Exploring the Function Space of Deep-Learning Machines

Bo Li[1, *] and David Saad[2, †]

[1] *Department of Physics, The Hong Kong University of Science and Technology, Hong Kong*
[2] *Non-linearity and Complexity Research Group, Aston University, Birmingham B4 7ET, United Kingdom*

The function space of deep-learning machines is investigated by studying growth in the entropy of functions of a given error with respect to a reference function, realized by a deep-learning machine. Using physics-inspired methods we study both sparsely and densely-connected architectures to discover a layer-wise convergence of candidate functions, marked by a corresponding reduction in entropy when approaching the reference function, gain insight into the importance of having a large number of layers, and observe phase transitions as the error increases.

Deep-learning machines (DLM) have both fascinated and bewildered the scientific community and have given rise to an active and ongoing debate [1]. They are carefully structured layered networks of non-linear elements, trained on data to perform complex tasks such as speech recognition, image classification, and natural language processing. While their phenomenal engineering successes [2] have been broadly recognized, their scientific foundations remain poorly understood, particularly their good ability to generalize well from a limited number of examples with respect to the degrees of freedom [3–5] and the nature of the layer-wise internal representations [6, 7].

Supervised learning in DLM is based on the introduction of example pairs of input and output patterns, which serve as constraints on space of candidate functions. As more examples are introduced the function space monotonically decreases. Statistical physics methods have been successful in gaining insight into both pattern-storage [8] and learning scenarios, mostly in single layer machines [9] but also in simple two-layer scenarios [10, 11]. However, extending these methods to DLM is difficult due to the recursive application of non-linear functions in successive layers and the undetermined degrees of freedom in intermediate layers. While training examples determine both input and output patterns, the constraint imposed on hidden-layer representations are difficult to pin down. These constitute the main difficulties for a better understanding of DLM.

In this Letter, we propose a general framework for analyzing DLM by mapping them onto a dynamical system and by employing the Generating Functional (GF) approach to analyze their typical behavior. More specifically, we investigate the landscape in function space around a reference function by perturbing its parameters (weights in the DLM setting), and quantifying the entropy of the corresponding functions space for a given level of error with respect to the reference function. This provides a measure for the abundance of nearly-perfect solutions and hence an indication for the ability to obtain good approximations using DLM. The function error measure is defined as the expected difference (Hamming distance in the discrete case) between the perturbed and reference functions' outputs given the same input (additional explanation is provided in [12]). This setup is reminiscent of the teacher-student scenario, commonly used in the neural networks literature [13] where the average error serves as a measure of distance between the perturbed and reference network in function space. For certain classes of reference networks, we obtain closed form solutions of the error as a function of perturbation on each layer, and consequently the weight-space volume for a given level of function error. By the virtue of supervised learning and constraints imposed by the examples provided, high-error functions will be ruled out faster than those with low errors, such that the candidate function space is reduced and the concentration of low-error functions increases. A somewhat similar approach, albeit based on recursive mean field relations between each two consecutive layers separately, has been used to probe the expressivity of DLM [14].

Through the GF framework and entropy maximization, we analyze the typical behavior of different classes of models including networks with continuous and binary parameters (weights) and different topologies, both fully and sparsely connected. We find that as one lowers the error level, typical functions gradually better match the reference network *starting from earlier layers to later ones*. More drastically, for fully connected binary networks, weights in earlier layers of the perturbed functions will perfectly match those of the reference function, implying a possible successive layer by layer learning behavior. Sparsely connected topologies exhibit phase transitions with respect to the number of layers, by varying the magnitude of perturbation, similar to the phase transitions in noisy Boolean computation [15], which support the need of deep networks for improving generalization.

