OPEN ACCESS
IOP Publishing
Journal of Physics A: Mathematical and Theoretical
J. Phys. A: Math. Theor. **53** (2020) 104002 (24pp)
https://doi.org/10.1088/1751-8121/ab6a6f

Large deviation analysis of function sensitivity in random deep neural networks

Bo Li and David Saad

Non-linearity and Complexity Research Group, Aston University, Birmingham, B4 7ET, United Kingdom

E-mail: b.li10@aston.ac.uk and d.saad@aston.ac.uk

Received 14 October 2019, revised 21 December 2019
Accepted for publication 10 January 2020
Published 20 February 2020

Abstract

Mean field theory has been successfully used to analyze deep neural networks (DNN) in the infinite size limit. Given the finite size of realistic DNN, we utilize the large deviation theory and path integral analysis to study the deviation of functions represented by DNN from their typical mean field solutions. The parameter perturbations investigated include weight sparsification (dilution) and binarization, which are commonly used in model simplification, for both ReLU and sign activation functions. We find that random networks with ReLU activation are more robust to parameter perturbations with respect to their counterparts with sign activation, which arguably is reflected in the simplicity of the functions they generate.

Keywords: large deviation theory, path integral, deep neural networks, function sensitivity

(Some figures may appear in colour only in the online journal)

1. Introduction

Learning machines realized by deep neural networks (DNN) have achieved impressive success in performing various machine learning tasks, such as speech recognition, image classification and natural language processing [1]. While DNN typically have numerous parameters and their training comes at a high computational cost, their applications have been extended also to include devices with limited memory or computational resources, such as mobile devices, thanks to compressed networks and reduced parameter precision [2]. Most supervised learning scenarios are of DNN functions representing some input–output mapping, on the basis

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

of input–output example patterns. DNN parameter estimation (training) aims at obtaining a network that approximates well the underlying mapping. Despite their profound engineering success, a comprehensive understanding of the intrinsic working mechanism [3, 4] and the generalization ability [5–8] of DNN are still lacking. The difficulty in analyzing DNN is due to the recursive nonlinear mapping between layers they implement and the coupling to data and learning dynamics.

A recent line of research utilizes the mean field theory in statistical physics to investigate various DNN characteristics, such as expressive power [9], Gaussian process-like behaviors of wide DNN [10–12], dynamical stability in layer propagation and its impact on weight initialization [13–15] and function similarity and entropy in the function space [16]. By assuming large layer-width and random weights, such techniques harness the specific type of nonlinearity used and many degrees of freedom to provide valuable analytical insights. The Gaussian process perspectives of infinitely wide DNN also facilitates the analysis of training dynamics and generalization by employing established kernel methods [17, 18].

To study the entropy of functions realized by DNN [16], we adopted similar assumptions but employed the generating functional analysis [19, 20], which is more general and can be applied to sparse and weight-correlated networks. The analysis of function error incurred by weight perturbations exhibits an exponential growth in error for DNN with sign activation functions, while networks with ReLU activation function are more robust to perturbations. We have also found that ReLU activation induces correlations among variables in random convolution networks [16]. The robustness of random networks with ReLU activation is related to the simplicity of the functions they compute [21, 22], which may converge to a constant function in the large depth and width limit [15], although, in principle, they admit high capacity with arbitrary weights. However, DNN used in practice are of finite size and finite depth, therefore it is essential to analyze the deviation of finite-size systems with respect to the typical mean field behavior, and characterize its rate of convergence with increasing size. An example of a recent study along these lines [23] investigates the deviation in performance of finite size neural networks with a single hidden layer from the Gaussian process behavior.

In this work, we adopt the large deviation approach and the path integral formalism of [16] to derive the deviation of function sensitivity of finite systems from their infinite system counterparts, which is applicable to a range of DNN structures. We analyze the effect of sparsifying (diluting) and binarizing DNN weights, commonly used for model simplification [24–27]. Although the dependence on data and training are not considered, the analysis of random DNN provides valuable insights and baseline comparisons. We will also investigate the sensitivity of functions to input perturbation [9, 13], which is related to function complexity and generalization [21, 22, 28, 29]. The paper is organized as follows. In sections 2 and 3, we introduce the random DNN model and review the basic results of generating functional analysis,respectively. In sections 4 and 5, we derive the large deviation of function sensitivity to weight and input perturbations, respectively, based on the path integral formalism. Finally, in section 6, we discuss the results and their implications.

2. The model

Following [16], we consider two coupled fully-connected DNN. One of them serves as the reference function under consideration, and the other as its perturbed counterpart, either in the weights or input variables. As shown in figure 1, each network consists of $L+1$ layer; layer l has N^l neurons, which can be layer dependent. The reference network is parameterized

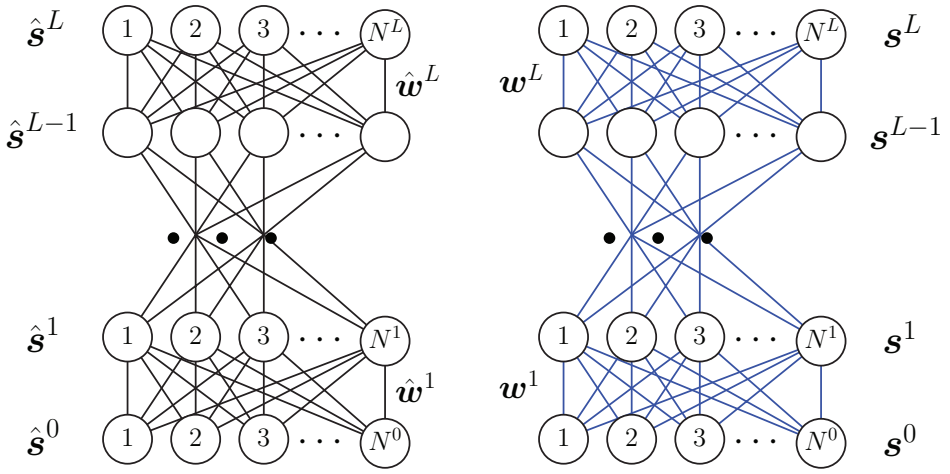

Figure 1. The reference and perturbed fully-connected DNN, parameterized by $\{\hat{\boldsymbol{w}}^l\}$ (black edges) and $\{\boldsymbol{w}^l\}$ (blue edges), respectively. Each layer l has $N^l = \alpha^l N$ nodes.

by the weight variables[1] $\{\hat{\boldsymbol{w}}^l\}_{l=1}^L$, while the perturbed network is parameterized with $\{\boldsymbol{w}^l\}_{l=1}^L$. Similarly, variables with a circumflex are associated with the reference network. In the following, $\boldsymbol{w}^l$ represents the $N^l \times N^{l-1}$ weight matrix at layer l, and $\boldsymbol{w}_i^l$ represents the N^{l-1} dimensional weight vector of the ith perceptron at layer l. Denoting the input dimension as $N = N^0$, we assume the sizes of all layers scale linearly with N as $N^l = \alpha^l N$.

A deterministic feed-forward network is defined by the recursive mapping $\forall\ 1 \leqslant l \leqslant L$

$$h_i^l = \frac{1}{\sqrt{N^{l-1}}} \sum_{j=1}^{N^{l-1}} w_{ij}^l s_j^{l-1}, \tag{1}$$

$$s_i^l = \phi^l\left(h_i^l\right), \tag{2}$$

where $\{w_{ij}^l\}$ are the weights, h_i^l and s_i^l are pre- and post-activation field and variable, respectively, and $\phi^l(\cdot)$ is the activation/transfer function at layer l. The scaling factor of $1/\sqrt{N^{l-1}}$ in equation (1) is introduced for normalization. We primarily focus on networks with either sign $[\phi_s(x) = \mathrm{sgn}(x)]$ or ReLU $[\phi_r(x) = \max(x, 0)]$ activation functions in the hidden layers, and consider binary input and output variables $s_i^0, s_i^L \in \{1, -1\}$ by applying the sign activation function at the output layer $s_i^L = \mathrm{sgn}(h_i^L)$ for a fair comparison across architectures. The resulting feed-forward DNN implements a Boolean mapping $f : \{1, -1\}^{N^0} \to \{1, -1\}^{N^L}$, where each output node $s_i^L(\boldsymbol{s}^0)$ computes a Boolean function. In the following, we call the two architectures sign-DNN and relu-DNN respectively, keeping in mind that sign activation function is always applied in the output layer.

To facilitate a path integral calculation, we consider stochastic dynamics between successive layers. For the layer with sign activation function, the activation s_i^l is disturbed by thermal noise according to the following probability

[1] The usual bias variables are omitted for simplicity, but it can be easily accommodated within the current framework.

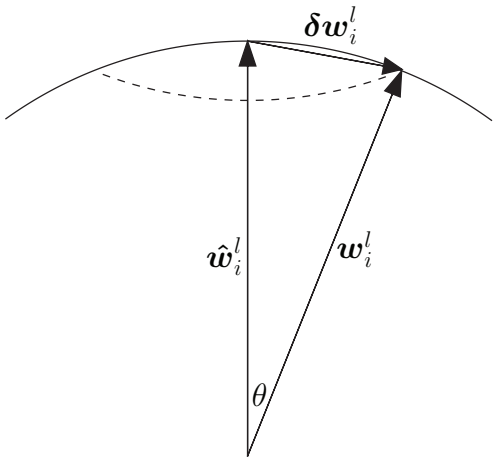

Figure 2. A geometric representation of perturbations on the parameter vector $\hat{\boldsymbol{w}}_i^l$ defined in equation (6), resulting in a rotated vector $\boldsymbol{w}_i^l$ at an angle $\theta^l = \sin^{-1}\eta^l$.

$$P\big(s_i^l|h_i^l(\boldsymbol{w}^l,\boldsymbol{s}^{l-1})\big) = \frac{\exp\big(\beta s_i^l h_i^l(\boldsymbol{w}^l,\boldsymbol{s}^{l-1})\big)}{2\cosh\big(\beta h_i^l(\boldsymbol{w}^l,\boldsymbol{s}^{l-1})\big)}, \tag{3}$$

while for relu activation function, s_i^l is disturbed by additive Gaussian noise

$$P\big(s_i^l|h_i^l(\boldsymbol{w}^l,\boldsymbol{s}^{l-1})\big) = \sqrt{\frac{\beta}{2\pi}}\exp\Big\{-\frac{\beta}{2}\Big[s_i^l - \phi\big(h_i^l(\boldsymbol{w}^l,\boldsymbol{s}^{l-1})\big)\Big]^2\Big\}. \tag{4}$$

In the limit $\beta\to\infty$, we recover the deterministic model. The evolution of the two systems follows the joint distribution

$$P(\{\hat{s}_i^l, s_i^l\}) = P(\hat{\boldsymbol{s}}^0,\boldsymbol{s}^0)\prod_{l=1}^{L}\prod_{i=1}^{N^l} P\big(\hat{s}_i^l|\hat{h}_i^l(\hat{\boldsymbol{w}}^l,\hat{\boldsymbol{s}}^{l-1})\big)P\big(s_i^l|h_i^l(\boldsymbol{w}^l,\boldsymbol{s}^{l-1})\big). \tag{5}$$

To probe the difference between the functions implemented by the two networks, we feed in the same *single* input $\boldsymbol{s}^0 = \hat{\boldsymbol{s}}^0$ to the two systems such that $P(\hat{\boldsymbol{s}}^0,\boldsymbol{s}^0) = P(\hat{\boldsymbol{s}}^0)\prod_{i=1}^{N^0}\delta_{\hat{s}_i^0,s_i^0}$, and study the resulting output difference due to parameter perturbation. For continuous weight variables, one useful choice for the weight perturbation is

$$w_{ij}^l = \sqrt{1-(\eta^l)^2}\,\hat{w}_{ij}^l + \eta^l\delta w_{ij}^l, \tag{6}$$

which ensures that w_{ij}^l has the same variance of $\hat{w}_{ij}^l$ as long as δw_{ij}^l follows the same distribution of $\hat{w}_{ij}^l$, and effectively rotates the high dimensional vector $\hat{\boldsymbol{w}}_i^l$ by an angle $\theta^l = \sin^{-1}\eta^l$ as demonstrated schematically in figure 2.

In probing the sensitivity of a function due to input perturbations, the weights of two networks are kept the same $\boldsymbol{w} = \hat{\boldsymbol{w}}$ and a fixed fraction of input variables are flipped randomly. The resulting output difference of the two systems reflects the sensitivity and complexity of the underlying DNN.

