On the Maximum Mutual Information Capacity of Neural Architectures

Anonymous Authors[1]

Abstract

We derive the closed-form expression of the maximum mutual information - the maximum value of $I(X;Z)$ obtainable via training - for a broad family of neural network architectures. The quantity is essential to several branches of machine learning theory and practice. Quantitatively, we show that the maximum mutual information for these families all stem from generalizations of a single catch-all formula. Qualitatively, we show that the maximum mutual information of an architecture is most strongly influenced by the width of the smallest layer of the network - the "information bottleneck" in a different sense of the phrase, and by any statistical invariances captured by the architecture.

1. Introduction

Information Theory is a powerful supplement to statistical machine learning theory. It provides bounds that are agnostic to parameter size, and yields interpretable results through its relation to physical fields of study. Yet rigorous inspections of information-theoretic properties relating to the architecture of neural networks remain undeveloped. In this paper, we will study the maximum mutual information (MMI) that the network's hidden representation can hold from the feature space.

That is, we are studying the supremum $\sup_\theta I(X;Z_\theta)$ where θ is the parameter space of our architecture, Z is the random variable denoting the output of our final hidden layer (our representation), and $I(X;Z)$ is the mutual information between the representation and the input. The reason for studying this quantity is due to the prevalence of the following concept in Information-Theoretic Machine Learning Theory (ITMLT): $I(X;Z)$ acts as a data-sensitive measure of complexity for the representation. Indeed, we can often obtain bounds on machine learning performance in a way that is highly dependent on $I(X;Z)$. In this sense, $\sup_\theta I(X;Z_\theta)$ is a data-sensitive measure of complexity for the network itself. Thus its study can shed light on why certain architectures perform well on certain datasets and be useful in finding architectures for new datasets.

In a much more practical sense, Information Theoretic Machine Learning Theory establishes the existence of a "best" value of $I(X;Z)$, which maximizes the potential of obtaining a good representation. While this target always exists, it typically will not be known entirely. A sub-branch of this field argues that neural networks will attempt to find this target naturally through training (Shwartz-Ziv & Tishby, 2017; Tishby & Zaslavsky, 2015; Achille & Soatto, 2018), but this hypothesis is still disputed. Even if this hypothesis is true, there is no guarantee that the training process will find it. However, it should be possible to approximate this target with bounds. We can then choose our architecture to enforce the upper component of these bound via its maximum mutual information - significantly reducing the search space for the target $I(X;Z)$ in the training process.

While the capacity of neural networks has been analytically studied in a different sense (via storage of 'patterns') (MacKay & Mac Kay, 2003) (chapter 5), (Prados & Kak, 1989), and numerical methods of estimation of mutual information in deep networks exist (Gabrié et al., 2018; Paninski, 2003; Belghazi et al., 2018; Hjelm et al., 2018; Gao et al., 2017), no analytical studies of this type exist in literature.

2. Notation, Background, and Motivation

Our primary motivation for studying maximum mutual information comes from some modern theoretical work in ITMLT, which studies the classification of a discrete variable $Y \in \mathcal{Y}$ from an input $X \in \mathcal{X}$, jointly distributed as $p(x,y)$. The theory studies the potential losses in the quality of a learning machine's representation of its input, Z, when trained from just a sample of data $S = ((X_1,Y_1),\cdots,(X_m,Y_m))$.

A quality measure for a representational variable is its mutual information with the classification variable, $I(Y;Z)$ (Cover & Thomas, 2012). This is a measure of how well Z predicts the class variable Y, and is a component in the objective functions of several deep learning methods (Chen et al., 2016; Foggo & Yu, 2019b; Alemi et al., 2016;

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Kolchinsky et al., 2019; Bang et al., 2019). We can then conceive of a "best possible" representation being that which maximizes $I(Y;Z)$ when the true distribution $p(x,y)$ is known perfectly. This "best possible" representation is denoted Z^*. Of course, we typically don't know $p(x,y)$ perfectly, but we can still obtain a representation $\hat{Z}(S)$ from our samples. If we constrain both variables, Z^* and $\hat{Z}$, to have a fixed mutual information with X (i.e. $I(X;Z^*) = I(X;\hat{Z}) \triangleq I(X;Z)$), then we then have the following bound on the quality loss of $\hat{Z}(S)$ against Z^* (Foggo et al., 2019):

$$\underbrace{\left|I(Y;\hat{Z}(S)) - I(Y;Z^*)\right|}_{\text{quality loss}} \leq 2\,\underbrace{\bar{\delta}_{tv}}_{\text{data}}(S)\,\underbrace{I(X;Z)}_{\text{architecture}} + 2h_2(\bar{\delta}_{tv}(S)) \quad (1)$$

