

Geometry of energy landscapes and the optimizability of deep neural networks

Simon Becker,¹ Yao Zhang,² 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*

Deep neural networks are workhorse models in machine learning with multiple layers of non-linear functions composed in series. Their loss function is highly non-convex, yet empirically even gradient descent minimisation is sufficient to arrive at accurate and predictive models. It is hitherto unknown why are deep neural networks 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 tradeoff 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.

Nonlinear multiparameter fitting is an ubiquitous in science, from cosmology [1] to biophysics [2]. The key challenge is non-convexity: Typically fitting is done by finding parameters that minimise the discrepancy between model prediction and data, known as the loss function. The loss function of non-linear models often have many minima and minimisation 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 non-linear 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 deep neural network raises two fundamental unsolved puzzles: First, industrial models have millions of parameters [6] and the loss function is highly non-convex, yet surprisingly even 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 non-linear 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 renormalisation 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 the trained deep neural network is used augment the dataset by predicting labels of unlabelled data [14]. This obser-

vation suggests that deep and shallow networks are comparable in expressivity. 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]. Modelling a neural network as a Gaussian random function, some argue that the value of the loss function 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 network. Pioneering and seminal 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 [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 tradeoff 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 0-th layer and returns scalar outputs Y from the H -th layer

$$Y(\mathbf{X}, \mathbf{w}) = q\theta(\mathbf{W}_H^T \theta(\mathbf{W}_{H-1}^T \dots \theta(\mathbf{W}_1^T \mathbf{X}))) \quad (1)$$

where the matrices \mathbf{W}_k contain the weights \mathbf{w} and the

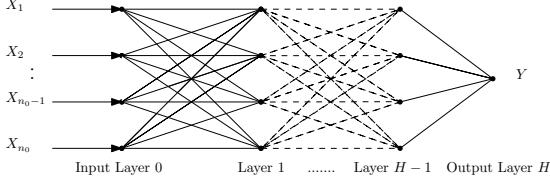


FIG. 1. A schematic of the the feedforward network architecture studied in this paper.

functions θ are the activation functions. We restrict the analysis to the commonly used rectified linear units (ReLUs) $\theta(x) = \max(x, 0)$. The normalising 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 labelling scenario where the ground truth Y_{true} takes values $\pm\zeta$ independent of input \mathbf{X} . Our goal is to characterise the loss function $\mathcal{L}(\mathbf{w}) = \mathbb{E}_A |Y_{\text{true}} - Y(\mathbf{X}, \mathbf{w})|$ for this randomly labelled dataset.

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

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

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 H -th 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 $\frac{1}{\Lambda} \sum_{n=1}^{\Lambda} w_n^2 = 1$. This spherical constraint models regularisation methods commonly used in the literature that penalises 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

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

and Z_{i_1, \dots, 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 . All layers aside from the scalar output layer shall be assumed to be of equal size $n_0 = \dots = n_{H-1}$ as in Fig. 1. The number of network parameters $N_e = (H-1)n_0^2 + n_0$ and

$$\Lambda = \frac{\sqrt{4N_e(H-1)+1}+1}{2(H-1)}. \quad (4)$$

Number of critical points: The spin glass Hamiltonian (3) is evidently non-convex. Thus a natural question to ask is how does the number of critical point varies as the function of number of layers. The number of critical points \mathcal{N} satisfies a remarkably simple theorem

$$\mathcal{N} = \frac{(H-1)^{\Lambda} - 1}{H-2}. \quad (5)$$

Proof: The loss function can be represented by a homogeneous symmetric random polynomial. To fix ideas we illustrate the link between the two for $H = 2$ when the Hamiltonian is just $\mathcal{H}_{\Lambda}(\mathbf{w}) = \sum_{i_1=1}^{\Lambda} \frac{X_{i_1, i_1}}{\sqrt{\Lambda}} w_{i_1}^2 + \sum_{i_1 < i_2}^{\Lambda} \frac{(X_{i_1, i_2} + X_{i_2, i_1})}{\sqrt{\Lambda}} w_{i_1} w_{i_2}$. 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 $Y_{i_1, i_2} = \frac{X_{i_1, i_2} + X_{i_2, i_1}}{2} \sim \mathcal{N}(0, 1/2)$. Critical weights \mathbf{w} of $\mathcal{H}_{\Lambda}(\mathbf{w})$ are precisely the generalized eigenvectors satisfying for $j \in \{1, \dots, \Lambda\}$ the eigenvalue equation $\frac{1}{\Lambda^{(H-1)/2}} \sum_{i_2, \dots, i_H=1}^{\Lambda} Y_{j, \dots, i_H} \prod_{k=2}^H w_{i_k} = \lambda w_j$ 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 Λ -many homogeneous equations of degree $H-1$ in $\Lambda+1$ many variables $\lambda, w_1, \dots, w_{\Lambda}$. The multi-homogeneous Bézout's theorem [27, Ch. 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 Equation (5). This has been obtained using methods from toric geometry in [28, Theorem 1.2] (see Supplemental Materials (SM)).