3. Generating functional analysis for typical behavior

Viewing the weights $\{\hat{w}_{ij}^l, w_{ij}^l\}$ as quenched random variables, a generating functional analysis has been proposed [16] to derive the typical behavior of DNN. It starts with computing the disorder-averaged generating functional

$$\overline{\Gamma}(\hat{\boldsymbol{\psi}}, \boldsymbol{\psi}) = \mathbb{E}_{\hat{\boldsymbol{w}},\boldsymbol{w}} \mathbb{E}_{\hat{\boldsymbol{s}},\boldsymbol{s}} \exp\left(-\mathrm{i} \sum_{l,i} (\hat{\psi}_i^l \hat{s}_i^l + \psi_i^l s_i^l) \right), \tag{7}$$

where the average $\mathbb{E}_{\hat{\boldsymbol{s}},\boldsymbol{s}}$ is taken with respect to the joint probability equation (5). Assume the layer widths are the same $N^l = N$ for all l. Upon averaging over the disorder $\hat{\boldsymbol{w}}, \boldsymbol{w}$, the generating functional can be expressed through a set of macroscopic order parameters such as the overlaps $q^l = 1/N^l \sum_i \langle \hat{s}_i^l s_i^l \rangle$ and magnetizations $\hat{m}^l = 1/N^l \sum_i \langle \hat{s}_i^l \rangle, m^l = 1/N^l \sum_i \langle s_i^l \rangle$ as

$$\overline{\Gamma} = \int \{ \mathrm{d}\boldsymbol{q} \mathrm{d}\boldsymbol{Q} \ldots \} \exp\left[N \Psi(\boldsymbol{q}, \boldsymbol{Q}, \ldots) \right] \tag{8}$$

where $\boldsymbol{Q}$ is the conjugate variable of the order parameter $\boldsymbol{q}$. In the large system size limit $N \to \infty$, the generating functional $\overline{\Gamma}$ is dominated by the saddle point of the potential function $\Psi(\boldsymbol{q}, \boldsymbol{Q}, \ldots)$. It gives rise to typical overlaps that dominate in probability, which facilitates analytical studies of random DNN.

Assume the weight perturbation follows the form of equation (6), and both weight and perturbation are independent of each other and follow a Gaussian distribution $\hat{w}_{ij}^l, \delta w_{ij}^l \sim \mathcal{N}(0, \sigma_w^2)$. It is found that for the layer with sign activation function in the limit $\beta \to \infty$, the overlap evolves as [16]

$$q^l = \frac{2}{\pi} \sin^{-1}\left(\sqrt{1 - (\eta^l)^2} q^{l-1} \right), \quad 1 \leqslant l \leqslant L. \tag{9}$$

Similarly, for ReLU activation function in the deterministic limit, if the weight standard deviation is chosen as $\sigma_w = \sqrt{2}$, the magnitude of the activations remains stable and the overlap evolves as

$$\begin{aligned} q^l = \frac{1}{\pi} \Bigg\{ & \sqrt{1 - \left[1 - (\eta^l)^2\right] (q^{l-1})^2} \\ & + \sqrt{1 - (\eta^l)^2} q^{l-1} \left[\frac{\pi}{2} + \sin^{-1}\left(\sqrt{1 - (\eta^l)^2} q^{l-1} \right) \right] \Bigg\}, \end{aligned} \tag{10}$$

while the output layer L follows equation (9) due to the use of the sign activation function. The restriction $\boldsymbol{s}^0 = \hat{\boldsymbol{s}}^0$ leads to $q^0 = 1$ in both cases.

4. Large deviations in parameter sensitivity of functions

The generating functional analysis above gives typical behaviors of random DNN in the limit $N \to \infty$. However, practical DNN always have finite sizes. Therefore, it is worthwhile to understand the deviation to the most probable behaviors under finite N. In the following, we adopt the large deviation analysis to tackle this problem. An introduction of large deviation theory and its application to statistical mechanics can be found in [30]. In essence, a continuous observable $\mathcal{O}$ in a system of size N (assumed to be large) is said to satisfy the large deviation principle if the probability of finding $\mathcal{O}$ follows

$$\text{Prob}_N(\mathcal{O} \in [x, x+\mathrm{d}x]) \simeq \mathrm{e}^{-NI(x)}\mathrm{d}x, \tag{11}$$

where $I(x)$ is the rate function of the observable. It implies that the probability density of $\mathcal{O}$ scales as $P_N(\mathcal{O}=x) \simeq \mathrm{e}^{-NI(x)}$, which is concentrated at the minimum of the rate function $x^* = \text{argmin}_x I(x)$ in large systems and the profile of $I(x)$ quantifies the fluctuation of the observable.

In this work the overlap of the output layer $q^L := 1/N^L \sum_i \hat{s}_i^L s_i^L$ is at the focus of our study. The path integral techniques adopted in the generating functional framework [16] can be adapted to tackle the large deviation analysis. We start with computing the probability density[2]

$$\begin{aligned} P(q^L) &= \left\langle \delta\left(\frac{1}{N^L}\sum_i \hat{s}_i^L s_i^L - q^L\right)\right\rangle \\ &= \mathbb{E}_{\hat{\boldsymbol{w}},\boldsymbol{w}}\mathrm{Tr}_{\hat{\boldsymbol{s}},\boldsymbol{s}} P(\hat{\boldsymbol{s}}^0)\prod_{i=1}^{N^0}\delta_{s_i^0,\hat{s}_i^0}\prod_{l=1}^{L}P(\hat{\boldsymbol{s}}^l|\hat{\boldsymbol{w}}^l,\hat{\boldsymbol{s}}^{l-1})P(\boldsymbol{s}^l|\boldsymbol{w}^l,\boldsymbol{s}^{l-1})\delta\left(\frac{1}{N^L}\sum_i \hat{s}_i^L s_i^L - q^L\right), \end{aligned} \tag{12}$$

where the operation $\mathrm{Tr}_{\hat{\boldsymbol{s}},\boldsymbol{s}}$ is understood as an integration or summation depending on the nature of variables. The input distribution follows $P(\hat{\boldsymbol{s}}^0) = \prod_i P(\hat{s}_i^0) = \prod_i (\frac{1}{2}\delta_{\hat{s}_i^0,1} + \frac{1}{2}\delta_{\hat{s}_i^0,-1})$. To deal with the non-linearity of the pre-activation fields in the conditional probability, we introduce auxiliary fields $\{\hat{x}_i^l, x_i^l\}$ through the integral representation of delta-function

$$1 = \int_{-\infty}^{\infty}\frac{\mathrm{d}\hat{h}_i^l\mathrm{d}\hat{x}_i^l}{2\pi}\mathrm{e}^{\mathrm{i}\hat{x}_i^l\left(\hat{h}_i^l - \frac{1}{\sqrt{N^{l-1}}}\sum_j \hat{w}_{ij}^l\hat{s}_j^{l-1}\right)}, \quad 1 = \int_{-\infty}^{\infty}\frac{\mathrm{d}h_i^l\mathrm{d}x_i^l}{2\pi}\mathrm{e}^{\mathrm{i}x_i^l\left(h_i^l - \frac{1}{\sqrt{N^{l-1}}}\sum_j w_{ij}^l s_j^{l-1}\right)}, \tag{13}$$

which allows us to express the quenched random variables $\hat{w}_{ij}^l$ and w_{ij}^l linearly in the exponents, leading to

$$\begin{aligned} P(q^L) = \mathbb{E}_{\hat{\boldsymbol{w}},\boldsymbol{w}}\mathrm{Tr}_{\hat{\boldsymbol{s}},\boldsymbol{s}}\delta\left(\frac{1}{N^L}\hat{s}_i^L s_i^L - q^L\right)\prod_{i=1}^{N^0}P(\hat{s}_i^0)\delta_{s_i^0,\hat{s}_i^0}\int\prod_{l=1}^{L}\prod_{i=1}^{N^l}\frac{\mathrm{d}\hat{h}_i^l\mathrm{d}\hat{x}_i^l}{2\pi}\frac{\mathrm{d}h_i^l\mathrm{d}x_i^l}{2\pi} \\ \times\exp\left[\sum_{l=1}^{L}\sum_{i=1}^{N^l}\left(\log P(\hat{s}_i^l|\hat{h}_i^l) + \log P(s_i^l|h_i^l) + \mathrm{i}\hat{x}_i^l\hat{h}_i^l + \mathrm{i}x_i^l h_i^l\right)\right] \\ \times\exp\left[-\sum_{l=1}^{L}\frac{\mathrm{i}}{\sqrt{N^{l-1}}}\sum_{i=1}^{N^l}\sum_{j=1}^{N^{l-1}}\left(\hat{w}_{ij}^l\hat{x}_i^l\hat{s}_j^{l-1} + w_{ij}^l x_i^l s_j^{l-1}\right)\right]. \end{aligned} \tag{14}$$

Assuming self-averaging [31] we exchange the order of summation and integration, and first carry out the average over the disorder variables. Specifically, we consider the weights of the reference network to be independent and follow a Gaussian distribution $\hat{w}_{ij}^l \sim \mathcal{N}(0, \sigma_w^2)$ as before, and three types of perturbations

[2] Here we assume $q^L = 1/N^L\sum_{i=1}^{N^L}\hat{s}_i^L s_i^L$ to be a continuous variable by considering large N^L. Instead, one can view q^L as a discrete variable by definition (since the inputs are binary variables), where $\delta(\cdot)$ should be understood as the Kronecker delta function.

(i) rotation of the weight vector $\hat{\boldsymbol{w}}_i^l$ following equation (6);
(ii) sparsification of the weight matrix $\hat{\boldsymbol{w}}^l$ by randomly dropping connections with probability p^l and rescaling the remaining weights by $1/\sqrt{1-p^l}$ to ensure the same weight strength

$$w_{ij}^l = \begin{cases} 0, & \text{with probability } p^l, \\ \frac{1}{\sqrt{1-p^l}}\hat{w}_{ij}^l, & \text{with probability } 1-p^l, \end{cases} \tag{15}$$

(iii) binarization of weight element $\hat{w}_{ij}^l$

$$w_{ij}^l = \mathrm{sgn}(\hat{w}_{ij}^l)\sigma_w, \tag{16}$$

where σ_w is introduced for keeping the variance of w_{ij}^l the same as $\hat{w}_{ij}^l$.

4.1. Macroscopic order parameters

For perturbation of type (i), the disorder average of the third line of equation (14) yields

$$\prod_{l,i} \exp\left\{-\sigma_w^2\left[\frac{1}{2}(\hat{x}_i^l)^2\frac{\sum_j(\hat{s}_j^{l-1})^2}{N^{l-1}} + \frac{1}{2}(x_i^l)^2\frac{\sum_j(s_j^{l-1})^2}{N^{l-1}} + \sqrt{1-(\eta^l)^2}\hat{x}_i^l x_i^l\frac{\sum_j \hat{s}_j^{l-1}s_j^{l-1}}{N^{l-1}}\right]\right\}. \tag{17}$$

To decouple equations (14) and (17) over sites we introduce three sets of order parameters by inserting the identity

$$\begin{aligned} 1 &= \int \frac{\mathrm{d}\hat{V}^l\mathrm{d}\hat{v}^l}{2\pi/N^l}\mathrm{e}^{\mathrm{i}N^l\hat{V}^l\left[\hat{v}^l - \frac{1}{N^l}\sum_j(\hat{s}_j^l)^2\right]}, \quad 1 = \int \frac{\mathrm{d}V^l\mathrm{d}v^l}{2\pi/N^l}\mathrm{e}^{\mathrm{i}N^lV^l\left[v^l - \frac{1}{N^l}\sum_j(s_j^l)^2\right]}, \\ 1 &= \int \frac{\mathrm{d}Q^l\mathrm{d}q^l}{2\pi/N^l}\mathrm{e}^{\mathrm{i}N^lQ^l\left[q^l - \frac{1}{N^l}\sum_j \hat{s}_j^l s_j^l\right]}, \quad \forall\, l \neq L, \end{aligned} \tag{18}$$

and by expressing the output constraint as

$$\delta\left(\frac{1}{N^L}\sum_{i=1}^{N^L}\hat{s}_i^L s_i^L - q^L\right) = \int \frac{\mathrm{d}Q^L}{2\pi/N^L}\mathrm{e}^{\mathrm{i}N^LQ^L\left[q^L - \frac{1}{N^L}\sum_j \hat{s}_j^L s_j^L\right]}. \tag{19}$$

Upon introducing these macroscopic order parameters, equation (17) becomes $\prod_{l,i}\exp\{-1/2[\hat{x}_i^l, x_i^l]\cdot\Sigma_l\cdot[\hat{x}_i^l, x_i^l]^\top\}$ with the covariance matrix Σ_l

$$\Sigma_l := \sigma_w^2\begin{bmatrix} \hat{v}^{l-1} & \sqrt{1-(\eta^l)^2}q^{l-1} \\ \sqrt{1-(\eta^l)^2}q^{l-1} & v^{l-1} \end{bmatrix}. \tag{20}$$