Note that the constraint $I(X;Z^*) = I(X;\hat{Z}) \triangleq I(X;Z)$ is a traditional choice first made in the early literature on the Information Bottleneck method (Tishby et al., 2000; Slonim & Tishby, 2000; Friedman et al., 2013) - a method which helped spark the existence of ITMLT. The constraint just exists to normalize the two representations to a fixed level of "forgetfullness" from the input. The under-braces on the right hand side of this inequality - implying that $\bar{\delta}_{tv}$ is data dependent and $I(X;Z)$ is architecture dependent in a decoupled way - is justified by the work introducing that bound (Foggo et al., 2019). In that work, $\bar{\delta}_{tv}$ was upper bounded in a way that did not depend on the architecture's complexity by terms on the order of $O\left(\sqrt{\frac{|\mathcal{Y}|m' log(2)}{2m}}\right)$ where m' is a small integer, and m is the size of the training sample. Further research has verified similar data dependent bounds (Foggo & Yu, 2019a). Since these bound did not depend on architectural complexity, all of the dependence on architecture must be contained in the other term: $I(X;Z)$.

On the other side of the coin, we have strong data processing inequalities (Polyanskiy & Wu, 2017; 2015; Anantharam et al., 2013) such as $I(Y;Z^*) \leq s^*(X,Y)I(X;Z)$ where $s^*(X,Y) = \sup_{r(x)\neq p(x)} \frac{D_{KL}(r(y)||p(y))}{D_{KL}(r(x)||p(x))}$.

A qualitative combination of these two contrasting inequalities is summarized as follows: $I(Y;Z^*)$ increases quickly with $I(X;Z)$ up to its maximum of $I(Y;X)$ [Strong Data Processing Inequality], and $I(Y;\hat{Z})$ follows this trend due to the low information losses at low $I(X;Z)$. But eventually the risk of information loss by inequality (1) forces our best estimate of $I(Y;\hat{Z})$ to decrease linearly from this maximum value. The location of this behavioral change is problem-specific. But since $\eta_{X,Y}$ can be estimated via a small sample of data, these equations can be used to approximate a target $I(X;Z)$ - yielding an interval in which it is likely to be contained. Setting our architecture to have a maximum mutual information near the supremum of this estimated interval is desirable.

3. The MMI Bottleneck & Parallel Structures

We begin with a key takeaway in terms of series and parallel components. The MMI over channels in series is the smallest MMI of the series, and the MMI over channels in parallel is the sum of the parallel MMI values. One consequence of this takeaway is that the information theoretic properties of fully connected architectures are strongly dominated by the dimension of the smallest layer. We will often be able to identify a dominant structural parameter that limits the MMI of a series calculation. When we identify the parameter, we will call it the *MMI bottleneck*. In a fully connected network, this will be given by the dimension of the smallest hidden layer. Special focus should be given to the MMI bottleneck when attempting to control $I(X;Z)$.

4. Single Layer Linear Networks

We will first study networks consisting of a single layer, and with no activation function. While this is a highly specialized case, the results and methods of obtaining those results generalize quite well to other cases. This section will consider fully connected architectures.

4.1. Fully Connected Case

We begin by deriving the MMI of a linear network with a standardized Gaussian input. This is a highly specialized case, but we will see that it generalizes quite nicely to a large family of architectures including architectures with relu activation functions. We consider the constrained problem in which the weight matrices W are constrained by Frobenius norm. We will see that the Maximum Mutual Information of this family of architectures is discontinuous in the Frobenius norm constraint with at most $dim(\mathcal{X})$ points of discontinuity. We will thus first specialize to the case where our Frobenius norm constraint is larger than the largest discontinuity point before moving to the more general scenarios. This is expressed in the following lemma. The proof of this lemma is very similar to the argument of (Cover, 1999) (page 253, section 10.5) for the mutual information of gaussian channels with colored noise, and so we will not repeat it here. However, we will provide this proof in the appendix of this paper.