Densely connected network models–The model considered here comprises two coupled feed-forward DLM as illustrated in Fig. 1, one of which serves as the reference function and the other is obtained by perturbing the reference network parameters. We first consider the densely connected networks. Each network is composed of $L+1$ layers of N neurons each. The reference function is parameterized by $N^2 \times L$ weight variables $\hat{w}^l_{ij}, \forall\ l=1,2,...,L,\ \ i,j=1,2,...,N$ and maps an N-dimensional input $\hat{\boldsymbol{s}}^0 \in \{-1,1\}^N$ to an N-dimensional

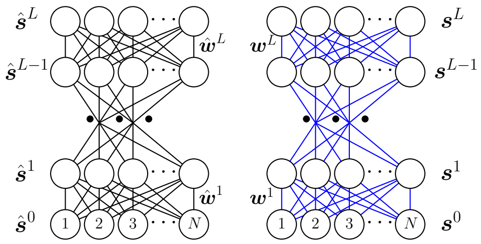

Figure 1. (Color online). The model of two coupled DLMs. The reference and perturbed functions are denoted by $\{\hat{\boldsymbol{w}}^l\}$ (black edges) and $\{\boldsymbol{w}^l\}$ (blue edges), respectively.

output $\hat{\boldsymbol{s}}^L \in \{-1,1\}^N$, through intermediate-layer internal representations and according to the stochastic rule

$$P(\hat{\boldsymbol{s}}^L|\hat{\boldsymbol{w}},\hat{\boldsymbol{s}}^0) = \prod_{l=1}^{L} P(\hat{\boldsymbol{s}}^l|\hat{\boldsymbol{w}}^l,\hat{\boldsymbol{s}}^{l-1}). \tag{1}$$

The i-th neuron in the l-th layer experiences a local field $\hat{h}_i^l(\hat{\boldsymbol{w}}^l,\hat{\boldsymbol{s}}^{l-1}) = \frac{1}{\sqrt{N}}\sum_j \hat{w}_{ij}^l \hat{s}_j^{l-1}$, and its state is determined by the conditional probability

$$P(\hat{s}_i^l|\hat{\boldsymbol{w}}^l,\hat{\boldsymbol{s}}^{l-1}) = \frac{e^{\beta \hat{s}_i^l \hat{h}_i^l(\hat{\boldsymbol{w}}^l,\hat{\boldsymbol{s}}^{l-1})}}{2\cosh\left[\beta \hat{h}_i^l(\hat{\boldsymbol{w}}^l,\hat{\boldsymbol{s}}^{l-1})\right]}, \tag{2}$$

where the temperature β quantifies the strength of thermal noise. In the noiseless limit $\beta \to \infty$, node i represents a perceptron $\hat{s}_i^l = \mathrm{sgn}(\hat{h}_i^l)$ and Eq. (1) corresponds to a deterministic neural network with a *sign* activation function. The perturbed network operates in the same manner, but the weights w_{ij}^l are obtained by applying independent perturbation to each of the reference weights; the perturbed weights w_{ij}^l, give rise to a function that is correlated with the reference function.

We focus on the similarity between reference and perturbed functions outputs for randomly sampled input patterns $\boldsymbol{s}^0 = \hat{\boldsymbol{s}}^0$, drawn from some distribution $P(\hat{\boldsymbol{s}}^0)$. Considering the joint probability of the two systems

$$P\left[\{\hat{\boldsymbol{s}}^l\},\{\boldsymbol{s}^l\}\right] = P(\hat{\boldsymbol{s}}^0)\prod_{i=1}^{N}\delta_{s_i^0,\hat{s}_i^0} \tag{3}$$
$$\prod_{l=1}^{L} P(\hat{\boldsymbol{s}}^l|\hat{\boldsymbol{w}}^l,\hat{\boldsymbol{s}}^{l-1})P(\boldsymbol{s}^l|\boldsymbol{w}^l,\boldsymbol{s}^{l-1}),$$

where the weight parameters $\{\hat{w}_{ij}^l\}$ and $\{w_{ij}^l\}$ are quenched disordered variables. We consider two cases, where the weights are continuous or discrete variables drawn from the Gaussian and Bernoulli distributions, respectively. The quantity of interests are the overlaps between the two functions at the different layers $q^l(\hat{\boldsymbol{w}},\boldsymbol{w}) \equiv \frac{1}{N}\sum_i \langle \hat{s}_i^l s_i^l\rangle$, where angled brackets $\langle\cdots\rangle$ denote the average over the joint probability $P[\{\hat{\boldsymbol{s}}^l\},\{\boldsymbol{s}^l\}]$. The N outputs represent N weakly coupled Boolean functions of the same form of disordered, and thus share the same average behavior.

