Information-Bottleneck under Mean Field Initialization

V. Abrol [*1] **J. Tanner** [*1]

Abstract

This work explores the sensitivity of mutual information (MI) flow in hidden layers of very deep neural networks (DNNs) as a function of the initialization variance. Specifically, we demonstrate that information-bottleneck (IB) interpretations of DNNs are significantly affected by their choice of nonlinearity as well as weight/bias variance. Initialization on the network mean field (MF) edge of chaos (EOC) results in maximal information propagation through layers of even very deep networks; consequently their IB plots are effectively single points which do not vary and high accuracy is rapidly obtained with training. Alternatively, initialization away from EOC results in loss of MI through depth and the more characteristic IB plots observed in the literature. We also demonstrate that popular MI estimators give substantially different estimates, especially for sigmoidal nonlinearity and high weight variance.

1. Introduction

The choice of weight initialization and nonlinearity for a deep neural network (DNN) has a crucial impact on both the performance of the model and the overall training dynamics. Studies in (Schoenholz et al., 2017; Xiao et al., 2018) developed the mean field (MF) theory to understand the properties of untrained random DNNs with the aim to avoid the problem of vanishing/exploding gradients. Further, (Poole et al., 2016) showed that deep neural networks of arbitrary depth can be trained for specific choices of initial weight and bias variance chosen on a operating curve known as the 'Edge of Chaos' (EOC). Prior MF theory has as its focus on enhancing signal propagation and training in very deep networks. Here we show that MF theory also plays an important role in information propagation as viewed through information bottleneck (IB) theory (Tishby & Zaslavsky, 2015) whose focus is to explain generalization error of DNNs.

[*]Equal contribution [1]Mathematical Institute, University of Oxford, UK. Correspondence to: V. Abrol <abrol@maths.ox.ac.uk>.

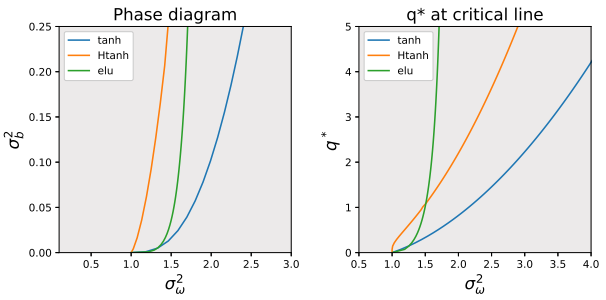
Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning.

The IB curve characterizes the set of 'bottleneck' hidden variables M that achieve maximal hidden-output MI $I(Y;M)$ while trying to achieve minimal input-hidden MI $I(X;M)$, for input and output variables X and Y (Shwartz-Ziv & Tishby, 2017). Assuming DNNs obeys the Markov condition $Y-X-M$, one can perform analysis using the trade-off between $I(Y;M)$ and $I(X,M)$ in the information plane (IP). Study in (Shwartz-Ziv & Tishby, 2017) showed that during training each layer in the network evolve from 1) an initial fitting phase maximizing $I(Y;M)$ to 2) a compression phase reducing $I(X;M)$. Study in (Saxe et al., 2018) demonstrated that this compression property occurs only for specific nonlinearities. Following this, various recent studies (Hodas & Stinis, 2018; Gabrié et al., 2018; Noshad et al., 2019; Wickstrøm et al., 2019) based on different MI estimators have supported the contradictory claims of either (Shwartz-Ziv & Tishby, 2017) or (Saxe et al., 2018).

Existing works have not study IB for very deep networks since estimation of MI for DNN is difficult and it is unclear if existing estimators are robust and/or computationally tractable in higher dimensions. In this paper we study the behaviour of popular MI estimators namely replica estimator (Gabrié et al., 2018), kernel-density estimator (KDE) (Kolchinsky & Tracey, 2017), and ensemble dependency graph estimator (EDGE) (Noshad et al., 2019), on DNNs under MF initialization. We demonstrate inconsistencies between different MI estimators; specifically, at large depths as well as high weight variance existing MI estimators suffers from MI overestimation which would be reflected in IB plots as artificially high MI flow through layers that are inconsistent with inability to train DNNs with such initialization. For practical regimes, regardless of the choice of nonlinearity, we advocate in favour of DNN initialization on EOC for which most MI estimators are well behaved and MI is maximized in most layers. The reason could be understood in terms of efficient signal propagation during both back-propagation (gradient flow) and forward-propagation (feature learning). We support our claims by first analysing IB for untrained random DNNs and then via experiments by training DNNs on real data.