The probability density in equation (14) involves N^l identical integration and summation at each layer l, which can be performed individually [16], yielding

$$P(q^L) = \int \frac{\mathrm{d}Q^L}{2\pi/N^L} \prod_{l=0}^{L-1} \frac{\mathrm{d}\hat{V}^l \mathrm{d}\hat{v}^l}{2\pi/N^l} \frac{\mathrm{d}V^l \mathrm{d}v^l}{2\pi/N^l} \frac{\mathrm{d}Q^l \mathrm{d}q^l}{2\pi/N^l}$$
$$\times \mathrm{e}^{\sum_{l=0}^{L-1} N^l \left(\mathrm{i}\hat{V}^l \hat{v}^l + \mathrm{i}V^l v^l + \mathrm{i}Q^l q^l\right) + N^L \mathrm{i}Q^L q^L} \mathrm{e}^{-N^0\left(\mathrm{i}\hat{V}^0 + \mathrm{i}V^0 + \mathrm{i}Q^0\right)}$$
$$\times \prod_{l=1}^{L-1} \left[\int \mathrm{d}H^l \frac{\mathrm{e}^{-\frac{1}{2}(H^l)^\top \Sigma_l^{-1} H^l}}{\sqrt{(2\pi)^2 |\Sigma_l|}} \mathrm{Tr}_{\hat{s}^l, s^l} P(\hat{s}^l|\hat{h}^l) P(s^l|h^l) \mathrm{e}^{-\mathrm{i}\hat{V}^l (\hat{s}^l)^2 - \mathrm{i}V^l (v^l)^2 - \mathrm{i}Q^l \hat{s}^l s^l} \right]^{N^l}$$
$$\times \left[\int \mathrm{d}H^L \frac{\mathrm{e}^{-\frac{1}{2}(H^L)^\top \Sigma_L^{-1} H^L}}{\sqrt{(2\pi)^2 |\Sigma_L|}} \mathrm{Tr}_{\hat{s}^L, s^L} P(\hat{s}^L|\hat{h}^L) P(s^L|h^L) \mathrm{e}^{-\mathrm{i}Q^L \hat{s}^L s^L} \right]^{N^L}, \tag{21}$$

where we have integrated out the auxiliary fields $\{\hat{x}^l, x^l\}$ and introduced the field doublet $H^l := [\hat{h}^l, h^l]^\top$. We further write $P(q^L)$ as

$$P(q^L) = \int \frac{\mathrm{d}Q^L}{2\pi/N^L} \prod_{l=0}^{L-1} \frac{\mathrm{d}\hat{V}^l \mathrm{d}\hat{v}^l}{2\pi/N^l} \frac{\mathrm{d}V^l \mathrm{d}v^l}{2\pi/N^l} \frac{\mathrm{d}Q^l \mathrm{d}q^l}{2\pi/N^l} \exp[-N\Phi(\boldsymbol{Q}, \boldsymbol{q}, \hat{\boldsymbol{V}}, \hat{\boldsymbol{v}}, \boldsymbol{V}, \boldsymbol{v}|q^L)], \tag{22}$$

where $-N\Phi(\boldsymbol{Q}, \boldsymbol{q}, \hat{\boldsymbol{V}}, \hat{\boldsymbol{v}}, \boldsymbol{V}, \boldsymbol{v}|q^L)$ is equal to the logarithm of the integrand in equation (21). Similar to the analysis in [16], the probability density $P(q^L)$ is dominated by the saddle point $(\boldsymbol{Q}^*, \boldsymbol{q}^*, \ldots)$ of the potential function $\Phi(\ldots)$ in the large N limit ($N^l = \alpha^l N$ with α^l as a constant)

$$P(q^L) \approx \exp[-N\Phi(\boldsymbol{Q}^*, \boldsymbol{q}^*, \ldots|q^L)], \tag{23}$$

where $I(q^L) = \Phi(\boldsymbol{Q}^*, \boldsymbol{q}^*, \ldots|q^L)$ is the desired rate function.

While this set-up is based on computing the deviation in function similarity with a single input $q^L = 1/N^L \sum_i \hat{s}_i^L s_i^L$, one may argue that it requires testing on more than one input for obtaining a robust estimation, e.g.

$$\tilde{q}^L := \frac{1}{N^L M} \sum_{\mu=1}^{M} \sum_{i=1}^{N^L} \hat{s}_i^{L,\mu} s_i^{L,\mu}, \tag{24}$$

where M is the number of independent patterns used. Assuming that representation of different patterns are uncorrelated, we show in appendix C that for small M, the rate function $I(\tilde{q}^L)$ is approximately related to the single input case through a simple scaling

$$I(\tilde{q}^L) \approx M\Phi(\boldsymbol{Q}^*, \boldsymbol{q}^*, \ldots|\tilde{q}^L). \tag{25}$$

This assumption is valid for sign-DNN but not for relu-DNN. We also confirm this scaling relation by numerical experiments (see below and in appendix C).

4.2. Unifying three types of weight perturbations

The other two types of perturbations can be treated similarly. For network sparsification (15), the disorder average of equation (14) has the following form in the large N^l limit (see appendix A for details)

$$\prod_{l,i} \exp\left\{ -\sigma_w^2 \left[\frac{1}{2} (\hat{x}_i^l)^2 \frac{\sum_j (\hat{s}_j^{l-1})^2}{N^{l-1}} + \frac{1}{2} (x_i^l)^2 \frac{\sum_j (s_j^{l-1})^2}{N^{l-1}} + \sqrt{1-p^l} \hat{x}_i^l x_i^l \frac{\sum_j \hat{s}_j^{l-1} s_j^{l-1}}{N^{l-1}} \right] \right\}, \tag{26}$$

which has the same form of equation (17) when p^l is replaced by $(\eta^l)^2$. Introducing the same order parameters, we obtain the covariance of the fields $\hat{h}^l$ and h^l in the form of

$$\Sigma_l^{\rm s} := \sigma_w^2 \begin{bmatrix} \hat{v}^{l-1} & \sqrt{1-p^l}q^{l-1} \\ \sqrt{1-p^l}q^{l-1} & v^{l-1} \end{bmatrix}. \tag{27}$$

Hence, diluting connections with probability p^l at layer l in a random DNN corresponds to rotating each of the weight vector $\hat{\boldsymbol{w}}_i^l$ by an angle $\theta^l = \sin^{-1}\sqrt{p^l}$.

Similarly, for network binarization in equation (16), the disorder average of equation (14) yields (see appendix B for details)

$$\prod_{l,i} \exp\left\{-\sigma_w^2\left[\frac{1}{2}(\hat{x}_i^l)^2\frac{\sum_j(\hat{s}_j^{l-1})^2}{N^{l-1}} + \frac{1}{2}(x_i^l)^2\frac{\sum_j(s_j^{l-1})^2}{N^{l-1}} + \sqrt{\frac{2}{\pi}}\hat{x}_i^l x_i^l\frac{\sum_j \hat{s}_j^{l-1}s_j^{l-1}}{N^{l-1}}\right]\right\}, \tag{28}$$

which corresponds to the covariance matrix of the fields $\hat{h}^l$ and h^l to be in the form

$$\Sigma_l^{\rm b} := \sigma_w^2 \begin{bmatrix} \hat{v}^{l-1} & \sqrt{\frac{2}{\pi}}q^{l-1} \\ \sqrt{\frac{2}{\pi}}q^{l-1} & v^{l-1} \end{bmatrix}. \tag{29}$$

Comparing to type (i) perturbation, one finds that binarizing weight elements in a random DNN corresponds to rotating each of the weight vectors $\hat{\boldsymbol{w}}_i^l$ by a fixed angle $\theta^l = \cos^{-1}\sqrt{\frac{2}{\pi}} \approx 37°$. This phenomenon has been observed in [32] and is linked to the practical success of binary DNN. It is argued [32] that 37° is a very small angle in high dimensional spaces where two randomly sampled vectors are typically orthogonal to each other; therefore weight binarization approximately preserves the directions of the high dimensional weight vectors, which contributes to the success of binary DNN.

Therefore, we establish that the three types of perturbations on random DNN can be unified in the same framework developed in section 4.1.

4.3. Saddle point equations

For networks with a generic activation function, the large deviation potential function $\Phi(\ldots)$ can be express as

$$\begin{aligned}\Phi = &-\alpha^0\left[\mathrm{i}\hat{V}^0(\hat{v}^0-1) + \mathrm{i}V^0(v^0-1) + \mathrm{i}Q^0(q^0-1)\right] - \sum_{l=1}^{L-1}\alpha^l(\mathrm{i}\hat{V}^l\hat{v}^l + \mathrm{i}V^l v^l + \mathrm{i}Q^l q^l)\\ &- \mathrm{i}Q^L q^L - \sum_{l=1}^{L}\alpha^l\log\int \mathrm{d}\hat{h}^l\mathrm{d}h^l \mathrm{Tr}_{\hat{s}^l,s^l}\mathcal{M}^l(\hat{s}^l,s^l,\hat{h}^l,h^l),\end{aligned} \tag{30}$$

$$\mathcal{M}^l(\hat{s}^l,s^l,\hat{h}^l,h^l) := \frac{\mathrm{e}^{-\frac{1}{2}(H^l)^\top\Sigma_l^{-1}H^l}}{\sqrt{(2\pi)^2|\Sigma_l|}}P(\hat{s}^l|\hat{h}^l)P(s^l|h^l)\mathrm{e}^{-\mathrm{i}\hat{V}^l(\hat{s}^l)^2-\mathrm{i}V^l(v^l)^2-\mathrm{i}Q^l\hat{s}^l s^l},\ 1\leqslant l<L, \tag{31}$$

$$\mathcal{M}^L(\hat{s}^L,s^L,\hat{h}^L,h^L) := \frac{\mathrm{e}^{-\frac{1}{2}(H^L)^\top\Sigma_L^{-1}H^L}}{\sqrt{(2\pi)^2|\Sigma_L|}}\frac{\mathrm{e}^{\beta\hat{s}^L\hat{h}^L}}{2\cosh(\beta\hat{h}^L)}\frac{\mathrm{e}^{\beta s^L h^L}}{2\cosh(\beta h^L)}\mathrm{e}^{-\mathrm{i}Q^L\hat{s}^L s^L}, \tag{32}$$

where $\alpha^0 = 1$ since $N^0 = N$.

Setting the derivatives with respect to the conjugate order parameters $\partial\Phi/\partial \mathrm{i}\hat{V}^l$, $\partial\Phi/\partial \mathrm{i}V^l$, $\partial\Phi/\partial \mathrm{i}Q^l$ to zero yields the saddle point equations

$$\hat{v}^0 = v^0 = 1, \quad q^0 = 1, \tag{33}$$

$$\hat{v}^l = \frac{\int \mathrm{d}\hat{h}^l \mathrm{d}h^l \mathrm{Tr}_{\hat{s}^l,s^l} \left(\hat{s}^l\right)^2 \mathcal{M}^l(\hat{s}^l, s^l, \hat{h}^l, h^l)}{\int \mathrm{d}\hat{h}^l \mathrm{d}h^l \mathrm{Tr}_{\hat{s}^l,s^l} \mathcal{M}^l(\hat{s}^l, s^l, \hat{h}^l, h^l)} = \langle \left(\hat{s}^l\right)^2 \rangle_{\mathcal{M}^l}, \quad v^l = \langle \left(s^l\right)^2 \rangle_{\mathcal{M}^l}, \quad 1 \leqslant l < L, \tag{34}$$

$$q^l = \frac{\int \mathrm{d}\hat{h}^l \mathrm{d}h^l \mathrm{Tr}_{\hat{s}^l,s^l} \left(\hat{s}^l s^l\right) \mathcal{M}^l(\hat{s}^l, s^l, \hat{h}^l, h^l)}{\int \mathrm{d}\hat{h}^l \mathrm{d}h^l \mathrm{Tr}_{\hat{s}^l,s^l} \mathcal{M}^l(\hat{s}^l, s^l, \hat{h}^l, h^l)} = \langle \hat{s}^l s^l \rangle_{\mathcal{M}^l}, \quad 1 \leqslant l \leqslant L, \tag{35}$$

in which $\mathcal{M}^l(\hat{s}^l, s^l, \hat{h}^l, h^l)$ bears the meaning of an effective measure [33]. Notice that q^L is an *input parameter* imposing a nonlinear end point constraint on $\mathrm{i}Q^L$, which differs from the generating functional analysis calculation of typical behaviors [16], where q^L is a dynamical variable and $\mathrm{i}Q^L = 0$ at the saddle point.

Setting $\partial\Phi/\partial q^l$ to zero yields the saddle point equations for the conjugate order parameters $\mathrm{i}Q^l$

$$\mathrm{i}Q^{l-1} = \frac{\alpha^l}{\alpha^{l-1}} \frac{\int \mathrm{d}\hat{h}^l \mathrm{d}h^l \mathrm{Tr}_{\hat{s}^l,s^l} \frac{\partial}{\partial q^{l-1}} \mathcal{M}^l(\hat{s}^l, s^l, \hat{h}^l, h^l)}{\int \mathrm{d}\hat{h}^l \mathrm{d}h^l \mathrm{Tr}_{\hat{s}^l,s^l} \mathcal{M}^l(\hat{s}^l, s^l, \hat{h}^l, h^l)}, \quad 1 \leqslant l \leqslant L. \tag{36}$$

Similar relations holds for $\mathrm{i}\hat{V}^l$ and $\mathrm{i}V^l$. While the conjugate order parameters $\{\hat{V}^l, V^l, Q^l\}$ are defined on the real axis, they can be extended to the complex plane and evaluated on the imaginary axis in the saddle point approximation, in which case $\{\mathrm{i}\hat{V}^l, \mathrm{i}V^l, \mathrm{i}Q^l\}$ are real variables. Other observables can be computed by resorting to the effective measure $\mathcal{M}^l$ once the saddle point is obtained, e.g. the mean activations are given by [33]

$$\hat{m}^l = \langle \hat{s}^l \rangle_{\mathcal{M}^l}, \quad m^l = \langle s^l \rangle_{\mathcal{M}^l}. \tag{37}$$

Since the covariance matrix $\Sigma_l(q^{l-1}, \ldots)$ depends on the order parameters of layer $l-1$, the effective measure $\mathcal{M}^l$ at layer l depends on the order parameters $\{q^{l-1}, \ldots\}$ of the previous layer, while it depends on the conjugate order parameters $\{\mathrm{i}Q^l, \ldots\}$ of the current layer. We then observe that the order parameters $\{q^l, \ldots\}$ propagate forward in layers, while $\{\mathrm{i}Q^l, \ldots\}$ encoding the randomness leading to the desired deviation propagate backward, which resembles the structure in optimal control problem [34]. Therefore, we solve the saddle point equations in a forward-backward iteration manner until convergence. Another feature to notice in equation (36) is the dependence of the saddle point solution on the layer-shape parameters $\{\alpha^l\}$, which does not play a role in the mean field solutions where all the conjugate order parameters $\{\mathrm{i}Q^l, \ldots\}$ vanish [16].