The form of probability distribution (3) is analogous to the dynamical evolution of disordered Ising spin systems [16] if the layers are viewed as discrete time steps of parallel dynamics. We therefore apply the GF formulation from statistical physics to these deep feed-forward functions similarly to the approach used to investigate random Boolean formulae [15]. We compute the GF $\Gamma[\hat{\boldsymbol{\psi}},\boldsymbol{\psi}] = \left\langle e^{-\mathrm{i}\sum_{l,i}(\hat{\psi}_i^l\hat{s}_i^l + \psi_i^l s_i^l)}\right\rangle$, from which the moments can be calculated, e.g., $q^l(\hat{\boldsymbol{w}},\boldsymbol{w}) = -\frac{1}{N}\sum_i \lim_{\hat{\boldsymbol{\psi}},\boldsymbol{\psi}\to 0}\frac{\partial^2}{\partial\hat{\psi}_i^l\partial\psi_i^l}\Gamma[\hat{\boldsymbol{\psi}},\boldsymbol{\psi}]$. Assuming the systems are self-averaging for $N\to\infty$ and computing the disorder average (denoted by the upper line) $\overline{\Gamma[\hat{\boldsymbol{\psi}},\boldsymbol{\psi}]}$, the disorder-averaged overlaps can be obtained $q^l = \frac{1}{N}\sum_{i=1}\overline{\langle \hat{s}_i^l s_i^l\rangle}$. For convenience, we introduce the field doublet $H^l \equiv [\hat{h}^l, h^l]^T$. Expressing the GF $\Gamma[\hat{\boldsymbol{\psi}},\boldsymbol{\psi}]$ by macroscopic order parameters and averaging over the disorder yields the saddle-point integral $\overline{\Gamma} = \int\{d\boldsymbol{q}d\boldsymbol{\mathcal{Q}}\}e^{N\Psi[\boldsymbol{q},\boldsymbol{\mathcal{Q}}]}$ where $\Psi[...]$ is [12]

$$\Psi = \mathrm{i}\sum_{l=0}^{L}\mathcal{Q}^l q^l + \log\int\prod_{l=1}^{L}d\hat{h}^l dh^l \sum_{\{\hat{s}^l,s^l\}} M[\hat{s},s,\hat{h},h], \tag{4}$$

and the effective single site measure $M[...]$ has the following form for both continuous and binary weights

$$M[\hat{s},s,\hat{h},h] = P(\hat{s}^0)\delta_{\hat{s}^0,s^0}e^{-\mathrm{i}\sum_{l=0}^{L}\mathcal{Q}^l\hat{s}^l s^l}$$
$$\times\prod_{l=1}^{L}\left\{\frac{e^{\beta\hat{s}^l\hat{h}^l}}{2\cosh\beta\hat{h}^l}\frac{e^{\beta s^l h^l}}{2\cosh\beta h^l}\frac{e^{-\frac{1}{2}(H^l)^T\cdot\Sigma_l^{-1}\cdot H^l}}{\sqrt{(2\pi)^2|\Sigma_l(q^{l-1})|}}\right\}. \tag{5}$$

