Mutual Information of Neural Network Initialisations: Mean Field Approximations

Jared Tanner and Giuseppe Ughi
University of Oxford, Mathematical Institute
Andrew Wiles Building, Radcliffe Observatory Quarter
OX2 6GG, Oxford, UK
Email: {tanner, ughi}@maths.ox.ac.uk

***Abstract*—The ability to train randomly initialised deep neural networks is known to depend strongly on the variance of the weight matrices and biases as well as the choice of nonlinear activation. Here we complement the existing geometric analysis of this phenomenon [1] with an information theoretic alternative. Lower bounds are derived for the mutual information between an input and hidden layer outputs. Using a mean field analysis we are able to provide analytic lower bounds as functions of network weight and bias variances as well as the choice of nonlinear activation. These results show that initialisations known to be optimal from a training point of view are also superior from a mutual information perspective.**

I. Introduction

Randomly initialised deep neural networks (DNNs) are random nonlinear functions which are drawn and subsequently trained to map a training set of inputs to known outputs. The work in [1] showed that DNN initialisations that preserve information about the inputs are typically easier to train. This conclusion was based on geometric considerations of input signal dynamics in a random feed-forward DNN, measured with the distributions of intermediate hidden layers. This was possible as for a DNN denoted by

$$\mathbf{h}^{(\ell)} = \mathbf{W}^{(\ell)}\phi\left(\mathbf{h}^{(\ell-1)}\right) + \mathbf{b}^{(\ell)} \tag{1}$$

where $\mathbf{h}^{(1)} = \mathbf{W}^{(1)}\mathbf{X} + \mathbf{b}^{(1)}$ with input $\mathbf{X} \in \mathbb{R}^n$, and with $\mathbf{W}^{(\ell)} \in \mathbb{R}^{n\times n}$, [2] determined as a function of the DNN parameters $(\sigma_w, \sigma_b, \phi(\cdot))$ the dynamics of $q^{(\ell)} := \|\mathbf{h}^{(\ell)}\|_2^2$ to its large depth limit q^* in the mean field infinite width limit for n for Gaussian initialisation

$$\mathbf{W}_{ij}^{(\ell)} \sim \mathcal{N}(0, \sigma_w^2/n) \quad \text{and} \quad \mathbf{b}_i^{(\ell)} \sim \mathcal{N}(0, \sigma_b^2). \tag{2}$$

This allowed us to consider the geometric stability of the DNN when applied to two nearby points. Specifically, for a given nonlinear activation $\phi(\cdot)$ they derived the set of initialisation parameters (σ_w, σ_b), denoted the edge of chaos (EoC), which separates the parameter space where nearby points converge to one another (ordered phase) from the domain where nearby points diverge (chaotic phase); see for example Figure 1 for $\phi(\cdot) = \tanh(\cdot)$. The EoC conditions were later shown by [1] and [3] to similarly control the size of entries in the gradients used to train DNNs and are essential for training DNNs with many layers.

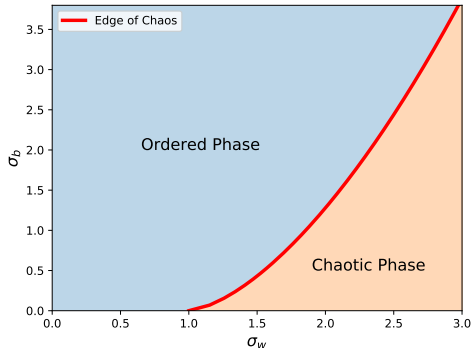


Fig. 1: Relation between σ_b and σ_w at the Edge of Chaos for a feed-forward DNN with $\phi(\cdot) = tanh(\cdot)$. For combinations within the ordered phase, the similar inputs converge to the same output and the gradients vanish with depth, while within the chaotic phase similar inputs diverge and the gradients explode.

Here we conduct an alternative information theoretic investigation of random feed-forward DNNs in order to determine how the DNN parameters $(\sigma_w, \sigma_b, \phi(\cdot))$ impact the flow of information through the DNN. We derive the lower bound of the mutual information between the input $\mathbf{X}$ and its associated hidden layer value $\mathbf{h}^{(\ell)}$, denoted $I(\mathbf{X}; \mathbf{h}^{(\ell)})$. Mutual Information (MI) is a measure of the dependence of two random variables. Given two variables $(\mathbf{X}, \mathbf{Y}) \in \mathcal{X} \times \mathcal{Y}$, MI is defined as the Kullback–Leibler divergence between the joint distribution $P_{(\mathbf{X},\mathbf{Y})}$ and the marginal distributions $P_\mathbf{X}$ and $P_\mathbf{Y}$

$$I(\mathbf{X};\mathbf{Y}) = D_{KL}(P_{(\mathbf{X},\mathbf{Y})} || P_\mathbf{X} \otimes P_\mathbf{Y}) \tag{3}$$

$$= \int_\mathcal{X}\int_\mathcal{Y} p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) \log\left(\frac{p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})}{p_\mathbf{X}(\mathbf{x})p_\mathbf{Y}(\mathbf{y})}\right) d\mathbf{x}\, d\mathbf{y}. \tag{4}$$

Our main contribution are lower bounds on $I(\mathbf{X}; \mathbf{h}^{(\ell)})$ which, similarly to [1], are functions of the DNN parameters $(\sigma_w, \sigma_b, \phi(\cdot))$. We include the mean field analysis of [3] in order to obtain an analytic approximation of the lower bounds on $I(\mathbf{X}; \mathbf{h}^{(\ell)})$ and observe that $I(\mathbf{X}; \mathbf{h}^{(\ell)})$ is maximised on the EoC, thus suggesting that initialisations which are optimal for geometric training analysis are also preferable from an MI perspective.