2. Review of Signal Propagation in DNNs

Consider an untrained feed-forward neural network of depth L with weight matrices $\mathbf{W}^l \in \mathbb{R}^{N\times N}$, bias vectors $\mathbf{b}^l \in$



(a) (b)

Figure 1. (a) EOC phase transition with critical line $\chi = 1$ for different nonlinearities. Gradients either explodes and vanishes for (σ_w, σ_b) on right and left side of each EOC curve, respectively. For ReLU EOC is a singleton $(\sqrt{2}, 0)$. (b) σ_w as a function of q^*.

$\mathbb{R}^N$, pre-activations $\mathbf{h}^l \in \mathbb{R}^N$, and post-activations $\mathbf{x}^l \in \mathbb{R}^N$. The signal propagation in DNN is described by:

$$\mathbf{h}^l = \mathbf{W}^l \mathbf{x}^{l-1} + \mathbf{b}^{l-1}, \; \mathbf{x}^l = \phi(\mathbf{h}^l), \tag{1}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a pointwise nonlinearity and the input is $\mathbf{x}^0$ (Schoenholz et al., 2017). The biases $\mathbf{b}_i^l$ are drawn i.i.d. from a zero-mean Gaussian with variance σ_b^2, and weights $\mathbf{W}_{ij}^l$ are either: (1) drawn i.i.d from a zero-mean Gaussian with variance σ_w^2/N, or (2) drawn from a uniform distribution over scaled orthogonal matrices i.e., $\mathbf{W}^{l^T}\mathbf{W}^l = \sigma_w \mathbf{I}$.

2.1. MF Theory: Limiting Behaviour of Variance

Study in (Poole et al., 2016; Sirignano & Spiliopoulos, 2019) established that in large N limit the empirical distribution of pre-activations converges to a zero mean Gaussian and fixed point variance $q^* = \sigma_w^2 \mathbb{E}[\phi(\sqrt{q^*}z)^2] + \sigma_b^2$. The propagation of a pair of signals through such a network can be understood in a similar way where the covariance between different pre-activations follows a recursive relation (Schoenholz et al., 2017). The covariance has a fixed point correlation $c^* = 1$ which is stable when the quantity $\chi = \sigma_w^2 \mathbb{E}[\phi'(\sqrt{q^*}z)^2] < 1$, where ϕ' is the derivative of ϕ. Thus $\chi(\sigma_w, \sigma_b) = 1$ separates the (σ_w, σ_b) plane into chaotic $(\chi > 1)$ or ordered $(\chi < 1)$ regimes, where gradients either exponentially explode or vanish, respectively, and all inputs end up asymptotically correlated or decorrelated, respectively (Poole et al., 2016; Schoenholz et al., 2017). Interestingly χ is also the mean singular value of the input-output Jacobian when the pre-activations are at q^*. Fig.1(a-b) shows the phase diagrams for few nonlinearities for different values of q^*. Further, for a given q^* we have different choices of (σ_w, σ_b), and an optimal point on EOC is the one which results in a well conditioned Jacobian. This translates to singular values of the Jacobian to concentrate around 1, a property known as dynamical isometry in DNNs (Pennington et al., 2018; Tarnowski et al., 2019), which is achieved for smaller values of q^* and orthogonal weights. From Fig.1(b) it is evident that as $q^* \to 0$, $\sigma_b \to 0$ i.e., with a smaller bias the information propagates deeper in DNNs.

3. Information-bottleneck Principle

IB aims to find a bottleneck or hidden variable M to quantify the information flow in a DNN via input-hidden MI $I(X;M)$ and hidden-output MI $I(Y;M)$, which reflects how much a given hidden layer M compresses or forgets about X, and how well it predicts Y (Tishby et al., 1999; Tishby & Zaslavsky, 2015). In practice, one can observe the training dynamics using the trade-off between $I(Y;M)$ and $I(X,M)$ in the information plane (IP) for all layers in each epoch. In (Shwartz-Ziv & Tishby, 2017) authors used Tanh nonlinearity and MI estimation was performed by binning of the hidden neurons activities. The transformation via DNNs is usually deterministic and in order to obtain a finite MI estimate one needs to either use binning or add noise in output of the nonlinearity (Kolchinsky et al., 2019). Since, binning/noise is not inherent in the network architecture, different MI estimators can result in different IB behaviour.