Figure 2 show that Equation (5) implies that the number of critical points is a non-monotonic 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 optimisable because there are less critical points that traps the optimiser. Figure 2 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

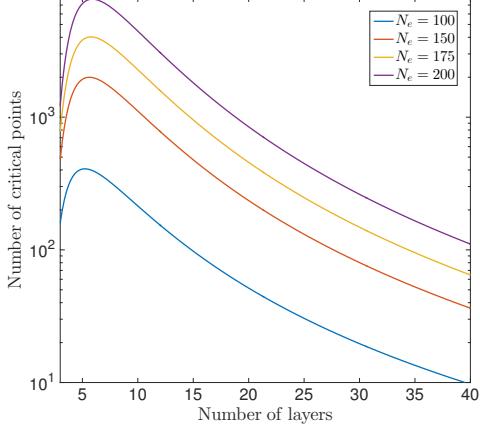


FIG. 2. The number of critical points \mathcal{N} in a deep neural network decreases as a function of depth for fixed number of parameters N_e .

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 non-optimisable [8]. The deep learning boon began when there were sufficient computational resources to train a very deep neural network.

Location of minima: Having considered how many critical points are there in a deep neural network, we next consider where are critical points located in weight space. Intuitively, the more clustered they are, the easier it is for an optimiser to search for minima. Let $\text{Crt}(-\infty, \mathcal{E})$ 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 $\text{Crt}(-\infty, \mathcal{E})$ with relative angle $\mathbf{w} \cdot \mathbf{w}'/\Lambda$ contained in I . This set will be denoted by $[\text{Crt}((-\infty, \mathcal{E}), I)]_2$. 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 upper bound to distance between minima is

$$\begin{aligned} & \limsup_{\Lambda \rightarrow \infty} \frac{1}{\Lambda} \log \left(\frac{\mathbb{E} |[\text{Crt}((-\infty, \mathcal{E}), I)]_2|}{\mathbb{E} |\text{Crt}(-\infty, \mathcal{E})|} \right) \\ & \leq \sup_{r \in I} \sup_{v \in (-\infty, \mathcal{E}/p)} \Psi_H(r, v, \mathcal{E}) \end{aligned} \quad (6)$$

where

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

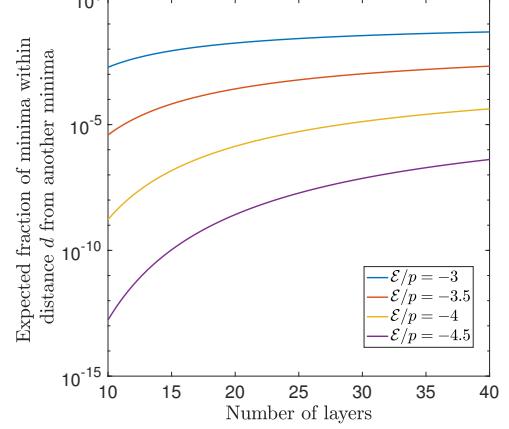


FIG. 3. Minima are more clustered for deeper networks. The figure shows the relative expected number of critical points (6) that attains a loss function value in the interval $(-\infty, \Lambda\mathcal{E})$ with $\|\mathbf{w} - \mathbf{w}'\|_2^2 \leq d\Lambda$ with $d = 0.02$ for fixed number of network parameters $N_e = 400$.