The Gaussian density of the local field $\{\hat{h}^l, h^l\}$ in (5) comes from summing a large number of random variables in $\hat{h}^l$ and h^l. The precision matrix Σ_l^{-1}, linking the effective field $\hat{h}^l$ and h^l, measures the correlation between internal fields of the two systems and depends on the overlap q^{l-1} of the previous layer. In the limit $N\to\infty$ the GF $\overline{\Gamma}$ is dominated by the extremum of Ψ. Variation with respect to $\mathcal{Q}^l$ gives rise to saddle-point equations of the order parameters $q^l = \langle \hat{s}^l s^l\rangle_{M[...]}$, where the average is taken over the measure $M[...]$ of (5). The conjugate order parameter $\mathcal{Q}^l$, ensuring the normalization of the measure, vanishes identically. It leads to the evolution equation [12]

$$q^l = \int d\hat{h}^l dh^l \tanh(\beta\hat{h}^l)\tanh(\beta h^l)\frac{e^{-\frac{1}{2}(H^l)^T\cdot\Sigma_l^{-1}\cdot H^l}}{\sqrt{(2\pi)^2|\Sigma_l|}}. \tag{6}$$

The overlap evolution is somewhat similar to dynamical mean field relation in [14], but the objects investigated and the remainder of the study are different. We focus on the function-space landscape rather than the sensitivity of function to input perturbations.

Densely connected continuous weights–In the first scenario, we assume weight variables $\hat{w}_{ij}^l$ to be independently drawn from a Gaussian density $\mathcal{N}(0,\sigma^2)$ and the perturbed weights to have the form $w^l_{ij} = \sqrt{1-(\eta^l)^2}\hat{w}^l_{ij} + \eta^l \delta w^l_{ij}$, where δw^l_{ij} are drawn from $\mathcal{N}(0,\sigma^2)$ independently of $\hat{w}^l_{ij}$. It ensures that w^l_{ij} has the same variance σ^2. The parameter η^l quantifies the strength of perturbation introduced in layer l. In this case the covariance matrix between the local fields $\hat{h}^l$ and h^l takes the form

$$\Sigma_l(\eta^l, q^{l-1}) = \sigma^2 \begin{bmatrix} 1 & \sqrt{1-(\eta^l)^2}q^{l-1} \\ \sqrt{1-(\eta^l)^2}q^{l-1} & 1 \end{bmatrix}, \tag{7}$$

leading to the close form solution of the overlap as $\beta \to \infty$,

$$q^l = \frac{2}{\pi}\sin^{-1}\left(\sqrt{1-(\eta^l)^2}q^{l-1}\right). \tag{8}$$

Of particular interest is the final-layer overlap given the same input for the two system under specific perturbations $q^L(\{\eta^l\}, q^0 = 1)$. The average error $\varepsilon = \frac{1}{2}(1-q^L)$ measures the typical distance between the two mappings.

The number of solutions at a given distance (error) ε away from the reference function is indicative of how difficult it is to obtain this level of approximation at the vicinity of the exact function. Let the N-dimensional vectors $\hat{\boldsymbol{w}}^{l,i}$ and $\boldsymbol{w}^{l,i}$ denote the weights of the i-th perceptron of the reference and perturbed systems at layer l, respectively; the expected angle between them is $\theta^l = \sin^{-1}\eta^l$. Then the perceptron $\boldsymbol{w}^{l,i}$ occupies on average an angular volume around $\hat{\boldsymbol{w}}^{l,i}$ as $\Omega(\eta^l) \sim \sin^{N-2}\theta^l = (\eta^l)^{N-2}$ [17, 18]. The total weight-space volume of the perturbed system is $\Omega_{\mathrm{tot}}(\{\eta^l\}) = \prod_{l=1}^{L}\prod_i(\eta^l)^{N-2}$, and the corresponding entropy density is

$$S_{\mathrm{con}}(\{\eta^l\}) = \frac{1}{LN^2}\log\Omega_{\mathrm{tot}}(\{\eta^l\}) \approx \frac{1}{L}\sum_{l=1}^{L}\log\eta^l. \tag{9}$$