In Section II, we prove the existence of a lower bound on the MI problem by introducing the Gaussian model of the DNN. In Section III, we approximate this lower bound via the mean field approach. Finally in Section IV, we compare our lower bound approximation to a state-of-the-art MI approximator and see that the initialisation at edge of chaos similarly increases the MI at deep layers.

II. Gaussian Lower Bound on the MI

A MI analysis of DNNs was previously conducted in the unsupervised setting [4] and furthermore in the analysis of deep learning architectures [5]–[9]. For example, the work in [5] proposed the use of MI to describe the state of the training of DNNs by plotting the MI between an input $\mathbf{X}$ and the hidden layer $\mathbf{h}^{(\ell)}$, $I(\mathbf{X};\mathbf{h}^{(\ell)})$ against the MI between the hidden layer $\mathbf{h}^{(\ell)}$ and the output $\mathbf{Y}$, $I(\mathbf{h}^{(\ell)};\mathbf{Y})$. The focus of [5] is on the dynamics of the training process, as opposed to the dependence on DNN parameters $(\sigma_w, \sigma_b, \phi(\cdot))$ at initialisation. In [10], it was observed experimentally that the MI planes advocated by [5] depend strongly on the parameters $(\sigma_w, \sigma_b, \phi(\cdot))$. The mathematical analysis here compliments the observations in [10].

We compute the MI $I(\mathbf{X};\mathbf{h}^{(\ell)})$ between the input $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma_x^2\mathbf{I})$, as also modelled in [11], and the hidden layer $\mathbf{h}^{(\ell)}$ by conditioning on a realisation of the random weights $\mathcal{W}^{:\ell} = \{\mathbf{W}^{(i)}, \mathbf{b}^{(i)}\}_{i=1}^{\ell}$

$$I(\mathbf{X};\mathbf{h}^{(\ell)}) = \mathbb{E}_{\mathcal{W}^{:l}}\left[I(\mathbf{X};\mathbf{h}^{(\ell)}|\mathcal{W}^{:\ell})\right] \tag{5}$$

where $\mathbf{n}^{(\ell)}$ is a Gaussian noise variable that we add before the activation function as shown in Figure 2. This noise term is necessary because once the the DNN map is sampled, $\mathbf{X} \mapsto \mathbf{h}^{(\ell)}$ is completely deterministic and consequently, the MI $I(\mathbf{X};\mathbf{h}^{(\ell)}|\mathcal{W}^{:\ell})$ between the hidden layers and the input is infinite. Initially, this was solved by considering binning the values in the hidden variables [5], but the results were found to be too dependent on the choices of the bins [7]. The research in [6] and [9] showed that it is more appropriate to consider a random noise $\mathbf{n}^{(\ell)} \sim \mathcal{N}(0, \sigma_n^2)$ added at each layer.

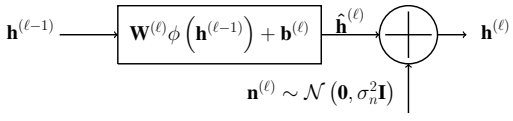


Fig. 2: Signal Propagation in the considered perturbed DNN.

For two random variables with a generic distribution, the MI is not known explicitly and its approximation is a challenging task that has been attempted primarily with non-parametric models [8], [12], [13]. However, if two random variables $\mathbf{x}$ and $\mathbf{y}$ are Gaussian with marginal covariances, $\mathbf{\Lambda}_x$ and $\mathbf{\Lambda}_y$, and joint covariance, $\mathbf{\Lambda}_{xy}$, the MI is

$$GMI(\mathbf{X};\mathbf{Y}) = \frac{1}{2}\log\left(\frac{|\mathbf{\Lambda}_x|\,|\mathbf{\Lambda}_y|}{|\mathbf{\Lambda}_{xy}|}\right) \tag{6}$$

While the distribution of $\mathbf{h}^{(\ell)}$ is known to converge to a Gaussian distribution with large depth ℓ, in order to lower bound the MI throughout the layers we note that the MI for a general distribution is lower bounded by that of a Gaussian distribution.