$\Sigma(r) = -\frac{1}{H} \begin{pmatrix} b_1(r) & b_2(r) \\ b_2(r) & b_1(r) \end{pmatrix}$ is a matrix defined by

$$\alpha_H(r) = (H - H(r^H - (H-1)(r^{H-2} - r^H))^2)^{-1},$$

$$b_1(r) = -H + \alpha_H(r)H^3(r^{2H-2} - r^{2H}), \text{ and}$$

$$b_2(r) = -Hr^H - \alpha_H(r)H^3r^{3H-4}(r^2 + H(r^2 - 1)^2 - 1).$$

Proof: The full proof is in the SM. Our proof strategy combines the asymptotics for the minima of the Hamiltonian [29, Theorem 10] with the upper bound on the angle between minima [29, Theorem 5 and Lemma 6].

Figure 3 shows that the number of minima, relative to the total number of minima, that are close to other minima (c.f. Equation (6)) increases as the number of layers increases. In other words, minima are more clustered for deeper networks, thus deep networks are more optimisable 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: Having shown that there are less minima in deep networks and the minima are closer together, we 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 $S(\mathbf{W}_q) = -\log \det(\text{Hess}(\mathcal{L}(\mathbf{W}_q)))$, with Hess being the Hessian matrix [23, 24, 30]. Within the harmonic approximation, larger entropy corresponds to larger basin volume. Intuitively, if wider minima are also deeper, then the function is easy to optimise, whereas functions with deep and narrow minima are difficult to optimise.

The expected entropy of the Hessian of the minima of

loss function that takes value $\Lambda\mathcal{E}$ satisfies asymptotically

$$\begin{aligned} \mathbb{E}(S(\text{Hess } \mathcal{L})|\Lambda\mathcal{E}) &\simeq \\ &= (\Lambda - 1) \log(p) + \frac{\Lambda - 1}{2} \log\left(\frac{\Lambda}{2(\Lambda - 1)H(H - 1)}\right) \\ &\quad - \frac{\Lambda - 1}{\pi} \int_{-\sqrt{2}}^{\sqrt{2}} \log\left|\sigma\sqrt{\frac{\Lambda}{\Lambda - 1}}\frac{\mathcal{E}}{p} - t\right| \sqrt{2 - t^2} dt. \end{aligned} \quad (7)$$

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 $G = \sigma\sqrt{\frac{\Lambda}{\Lambda - 1}}E/p$ 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 $X = \frac{\sigma\mathcal{H}_\Lambda}{\sqrt{\Lambda(\Lambda - 1)}}$ 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]

$$\begin{aligned} \mathbb{E}(S(\text{Hess } \mathcal{H}_\Lambda)|M_{\mathcal{H}_\Lambda}(\Lambda E/p)) \\ = \frac{\mathbb{E}\left(\mathbb{E}\left(S(\text{Hess } \mathcal{H}_\Lambda)1_{M_{\mathcal{H}_\Lambda}(\Lambda E/p)}|\{\mathcal{H}_\Lambda\}\right)\right)}{\mathbb{E}(\mathbb{P}(M^{\Lambda-1} \geq X, X \in G|\{X\}))}. \end{aligned} \quad (8)$$

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

$$\begin{aligned} \mathbb{E}\left(\mathbb{E}\left(S(\text{Hess } \mathcal{H}_\Lambda)1_{M_{\mathcal{H}_\Lambda}(\Lambda E/p)}|\{\mathcal{H}_\Lambda\}\right)\right) \\ \simeq -\frac{\Lambda - 1}{\pi} \int_{-\sqrt{2}}^{\sqrt{2}} f_{-\sqrt{2}}(t) \sqrt{2 - t^2} dt \\ + \frac{\Lambda - 1}{2} \log\left(\frac{\Lambda}{2(\Lambda - 1)H(H - 1)}\right) \mathbb{P}(M_{\mathcal{H}_\Lambda}(\Lambda E/p)). \end{aligned} \quad (9)$$

For the denominator in (8), 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 [31, Theorem 1]. Hence, it follows that $\mathbb{E}(\mathbb{P}(M^{\Lambda-1} \geq X, X \in G|\{X\})) = \sqrt{\frac{\Lambda - 1}{2\pi\sigma^2}} \int_G e^{-\frac{t^2(\Lambda - 1)}{2\sigma^2}} dt$. Having obtained asymptotic expressions for both the numerator and denominator in (8), we take the limit $\varepsilon \downarrow 0$ such that the energy interval E shrinks down to a single energy value \mathcal{E} such that (7) follows immediately.