In the thermodynamic limit $N \to \infty$, the set of perturbed functions at distance ε away from the reference function is dominated by those with perturbation vector $\{\eta^{*l}\}$, which maximizes the entropy $S_{\mathrm{con}}(\{\eta^l\})$ subject to the constraint $q^L(\{\eta^l\}) = 1-2\varepsilon$. The result of $\{\eta^{*l}\}$ for a four-layer network, shown in Fig. 2(a), reveals that the dominant perturbation η^{*l} to the reference network decays faster for smaller l values; this indicates that closer to the reference function, solutions are dominated by functions where early-layer weights match better the reference network. Consequently, high-ε function are ruled out faster during training through the successful alignment of earlier layers, resulting in the increasing concentration of low-ε functions and better generalization. We denote the maximal weight-space volume at distance ε away from the reference function as $\Omega_0(\varepsilon) \equiv \Omega_{\mathrm{tot}}(\{\eta^{*l}\})$.

Supervised learning is based on the introduction of input-output example pairs. Introducing constraints, in the form of $P \equiv \alpha L N^2$ examples provided by the reference function, the weight-space volume at small distance ε away from the reference function is re-shaped as $\Omega_\alpha(\varepsilon) = \Omega_0(\varepsilon)(1-\varepsilon)^P$ in the annealed approximation [17, 18]; details of the derivation can be found in [12]. The typical distance $\varepsilon^*(\alpha) = \mathrm{argmax}_\varepsilon \Omega_\alpha(\varepsilon)$ can be interpreted as the generalization error in the presence of P examples, giving rise to an approximate generalization curves shown in Fig. 2(c). These are expected to be valid in the small ε (large α) limit on which the perturbation analysis is based. It is observed that typically a large number of examples ($\alpha \gg 10$) are needed for good generalization. This may imply that DLMs trained on realistic data sets (usually $\alpha \ll 1$) occupy a small, highly-biased subspace, different from the typical function space analyzed here (e.g., the handwritten digit MNIST database represents highly biased inputs that occupy a very small fraction of the input space). Note that the results correspond to a typical generalization performance under the assumption of self-averaging, potentially with unlimited computational resources and independently of the training rule used.

Densely connected binary weights–Once trained, networks with binary weights are highly efficient computationally, which is especially useful in devices with limited memory or computational resources [19, 20]. Here we consider a reference network with binary weight variables drawn from the distribution $P(\hat{w}^l_{ij}) = \frac{1}{2}\delta_{\hat{w}^l_{ij},1} + \frac{1}{2}\delta_{\hat{w}^l_{ij},-1}$, while the perturbed network weights follow the distribution $P(w^l_{ij}) = (1-p^l)\delta_{w^l_{ij},\hat{w}^l_{ij}} + p^l\delta_{w^l_{ij},-\hat{w}^l_{ij}}$, where p^l is the flipping probability at layer l. The covariance matrix

$$\Sigma_l(p^l, q^{l-1}) = \begin{bmatrix} 1 & (1-2p^l)q^{l-1} \\ (1-2p^l)q^{l-1} & 1 \end{bmatrix}, \tag{10}$$

gives rise to overlaps q^l as $\beta \to \infty$ of the form

$$q^l = \frac{2}{\pi}\sin^{-1}\left((1-2p^l)q^{l-1}\right). \tag{11}$$

The entropy density of the perturbed system is given by

$$S_{\mathrm{bin}}(\{p^l\}) = \frac{1}{L}\sum_{l=1}^{L} -p^l\log p^l - (1-p^l)\log(1-p^l). \tag{12}$$

Similarly, the entropy $S_{\mathrm{bin}}(\{p^l\})$ is maximized by the perturbation vector $\{p^{*l}\}$ subject to $q^L(\{p^l\}) = 1-2\varepsilon$ at a distance ε away from the reference function. The result of $\{p^{*l}\}$ for a four-layer binary neural network is shown in Fig. 2(b). Surprisingly, as ε decreases, the first-layer weights are first to align perfectly with those of the reference function followed by the second-layer weights and so on. The discontinuities come from the non-convex nature of the entropy landscape $S_{\mathrm{bin}}(\{p^l\})$ when one restricts the perturbed system to the nonlinear ε-error surface satisfying $q^L(\{p^l\}) = 1-2\varepsilon$. Nevertheless, there exists many more high-ε than low-ε functions for densely-connected binary networks (as indicated by the entropy shown in the inset of Fig. 2(b)), and it remains to explore how low generalization error functions could be identified.