4.4. Explicit solutions for sign and ReLU activation functions

For networks with sign activation function the order parameters satisfy $\hat{v}^l = v^l = 1$, such that the only meaningful order parameters are $\{q^l, Q^l\}$. The potential function Φ can be computed analytically, taking the form

$$\Phi(\boldsymbol{Q},\boldsymbol{q}|q^L) = -\alpha^0 \mathrm{i}Q^0(q^0-1) - \sum_{l=1}^{L}\alpha^l \mathrm{i}Q^l q^l \\ - \sum_{l=1}^{L}\alpha^l \log\left[\cosh(\mathrm{i}Q^l) - \sinh(\mathrm{i}Q^l)\frac{2}{\pi}\sin^{-1}(\sqrt{1-(\eta^l)^2}q^{l-1})\right], \tag{38}$$

while the saddle point equations become

$$q^0 = 1, \tag{39}$$

$$q^l = \frac{-\sinh(\mathrm{i}Q^l) + \cosh(\mathrm{i}Q^l)\frac{2}{\pi}\sin^{-1}(\sqrt{1-(\eta^l)^2}q^{l-1})}{\cosh(\mathrm{i}Q^l) - \sinh(\mathrm{i}Q^l)\frac{2}{\pi}\sin^{-1}(\sqrt{1-(\eta^l)^2}q^{l-1})}, \quad \forall 1 \leqslant l \leqslant L, \tag{40}$$

$$\mathrm{i}Q^{l-1} = \frac{\frac{2}{\pi}\sinh(\mathrm{i}Q^l)}{\cosh(\mathrm{i}Q^l) - \sinh(\mathrm{i}Q^l)\frac{2}{\pi}\sin^{-1}(\sqrt{1-(\eta^l)^2}q^{l-1})} \\ \times \frac{\alpha^l\sqrt{1-(\eta^l)^2}}{\alpha^{l-1}\sqrt{1-[1-(\eta^l)^2](q^{l-1})^2}}, \quad \forall 1 \leqslant l \leqslant L. \tag{41}$$

Note that q^L in equation (40) is an input parameter.

For networks with ReLU activation function the potential function Φ also admits an explicit expression

$$\Phi(\boldsymbol{Q},\boldsymbol{q},\hat{\boldsymbol{V}},\hat{\boldsymbol{v}},\boldsymbol{V},\boldsymbol{v}|q^L) = -\alpha^0\left[\mathrm{i}\hat{V}^0(\hat{v}^0-1) + \mathrm{i}V^0(v^0-1) + \mathrm{i}Q^0(q^0-1)\right] \\ - \sum_{l=1}^{L-1}\alpha^l(\mathrm{i}\hat{V}^l\hat{v}^l + \mathrm{i}V^l v^l + \mathrm{i}Q^l q^l) - \mathrm{i}Q^L q^L \\ - \sum_{l=1}^{L-1}\alpha^l \log\left\{\frac{1}{2\pi\sqrt{|\Sigma_l|}}\left[\frac{1}{\sqrt{|A^l|}}\left(\frac{\pi}{2} - \tan^{-1}\left(\frac{A^l_{12}}{\sqrt{|A^l|}}\right)\right) + \frac{1}{\sqrt{|B^l|}}\left(\frac{\pi}{2} + \tan^{-1}\left(\frac{B^l_{12}}{\sqrt{|B^l|}}\right)\right)\right.\right. \\ \left.\left. + \frac{1}{\sqrt{|\Sigma_l^{-1}|}}\left(\frac{\pi}{2} - \tan^{-1}\left(\frac{\Sigma^{-1}_{l,12}}{\sqrt{|\Sigma_l^{-1}|}}\right)\right) + \frac{1}{\sqrt{|C^l|}}\left(\frac{\pi}{2} + \tan^{-1}\left(\frac{C^l_{12}}{\sqrt{|C^l|}}\right)\right)\right]\right\} \\ - \alpha^L \log\left[\cosh(\mathrm{i}Q^L) - \sinh(\mathrm{i}Q^L)\frac{2}{\pi}\tan^{-1}\left(\frac{\Sigma_{L,12}}{\sqrt{|\Sigma_L|}}\right)\right], \tag{42}$$

where A^l, B^l, C^l are 2×2 matrices defined as

$$A^l = \Sigma_l^{-1} + \begin{bmatrix} 2\mathrm{i}\hat{V}^l & \mathrm{i}Q^l \\ \mathrm{i}Q^l & 2\mathrm{i}V^l \end{bmatrix}, \quad B^l = \Sigma_l^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & 2\mathrm{i}V^l \end{bmatrix}, \quad C^l = \Sigma_l^{-1} + \begin{bmatrix} 2\mathrm{i}\hat{V}^l & 0 \\ 0 & 0 \end{bmatrix}. \tag{43}$$

The saddle point equations also admit a close-form expression accordingly.

5. Large deviations in input sensitivity of functions

In probing the sensitivity of a function to the flipping of input variables, the weights of two networks considered are taking the same values $\boldsymbol{w} = \hat{\boldsymbol{w}}$, which is done by setting $\eta^l = 0$ in equation (6). We constrain the input $\boldsymbol{s}^0$ of the perturbed system to have a pre-defined overlap q^0 (or Hamming distance $N^0(1-q^0)/2$) with the input $\hat{\boldsymbol{s}}^0$ of the reference system. The

sensitivity of the output overlaps to input perturbations is investigated through the conditional probability

$$P(q^L|q^0)=\frac{P(q^L,q^0)}{P(q^0)}=\frac{\left\langle\delta\left(\frac{1}{N^L}\sum_i\hat{s}_i^Ls_i^L-q^L\right)\delta\left(\frac{1}{N^0}\sum_i\hat{s}_i^0s_i^0-q^0\right)\right\rangle}{\left\langle\delta\left(\frac{1}{N^0}\sum_i\hat{s}_i^0s_i^0-q^0\right)\right\rangle}. \quad (44)$$

Without loss of generality, we choose a decoupled input distribution $P(\hat{\boldsymbol{s}}^0,\boldsymbol{s}^0)=\prod_i P(\hat{s}_i^0)P(s_i^0)=\prod_i(\frac{1}{2}\delta_{\hat{s}_i^0,1}+\frac{1}{2}\delta_{\hat{s}_i^0,-1})(\frac{1}{2}\delta_{s_i^0,1}+\frac{1}{2}\delta_{s_i^0,-1})$ while the delta function involving q^0 in equation (44) constrains the systems to have the desired input correlation. The probability of input overlap $P(q^0)$ can be computed as

$$\begin{aligned}P(q^0)&=\mathrm{Tr}_{\hat{\boldsymbol{s}}^0,\boldsymbol{s}^0}\prod_i P(\hat{s}_i^0)P(s_i^0)\int\frac{\mathrm{d}Q^0}{2\pi/N^0}\mathrm{e}^{\mathrm{i}N^0Q^0\left(q^0-\frac{1}{N^0}\sum_i\hat{s}_i^0s_i^0\right)}\\&=\int\frac{\mathrm{d}Q^0}{2\pi/N^0}\exp\left[N^0\left(\mathrm{i}Q^0q^0+\log\cosh(\mathrm{i}Q^0)\right)\right]\\&\approx\exp\left[N^0\left(\mathrm{i}Q^{0*}q^0+\log\cosh(\mathrm{i}Q^{0*})\right)\right]\\&=:\exp\left[-N\Phi_{\mathrm{P}}(\mathrm{i}Q^{0*}|q^0)\right],\end{aligned} \quad (45)$$

$$\Phi_{\mathrm{P}}(\mathrm{i}Q^0|q^0):=-\alpha^0\left(\mathrm{i}Q^0q^0+\log\cosh(\mathrm{i}Q^0)\right), \quad (46)$$

$$\mathrm{i}Q^{0*}:=-\tanh^{-1}(q^0), \quad (47)$$

where we have made use of the saddle point approximation of $P(q^0)$ in the large N^0 limit, with the corresponding potential function defined in equation (46) and the saddle point solution $\mathrm{i}Q^{0*}$ given in equation (47).

The computation of the joint probability $P(q^L,q^0)$ is analogous to that of $P(q^L)$ in earlier sections,

$$\begin{aligned}P(q^L,q^0)&=\mathbb{E}_{\hat{\boldsymbol{w}},\boldsymbol{w}}\mathrm{Tr}_{\hat{\boldsymbol{s}},\boldsymbol{s}}P(\hat{\boldsymbol{s}}^0)\prod_{i=1}^{N^0}\delta_{s_i^0,\hat{s}_i^0}\prod_{l=1}^{L}P(\hat{\boldsymbol{s}}^l|\hat{\boldsymbol{w}}^l,\hat{\boldsymbol{s}}^{l-1})P(\boldsymbol{s}^l|\boldsymbol{w}^l,\boldsymbol{s}^{l-1})\\&\quad\times\int\frac{\mathrm{d}Q^0}{2\pi/N^0}\frac{\mathrm{d}Q^L}{2\pi/N^L}\mathrm{e}^{\mathrm{i}N^0Q^0\left(q^0-\frac{1}{N^0}\sum_i\hat{s}_i^0s_i^0\right)+\mathrm{i}N^LQ^L\left(q^L-\frac{1}{N^L}\sum_i\hat{s}_i^Ls_i^L\right)}\\&=\int\{\mathrm{d}\boldsymbol{Q}\mathrm{d}\boldsymbol{q}\ldots\}\exp[-N\Phi_{\mathrm{J}}(\boldsymbol{Q},\boldsymbol{q},\ldots|q^L,q^0)],\end{aligned} \quad (48)$$

$$\begin{aligned}\Phi_{\mathrm{J}}=&-\alpha^0\left[\mathrm{i}\hat{V}^0(\hat{v}^0-1)+\mathrm{i}V^0(v^0-1)+\left(\mathrm{i}Q^0q^0+\log\cosh(\mathrm{i}Q^0)\right)\right]-\mathrm{i}Q^Lq^L\\&-\sum_{l=1}^{L-1}\alpha^l(\mathrm{i}\hat{V}^l\hat{v}^l+\mathrm{i}V^lv^l+\mathrm{i}Q^lq^l)-\sum_{l=1}^{L}\alpha^l\log\int\mathrm{d}\hat{h}^l\mathrm{d}h^l\mathrm{Tr}_{\hat{s}^l,s^l}\mathcal{M}^l(\hat{s}^l,s^l,\hat{h}^l,h^l).\end{aligned} \quad (49)$$

The saddle point of $\mathrm{i}Q^0$ satisfies $\mathrm{i}Q^{0*}=-\tanh^{-1}(q^0)$, which coincides with the one of $P(q^0)$ in equation (47). So the conditional distribution satisfies