Proposition 1. *Let g(x,y) be an n-dimensional Gaussian distribution , $\mathcal{N}(\mu, \Lambda)$, with mean μ and covariance matrix Λ. If f(x,y) is an arbitrary distribution with the same mean and covariance matrix, and with f(x) Gaussian, then*

$$I_{f_{(xy)}}[X,Y] \geq I_{g_{(xy)}}[X,Y] \tag{7}$$

Proof.

$$\begin{aligned}
I_f[X,Y] - I_g[X,Y] &= \int_{\mathcal{X},\mathcal{Y}} f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) \log\left(\frac{f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})}{f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}}(\mathbf{y})}\right) d\mathbf{x}d\mathbf{y} \\
&\quad - \int_{\mathcal{X},\mathcal{Y}} g_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) \log\left(\frac{g_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})}{g_{\mathbf{X}}(\mathbf{x}) g_{\mathbf{Y}}(\mathbf{y})}\right) d\mathbf{x}d\mathbf{y} \\
&= -\int_{\mathcal{X}} f_{\mathbf{X}}(\mathbf{x}) \log\left(f_{\mathbf{X}}(\mathbf{x})\right) d\mathbf{x} - \int_{\mathcal{Y}} f_{\mathbf{Y}}(\mathbf{y}) \log\left(f_{\mathbf{Y}}(\mathbf{y})\right) d\mathbf{y} \\
&\quad + \int_{\mathcal{X},\mathcal{Y}} f_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) \log\left(f_{\mathbf{XY}}(\mathbf{x},\mathbf{y})\right) d\mathbf{x}d\mathbf{y} + \int_{\mathcal{X}} g_{\mathbf{X}}(\mathbf{x}) \log\left(g_{\mathbf{X}}(\mathbf{x})\right) d\mathbf{x} \\
&\quad + \int_{\mathcal{Y}} g_{\mathbf{Y}}(\mathbf{y}) \log\left(g_{\mathbf{Y}}(\mathbf{y})\right) d\mathbf{y} - \int_{\mathcal{X},\mathcal{Y}} g_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) \log\left(g_{\mathbf{XY}}(\mathbf{x},\mathbf{y})\right) d\mathbf{x}d\mathbf{y} \\
&= -\int_{\mathcal{Y}} f_{\mathbf{Y}}(\mathbf{y}) \log\left(f_{\mathbf{Y}}(\mathbf{y})\right) d\mathbf{y} + \int_{\mathcal{X},\mathcal{Y}} f_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) \log\left(f_{\mathbf{XY}}(\mathbf{x},\mathbf{y})\right) d\mathbf{x}d\mathbf{y} \\
&\quad + \int_{\mathcal{Y}} g_{\mathbf{Y}}(\mathbf{y}) \log\left(g_{\mathbf{Y}}(\mathbf{y})\right) d\mathbf{y} - \int_{\mathcal{X},\mathcal{Y}} g_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) \log\left(g_{\mathbf{XY}}(\mathbf{x},\mathbf{y})\right) d\mathbf{x}d\mathbf{y} \\
&= \int_{\mathcal{Y}} f_{\mathbf{Y}}(\mathbf{y}) \log\left(\frac{g_{\mathbf{Y}}(\mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})}\right) d\mathbf{y} \\
&\quad - \int_{\mathcal{X},\mathcal{Y}} f_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) \log\left(\frac{g_{\mathbf{XY}}(\mathbf{x},\mathbf{y})}{f_{\mathbf{XY}}(\mathbf{x},\mathbf{y})}\right) d\mathbf{x}d\mathbf{y} \\
&= \int_{\mathcal{X},\mathcal{Y}} f_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) \log\left(\frac{g_{\mathbf{Y}}(\mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})}\right) d\mathbf{x}d\mathbf{y} \\
&\quad - \int_{\mathcal{X},\mathcal{Y}} f_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) \log\left(\frac{g_{\mathbf{XY}}(\mathbf{x},\mathbf{y})}{f_{\mathbf{XY}}(\mathbf{x},\mathbf{y})}\right) d\mathbf{x}d\mathbf{y} \\
&= \int_{\mathcal{X},\mathcal{Y}} f_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) \log\left(\frac{g_{\mathbf{Y}}(\mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})} \frac{f_{\mathbf{XY}}(\mathbf{x},\mathbf{y})}{g_{\mathbf{XY}}(\mathbf{x},\mathbf{y})}\right) d\mathbf{x}d\mathbf{y} \\
&= \int_{\mathcal{Y}} f_{\mathbf{Y}}(\mathbf{y}) \int_{\mathcal{X}} f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{Y}=\mathbf{y}) \log\left(\frac{f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{Y}=\mathbf{y})}{g_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{Y}=\mathbf{y})}\right) d\mathbf{x}d\mathbf{y} \\
&= \int_{\mathcal{Y}} f_{\mathbf{Y}}(\mathbf{y}) D_{KL}(f_{\mathbf{X}|\mathbf{Y}=\mathbf{y}} || g_{\mathbf{X}|\mathbf{Y}=\mathbf{y}}) d\mathbf{y} \geq 0
\end{aligned}$$

where the third equality is due to f_x being Gaussian, hence $f_x = g_x$; the fourth is due to [14] where it is shown that for the distributions f and g considered here

$$\int_{\mathcal{X}} f(\mathbf{x}) \log(g(\mathbf{x}) d\mathbf{x} = \int_{\mathcal{X}} g(\mathbf{x}) \log(g(\mathbf{x})) d\mathbf{x}$$