3.1. Mutual Information vs Signal Propagation

Under the data processing inequality (DPI) (Cover & Thomas, 1991) we have $I(X;M) \geq I(Y;M)$. Also for a deterministic mapping $(X - M_{l-1} - M_l)$, we have $I(M_l; M_{l-1}) = I(X; M_l)$ and $I(X; M_{l-1}) \geq I(X; M_l)$, where M_l is the variable at layer l. In other words, as input propagates the hidden representation forgets about input X and learns more about output Y. MF initialization suggest another regime where M forgets very slowly, and is in-fact necessary if one wishes to train a very deep network. This is not surprising as the bottleneck variable now has to propagate the information about inputs and gradients through its full depth in order to be able to train a network, while at the same time learning meaningful features for a task. Study in (Schoenholz et al., 2017) showed that only for a specific choice of hyperparameters (initialization at EOC) the information stored in the correlation between inputs can propagate infinitely far in random networks, and some nonlinearities like Tanh and ELU are much better than others e.g., ReLU in signal propagation.

IB for Different Nonlinearities

IB trajectory is a function of nonlineary and weight/bias variance. In particular, double-sided saturating nonlinearities like Tanh yield a reduction in MI for larger weights, whereas unbounded nonlinearities such as ReLU/ELU do not show this phenomena (Saxe et al., 2018). While this provides an important insight about the behaviour of a nonlinearity, it is still not evident what is the impact of depth on MI estimation (and higher dimension) as most of the existing works have not study IB for very deep networks.

4. MI Estimation for Deep Random Networks

The replica estimator recently proposed in (Gabrié et al., 2018), see Fig.2(a-b), is specifically designed to compute MI in stochastic DNNs. The analytical MI computation for a three neuron network presented in (Saxe et al., 2018) (see Fig.5(a) in section A.1) is consistent with the replica estimator (Gabrié et al., 2018) at layer 2 in Fig.2(a-b). As the depth of the network increases, Fig.2(a) shows the MI for Htanh DNN converging to a curve which increases initially with σ_w up to the value $\sigma_w = 1$ which coincides with value indicated for stable MF information propagation for Htanh; and the MI decreases monotonically as σ_w is increased beyond 1. Similarly, Fig.2(b) shows the MI for the ReLU activation whose limiting distribution is a stable fixed point near $\sigma_w = \sqrt{2}$, the value indicated for stable mean field information propagation for ReLU, and converges to 0 below this value and increases as value of σ_w increases.

4.1. Non-parametric estimates of MI

The MI estimator (Gabrié et al., 2018) gives robust estimate but is restricted to a specific class of models, and computationally expensive to apply in practice. In contrast, while non-parametric estimators can approximate MI just from samples, study in (Gabrié et al., 2018) showed that they tend to over estimate MI. Various existing works based on IB theory employed non-parametric estimators to compute IB trajectory. Popular methods include kernel-density estimation (KDE) (Kolchinsky & Tracey, 2017), nearest neighbour (KNN) (Kraskov et al., 2004), ensemble dependency graph estimator (EDGE) (Noshad et al., 2019) and differential entropy estimator (Goldfeld et al., 2019). However, the behaviour of such estimators can be unstable in higher dimension as even for a known distribution the entropy computation is intractable in most cases. Also, in higher dimension as depth increases there are estimation errors, which are more prominent for large values of σ_w as entropy diverges and one needs to increase the noise to avoid such divergence. Another source of error is the choice of nonlinearity e.g., for ReLU MI increases without bound and most estimators are incapable of providing a stable estimate. As an illustration, Fig.2(b) shows the MI estimate for a ReLU network at large depths with replica estimator. Interestingly, the increase in MI is near exponential as depth increases from 2 to 6. However, at large values of σ_w one can also observe estimation errors. We believe this explains the inconstancy among arguments linking generalization and MI flow in DNNS in recent works of (Shwartz-Ziv & Tishby, 2017; Hodas & Stinis, 2018; Gabrié et al., 2018; Noshad et al., 2019; Wickstrøm et al., 2019), due to MI estimation errors in contrast to what theory suggests.

These observations are consistent with non-parametric entropy estimators which are the usual practical choice for