Figure 4 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 non-trivial and unlike many

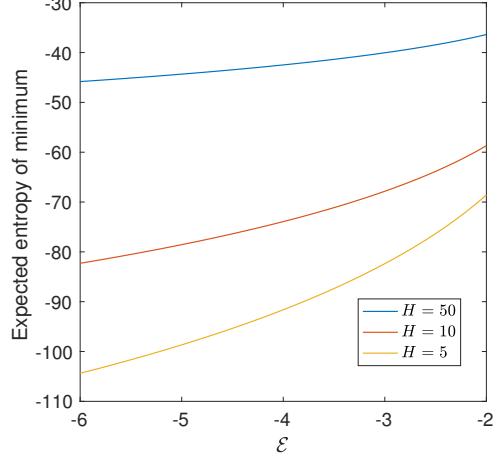


FIG. 4. Energy-entropy competition is eased by increasing network depth. The expected entropy at minima of the loss function as a function of minima depth for $N_e = 400$ network parameters and $p = 0.8$.

atomic cluster systems analysed in the literature [32–34], where the lower minima have larger basins of attraction. However, this competition is smoothed 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 minimisation algorithms getting trapped in wide but very suboptimal local minima.

To verify our analytical results, we consider a classical set of 10 benchmark datasets [35, 36]. Figure 5 shows the results for one dataset (results for the remaining datasets, shown in the SM, 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 [37, 38], 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 22 nodes each and the deep network comprises 6 hidden layers with 22 nodes each, such that the total number of parameters is 726. Further details are discussed in the SM.

In summary, we derived a series of analytical results showing that deep networks are more optimisable than shallow networks because there are less critical points, the minima are more clustered, and the energy-entropy tradeoff 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

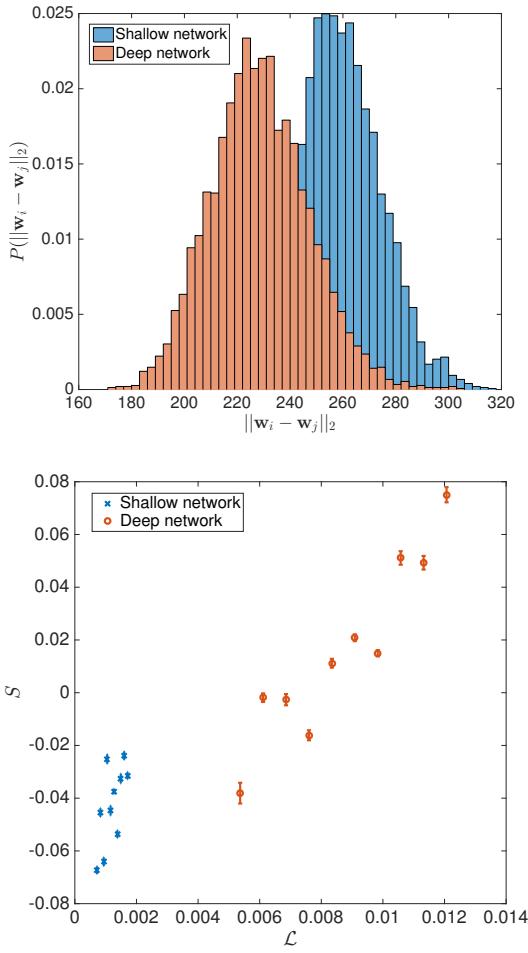


FIG. 5. Numerical experiments agree qualitatively with the analytical predictions. Top: The histogram of distances between minima. Minima in deeper networks are closer together. Bottom: The loss function at minima plotted against the expected entropy of minima. Lower minima are narrower but this energy-entropy tradeoff is less severe for deep networks. The figures are plotted for the “Boston Housing” dataset, c.f. [35, 36]. To compute the expected value of entropy, we discretise the distribution of values that that loss function takes into 10 bins.

of optimisation, as well as suggests new design principles. For example, the most optimisable 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 [32–34] holds the key to engineering such architectures.

This work was supported by the EPSRC grant EP/L016516/1 for the University of Cambridge CDT, the CCA (S.B.). AAL acknowledges support from the Winton Programme for the Physics of Sustainability.