□

The analysis herein relies on the first two moments of $\left[\mathbf{X}^\top,\ \mathbf{h}^{(\ell)\top}\right]^T$ which for DNNs drawn according to (2) with $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma_x^2\mathbf{I})$ are:

$$\mathbb{E}_{\mathbf{X}|\mathcal{W}^{:\ell}} \begin{bmatrix} \mathbf{X} \\ \mathbf{h}^{(\ell)} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\mu} \end{bmatrix} \tag{8}$$

$$\boldsymbol{\Lambda}_{xh^{(\ell)}} := \mathbb{V}ar_{\mathbf{X}|\mathcal{W}^{:\ell}} \begin{bmatrix} \mathbf{X} \\ \mathbf{h}^{(\ell)} \end{bmatrix} = \begin{bmatrix} \sigma_x^2\mathbf{I} & \boldsymbol{\Sigma}_{xh^{(\ell)}} \\ \boldsymbol{\Sigma}_{xh^{(\ell)}}^\top & \boldsymbol{\Lambda}_{h^{(\ell)}} \end{bmatrix}. \tag{9}$$

In order to compute the GMI in (6), we reformulate the determinant of the block covariance matrix (9) using the decomposition [15]:

$$\begin{vmatrix} \boldsymbol{\Lambda}_x & \boldsymbol{\Lambda}_{xy} \\ \boldsymbol{\Lambda}_{xy}^\top & \boldsymbol{\Lambda}_{\boldsymbol{y}} \end{vmatrix} = |\boldsymbol{\Lambda}_x| \left|\boldsymbol{\Lambda}_y - \boldsymbol{\Lambda}_{xy}^\top \boldsymbol{\Lambda}_x^{-1} \boldsymbol{\Lambda}_{xy}\right|. \tag{10}$$

Incorporating (10) for the determinant of $\boldsymbol{\Lambda}_{xh^{(l)}}$ in (9) we define the following lower bound for the MI for a DNN conditional on a set of weights $\mathcal{W}^{:\ell}$ with Gaussian input

$$I(\mathbf{X};\mathbf{h}^{(\ell)}|\mathcal{W}^{:\ell}) \quad \geq \frac{1}{2} \log \left(\frac{|\boldsymbol{\Lambda}_{h^{(\ell)}}|}{\left|\boldsymbol{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2} \boldsymbol{\Sigma}_{xh^{(\ell)}}^\top \boldsymbol{\Sigma}_{xh^{(\ell)}}\right|} \right). \tag{11}$$

Thus, from (5) the MI of a DNN is bounded as follows

$$I(\mathbf{X};\mathbf{h}^{(\ell)}) \geq \frac{1}{2} \underbrace{\mathbb{E}_{\mathcal{W}^{:\ell}} \left[\log\left(|\boldsymbol{\Lambda}_{h^{(\ell)}}|\right)\right]}_{(E_1)} \tag{12}$$

$$\underbrace{-\mathbb{E}_{\mathcal{W}^{:\ell}} \left[\frac{1}{2} \log\left(\left|\boldsymbol{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2} \boldsymbol{\Sigma}_{xh^{(\ell)}}^\top \boldsymbol{\Sigma}_{xh^{(\ell)}}\right|\right)\right]}_{(E_2)}. \tag{13}$$

A. Numerical Evaluation of the Gaussian Lower Bound

The quantities (E_1) and (E_2) in (12) and (13) respectively can be computed by sampling different realisations of the weights $\mathcal{W}^{:\ell}$ and by then averaging over the log-determinant of the matrices $\boldsymbol{\Lambda}_{h^{(\ell)}}$ and $\boldsymbol{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2} \boldsymbol{\Sigma}_{xh^{(\ell)}}^\top \boldsymbol{\Sigma}_{xh^{(\ell)}}$. Here we compute these estimates by sampling 10^5 different inputs $\mathbf{X}$ and noises $\mathbf{n}$ with $\sigma_x = 1$ and $\sigma_n = 0.1$. Figure 3 shows how the Gaussian lower bounds changes as a function of σ_w for DNNs with square weight matrices of size $n \times n$ with $n = 90$, with the $tanh$ activation function, and with σ_b chosen such that (σ_w, σ_b) lies on the EoC.

These numerical experiments show how the MI between the input and the hidden layers decreases at each layer and is maximised for a value of (σ_w, σ_b) close to $(1, 0)$, as advocated in [2].