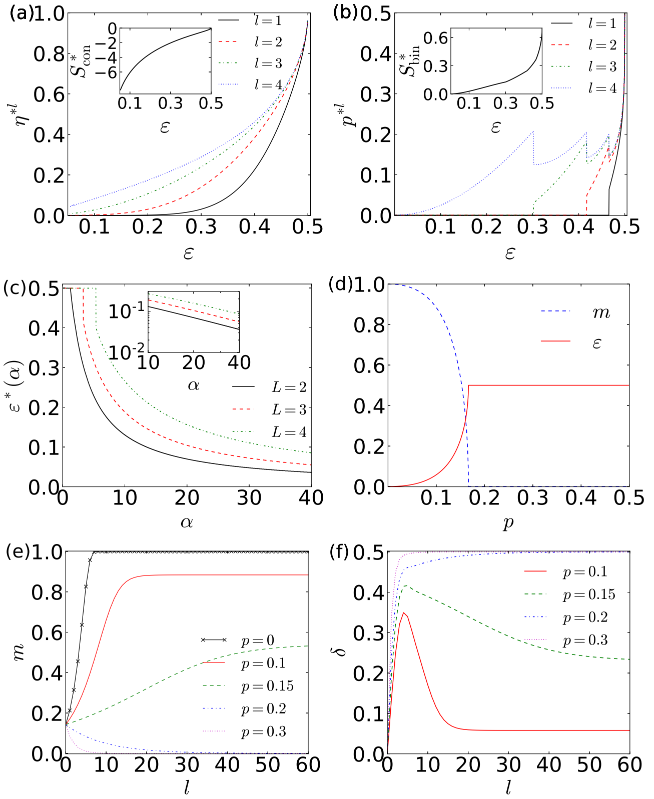

Figure 2. (Color online) Maximal-entropy perturbations as a function of output error ε for a four-layer densely-connected networks with (a) continuous weights and (b) binary weights. Inset represents the growth in entropy with respect to ε. (c) Generalization curves of densely-connected networks with continuous weights by using the annealed approximation. Inset demonstrates the classical asymptotic behavior of $\varepsilon^* \sim \alpha^{-1}$ in the large α limit [18]. (d) Stationary magnetization m and function error ε for sparsely connected MAJ-3 based DLM as a function of perturbation probability p in networks with binary weights. We show the evolution of (e) magnetization and (f) internal activation error δ over layers. Note that $p=0$ corresponds to the reference network. All results are obtained in the deterministic limit $\beta \to \infty$.

Sparsely connected binary weights–Lastly, we consider the sparsely connected DLM with binary weights; these topologies are of interest to practitioners due to the reduction in degrees of freedom and their computational and energy efficiency. The layered setup is similar to the previous case, except that unit i at layer l is randomly connected to a small number k of units in layer $(l–1)$ and its local field is given by $\hat{h}_i^l(\hat{\boldsymbol{w}}^l, \hat{\boldsymbol{s}}^{l-1}) = \frac{1}{\sqrt{k}} \sum_j A_{ij}^l \hat{w}_{ij}^l \hat{s}_j^{l-1}$ where the adjacency matrix A^l represents the connectivity between the two layers. The perturbed network has the same topology but its weights are randomly flipped $P(w_{ij}^l) = (1–p^l)\delta_{w_{ij}^l, \hat{w}_{ij}^l} + p^l \delta_{w_{ij}^l, -\hat{w}_{ij}^l}$; the activation and the joint probability of the two systems follow from (2) and (3). Unlike the case of densely-connected networks, the magnetization $m^l \equiv \frac{1}{N} \sum_i s_i^l$ also plays at important role in the evolution of sparse networks. The GF approach gives rise to the order parameter $\mathcal{P}^l(\hat{s}, s) \equiv \frac{1}{N} \sum_i \delta_{\hat{s}_i^l, \hat{s}} \delta_{s_i^l, s}$ relating to the magnetization and overlap by $\mathcal{P}^l(\hat{s}, s) = \frac{1}{4}(1 + \hat{s}\hat{m}^l + s m^l + \hat{s} s q^l)$.