* aal44@cam.ac.uk

- [1] F. Leclercq, A. Pisani, and B. D. Wandelt, *New horizons for observational cosmology*, 189 (2014).
- [2] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna, *PLoS computational biology* **3**, e189 (2007).
- [3] Y. LeCun, Y. Bengio, and G. Hinton, *nature* **521**, 436 (2015).
- [4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, 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 preprint arXiv:1609.08144 (2016).
- [6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., *International Journal of Computer Vision* **115**, 211 (2015).
- [7] G. Cybenko, *Mathematics of control, signals and systems* **2**, 303 (1989).
- [8] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, Vol. 1 (MIT press Cambridge, 2016).
- [9] G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio, in *Advances in neural information processing systems* (2014) pp. 2924–2932.
- [10] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, in *Advances in neural information processing systems* (2016) pp. 3360–3368.
- [11] T. Poggio, H. Mhaskar, L. Rosasco, B. Miranda, and Q. Liao, *International Journal of Automation and Computing* **14**, 503 (2017).
- [12] P. Mehta and D. J. Schwab, arXiv preprint arXiv:1410.3831 (2014).
- [13] H. W. Lin, M. Tegmark, and D. Rolnick, *Journal of Statistical Physics* **168**, 1223 (2017).
- [14] J. Ba and R. Caruana, in *Advances in neural information processing systems* (2014) pp. 2654–2662.
- [15] Y. V. Fyodorov, *Physical review letters* **92**, 240601 (2004).
- [16] A. J. Bray and D. S. Dean, *Physical review letters* **98**, 150201 (2007).
- [17] Y. V. Fyodorov and I. Williams, *Journal of Statistical Physics* **129**, 1081 (2007).
- [18] A. Auffinger, G. B. Arous, et al., *The Annals of Probability* **41**, 4214 (2013).
- [19] A. Auffinger, G. B. Arous, and J. Černý, *Communications on Pure and Applied Mathematics* **66**, 165 (2013).
- [20] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, in *Advances in neural information processing systems* (2014) pp. 2933–2941.
- [21] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, in *Artificial Intelligence and Statistics* (2015) pp. 192–204.
- [22] A. Choromanska, Y. LeCun, and G. B. Arous, in *Conference on Learning Theory* (2015) pp. 1756–1760.
- [23] R. Das and D. J. Wales, *Physical Review E* **93**, 063310 (2016).
- [24] A. J. Ballard, J. D. Stevenson, R. Das, and D. J. Wales, *The Journal of chemical physics* **144**, 124119 (2016).
- [25] A. J. Ballard, R. Das, S. Martiniani, D. Mehta, L. Sagun, J. D. Stevenson, and D. J. Wales, *Physical Chemistry*

- Chemical Physics **19**, 12585 (2017).
- [26] D. Mehta, X. Zhao, E. A. Bernal, and D. J. Wales, Physical Review E **97**, 052307 (2018).
- [27] I. Shafarevich, *Basic algebraic geometry* (Springer-Verlag, 1977).
- [28] D. Cartwright and B. Sturmfels, Linear Algebra and its Applications **438**, 942 (2013).
- [29] E. Subag, The Annals of Probability **45**, 3385 (2017).
- [30] Y. Zhang, A. M. Saxe, M. S. Advani, and A. A. Lee, Molecular Physics , 1 (2018).
- [31] M. Ledoux and B. Rider, Institute of Mathematical Statistics **15**, 942 (2010).
- [32] J. P. Doye, D. J. Wales, and M. A. Miller, The Journal of Chemical Physics **109**, 8143 (1998).
- [33] J. P. Doye and C. P. Massen, The Journal of chemical physics **122**, 084105 (2005).
- [34] C. P. Massen and J. P. Doye, Physical Review E **75**, 037101 (2007).
- [35] J. M. Hernández-Lobato and R. Adams, in *International Conference on Machine Learning* (2015) pp. 1861–1869.
- [36] Y. Gal and Z. Ghahramani, in *international conference on machine learning* (2016) pp. 1050–1059.
- [37] S. Martiniani, K. J. Schrenk, J. D. Stevenson, D. J. Wales, and D. Frenkel, Physical Review E **93**, 012906 (2016).
- [38] S. Martiniani, K. J. Schrenk, J. D. Stevenson, D. J. Wales, and D. Frenkel, Physical Review E **94**, 031301 (2016).