III. An Analytic Gaussian Lower Bound

The numerical computation of (E_1) and (E_2) based on sampling the covariance matrix is computationally expensive for large layer width n and depth L. Moreover, it less directly shows how this observed MI lower bound compares with the mean field analysis in [2] and the corresponding EoC analysis. For easier computation and to better link these analyses, we approximate the matrices $\boldsymbol{\Lambda}_{h^{(l)}}$ and $\boldsymbol{\Lambda}_{h^{(l)}} - \frac{1}{\sigma_x^2} \boldsymbol{\Sigma}_{xh^{(l)}}^\top \boldsymbol{\Sigma}_{xh^{(l)}}$ based on the mean field assumption of [2].

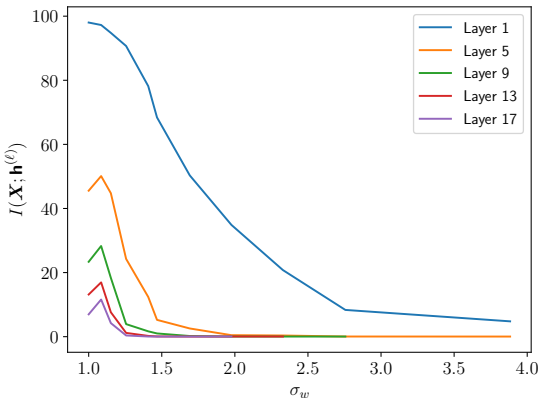


Fig. 3: Numerical approximation by sampling of the Gaussian lower bound in (12)-(13) for the conditional MI of a feed-forward DNN when σ_w varies along the EoC with $n = 90$ and $\phi(\cdot) = tanh(\cdot)$.

A. Mean Field Approximation

By the mean field analysis in [2], the hidden layers are normally distributed in the large limit, i.e. $n \gg 1$, with the mean and variance given in (8) and (9) respectively. In order to compute (12) and (13), we model the factors $\boldsymbol{\Lambda}_{h^{(l)}}$ and $\boldsymbol{\Lambda}_{h^{(l)}} - \frac{1}{\sigma_x^2} \boldsymbol{\Sigma}_{xh^{(l)}}^\top \boldsymbol{\Sigma}_{xh^{(l)}}$ respectively by their expectation over the weights $\mathcal{W}^{:\ell}$.

1) Expectation of $\boldsymbol{\Lambda}_{h^{(\ell)}}$: The expectation of $\boldsymbol{\Lambda}_{h^l}$ in (12) was studied in [2], [3] and was shown that for large DNN width $n \gg 1$

$$\mathbb{E}_{\mathcal{W}^{:\ell}} \left[\boldsymbol{\Lambda}_{h^{(\ell)}}\right] = q^{(\ell)}\mathbf{I}, \tag{14}$$

where $q^{(\ell)}$ corresponds to the variance of the hidden layer signal and is defined recursively via

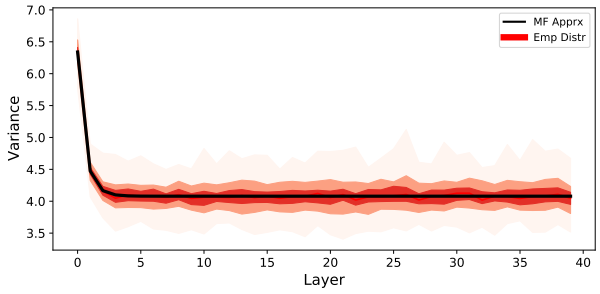
$$q^{(\ell)} = \sigma_w^2 \int \phi(\sqrt{q^{(\ell-1)}}z)\mathcal{D}z + \sigma_b^2 + \sigma_n^2 \tag{15}$$

$$q^{(1)} = \sigma_w^2\sigma_x^2 + \sigma_b^2 + \sigma_n^2. \tag{16}$$

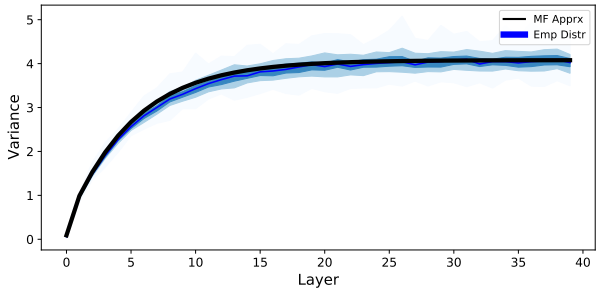
In Figure 4a we demonstrate the validity of this approximation by plotting $q^{(\ell)}$ as the solid black line along with the empirical distributions obtained by generating 100 realisations of (1) for $(\sigma_w, \sigma_b) = (2.5, 0.3)$, $\phi(\cdot) = tanh(\cdot)$ and $n = 50$ and showing the distribution over 10^4 randomly drawn inputs $\mathbf{X}$. Note the good agreement of $q^{(l)}$ and the empirical values, as well as the limiting value for large depth, denoted q^* in [1]. The striking agreement is notable given the relatively small DNN width $n = 50$ and the mean field approximation following from the $n \to \infty$ limit.

