williams, J. Stat. Phys. **129**, 1081 (2007).
[18] A. Auffinger, G. B. Arous *et al.*, Ann. Probab. **41**, 4214 (2013).
[19] A. Auffinger, G. B. Arous, and J. Černý, Commun. Pure Appl. Math. **66**, 165 (2013).
[20] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, in *Advances in Neural Information Processing Systems* (2014), pp. 2933–2941.
[21] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, The loss surfaces of multilayer networks, in *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)* (2015), pp. 192–204.
[22] A. Choromanska, Y. LeCun, and G. B. Arous, Open problem: The landscape of the loss surfaces of multilayer networks, in *Proceedings of The 28th Conference on Learning Theory* (2015), pp. 1756–1760.
[23] R. Das and D. J. Wales, Phys. Rev. E **93**, 063310 (2016).
[24] A. J. Ballard, J. D. Stevenson, R. Das, and D. J. Wales, J. Chem. Phys. **144**, 124119 (2016).
[25] A. J. Ballard, R. Das, S. Martiniani, D. Mehta, L. Sagun, J. D. Stevenson, and D. J. Wales, Phys. Chem. Chem. Phys. **19**, 12585 (2017).
[26] D. Mehta, X. Zhao, E. A. Bernal, and D. J. Wales, Phys. Rev. E **97**, 052307 (2018).
[27] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.124.108301 for mathematical proofs of results presented in this Letter and numerical experiments, which includes Refs. [28–32].
[28] M. Denil, B. Shakibi, L. Dinh, M. Ranzato, and N. Freitas, Predicting parameters in deep learning, in *Advances in Neural Information Processing Systems* (2013), pp. 2148–2156.
[29] Z. W. B. J. L. Y. Denton, E., and R. Fergus, Exploiting linear structure within convolutional networks for efficient evaluation, in *Advanced in Neural Information Processing Systems* (2014), pp. 1269–1277.
[30] A. Edelman, E. Kostlan, and M. Shub, J. Am. Math. Soc. **7**, 247 (1994).
[31] L. Qi, J. Symb. Comput. **40**, 1302 (2005).
[32] L. Qi, J. Math. Anal. Appl. **325**, 1363 (2007).
[33] P. Breiding, SIAM J. Appl. Algebra Geom. **1**, 254 (2017).
[34] I. Shafarevich, *Basic Algebraic Geometry* (Springer-Verlag, Berlin, 1977).
[35] D. Cartwright and B. Sturmfels, Linear Algebra Appl. **438**, 942 (2013).
[36] D. J. Wales, *Energy Landscapes* (Cambridge University Press, Cambridge, England, 2003).
[37] D. J. Wales and J. P. K. Doye, J. Chem. Phys. **119**, 12409 (2003).
[38] P. A. Braier, R. S. Berry, and D. J. Wales, J. Chem. Phys. **93**, 8745 (1990).
[39] E. Subag, Ann. Probab. **45**, 3385 (2017).
[40] Y. Zhang, A. M. Saxe, M. S. Advani, and A. A. Lee, Mole. Phys. **116**, 3214 (2018).
[41] M. Ledoux and B. Rider, Electron. J. Pro **15**, 1319 (2010).
[42] J. P. Doye, D. J. Wales, and M. A. Miller, J. Chem. Phys. **109**, 8143 (1998).
[43] J. P. Doye and C. P. Massen, J. Chem. Phys. **122**, 084105 (2005).
[44] C. P. Massen and J. P. K. Doye, Phys. Rev. E **75**, 037101 (2007).
[45] C. J. Pickard and R. J. Needs, J. Phys. Condens. Matter **23**, 053201 (2011).
[46] J. M. Hernández-Lobato and R. Adams, Probabilistic backpropagation for scalable learning of bayesian neural networks, in *International Conference on Machine Learning* (2015), pp. 1861–1869.
[47] Y. Gal and Z. Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in *International Conference on Machine Learning* (2016), pp. 1050–1059.
[48] S. Martiniani, K. J. Schrenk, J. D. Stevenson, D. J. Wales, and D. Frenkel, Phys. Rev. E **93**, 012906 (2016).
[49] S. Martiniani, K. J. Schrenk, J. D. Stevenson, D. J. Wales, and D. Frenkel, Phys. Rev. E **94**, 031301(R) (2016).