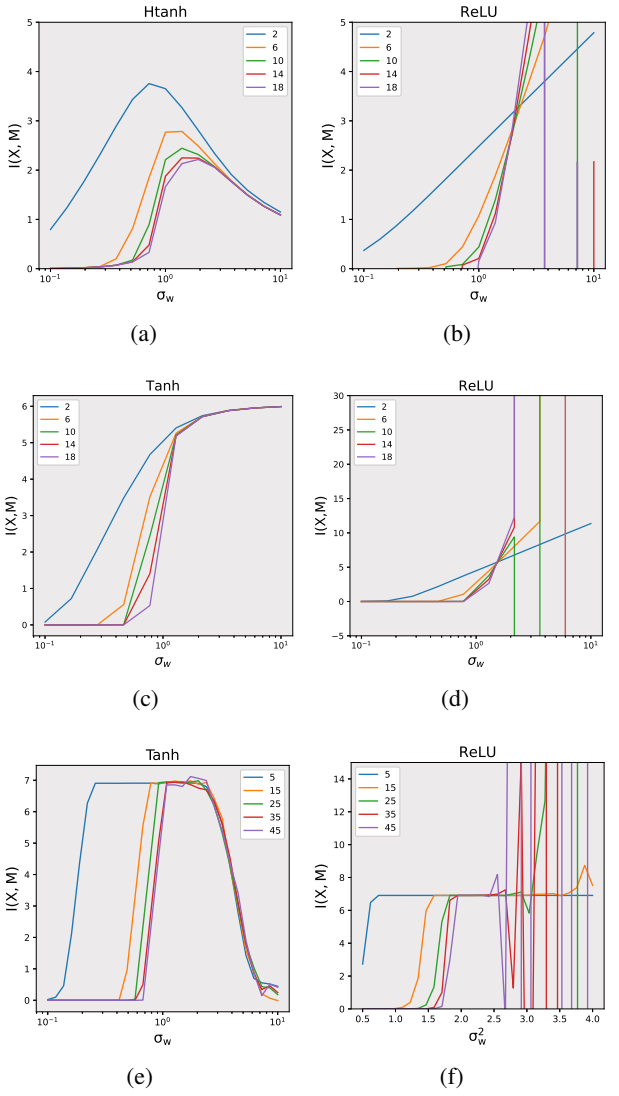
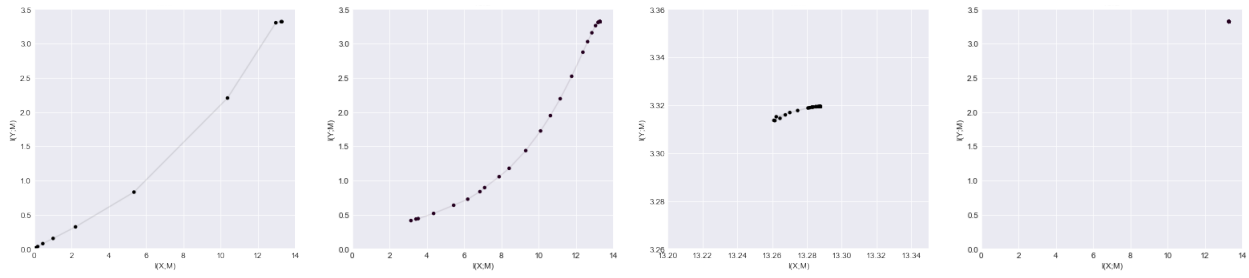


Figure 2. Mutual Information $I(X, M)$ with Gaussian input ($N = 1000$) as a function of weight scale σ_w for a random DNN without bias. (a-b) Replica; (c-d) KDE and (e-f) EDGE estimator. Each curve corresponds to a DNN with fixed number of layers. Vertical lines in each plots corresponds to estimation errors.

IB analysis of DNNs. For instance, Fig.2(c-d) and (e-f) shows the MI estimates for a network with Tanh and ReLU nonlinearity as a function of depth L using KDE and EDGE methods. It can be observed that for Tanh MI attains a maximum around $\sigma_w = 1$ and the behaviour converges as depth increases. However, for $\sigma_w > 1$ MI estimates of replica, KDE and EDGE method in Fig.2 (a), (c) and (e), respectively are very different. This observation is interesting as in contrast to the argument made in (Saxe et al., 2018) that saturating nonlinearities like Tanh yield a reduction in MI for larger weights; the actual MI estimation using KDE approach (also employed in their study) results



(a) $\sigma_w^2 = 1.2$; ACC=10.4% (b) $\sigma_w^2 = 1.7$; ACC=93.5% (c) $\sigma_w^2 = 2$; ACC=95.6% (d) $\sigma_w^2 = 2.3$; ACC=95.1%

Figure 3. IB dynamics using KDE method at initialization (epoch 0) for a ReLU network of depth 30, width 300 with different values of σ_w^2 and σ_b^2=0. Plot (c) is zoomed and MI for last layer in each plot is discarded for better visualization.