2) Expectation of $\boldsymbol{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2} \boldsymbol{\Sigma}_{xh^{(\ell)}}^\top \boldsymbol{\Sigma}_{xh^{(\ell)}}$: To compute the expectation of the argument in the logarithm in (E_2) from (13), we can make use of (14) and then compute $\mathbb{E}_{\mathcal{W}^{:\ell}} \left[\boldsymbol{\Sigma}_{xh^{(\ell)}}^\top \boldsymbol{\Sigma}_{xh^{(\ell)}}\right]$. Since

$$\mathbb{E}_{\mathcal{W}^{:\ell}} \left[\boldsymbol{\Sigma}_{xh^{(\ell)}}^\top \boldsymbol{\Sigma}_{xh^{(\ell)}}\right]_{ij} = \mathbb{E}_{\mathcal{W}^{:\ell}} \left[\sum_k \mathbb{E}\left[\mathbf{x}_k \mathbf{h}_i^{(\ell)}\right] \mathbb{E}\left[\mathbf{x}_k \mathbf{h}_j^{(\ell)}\right]\right]$$



(a)



(b)

Fig. 4: In (a) and (b) we compare mean field approximation (MF Apprx) of the mean value on the diagonal of the matrices $\mathbf{\Lambda}_{h^l}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}^{\top}_{xh^{(\ell)}}\mathbf{\Sigma}_{xh^{(\ell)}}$ respectively with the empirical distribution (Emp. Distr) in function of the layers. The empirical distribution is obtained by considering 100 different set of weights $\mathcal{W}^{:\ell}$ for the DNN when $n = 50$ and with 10^5 different random inputs.

and as the weights in $\mathbf{W}^{(\ell)}$ are independent, the expected covariance matrix is diagonal with values

$$\mathbb{E}_{\mathcal{W}^{:\ell}}\left[\mathbf{\Sigma}^{\top}_{xh^{(\ell)}}\mathbf{\Sigma}_{xh^{(\ell)}}\right]_{ii} = \mathbb{E}_{\mathcal{W}^{:\ell}}\left[\sum_k \mathbb{E}\left[\mathbf{x}_k\mathbf{h}_i^{(\ell)}\right]^2\right]. \quad (17)$$

With the mean field analysis, we consider any elements i and j of respectively the input and the hidden layer to be jointly normally distributed according to

$$\begin{bmatrix}\mathbf{x}_i\\ \mathbf{h}_j^{(\ell)}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma_x^2 & \rho^{(\ell)}\sigma_x\sqrt{q^{(\ell)}}\\ \rho^{(\ell)}\sigma_x\sqrt{q^{(\ell)}} & q^{(\ell)}\end{bmatrix}\right) \quad (18)$$

with $\rho^{(\ell)}$ being the correlation coefficient. The correlation is given at each layer by solving

$$\rho^{(\ell)}\sigma_x\sqrt{q^{(\ell)}} = \mathbb{E}_{\mathcal{W}^{:\ell}}\left[\mathbf{x}_i\mathbf{h}_j^{(\ell)}\right] = \int\int u_1u_2\mathcal{D}u_1\mathcal{D}u_2 \quad (19)$$

$$= \underbrace{\int\int \sigma_x z_1 \frac{\sigma_w}{\sqrt{n}}\phi\left(\sqrt{q^{(\ell)}}\left(\rho^{(\ell-1)}z_1 + \sqrt{1-\rho^{(\ell-1)2}}z_2\right)\right)}_{(E_3)} \quad (20)$$

The expected squared covariance is then given by

$$\mathbb{E}_{\mathcal{W}^{:\ell}}\left[\sum_k \mathbb{E}\left[\mathbf{x}_k\mathbf{h}_i^{(\ell)}\right]^2\right] = n(E_3)^2\mathbf{I} \quad (21)$$

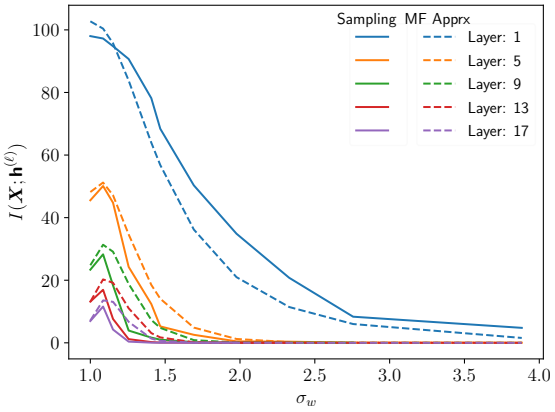


Fig. 5: Comparison of the analytical lower bound (23) from the mean field approximation to the sample MI obtained for a DNN with $n = 90$ and $tanh(\cdot)$ activation function when the variance σ_w is changed. Here σ_b is update so that (σ_w, σ_b) satisfy the EoC condition.

and consequently

$$\mathbb{E}\left[\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}^{\top}_{xh^{(\ell)}}\mathbf{\Sigma}_{xh^{(\ell)}}\right] = (q^{(\ell)} - n*(E_3)^2/\sigma_x^2)\mathbf{I} = q_c^{(\ell)}\mathbf{I}. \quad (22)$$

The validity of this approximation is shown in Figure 4b were as before we note the excellent agreement of the mean field limit (22) corresponding to width $n \to \infty$ and the observed distribution for (1) with width $n = 50$.