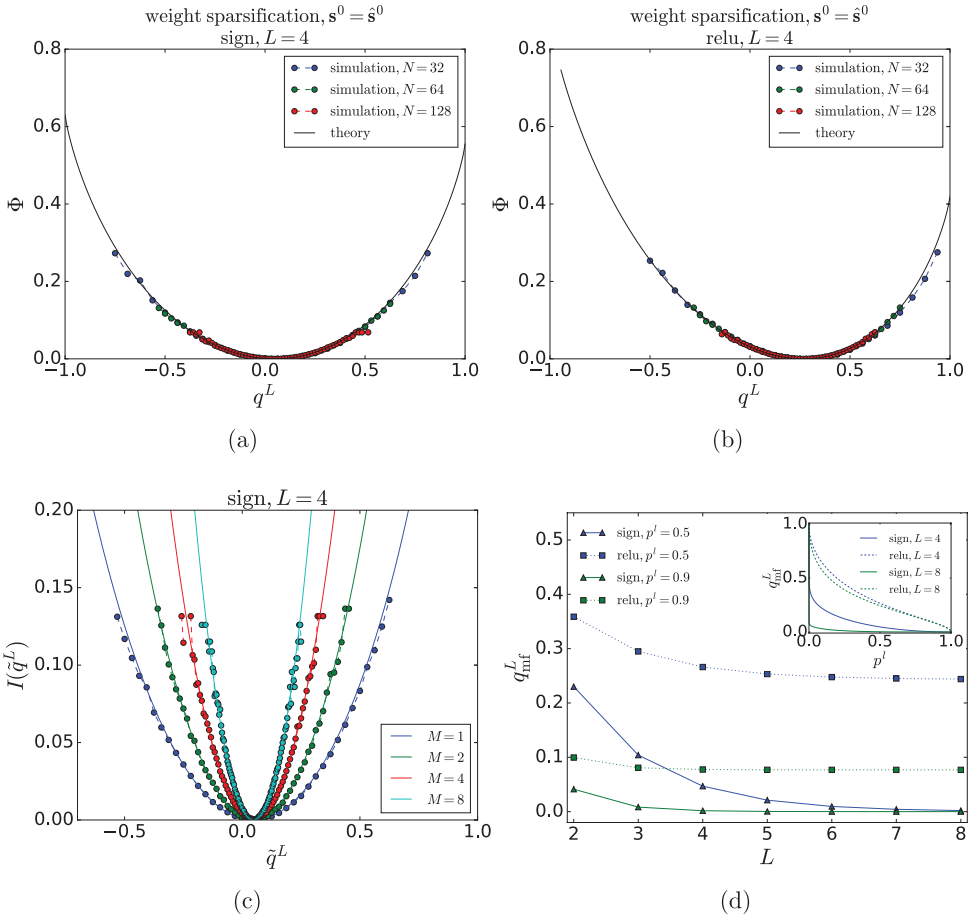

Figure 3. Weight sparsification of random DNN. In (a)–(c), we set $L = 4$ and $p^l = 1/2$; solid lines correspond to theory while dashed lines with circle markers correspond to estimation from simulation. The estimation of the rate function from simulations are obtained by 100 000 samples and the corresponding curve has been shifted such that the minimum is at zero. (a) The rate function Φ versus q^L for sign activation function. (b) The rate function Φ versus q^L for ReLU activation function. (c) The rate function $I(\tilde{q}^L)$ of output overlap $\tilde{q}^L$ defined by M patterns; the theoretical results are given by equation (25), while the simulation results are obtained on systems with $N = 64$. (d) Mean field solutions of output overlap $q^L_{\rm mf}$ as a function of system depth L. Inset: $q^L_{\rm mf}$ versus p^l for different depths.

$$
\begin{aligned}
&P(q^L|q^0) \approx \exp\left[-N\Phi(\boldsymbol{Q}^*, \boldsymbol{q}^*, \ldots|q^L, q^0)\right] = \exp\left[-N(\Phi^*_{\rm J} - \Phi^*_{\rm P})\right] \\
&\Phi(\boldsymbol{Q}, \boldsymbol{q}, \ldots|q^L, q^0) = -\alpha^0\left[{\rm i}\hat{V}^0(\hat{v}^0 - 1) + {\rm i}V^0(v^0 - 1)\right] - {\rm i}Q^L q^L \\
&\qquad - \sum_{l=1}^{L-1} \alpha^l({\rm i}\hat{V}^l\hat{v}^l + {\rm i}V^l v^l + {\rm i}Q^l q^l) - \sum_{l=1}^{L} \alpha^l \log \int {\rm d}\hat{h}^l {\rm d}h^l {\rm Tr}_{\hat{s}^l, s^l} \mathcal{M}^l(\hat{s}^l, s^l, \hat{h}^l, h^l),
\end{aligned}
\tag{50}
$$

where the saddle point solution $\{\boldsymbol{Q}^*, \boldsymbol{q}^*, \ldots\}$ have the same form as those in section 4.3, except that $q^0 = 1$ in equation (33) is replaced by the pre-defined value q^0 under investigation.

6. Results

6.1. Weight sparsification

We first consider the effect of weight perturbation by sparsifying connections as in equation (15). For a concrete example, we consider DNN with $L = 4$, uniform layer width $\alpha^l = 1$ and disconnection probability $p^l = 1/2$, for which we compute the large deviation rate function $I(q^L) = \Phi(\boldsymbol{Q}^*, \boldsymbol{q}^*, \ldots | q^L)$ by solving the saddle point equation in section 4.3 and compare it to numerical experiments. For relu-DNN, we always set $\sigma_w = \sqrt{2}$. The results are shown in figures 3(a) and (b), which exhibit a perfect match between the theory and simulation. The most probable q^L, located at the minimum of Φ corresponds to the mean field solution, where $q^L_{\rm mf} \approx 0.047$ for sign-DNN and $q^L_{\rm mf} \approx 0.266$ for the relu-DNN. However, in finite systems they have a non-zero probability of admitting a higher value of q^L due to fluctuations. We can compute the probability from the rate function by $P(q^L) = \exp(-N\Phi^*(q^L))/Z$[3] and estimate the tail probability of output mismatch. As an example we consider $N = 64$ and find that $P(q^L > 1/2) \approx 0.055\%$ for sign-DNN and $P(q^L > 1/2) \approx 3.8\%$ for relu-DNN, which is non-negligible especially for ReLU activation[4].

In figure 3(c), we also demonstrate that the approximation of rate function $I(\tilde{q}^L)$ of output overlap $\tilde{q}^L$, estimated for M patterns by employing equation (25), is accurate for DNN with sign activation, while the approximation does not hold for deep ReLU networks (see appendix C). Therefore in sign-DNN, the probability of finding perturbed DNN agreeing on all M patterns with the reference DNN decays exponentially with M (at least for small M values). This may not be the case in relu-DNN which requires further exploration in a future study.

In figure 3(d), we compare the mean field output overlaps $q^L_{\rm mf}$ between DNN with sign and ReLU activations for different system depths and disconnection probability p^l. It is shown that relu-DNN are more robust to weight sparsification perturbation, as expected; the perturbed relu-DNN have residual correlations with the reference networks even after removing 90% of the weights. The robustness of relu-DNN to weight dilution was also observed and theoretically analysed in [35]. Finally, we remark that our scenario is different from the practical methods used to prune networks trained on specific data; in this case particular heuristic rules have been developed to disconnect weights instead of the random removal used here. The success of weight pruning in practice hightlights the weight-redundancy in real trained networks [24, 35] but may also be influenced by properties of the data used and training methods. This behaviour is absent in random networks with random data, as indicated in the inset of figure 3(d), where even a small dilution probability can deteriorate the overlap. Additional modelling considerations are needed to address practical scenarios.

6.2. Weight binarization

We then consider the effect of perturbation by binarization of weight variables as in equation (16). Also here we consider uniform layer width $\alpha^l = 1$. The results shown in figure 4,

[3] For finite N^L, the output overlap is a discrete variable $q^L \in \{1, 1 - \frac{2}{N^L}, 1 - \frac{4}{N^L}, \ldots, -1\}$, so it is convenient to consider the discretized probability distribution of q^L as $\text{Prob}(q^L) = P(q^L)\Delta q^L = \exp(-N\Phi^*(q^L))/Z$; the normalization constant is computed as $Z = \sum_k \exp(-N\Phi^*(q^L_k))\Delta q^L$, where the summation runs over all possible values of q^L and $\Delta q^L = \frac{2}{N^L}$. Although we could not find the saddle point solution of $\Phi(\ldots|q^L)$ in the vicinity of $q^L = -1$ for relu-DNN (see figure 3(b)), the contribution from that region to the cumulative probability of the overlap is negligible .

[4] Notice that such estimation is obtained by saddle point approximation in equation (22) and by keeping the leading order contribution, which may be slightly biased for small N.

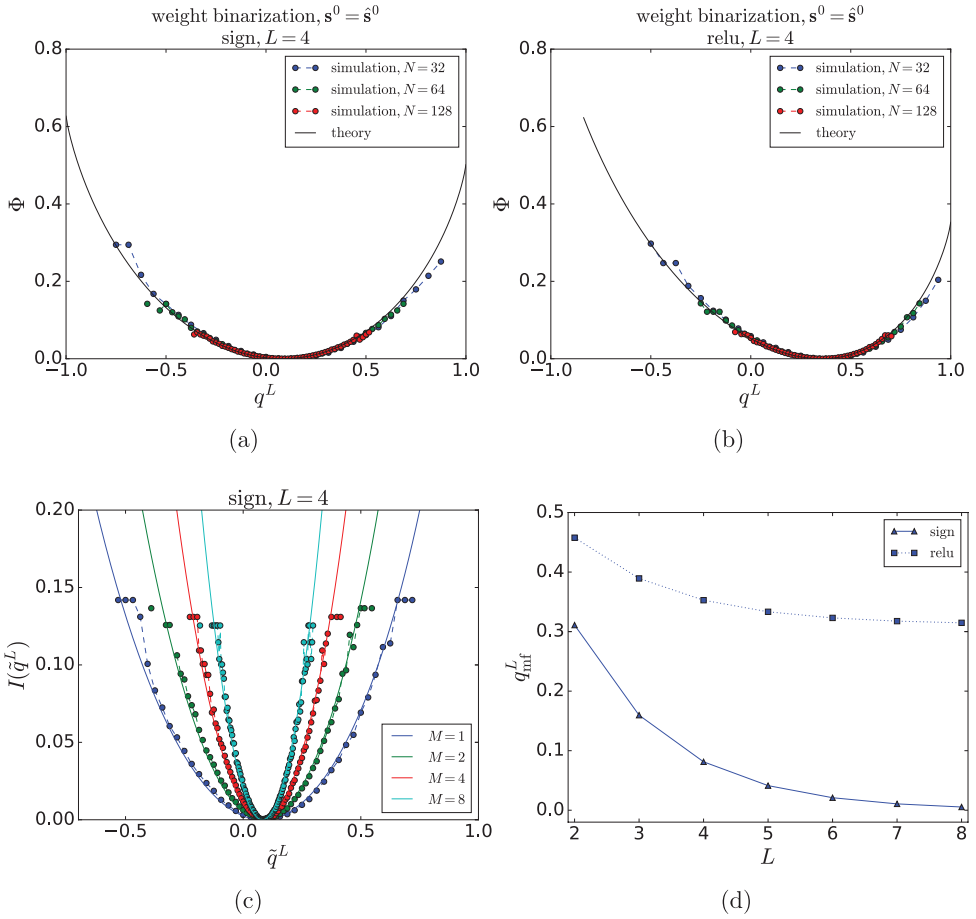

Figure 4. Weight binarization of random DNN. (a) Φ versus q^L for sign activation function. (b) Φ versus q^L for ReLU activation. (c) The rate function $I(\tilde{q}^L)$ of output overlap $\tilde{q}^L$ defined by M patterns; solid lines are theoretical results while dashed lines with circle markers are estimated by simulation. (d) Mean field solutions of output overlap $q^L_{\rm mf}$ as a function of system depth L.

are very similar to the effect of weight sparsification. As pointed out in section 4.2, binarizing weights of random DNN corresponds to rotating the weight vector $\hat{\boldsymbol{w}}^l_i$ by an angle $\theta^l = \cos^{-1}\sqrt{\frac{2}{\pi}}$ [32], or equivalently, disconnecting weights with a particular probability $p^l = 1 - \frac{2}{\pi}$. The matches between theory and simulation in figures 4(a)–(c) validates the large deviation-based analysis in both sign and relu-DNN and the scaling relation of equation (25) in sign-DNN. The relu-DNN are more biased to the regime of positive correlation and more robust to binarizing perturbation as seen in figure 4(d).

6.3. Sensitivity to input perturbation

We have shown that relu-DNN with random weights are robust to parameter perturbations such as weight sparsification and weight binarization, which is a desired property for better

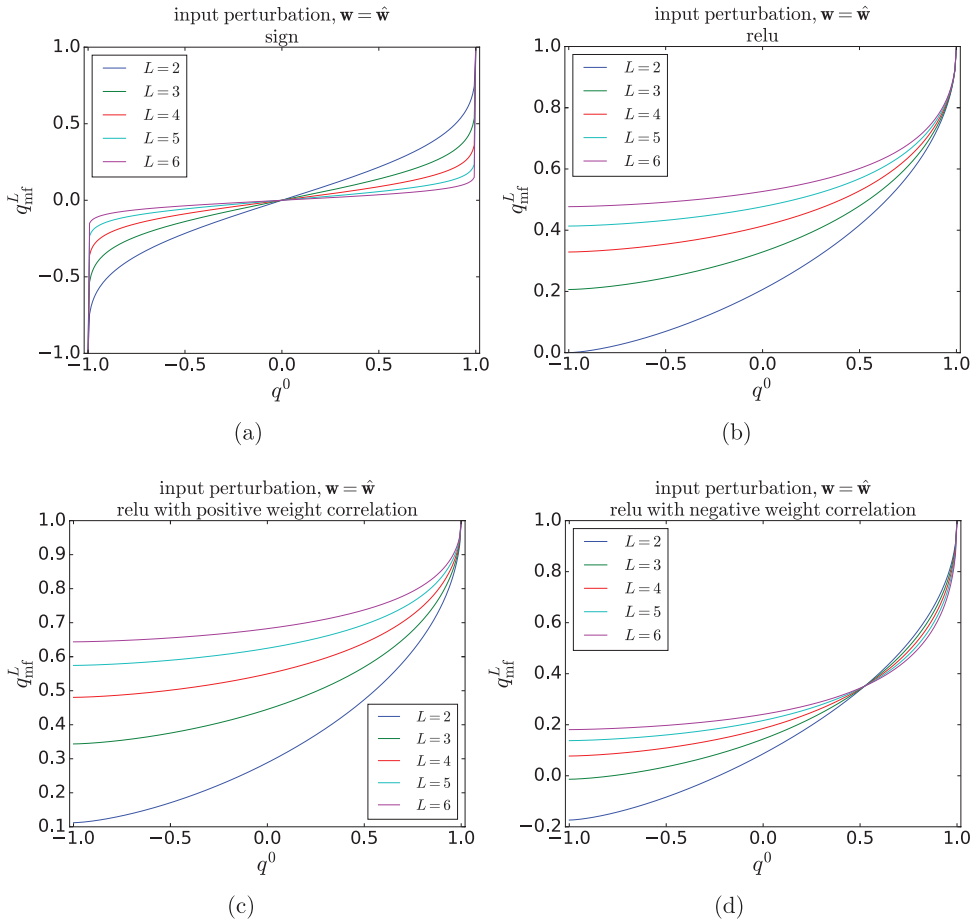

Figure 5. Mean field solutions $q^L_{\rm mf}$ versus q^0 in the scenario of input perturbation where $\boldsymbol{w} = \hat{\boldsymbol{w}}$. In all architectures, sign activation function is applied at the output layer. (a) DNN with sign activation functions and uncorrelated random weights. (b) DNN with ReLU activation at the hidden layers, with uncorrelated random weights, and sign activation at the output layer. (c) Relu-DNN with positive weight correlation $c = 2/(3N)$. (d) Relu-DNN with negative weight correlation $c = -2/(3N)$.

generalization. On the other hand, such network ensembles typically represent simple functions as studied in [21, 22]. The simplicity of the functions generated is one reason accounting for the observed robustness to parameter perturbation.