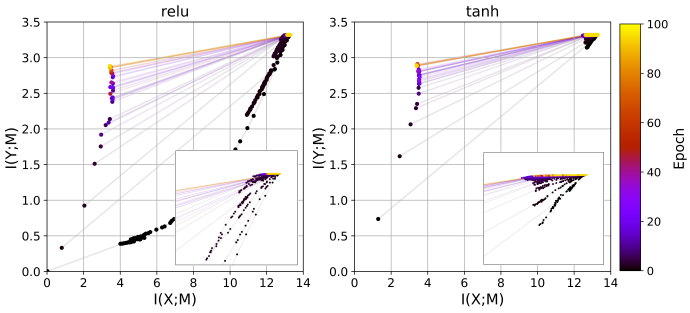
in saturated outputs. The MI estimation is also difficult for non-saturating nonlinearities like ReLU as shown in Fig.2(d) and (f), respectively. It can observed both KDE and EDGE methods result in large estimation errors for large values of σ_w, where now estimates in case of EDGE enters a saturation zone. This explains why contradictory IB plots one obtains when switching from the binning method used in (Shwartz-Ziv & Tishby, 2017) to KDE or KNN in (Saxe et al., 2018) and EDGE in (Noshad et al., 2019).

4.2. Experiments on MNIST

Reading IB plots: IB plots show MI between input, hidden and output layer. On x-axis we plot MI between each layer and the input, while on y-axis we plot MI between each layer and the output. Each layer produces a curve with the input layer at far right, output layer at far left, and different layers at the same epoch connected by fine lines. For instance, in Fig.3(b) points (3,0.45) and (13.8,3.4) corresponds to MI values at last and the first layer, respectively.

Impact of initialization on EOC: As demonstrated earlier initialization on EOC should maximise MI in DNN i.e., MI estimates in most layers should be close enough even without training and one should see closely concentrated values in IB plot. As an illustration, we visualize IB plots for a fixed σ_b and σ_w chosen to be on the EOC, as well as values of σ_w greater than and less than advocated by EOC. Also reported is the classification accuracy on MNIST test set after 100 epochs for each initialization.

Fig.3 shows the results for a 30 layer ReLU network with EOC being singleton at $(\sigma_w, \sigma_b = \sqrt{2}, 0)$. It can be observed that as expected, MI values with initialization on EOC are very close across layers. These MI estimates becomes nearly indistinguishable for larger values of σ_w. At value of σ_w below EOC, we observe a MI curve demonstrating different MI between input, hidden and output layers. Finally note that only the network trained on EOC achieves the best classification accuracy. Further, in Fig.4 we demonstrate IB plot for a very deep 200-layer network initialized



Gaussian Initialization

Figure 4. IB trajectories using KDE method for networks initialized on EOC and trained on MNIST. Each plot is overlaid with a zoomed version around coordinate (13,3.2) in IB plot.

on EOC with Gaussian initialization. See supplementary section for experimental details and results with orthogonal initialization. It can be observed that for ReLU except the last layer, MI for most layers quickly converges to a maximum attainable limit. In case of Tanh, the IB dynamics of the last layer rapidly increases from approximately (1.7,0.75) to (3.2,2.8), while the remaining layers start and remain at nearly maximum values of $I(X, M)$ and $I(Y, M)$.

5. Conclusion

Regardless of the choice of nonlinearity, in very deep networks initialization on EOC maximizes the MI in most layers which even prevails during actual training in practice. Thus, the link between generalization and mutual information in hidden representations is still elusive. For instance, ReLU and ELU has similar MI dynamics. Hence, while trainability issue due to ReLU at large depths can be solved for nonlinearities like ELU, IB theory can not explain its better generalization behaviour. We argue the reason for this is the inability of various MI estimators to provide reliable MI estimates in higher dimensions, large weight variance and large network depths.

A. Supplementary Material

A.1. Replication of Experiment from (Saxe et al., 2018)

This section presents a replication of three neuron network experiment from (Saxe et al., 2018), where MI can be estimated exactly using cumulative density function for monotonic nonlinearities. It can be observed from Fig.5(a) that for Tanh nonlinearity MI increases for small weights while it decreases for large ones, because outputs starts to saturate for large inputs. Hence, extreme bins concentrate more and more probability mass resulting in information loss. In contrast, MI for ReLU/ELU increases as weight variance increases.

The above result is validated using replica estimator from (Gabrié et al., 2018) for a two layer stochastic DNN without bias. Fig.5(b), compares the entropy estimates $H(M)$ of hidden representation for a DNN with linear, Htanh or ReLU nonlinearities. As observed in Fig.5(a), the entropy of hidden representation in case of Htanh nonlinearity increases with σ_w till reaching a maximum, whereas it always increases for ReLU. Similar observations can be made for estimates of $I(X;M)$ as shown in Fig. 5(c).

A.2. MI estimation for HTanh and ELU nonlinearity using KDE

We observe similar MI behaviour for a random DNN with Htanh and ELU nonlinearity as in case of Tanh and ReLU nonlinearity, respectively using KDE approach[1] as shown in Fig.6.

[1]Similar results are obtained using KNN based MI estimator.