3) Lower Bound Approximation: Since the matrices $\mathbf{\Lambda}_{h^{(l)}}$ and $\mathbf{\Lambda}_{h^{(l)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}^{\top}_{xh^{(l)}}\mathbf{\Sigma}_{xh^{(l)}}$ are approximated as multiples of the identity matrix, their log-determinant can be easily computed; in particular

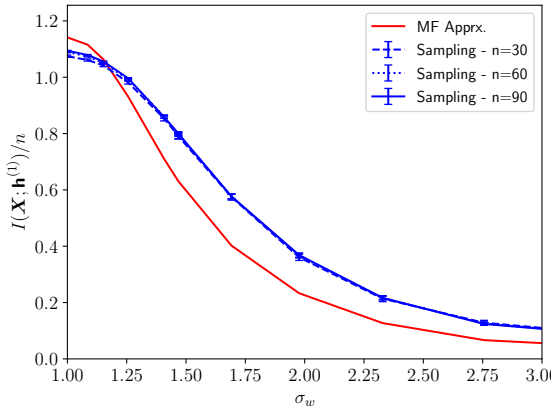
Proposition 2. *Under the mean field approximation the MI has the following lower bound*

$$I(\boldsymbol{X};\boldsymbol{h}^{(\ell)}) \geq \frac{n}{2}\log\left(\frac{q^{(\ell)}}{q_c^{(\ell)}}\right) \quad (23)$$

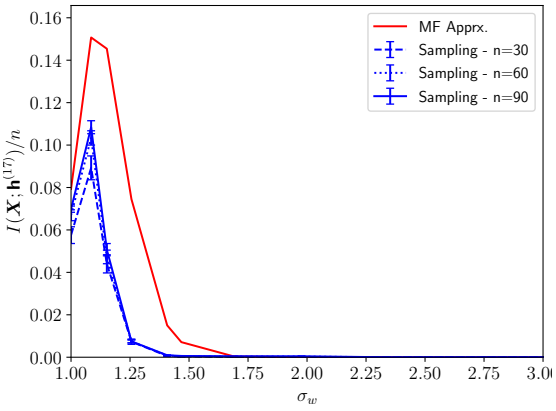
Figure 5 compares the mean field approximation from Prop. 2 with the direct sampling approach as described in Figure 3. We observe good general agreement, in particular at the locations where the MI is maximised. Figure 6 illustrates the sampling and mean field calculations for varying widths $n = 30, 60, 90$ as well as at layers $\ell = 1$ and 17; the error bars correspond to twice the standard deviation of 100 Monte Carlo approximations of $I(\mathbf{X};\mathbf{h}^{(\ell)})/n$ computed as the average of 100 samples of $I(\mathbf{X};\mathbf{h}^{(\ell)}|\mathcal{W}^{:\ell})$ with different weights where the covariance matrices were defined on 10,000 random inputs. Improved agreement is observed for increased DNN width as can be expected as the mean field analysis is the infinite width, $n \to \infty$ limit.

IV. Comparison of MI Approximation

The work in [10] showed that among different MI approximations as [12] and [13], the replica formula in [8] is the most



(a) Mutual Information at Layer 1



(b) Mutual Information at Layer 17

Fig. 6: Comparison of the analytic lower bound (23) of $I(\mathbf{X};\mathbf{h}^{(\ell)})/n$ and the sampled Gaussian lower bound in (12)-(13) for $n = 30, 60$, and 90 for layers 1 (a) and 17 (b) with two standard deviations error bars.

consistent measure with the arguments in [7], as it models with the decrease in MI for large enough variance σ_w^2 the loss in expressivity of the DNN due to the saturation of the $tanh$ activation function. Figure 7 compares the replica formula to the mean field lower bound for DNNs where the bias is null $\sigma_b = 0$, the noise has a variance $\sigma_n^2 = 10^{-5}$, and the input has i.i.d. normal elements, as the replica formula is applicable only on these DNNs. The results show that the replica formula and the approximated lower-bound are both maximised for a value of the standard deviation σ_w thus supporting the saturation argument in [7]. However, there is an inconsistency for low standard deviations σ_w due to the approximation of both methods.

Finally, in [10] the analysis with the replica formula suggested that the MI information converges to a non-trivial limit as the depth increases. Figure 8 shows a consistent behaviour where the mean field lower bound also converges to a maximum for $\sigma_w = 1$ as the depth increases, with $\sigma_b = 0$ and $\phi(\cdot) = tanh(\cdot)$. Since $(\sigma_w, \sigma_b) = (1, 0)$ is on the EoC for the $tanh$ activation function, these calculations suggests that the initialisations on the EoC, which considers a large depth limit, are preferable both for optimisation [3] and MI.