To probe the function complexity, we study the function sensitivity under input perturbation while keeping $\boldsymbol{w} = \hat{\boldsymbol{w}}$ [28]. Flipping n input variables corresponds to the input overlap $q^0 = 1 - \frac{2n}{N^0}$. In figures 5(a) and (b) we depict the overlap $q^L_{\rm mf}$ of the final output as a function of input overlap q^0 (keeping in mind that we always apply the sign activation in the output layer). While the outputs become more de-correlated in deeper layers of sign-DNN, the relu-DNN induce correlation at deeper layers. Therefore, random relu-DNN tend to forget the input structure at deeper layers, generating increasingly simpler functions that are robust to parameter perturbation. This phenomenon has been noticed in the Gaussian process-like analysis of DNN [10–12].

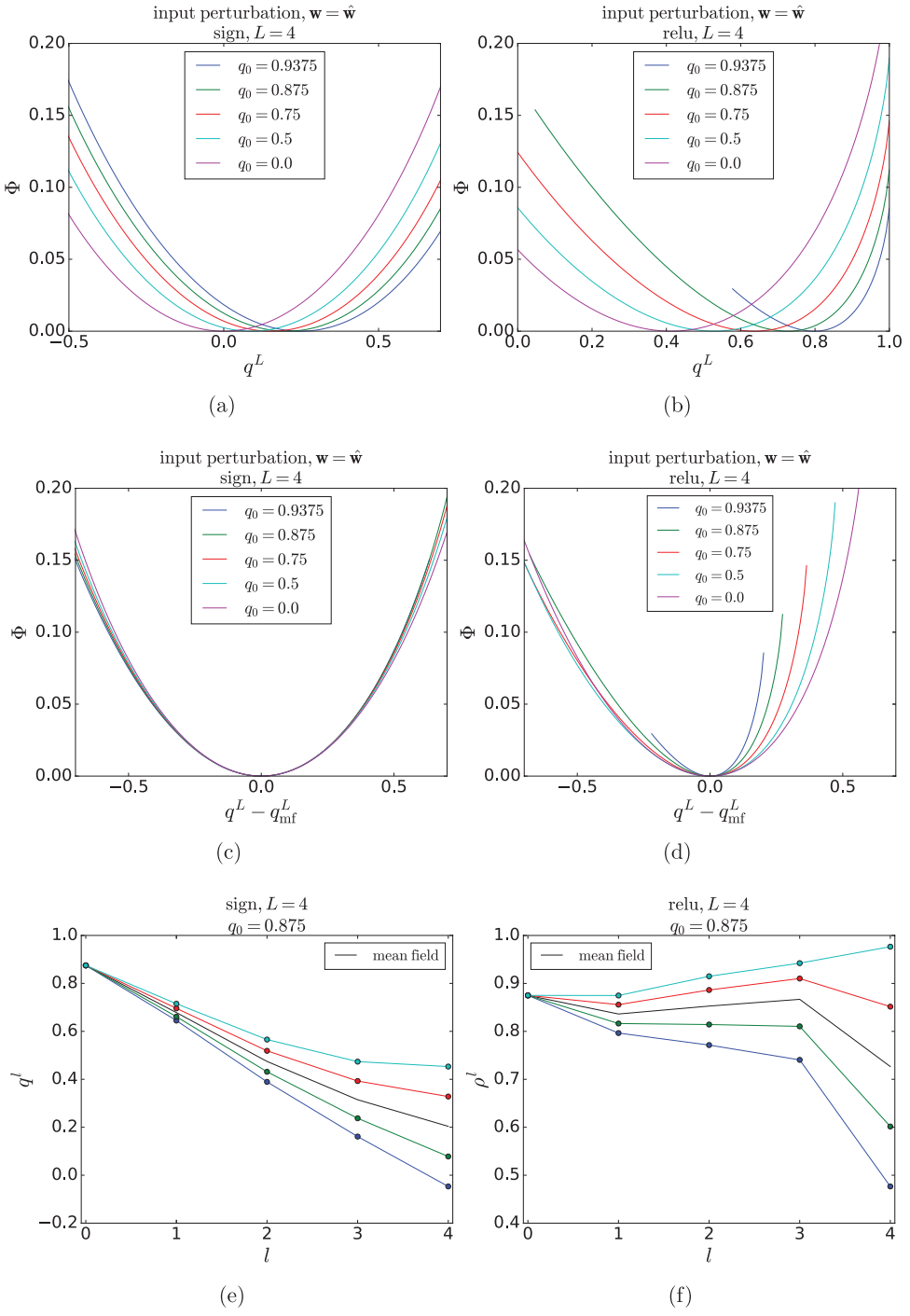

Figure 6. Large deviation of output similarity q^L under input perturbation where $\boldsymbol{w} = \hat{\boldsymbol{w}}$. Sub-figures (c) and (d) are the same as (a) and (b), except for the shifted x-coordinates. (a) and (b) Φ versus q^L for sign- and relu-DNN, respectively. (c) and (d) Φ versus $q^L - q^L_{\rm mf}$ for sign- and relu-DNN, respectively. (e) The dominant trajectories of overlap $\{q^l\}$ leading to particular deviation in sign-DNN. (f) The dominant trajectories of correlation coefficient $\{\rho^l\}$ leading to particular deviation in relu-DNN.

In [16], we investigated the effect of weight correlation in the form of $P(\hat{\boldsymbol{w}}_i^l) = \exp(-\frac{1}{2}(\hat{\boldsymbol{w}}_i^l)^\top A^{-1}\hat{\boldsymbol{w}}_i^l)/\sqrt{(2\pi)^{N^{l-1}}|A|}$, with $A = \sigma_w^2(I - cJ)$ where I is the identity matrix and J the all-one matrix. We found that DNN with ReLU activation functions and negative weight correlation $c < 0$ are more sensitive to parameter perturbation. Here we examine the sensitivity of relu-DNN to input perturbation by employing the same results developed in [16]. In figures 5 (c) and (d), we depict the mean field output overlap $q^L_{\rm mf}$ as a function of input overlap q^0. It is observed that negative weight correlation corresponds to a higher sensitivity to input perturbation, indicating that the relu-DNN with negatively correlated weights generate more complex functions than those with random or positively correlated weights. We conjecture that negative weight correlation develops in very deep ReLU networks when they are trained to performed complex task where a high expressive power is needed, a phenomenon that has been observed in [36].

In figure 6, we further investigate deviations from the typical behaviors in the presence of input perturbations for the specific example with $L = 4, \alpha^l = 1$. The rate functions $\Phi(q^L)$ depicted in figures 6(a) and (b) dictate the rate of convergence to the typical behaviors with increasing N by the large deviation principle, for both sign and ReLU activations, respectively. In figure 6(c), we observe that the rate functions have similar trends in the vicinity of the mean field solution $q^L_{\rm mf}$ for different levels of input perturbation (corresponding to different q^0) in sign-DNN, while they are more distinctive in relu-DNN as seen in figure 6(d). In relu-DNN, smaller input perturbation (larger q^0) leads to smaller variance of q^L around $q^L_{\rm mf}$. The rate function of relu-DNN is also more asymmetric around $q^L_{\rm mf}$, suggesting that large deviations will be more often observed below $q^L_{\rm mf}$ than above it. This indicates that random relu-DNN of finite size may produce functions that are slightly more complex than what would be expected by the mean field solutions, which remains to be verified.

We also examine the dominant trajectories across layers leading to particular deviations by monitoring the correlations of activations between the two systems across layers. The relevant quantity is the correlation coefficient

$$\rho^l = \frac{q^l - \hat{m}^l m^l}{\sqrt{\hat{v}^l - (\hat{m}^l)^2}\sqrt{v^l - (m^l)^2}}, \tag{51}$$

where the mean activations $\hat{m}^l$ and m^l are computed by equation (37). We find that sign-DNN satisfy $\hat{m}^l = m^l = 0, \hat{v}^l = v^l = 1$, such that $\rho^l = q^l$ in this case. The results are shown in figures 6(e) and (f), which suggest that the deviations of q^L from the typical value $q^L_{\rm mf}$ are mainly contributed by the deviations at later layers.

Lastly, we investigate the effect of DNN architecture on the deviation. In particular, we consider a single bottleneck layer at a particular hidden layer l' $(0 < l' < L)$ with $\alpha^{l'} = \frac{1}{8}$ while all other layers satisfy $\alpha^l = 1, \forall l \neq l'$. Placing the bottleneck at later layer introduces a higher variability of output overlap q^L by observing smaller values of the rate function in figure 7; this effect is more prominent in sign-DNN, while it is much less noticeable in relu-DNN.

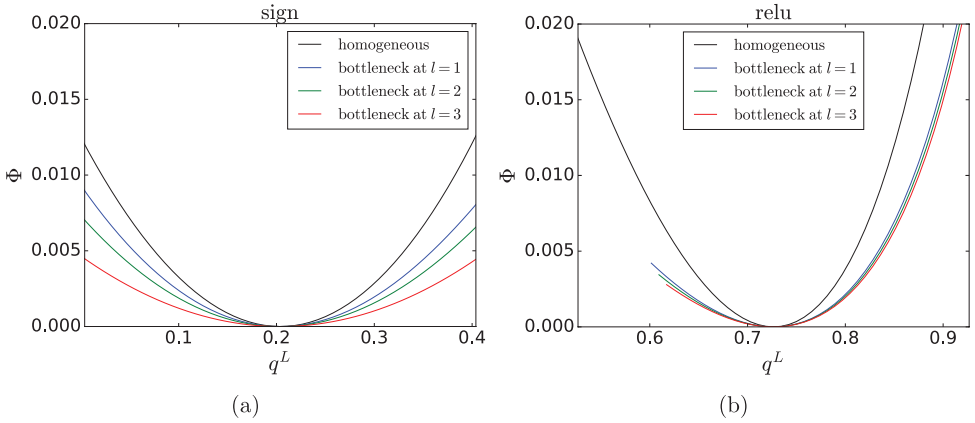

Figure 7. Effect of a single bottleneck layer on the rate function in the scenario of input perturbation. The bottleneck layer l' has width parameter $\alpha^{l'} = \frac{1}{8}$ while all other layers have $\alpha^l = 1$. (a) Sign-DNN. (b) relu-DNN.

7. Discussion

By utilizing the large deviation theory coupled with the path integral analysis, we derive the sensitivity of finite size random DNN under parameter and input perturbations. Random DNN with sign or ReLU activation function are shown to satisfy the large deviation principle, where the rate functions govern an exponential decay of the deviation to the mean field behaviors as the size of the system increases. We also investigate the effects of weight sparsification and binarization of random DNN, and uncover their equivalence to rotation of weight vector in high dimension. Random DNN with ReLU activation function are found to be robust to these parameter perturbations, which is caused by the low complexity of the corresponding function mappings. Random initializing the weights of ReLU DNN places a prior for simple functions, while they have the capacity to compute more complex functions with specifically trained weights. The next important question is how the networks adapt to perform complex tasks by the training processes.

Acknowledgments

BL and DS acknowledge support from the Leverhulme Trust (RPG-2018-092), European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant agreement No. 835913. DS acknowledges support from the EPSRC programme Grant TRANSNET (EP/R035342/1).