The random topology provides an additional disorder to average over. For simplicity, we assign the reference weights to $\hat{w}_{ij}^l = 1$, which in the limit $\beta \to \infty$ relate to the k-majority gate (MAJ-k) based Boolean formulas that provide all Boolean functions with uniform probability at the large L limit [21, 22]. For a uniform perturbation over layers $p^l = p$ we focus on functions generated in the deep regime $L \to \infty$, where the order parameters take the form

$$m^l = \sum_{\{s_j\}} \prod_{j=1}^{k} \frac{1}{2} \left[1 + s_j m^{l-1}(1-2p)\right] \mathrm{sgn}\left[\sum_{j=1}^{k} s_j\right], \quad (13)$$

$$q^l = \sum_{\{s_j, \hat{s}_j\}} \prod_{j=1}^{k} \frac{1}{4} \left[1 + \hat{s}_j \hat{m}^{l-1} + s_j m^{l-1}(1-2p) \right. \quad (14)$$

$$\left. + s_j \hat{s}_j q^{l-1}(1-2p)\right] \mathrm{sgn}\left[\sum_{j=1}^{k} \hat{s}_j\right] \mathrm{sgn}\left[\sum_{j=1}^{k} s_j\right].$$

For finite k, the macroscopic observables at layer l are polynomially dependent on the observables at layer $(l-1)$ up to order k. In the limit $L \to \infty$, the Boolean functions generated depend on the initial magnetization $m^0 = \frac{1}{N} \sum_i s_i^0$. Here, we consider biased case with initial conditions $\hat{m}^0 = m^0 > 0$ and $q^0 = 1$. The reference function admits a stationary solution $\hat{m}^\infty = 1$, computing a 1-bit information-preserving majority function [22]. Both magnetization of the perturbed function m^∞ and the function error $\varepsilon = \frac{1}{2}(1 - q^\infty)$ exhibit a transition from the ordered phase to the paramagnetic phase at some critical perturbation level p_c, below which the perturbed network computes the reference function with error $\varepsilon < \frac{1}{2}$. The results for $k=3$ are shown in Fig. 2(d). Interestingly, the critical perturbation p_c coincides with the location of the critical thermal noise $\epsilon_c = \frac{1}{2}(1 - \tanh \beta_c)$ for noisy k-majority gate-based Boolean formulas; for $k = 3$, the critical perturbation $p_c = \frac{1}{6}$ [15]. Below p_c, there exist two ordered states with $m^\infty = \pm\sqrt{(1-6p)/(1-2p)^3}$ and the overlap satisfies $q^\infty = m^\infty$ [12], which is also reminiscent of the thermal noise-induced solutions [15]. However, the underlying physical implications are drastically different. Here it indicates that even in the deep network regime, there exists a large number $\binom{Nk}{Nkp}^L$ of networks that can reliably represent the reference function when $p < p_c$. This function landscape is important for learning tasks to achieve a similar rule to the reference function. The propagation of internal error $\delta(l) \equiv \frac{1}{2}(1 - q^l)$, shown in Fig. 2(f), exhibits a stage of error-increase followed by a stage of error-decrease for $p < p_c$. Consequently *a successful sparse DLM requires more layers to reduce errors and provide a higher similarity to the reference function* *when we approach* p_c, indicating the need of deep networks in such models.