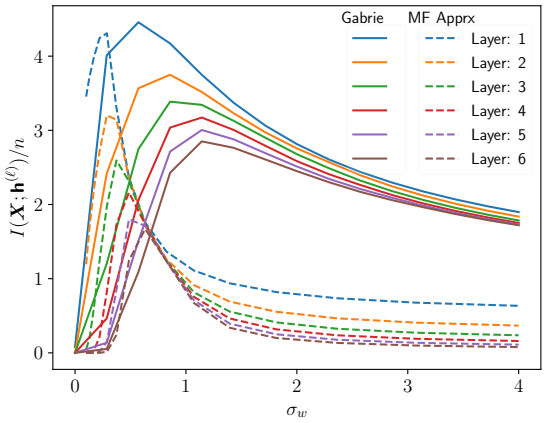


Fig. 7: Comparison between the [8] approximation of $I(\mathbf{X};\mathbf{h}^{(\ell)})/n$ and the lower bound obtained with mean field approximation for a DNN with $n = 1000$ when the σ_w is changed on the x axis and σ_b is kept fixed.

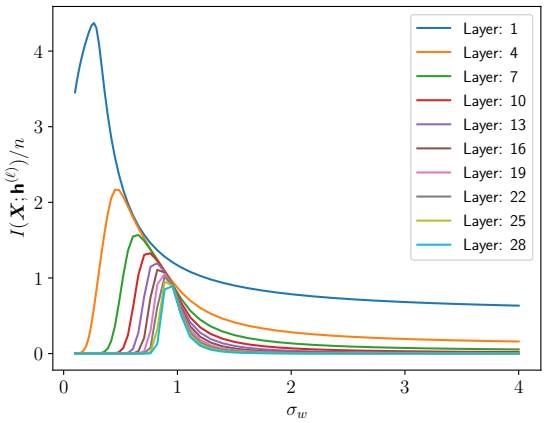


Fig. 8: Convergence of the lower bound of $I(\mathbf{X};\mathbf{h}^{(\ell)})/n$ obtained with the mean field approximation when we change the variance σ_w and keep $\sigma_b = 0$.

V. Conclusion

We have presented a lower bound of the MI for feed-forward DNNs and we derived an approximation with the mean field theory which numerical experiments showed to be consistent with the original measure. The analytic lower bound approximation allows direct investigation of how the MI of DNNs change for different initialisation parameters $(\sigma_w, \sigma_b, \phi(\cdot))$. In particular, we observe that with $\phi(\cdot) = tanh(\cdot)$ activation function, the MI is maximised for $(\sigma_w \sigma_b)$ on the EoC, which suggests the EoC initialisation are similarly optimal from a MI perspective.

Acknowledgment

This publication is based on work supported by the EPSRC Centre for Doctoral Training in Industrially Focused Mathematical Modelling (EP/L015803/1) in collaboration with New Rock Capital Management and by the Alan Turing Institute under the EPSRC grant EP/N510129/1.

REFERENCES

[1] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, "Deep information propagation," *International Conference on Learning Representations (ICLR)*, 2017.

[2] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, "Exponential expressivity in deep neural networks through transient chaos," *Neural Information Processing Systems (NeurIPS)*.

[3] J. Pennington, S. S. Schoenholz, and S. Ganguli, "The emergence of spectral universality in deep networks," *Artificial Intelligence and Statistics (AISTATS)*, 2 2018.

[4] R. Linsker, "Self-organization in a perceptual network," *Computer*, vol. 21, no. 3, pp. 105–117, 1988.

[5] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," 2017.

[6] Z. Goldfeld, E. van den Berg, K. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, and Y. Polyanskiy, "Estimating information flow in deep neural networks," *International Conference on Machine Learning (ICML)*, 2018.

[7] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, "On the information bottleneck theory of deep learning," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 12, p. 124020, 2019.

[8] M. Gabrié, A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, and L. Zdeborová, "Entropy and mutual information in models of deep neural networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 12, p. 124014, 2019.

[9] B. Foggo and N. Yu, "On the maximum mutual information capacity of neural architectures," 6 2020. [Online]. Available: https://arxiv.org/abs/2006.06037

[10] V. Abrol and J. Tanner, "Information-bottleneck under mean field initialization," *International Conference on Machine Learning (ICML) Workshop on Uncertainty and Robustness in Deep Learning*, 2020.

[11] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, "The loss surfaces of multilayer networks," in *Artificial intelligence and statistics*. PMLR, 2015, pp. 192–204.

[12] A. Kolchinsky, B. D. Tracey, and S. Van Kuyk, "Caveats for information bottleneck in deterministic scenarios," *International Conference on Learning Representations (ICLR)*, 2019.

[13] M. Noshad, Y. Zeng, and A. O. Hero, "Scalable mutual information estimation using dependence graphs," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2962–2966.

[14] S. Höt, *Information and Communication Theory*. Wiley, 9 2019.

[15] G. Strang, G. Strang, G. Strang, and G. Strang, *Introduction to linear algebra*. Wellesley-Cambridge Press Wellesley, MA, 1993, vol. 3.