A.3. Additional Experiments on MNIST

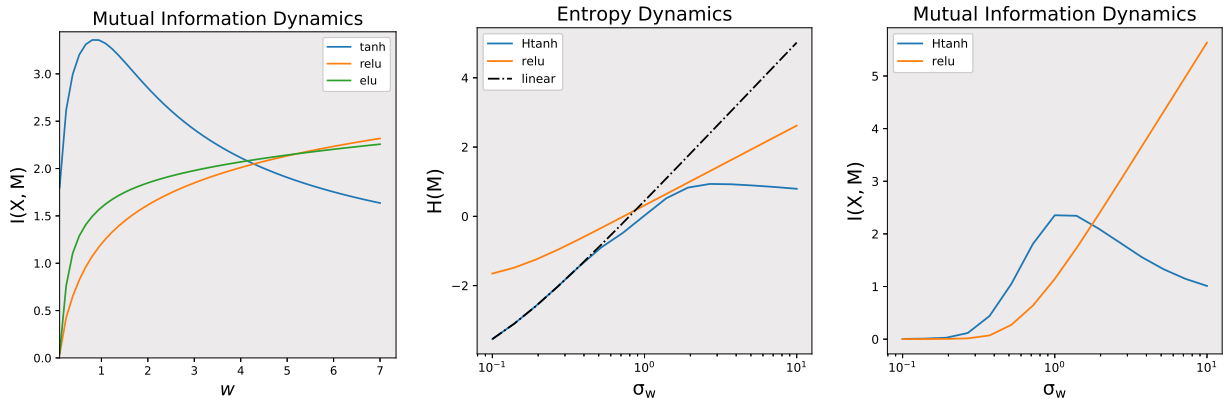
A.3.1. Impact of initialization on EOC for a Tanh Network

Similar to the case of ReLU (Fig.3), with initialization on EOC we expect MI estimates for most layers to be maximized. Fig.8 shows the IB plots for a 30 layer Tanh network with EOC chosen at $(\sigma_w, \sigma_b = 1.68, 0.038)$ corresponding to q^*=.5. It can observed that with values of σ_w on EOC and higher, all layers have the same MI in IB plots. Again the network trained on EOC achieves the best classification accuracy which decreases as σ_w increases. In line with observation in Fig.2 we observe the network doesn't train for $\sigma_w < 1$ due to approximately no flow of MI. Note that initialization on EOC (also value of q^* except in case of ReLU) is even more crucial for a sigmoidal type nonlinearity in order to train a very deep network.

A.3.2. Training at large depth

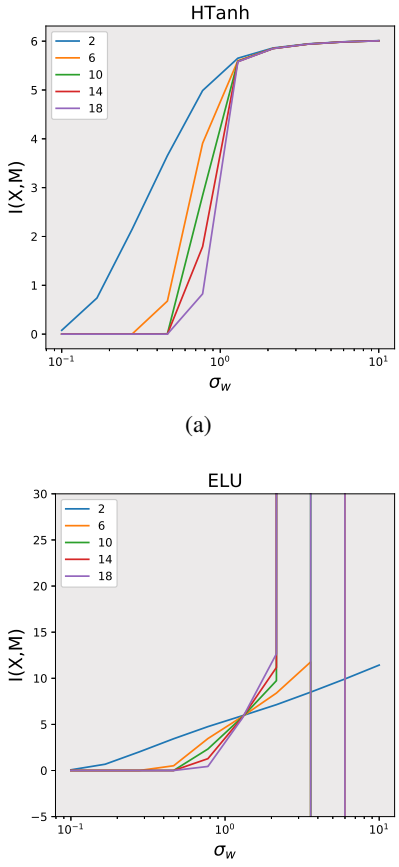
We train networks of depth L=200 and width N=400 for 100 epochs with a batch size of 64, and we set the optimal learning rate through grid search. For ReLU and Tanh nonlinearity we set q^*=1, and q^*=.5, respectively. The network achieved an accuracy of 93.60±.15%(25) and 95.21±.17%(20) with ReLU; 95.86±.24%(06) and 96.45±.23%(05) with Tanh nonlinearity for Gaussian and orthogonal initialization, respectively. To demonstrate training acceleration due to different initialization scheme, values in the bracket depict the average number of epochs required to achieve an accuracy of 90% on test set.

As expected, in Fig.9, we observe similar IB plots as in Fig.4. For lower values of q^* we have a better conditioned Jacobian in case of Tanh nonlinearity, and orthogonality



(a) (b) (c)

Figure 5. (a) Mutual information $I(X, M)$ with Gaussian input as a function of weight size 'w' in a three neuron network for different nonlinearities. (b-c) Entropy $H(M)$ and Mutual Information $I(X, M)$ with Gaussian input as a function of weight scale σ_w in a two layer random DNN ($N = 1000$) for different nonlinearities.