Appendix A. Disorder average for weight sparsification

For network sparsification (15), the disorder average in equation (14) can be computed as

$$
\begin{aligned}
&\mathbb{E}_{\hat{\boldsymbol{w}}} \prod_{l,i,j} \exp\left(\frac{-\mathrm{i}}{\sqrt{N^{l-1}}}\hat{w}_{ij}^{l}\hat{x}_{i}^{l}\hat{s}_{j}^{l-1}\right)\left[(1-p^{l})\exp\left(\frac{-\mathrm{i}}{\sqrt{N^{l-1}}\sqrt{1-p^{l}}}\hat{w}_{ij}^{l}x_{i}^{l}s_{j}^{l-1}\right)+p^{l}\right]\\
&=\prod_{l,i,j}\left[(1-p^{l})\exp\left[-\frac{\sigma_{w}^{2}}{2N^{l-1}}\left(\hat{x}_{i}^{l}\hat{s}_{j}^{l-1}+x_{i}^{l}s_{j}^{l-1}/\sqrt{1-p^{l}}\right)^{2}\right]+p^{l}\exp\left[-\frac{\sigma_{w}^{2}}{2N^{l-1}}\left(\hat{x}_{i}^{l}\hat{s}_{j}^{l-1}\right)^{2}\right]\right]\\
&=\prod_{l,i,j}\left\{(1-p^{l})\left[1-\frac{\sigma_{w}^{2}}{2N^{l-1}}\left(\hat{x}_{i}^{l}\hat{s}_{j}^{l-1}+x_{i}^{l}s_{j}^{l-1}/\sqrt{1-p^{l}}\right)^{2}\right]\right.\\
&\qquad\left.+p^{l}\left[1-\frac{\sigma_{w}^{2}}{2N^{l-1}}\left(\hat{x}_{i}^{l}\hat{s}_{j}^{l-1}\right)^{2}\right]+O\left(\frac{1}{(N^{l-1})^{2}}\right)\right\}\\
&\approx\prod_{l,i,j}\left\{1-\frac{\sigma_{w}^{2}}{N^{l-1}}\left[\frac{1}{2}(\hat{x}_{i}^{l})^{2}(\hat{s}_{j}^{l-1})^{2}+\frac{1}{2}(x_{i}^{l})^{2}(s_{j}^{l-1})^{2}+\sqrt{1-p^{l}}(\hat{x}_{i}^{l}x_{i}^{l})(\hat{s}_{j}^{l-1}s_{j}^{l-1})\right]\right\}\\
&\approx\prod_{l,i}\exp\left\{-\sigma_{w}^{2}\left[\frac{1}{2}(\hat{x}_{i}^{l})^{2}\frac{\sum_{j}(\hat{s}_{j}^{l-1})^{2}}{N^{l-1}}+\frac{1}{2}(x_{i}^{l})^{2}\frac{\sum_{j}(s_{j}^{l-1})^{2}}{N^{l-1}}+\sqrt{1-p^{l}}\hat{x}_{i}^{l}x_{i}^{l}\frac{\sum_{j}\hat{s}_{j}^{l-1}s_{j}^{l-1}}{N^{l-1}}\right]\right\},
\end{aligned}
\tag{A.1}
$$

where we have made use of the large N^{l} approximation.

Appendix B. Disorder average for weight binarization

For weight binarization in (16), the disorder average in equation (14) can be computed as

$$
\begin{aligned}
&\mathbb{E}_{\hat{\boldsymbol{w}}} \prod_{l,i,j} \exp\left[\frac{-\mathrm{i}}{\sqrt{N^{l-1}}}\left(\hat{w}_{ij}^{l}\hat{x}_{i}^{l}\hat{s}_{j}^{l-1}+\mathrm{sgn}(\hat{w}_{ij}^{l})\sigma_{w}x_{i}^{l}s_{j}^{l-1}\right)\right]\\
&=\prod_{l,i,j}\left\{\int_{-\infty}^{0}\mathrm{d}\hat{w}_{ij}^{l}\mathcal{N}(\hat{w}_{ij}^{l}|0,\sigma_{w}^{2})\exp\left[\frac{-\mathrm{i}}{\sqrt{N^{l-1}}}\left(\hat{w}_{ij}^{l}\hat{x}_{i}^{l}\hat{s}_{j}^{l-1}-\sigma_{w}x_{i}^{l}s_{j}^{l-1}\right)\right]\right.\\
&\qquad\left.+\int_{0}^{\infty}\mathrm{d}\hat{w}_{ij}^{l}\mathcal{N}(\hat{w}_{ij}^{l}|0,\sigma_{w}^{2})\exp\left[\frac{-\mathrm{i}}{\sqrt{N^{l-1}}}\left(\hat{w}_{ij}^{l}\hat{x}_{i}^{l}\hat{s}_{j}^{l-1}+\sigma_{w}x_{i}^{l}s_{j}^{l-1}\right)\right]\right\}\\
&=\prod_{l,i,j}\exp\left[-\frac{\sigma_{w}^{2}}{2N^{l-1}}(\hat{x}_{i}^{l})^{2}(\hat{s}_{j}^{l-1})^{2}\right]\frac{1}{2}\left\{\left[1+\mathrm{erf}\left(\frac{\mathrm{i}\hat{x}_{j}^{l}\hat{s}_{j}^{l-1}\sigma_{w}}{2\sqrt{N^{l-1}}}\right)\right]\exp\left(\frac{\mathrm{i}x_{i}^{l}s_{j}^{l-1}\sigma_{w}}{\sqrt{N^{l-1}}}\right)\right.\\
&\qquad\left.+\left[1-\mathrm{erf}\left(\frac{\mathrm{i}\hat{x}_{j}^{l}\hat{s}_{j}^{l-1}\sigma_{w}}{\sqrt{2N^{l-1}}}\right)\right]\exp\left(\frac{-\mathrm{i}x_{i}^{l}s_{j}^{l-1}\sigma_{w}}{\sqrt{N^{l-1}}}\right)\right\}\\
&=\prod_{l,i,j}\exp\left[-\frac{\sigma_{w}^{2}}{2N^{l-1}}(\hat{x}_{i}^{l})^{2}(\hat{s}_{j}^{l-1})^{2}\right]\frac{1}{2}\left\{\left(1+\frac{2}{\sqrt{\pi}}\frac{\mathrm{i}\hat{x}_{j}^{l}\hat{s}_{j}^{l-1}\sigma_{w}}{\sqrt{2N^{l-1}}}\right)\left[1+\frac{\mathrm{i}x_{i}^{l}s_{j}^{l-1}\sigma_{w}}{\sqrt{N^{l-1}}}-\frac{1}{2}\frac{(x_{i}^{l}s_{j}^{l-1}\sigma_{w})^{2}}{N^{l-1}}\right]\right.\\
&\qquad\left.+\left(1-\frac{2}{\sqrt{\pi}}\frac{\mathrm{i}\hat{x}_{j}^{l}\hat{s}_{j}^{l-1}\sigma_{w}}{\sqrt{2N^{l-1}}}\right)\left[1-\frac{\mathrm{i}x_{i}^{l}s_{j}^{l-1}\sigma_{w}}{\sqrt{N^{l-1}}}-\frac{1}{2}\frac{(x_{i}^{l}s_{j}^{l-1}\sigma_{w})^{2}}{N^{l-1}}\right]+O\left(\frac{1}{(N^{l-1})^{2}}\right)\right\}\\
&\approx\prod_{l,i,j}\exp\left[-\frac{\sigma_{w}^{2}}{2N^{l-1}}(\hat{x}_{i}^{l})^{2}(\hat{s}_{j}^{l-1})^{2}\right]\left\{1-\frac{\sigma_{w}^{2}}{N^{l-1}}\left[\frac{1}{2}(x_{i}^{l})^{2}(s_{j}^{l-1})^{2}+\sqrt{\frac{2}{\pi}}(\hat{x}_{i}^{l}x_{i}^{l})(\hat{s}_{j}^{l-1}s_{j}^{l-1})\right]\right\}\\
&\approx\prod_{l,i}\exp\left\{-\sigma_{w}^{2}\left[\frac{1}{2}(\hat{x}_{i}^{l})^{2}\frac{\sum_{j}(\hat{s}_{j}^{l-1})^{2}}{N^{l-1}}+\frac{1}{2}(x_{i}^{l})^{2}\frac{\sum_{j}(s_{j}^{l-1})^{2}}{N^{l-1}}+\sqrt{\frac{2}{\pi}}\hat{x}_{i}^{l}x_{i}^{l}\frac{\sum_{j}\hat{s}_{j}^{l-1}s_{j}^{l-1}}{N^{l-1}}\right]\right\},
\end{aligned}
\tag{B.1}
$$

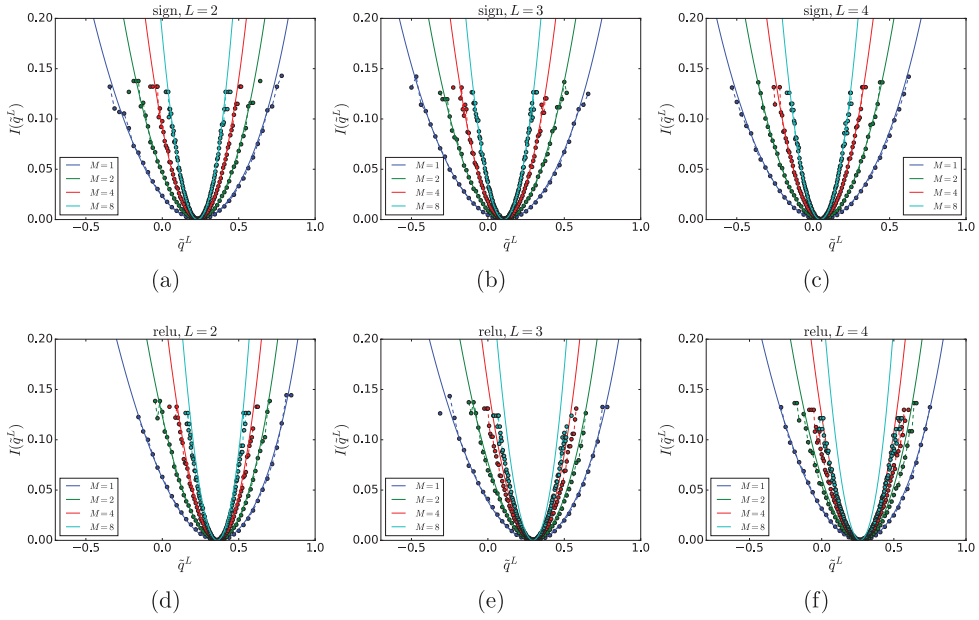

Figure C1. The rate function $I(\tilde{q}^L)$ of output overlap $\tilde{q}^L$ defined for M patterns and DNN with different activation functions and system depths, in the scenario of weight sparsification with disconnection probability $p^l = 1/2$. Solid lines correspond to theoretical results and dashed lines with circle markers correspond to estimation from simulation.

where the large N^l approximation has been employed.

Appendix C. Large deviation in the multiple-pattern scenario

Consider function similarity estimated for multiple patterns

$$\tilde{q}^L = \frac{1}{M}\sum_{\mu=1}^{M}\left(\frac{1}{N^L}\sum_{i=1}^{N^L}\hat{s}_i^{L,\mu}s_i^{L,\mu}\right) =: \frac{1}{M}\sum_{\mu=1}^{M}q^{L,\mu} \tag{C.1}$$

where $\hat{s}_i^{L,\mu}(\hat{\boldsymbol{s}}^{0,\mu})$ is the ith output of the reference network with the μth input $\hat{\boldsymbol{s}}^{0,\mu}$ drawn independently and identically from the input distribution $P(\boldsymbol{s}^0)$. In the small fluctuation regime, where each $q^{L,\mu}$ is close to the mean field solution $q^L_{\rm mf}$, we have $I(q^{L,\mu}) \approx 1/2I''(q^L_{\rm mf})(q^{L,\mu} - q^L_{\rm mf})^2$ (both $I(q^L_{\rm mf})$ and $I'(q^L_{\rm mf})$ vanish [30]), i.e. $P(q^{L,\mu})$ can be approximated by a Gaussian density

$$P(q^{L,\mu}) \sim \exp\left(-\frac{N}{2}I''(q^L_{\rm mf})(q^{L,\mu} - q^L_{\rm mf})^2\right), \tag{C.2}$$

where the corresponding variance is $1/(NI''(q^L_{\rm mf}))$. Since the M inputs are independent, we also assume the outputs are also approximately independent (which holds in sign-DNN but does not necessary for relu-DNN since ReLU non-linearity can induce correlations among variables), such that the variance of $\tilde{q}^L$ is $1/(MNI''(q^L_{\rm mf}))$. Therefore, in the vicinity of $q^L_{\rm mf}$ we have

$$P(\tilde{q}^L) \sim \exp\left(-\frac{MN}{2} I''(q^L_{\rm mf})(\tilde{q}^L - q^L_{\rm mf})^2 \right), \tag{C.3}$$

implying that the corresponding rate function differs from the one with single pattern by a factor of M.