In summary, we propose a GF analysis to probe the function landscapes of DLM, focusing on the entropy of functions, given their error with respect to a reference function. The entropy maximization of densely connected networks at fixed error to the reference function indicates that weights of earlier layers are the first to align with reference function parameters when the error decreases. It highlights the importance of early-layer weights for reliable computation [23] and sheds light on the parameter learning-dynamics in function space during the learning process. We also investigate the phase transitions behavior in sparsely-connected networks, which advocate the use of deeper machines for suppressing errors with respect to the reference function in these models. The suggested GF framework is very general and can accommodate other structures and computing elements, e.g., continuous variables, other activation functions (such as the commonly used ReLU activation function [12]) and more complicated weight ensembles. In [12], we also demonstrate the effect of negatively/positively correlated weight variables on the expressive power of networks with ReLU activation and their impact on the function space, and investigate the behavior of simple convolutional DLM. Moreover, the GF framework allows one to investigate other aspect as well, including finite size effects and the use of perturbative expansion to provide a systematic analysis of the interactions between network elements. This is a step towards a principled investigation of the typical behavior of DLM and we envisage follow up work on various aspects of the learning process.

We thank K. Y. Michael Wong for helpful discussions. Support from The Leverhulme Trust (RPG-2013-48) (DS) and Research Grants Council of Hong Kong (Grants No. 605813 and No. 16322616) (BL) is acknowledged.

* bliaf@connect.ust.hk

† d.saad@aston.ac.uk

[1] M. Elad, "Deep, Deep Trouble," siam news, https://sinews.siam.org/Details-Page/deep-deep-trouble (2017).

[2] Y. LeCun, Y. Bengio, and G. Hinton, Nature **521**, 436 (2015).

[3] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in 5th International Conference on Learning Representations, Toulon, France (2017).

[4] C. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina, "Entropy-sgd: Biasing gradient descent into wide valleys," in 5th International Conference on Learning Representations, Toulon, France (2017).

[5] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, in *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017) pp. 5949–5958.

[6] M. D. Zeiler and R. Fergus, in *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, edited by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (Springer International Publishing, Cham, 2014) pp. 818–833.

[7] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," in Deep Learning Workshop, International Conference on Machine Learning (2015).

[8] E. Gardner, Journal of Physics A: Mathematical and General **21**, 257 (1988).

[9] J. Hertz, A. Krogh, and R. Palmer, *Introduction To The Theory Of Neural Computation* (Addison-Wesley, 1991).

[10] T. L. H. Watkin, A. Rau, and M. Biehl, Rev. Mod. Phys. **65**, 499 (1993).

[11] D. Saad and S. A. Solla, Phys. Rev. Lett. **74**, 4337 (1995).

[12] See supplemental material for details.

[13] D. Saad, ed., *On-Line Learning in Neural Networks* (Cambridge University Press, 1998).

[14] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, in *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., 2016) pp. 3360–3368.

[15] A. Mozeika, D. Saad, and J. Raymond, Phys. Rev. Lett. **103**, 248701 (2009).

[16] J. P. L. Hatchett, B. Wemmenhove, I. P. Castillo, T. Nikoletopoulos, N. S. Skantzos, and A. C. C. Coolen, Journal of Physics A: Mathematical and General **37**, 6201 (2004).

[17] H. S. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992).

[18] A. Engel and C. V. d. Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, New York, 2001).

[19] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, arXiv:1602.02830 (2016).

[20] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, in *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV*, edited by B. Leibe, J. Matas, N. Sebe, and M. Welling (Springer International Publishing, Cham, 2016) pp. 525–542.

[21] P. Savický, Discrete Mathematics **83**, 95 (1990).

[22] A. Mozeika, D. Saad, and J. Raymond, Phys. Rev. E **82**, 041112 (2010).

[23] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, in *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 70, edited by D. Precup and Y. W. Teh (PMLR, International Convention Centre, Sydney, Australia, 2017) pp. 2847–2854.