(a)

(b)

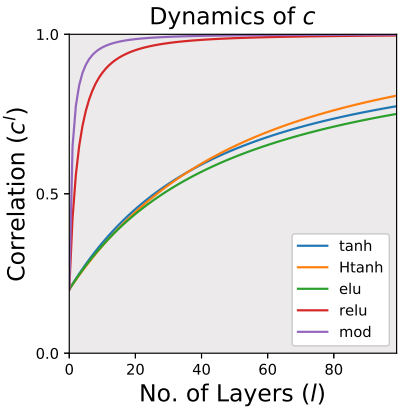
Figure 6. Mutual information $I(X;M)$ with Gaussian input as a function of weight scale σ_w for DNN ($N = 1000$) with different nonlinearity using KDE method. Each curve corresponds to a DNN with fixed number of layers.

bringing additional benefits of dynamical isometry too. Note that although orthogonality induces faster training times it does not achieves significantly better IB dynamics or training in case of ReLU nonlinearity. This is because first it is impossible to achieve dynamical isometry for ReLU even with orthogonal weights (Pennington et al., 2018), and secondly after subsequent updates, the network parameters deviates thus violating orthogonality criteria.

A.4. A note on differential entropy estimator

During submission we came across a recent study in (Goldfeld et al., 2019) which used differential entropy estimator to study information flow in stochastic DNNs. It argued that as the network trains, the clustering of the learned features at output layers is the underlying reason behind the reduction in MI across epochs. They demonstrated that prior works were in fact measuring this clustering through the lens of MI estimator based on binning. In terms of MF theory, as long as the information stored in the correlation between inputs propagates (without achieving the fixed point c^*), one can train a deep network i.e., learning useful clustered representations. As an illustration, Fig.9 shows the evolution of correlation c^l as inputs propagate through layers and a slow rate is desirable for deeper information propagation. Interestingly, while it seems the geometric phenomena described by (Goldfeld et al., 2019) relates to the limiting behaviour of correlations under MF initialization, we argue that this is tricky to observe in practice because lower dimensions may suppress it, yet most existing MI estimators are incapable of capturing it in higher dimension, especially when network depth is very large. Hence, in this work we restrict our study to the analysis of MI dynamics under MF initialization for specific choices of nonlinearities rather than the clustering behaviour.

Further, the authors of (Goldfeld et al., 2019) are still in process of publishing an implementation publicly and hence, a comparison with their proposed estimator could not be made.



(a)

Figure 7. Convergence of the correlations on EOC for different nonlinearities.

A.5. Reproducible Research

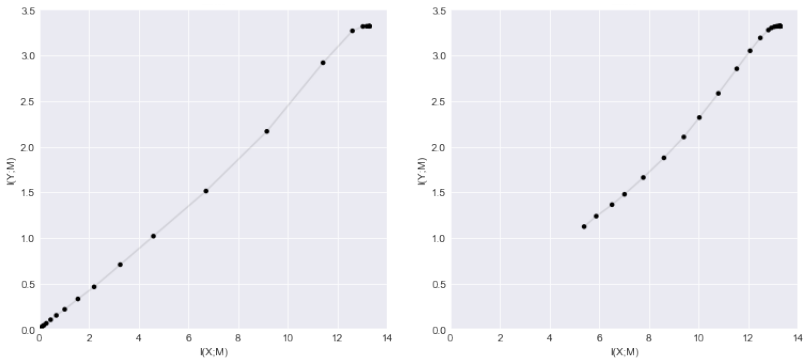
For reproducible-research purposes, a GPL Python implementation and related data to reproduce the figures in this work is available on request from authors. Alternatively, implementation of MI estimators proposed in existing works from respective authors is available at:

Replica: https://github.com/sphinxteam/dnner

Binning: https://github.com/ravidziv/IDNNs

KDE: https://github.com/artemyk/ibsgd/tree/iclr2018

EDGE: https://github.com/mrtnoshad/EDGE



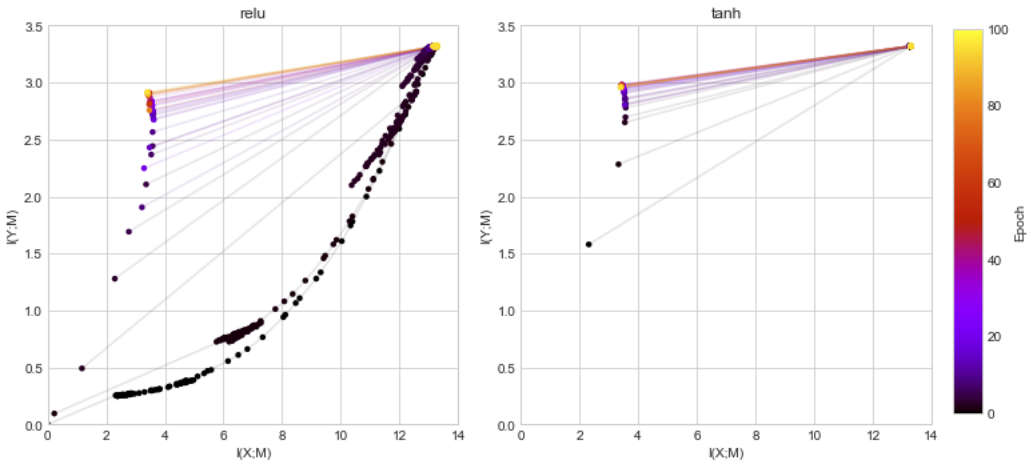
(a) $\sigma_w^2 = 0.85$; ACC=11.5%

(b) $\sigma_w^2 = 1.2$; ACC=11.5%

(c) $\sigma_w^2 = 1.68$; ACC=96.8%

(d) $\sigma_w^2 = 2$; ACC=94.2%

Figure 8. IB dynamics using KDE method at initialization (epoch 0) for a Tanh network of depth 30, width 300 with different values of σ_w^2, q^*=.5 and a fixed bias. Last layer is discarded for better visualization.



Orthogonal Initialization

Figure 9. IB trajectories using KDE method for networks initialized on EOC and trained on MNIST.

References

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. Wiley-Interscience, New York, NY, USA, 1991. ISBN 0-471-06259-6.

Gabrié, M., Manoel, A., Luneau, C., barbier, j., Macris, N., Krzakala, F., and Zdeborová, L. Entropy and mutual information in models of deep neural networks. In *Advances in Neural Information Processing Systems 31*, pp. 1821–1831, 2018.

Goldfeld, Z., Van Den Berg, E., Greenewald, K., Melnyk, I., Nguyen, N., Kingsbury, B., and Polyanskiy, Y. Estimating information flow in deep neural networks. In *International Conference on Machine Learning (ICML)*, pp. 2299–2308, June 2019.

Hodas, N. O. and Stinis, P. Doing the impossible: Why neural networks can be trained at all. *Frontiers in psychology*, 9, 2018. doi: 10.3389/fpsyg.2018.01185.

Kolchinsky, A. and Tracey, B. Estimating mixture entropy with pairwise distances. *Entropy*, 19(7):361, 2017.

Kolchinsky, A., Tracey, B. D., and Kuyk, S. V. Caveats for information bottleneck in deterministic scenarios. In *International Conference on Learning Representations*, 2019.

Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. *Physical Review E*, 69:066138, June 2004. doi: 10.1103/PhysRevE.69.066138.

Noshad, M., Zeng, Y., and Hero, A. O. Scalable mutual information estimation using dependence graphs. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2962–2966, May 2019. doi: 10.1109/ICASSP.2019.8683351.

Pennington, J., Schoenholz, S., and Ganguli, S. The emergence of spectral universality in deep networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1924–1932, April 2018.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. Exponential expressivity in deep neural networks through transient chaos. In *International Conference on Neural Information Processing Systems (NIPS)*, pp. 3368–3376, 2016.

Saxe, A. M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B. D., and Cox, D. D. On the information bottleneck theory of deep learning. In *International Conference on Learning Representations*, 2018.

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

Shwartz-Ziv, R. and Tishby, N. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Sirignano, J. and Spiliopoulos, K. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 2019. ISSN 0304-4149. doi: https://doi.org/10.1016/j.spa.2019.06.003.

Tarnowski, W., Warchol, P., Jastrzkebski, S., Tabor, J., and Nowak, M. A. Dynamical isometry is achieved in residual networks in a universal way for any activation function. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pp. 2221–2230, 2019.

Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pp. 1–5, April 2015. doi: 10.1109/ITW.2015.7133169.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. In *Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999.

Wickstrøm, K., Løkse, S., Kampffmeyer, M., Yu, S., Principe, J., and Jenssen, R. Information plane analysis of deep neural networks via matrix-based renyi's entropy and tensor kernels. *arXiv preprint arXiv:1909.11396*, 2019.

Xiao, L., Bahri, Y., Sohl-Dickstein, J., Schoenholz, S., and Pennington, J. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In *International Conference on Machine Learning (ICML)*, pp. 5393–5402, July 2018.