More formally, one can directly compute the probability density $P(\tilde{q}^L)$ as

$$\begin{aligned} P(\tilde{q}^L) &= \left\langle \delta\left(\frac{1}{MN^L} \sum_{\mu,i} \hat{s}_i^{L,\mu} s_i^{L,\mu} - \tilde{q}^L \right) \right\rangle \\ &= \mathbb{E}_{\hat{\boldsymbol{w}},\boldsymbol{w}} \mathrm{Tr}_{\hat{\boldsymbol{s}},\boldsymbol{s}} \delta\left(\frac{1}{MN^L} \sum_{\mu,i} \hat{s}_i^{L,\mu} s_i^{L,\mu} - \tilde{q}^L \right) \prod_{\mu,i} P(\hat{s}_i^{0,\mu}) \delta_{s_i^{0,\mu},\hat{s}_i^{0,\mu}} \int \prod_{\mu,l,i} \frac{\mathrm{d}\hat{h}_i^{l,\mu} \mathrm{d}\hat{x}_i^{l,\mu}}{2\pi} \frac{\mathrm{d}h_i^{l,\mu} \mathrm{d}x_i^{l,\mu}}{2\pi} \\ &\quad \times \exp\left[\sum_{\mu,l,i} \left(\log P(\hat{s}_i^{l,\mu}|\hat{h}_i^{l,\mu}) + \log P(s_i^{l,\mu}|h_i^{l,\mu}) + \mathrm{i}\hat{x}_i^{l,\mu}\hat{h}_i^{l,\mu} + \mathrm{i}x_i^{l,\mu}h_i^{l,\mu} \right) \right] \\ &\quad \times \exp\left[-\sum_{\mu,l} \frac{\mathrm{i}}{\sqrt{N^{l-1}}} \sum_{i,j} \left(\hat{w}_{ij}^l \hat{x}_i^{l,\mu} \hat{s}_j^{l-1,\mu} + w_{ij}^l x_i^{l,\mu} s_j^{l-1,\mu} \right) \right]. \end{aligned} \tag{C.4}$$

Since the weights $\{\hat{w}_{ij}^l, w_{ij}^l\}$ are shared among the M patterns, average over these variables on the last line of equation (C.4) leads to coupling between patterns on the pre-activation fields

$$\begin{aligned} \prod_{l,i} \exp\Bigg\{ -\sigma_w^2 \sum_{\mu,\nu} \Bigg[&\frac{1}{2} \hat{x}_i^{l,\mu} \hat{x}_i^{l,\nu} \frac{1}{N^{l-1}} \sum_j \hat{s}_j^{l-1,\mu} \hat{s}_j^{l-1,\nu} + \frac{1}{2} x_i^{l,\mu} x_i^{l,\nu} \frac{1}{N^{l-1}} \sum_j s_j^{l-1,\mu} s_j^{l-1,\nu} \\ &+ \sqrt{1-(\eta^l)^2}\, \hat{x}_i^{l,\mu} x_i^{l,\nu} \frac{1}{N^{l-1}} \sum_j \hat{s}_j^{l-1,\mu} s_j^{l-1,\nu} \Bigg] \Bigg\}. \end{aligned} \tag{C.5}$$

By introducing the following overlap matrices as macroscopic order parameters

$$q^{l,\mu\nu} = \frac{1}{N^l} \sum_j \hat{s}_j^{l,\mu} s_j^{l,\nu}, \quad \hat{v}^{l,\mu\nu} = \frac{1}{N^l} \sum_j \hat{s}_j^{l,\mu} \hat{s}_j^{l,\nu}, \quad v^{l,\mu\nu} = \frac{1}{N^l} \sum_j s_j^{l,\mu} s_j^{l,\nu}. \tag{C.6}$$

Equation (C.4) can be factorized over sites as before. However, we have $O(LM^2)$ order parameters here, while there are only $O(L)$ order parameters in the single pattern case. To further simplify the calculation, we assume a symmetric structure of the cross-pattern overlaps at the saddle point $q^{l,\mu\nu} = q^{l,\parallel}\delta_{\mu\nu} + q^{l,\perp}(1-\delta_{\mu\nu})$, where $q^{l,\parallel}, q^{l,\perp}$ are the diagonal and off-diagonal matrix elements respectively. Under this assumption, one can in principle evaluate the integral in (C.4), but the resulting calculation becomes rather involved.

Alternatively, since the M input patterns are independent, we expect the diagonal elements of the matrix $q^{l,\mu\nu}$ to be larger than the off-diagonal elements (sum of correlated variables versus sum of random variables). In particular, for sign activation we expect $q^{l,\parallel} \sim O(1), q^{l,\perp} \sim O(\frac{1}{\sqrt{N^l}})$ since $q^{l,\perp}$ involves a summation over weakly correlated positive and negative numbers. We therefore approximate the summation $\sum_{\mu\nu}[\ldots]$ in the exponential of equation (C.5) by $\sum_{\mu=\nu}[\ldots]$, which yields MN^l un-coupled identical integrals at each layer N^l. It eventually leads to the rate function of multiple-pattern overlap $\tilde{q}^L$ as $I(\tilde{q}^L) \approx M\Phi(\boldsymbol{Q}^*, \boldsymbol{q}^*, \ldots|\tilde{q}^L)$, where $\Phi(\boldsymbol{Q}^*, \boldsymbol{q}^*, \ldots|q^L)$ is the rate function of the single-pattern overlap q^L. While the off-diagonal elements of $q^{l,\mu\nu}$ have smaller values, there are more of

these terms ($M(M-1)$ off-diagonal terms compared to M diagonal terms in the summation $\sum_{\mu\nu}[\ldots]$ in the exponential of equation (C.5)), so we expect the above approximation to hold only for small M. The above argument may fail for ReLU activation, since $\hat{s}_j^{l,\mu}, s_j^{l,\mu}$ are always positive, and therefore $q^{l,\perp} \sim O(1)$.

In figure C1, we compare the approximate theoretical results $I(\tilde{q}^L) \approx M\Phi(\boldsymbol{Q}^*, \boldsymbol{q}^*, \ldots|\tilde{q}^L)$ to numerical simulations in the scenario of weight sparsification with disconnection probability $p^l = 1/2$. We observe a good match between the two approaches for sign-DNN, validating the de-correlation assumption of M patterns. For relu-DNN, the theory gives a good prediction on shallow networks with $L = 2$ but deteriorates for deeper networks; it suggests the importance of cross-pattern order parameters $q^{l,\perp}$ in this case, whose detailed treatment is beyond the scope of this work.

ORCID iDs

Bo Li https://orcid.org/0000-0001-9743-9447
David Saad https://orcid.org/0000-0001-9821-2623

References

[1] LeCun Y, Bengio Y and Hinton G 2015 *Nature* **521** 436–44
[2] Cheng Y, Wang D, Zhou P and Zhang T 2018 *IEEE Signal Process. Mag.* **35** 126–36
[3] Zeiler M D and Fergus R 2014 Visualizing and understanding convolutional networks *Proc., Part I. Computer Vision—ECCV 2014: 13th European Conf. (Zurich, Switzerland, 6–12 September 2014) ed* D Fleet *et al* (Cham: Springer) pp 818–33
[4] Yosinski J, Clune J, Nguyen A, Fuchs T and Lipson H 2015 Understanding neural networks through deep visualization *Proc. Deep Learning Workshop and Int. Conf. on Machine Learning* (arXiv:1506.06579)
[5] Zhang C, Bengio S, Hardt M, Recht B and Vinyals O 2017 Understanding deep learning requires rethinking generalization *Proc. 5th Int. Conf. on Learning Representations*
[6] Chaudhari P, Choromanska A, Soatto S, LeCun Y, Baldassi C, Borgs C, Chayes J, Sagun L and Zecchina R 2017 Entropy-sgd: biasing gradient descent into wide valleys *Proc. 5th Int. Conf. on Learning Representations*
[7] Neyshabur B, Bhojanapalli S, McAllester D and Srebro N 2017 Exploring generalization in deep learning *Advances in Neural Information Processing Systems* vol 30 ed I Guyon *et al* (New York: Curran Associates, Inc.) pp 5947–56
[8] Bartlett P L, Foster D J and Telgarsky M J 2017 Spectrally-normalized margin bounds for neural networks *Advances in Neural Information Processing Systems* vol 30 ed I Guyon *et al* (New York: Curran Associates, Inc.) pp 6240–9
[9] Poole B, Lahiri S, Raghu M, Sohl-Dickstein J and Ganguli S 2016 Exponential expressivity in deep neural networks through transient chaos *Advances in Neural Information Processing Systems* vol 29 ed D D Lee *et al* (New York: Curran Associates, Inc.) pp 3360–8
[10] Duvenaud D, Rippel O, Adams R and Ghahramani Z 2014 Avoiding pathologies in very deep networks *Proc. 17th Int. Conf. on Artificial Intelligence and Statistics* (*Proc. Machine Learning Research* vol 33) ed S Kaski and J Corander (Reykjavik: PMLR) pp 202–10
[11] Daniely A, Frostig R and Singer Y 2016 Toward deeper understanding of neural networks: the power of initialization and a dual view on expressivity *Advances in Neural Information Processing Systems* vol 29 ed D D Lee *et al* (New York: Curran Associates, Inc.) pp 2253–61
[12] Lee J, Sohl-dickstein J, Pennington J, Novak R, Schoenholz S and Bahri Y 2018 Deep neural networks as gaussian processes *Proc. 6th Int. Conf. on Learning Representations*

[13] Schoenholz S S, Gilmer J, Ganguli S and Sohl-Dickstein J 2017 Deep information propagation *Proc. 5th Int. Conf. on Learning Representations*
[14] Yang G and Schoenholz S 2017 Mean field residual networks: On the edge of chaos *Advances in Neural Information Processing Systems* vol 30 ed I Guyon *et al* (New York: Curran Associates, Inc.) pp 7103–14
[15] Pretorius A, van Biljon E, Kroon S and Kamper H 2018 Critical initialisation for deep signal propagation in noisy rectifier neural networks *Advances in Neural Information Processing Systems* vol 31 ed S Bengio *et al* (New York: Curran Associates, Inc.) pp 5717–26
[16] Li B and Saad D 2018 *Phys. Rev. Lett.* **120** 248301
[17] Jacot A, Gabriel F and Hongler C 2018 Neural tangent kernel: Convergence and generalization in neural networks *Advances in Neural Information Processing Systems* vol 31 ed S Bengio *et al* (New York: Curran Associates, Inc.) pp 8571–80
[18] Arora S, Du S, Hu W, Li Z and Wang R 2019 Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks *Proc. 36th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research* vol 97) ed K Chaudhuri and R Salakhutdinov (Long Beach, CA: PMLR) pp 322–32
[19] Mozeika A, Saad D and Raymond J 2009 *Phys. Rev. Lett.* **103** 248701
[20] Mozeika A, Saad D and Raymond J 2010 *Phys. Rev.* E **82** 041112
[21] Valle-Perez G, Camargo C Q and Louis A A 2019 Deep learning generalizes because the parameter-function map is biased towards simple functions *Proc. 7th Int. Conf. on Learning Representations*
[22] De Palma G, Kiani B and Lloyd S 2019 Random deep neural networks are biased towards simple functions *Advances in Neural Information Processing Systems* vol 32 ed H Wallach *et al* (New York: Curran Associates, Inc.) pp 1962–74
[23] Antognini J M 2019 Finite size corrections for neural network Gaussian processes *ICML 2019 Workshop on Theoretical Physics for Deep Learning* (arXiv:1908.10030)
[24] Le Cun Y, Denker J S and Solla S A 1990 Optimal brain damage *Advances in Neural Information Processing Systems* vol 2 ed D S Touretzky (Burlington, MA: Morgan–Kaufmann) pp 598–605
[25] Courbariaux M, Hubara I, Soudry D, El-Yaniv R and Bengio Y 2016 Binarized neural networks: training deep neural networks with weights and activations constrained to $+1$ or -1 *Advances in Neural Information Processing Systems* vol 29 ed D D Lee *et al* (New York: Curran Associates Inc.) pp 4107–15
[26] Rastegari M, Ordonez V, Redmon J and Farhadi A 2016 Xnor-net: imagenet classification using binary convolutional neural networks *Proc., Part IV. Computer Vision—ECCV 2016: 14th European Conf. (Amsterdam, The Netherlands, 11–4 October 2016)* ed B Leibe *et al* (Cham: Springer) pp 525–42
[27] Hou L, Yao Q and Kwok J T 2017 Loss-aware binarization of deep networks *Proc. 5th Int. Conf. on Learning Representations*
[28] Franco L 2006 *Neurocomputing* **70** 351–61
[29] Novak R, Bahri Y, Abolafia D A, Pennington J and Sohl-Dickstein J 2018 Sensitivity and generalization in neural networks: an empirical study *Proc. 6th Int. Conf. on Learning Representations*
[30] Touchette H 2009 *Phys. Rep.* **478** 1–69
[31] De Dominicis C 1978 *Phys. Rev.* B **18** 4913–9
[32] Anderson A G and Berg C P 2018 The high-dimensional geometry of binary neural networks *Proc. 6th Int. Conf. on Learning Representations*
[33] Coolen A 2001 Chapter 15 statistical mechanics of recurrent neural networks II—dynamics *Neuro-Informatics and Neural Modelling* (*Handbook of Biological Physics* vol 4) ed F Moss and S Gielen (Amsterdam: North-Holland) pp 619–84
[34] Grafke T and Vanden-Eijnden E 2019 *Chaos* **29** 063118
[35] Huang H and Goudarzi A 2018 *Phys. Rev.* E **98** 042311
[36] Shang W, Sohn K, Almeida D and Lee H 2016 Understanding and improving convolutional neural networks via concatenated rectified linear units *Proc. 33rd Int. Conf. on Machine Learning (New York)* (*Proc. Machine Learning Research* vol 48) ed M F Balcan and K Q Weinberger (PMLR) pp 2217–25