Theorem 1. *Let Σ_x be a positive definite matrix and let $\sigma^2 > 0$. Let N_0 and N_1 be natural numbers representing the input and hidden dimensions. Let $\mathcal{N}(\mu;A)$ denote the Gaussian distribution with mean μ and covariance matrix A. Let $X \sim \mathcal{N}(0;\Sigma_x)$, $X \in \mathbb{R}^{N_0}$, $Z|X,W,b \sim \mathcal{N}(WX+b;\ \sigma^2 Id_{N_1})$, $Z \in \mathbb{R}^{N_1}$, where $W \in \mathbb{R}^{N_1\times N_0}$, Id_{N_1} is the identity matrix in N_1 dimensions, and $b \in \mathbb{R}^{N_1}$ is the bias vector. Let $\tilde{N} = \min(N_0,N_1)$. Let*

$\Sigma_{x,\tilde{N}}$ denote $\tilde{N} \times \tilde{N}$ diagonal matrix containing the $\tilde{N}$ largest eigenvalues of Σ_x. Let $\lambda^x_{\tilde{N}}$ denote the smallest eigenvalue of $\Sigma_{x,\tilde{N}}$, and let $\rho_{\tilde{N}} \triangleq \sigma^2 \left(\frac{\tilde{N}}{\lambda^x_{\tilde{N}}} - Tr(\Sigma^{-1}_{x,\tilde{N}})\right)$. Let $F \geq \rho_{\tilde{N}}$, $I_{W,b}(X;Z) \triangleq \mathbb{E}_{\sim p(x,z|W,b)}\left[log\frac{p(z|x,W,b)}{p(z|W,b)}\right]$, and define MMI$(X;Z) \triangleq \sup_{Tr(W^TW)\leq F} I_{W,b}(X;Z)$. Then:

$$MMI(X;Z) = \frac{\tilde{N}}{2}log\left(\frac{F + \sigma^2 Tr(\Sigma^{-1}_{x,\tilde{N}})}{\sigma^2\tilde{N}}\right) + \frac{1}{2}log\,|\Sigma_{x,\tilde{N}}| \tag{2}$$

Some takeaways from Theorem 1 are in order. First, there is an MMI bottleneck given by $\min\{N_0, N_1\}$ - the minimum of the input dimension and the hidden dimension. Secondly, if $N_1 < N_0$, then N_1 controls the number of principal components used to maximize the mutual information. Thirdly, the largest discontinuity point of MMI(F) is dominated by the difference between the *largest* reciprocal eigenvalue and the *average* reciprocal eigenvalue of those principal components that are used. We also see that smaller principal components are removed first. Furthermore, if $\tilde{N}$ is large, then we have the following approximation MMI$(X;Z) \approx \frac{1}{2}\sum_{i=1}^{\tilde{N}} log(\lambda^x_i \cdot \bar{\lambda}^{-1})$ where $\bar{\lambda}^{-1}$ is the average reciprocal eigenvalue of the components used. We now move on to defining the rest of the discontinuity points. They are defined in the following Lemma.

Lemma 1. *Take all of the assumptions from Theorem 1 except for the assumption that $F \geq \rho$. Let the kth largest eigenvalue of Σ_x be denoted by λ^x_k. Let K be a natural number, $K < \tilde{N}$, and let $\Sigma_{x,\tilde{N}-K}$ denote the $(\tilde{N}-K)\times(\tilde{N}-K)$ diagonal matrix containing the $\tilde{N}-K$ largest eigenvalues of Σ_x. Now, let $\rho_{\tilde{N}-K} \triangleq \sigma^2\left(\frac{\tilde{N}-K}{\lambda^x_{\tilde{N}-K}} - Tr\left(\Sigma^{-1}_{x,\tilde{N}-K}\right)\right)$. Then:*

$$0 = \rho_1 \leq \cdots \leq \rho_{\tilde{N}-K} \leq \rho_{\tilde{N}-K+1} \leq \cdots \leq \rho_{\tilde{N}-1} \leq \rho_{\tilde{N}} \tag{3}$$

Note that each discontinuity point is calculated in the same way as the largest one, but with successive removals of the smallest principal components from our dataset. With all of the discontinuities defined, we can calculate the maximum mutual information for the case when our Frobenius norm constraint is contained in any of the corresponding intervals of continuity.

Theorem 2. *Take all of the assumptions from Theorem 1 except for the assumption that $F \geq \rho_{\tilde{N}}$ and take all definitions from lemma 1. Let $\rho_{\tilde{N}-K+1} \geq F \geq \rho_{\tilde{N}-K}$. Then MMI$(X;Z)$ is given by:*

$$\frac{\tilde{N}-K}{2}log\left(\frac{F + \sigma^2 Tr(\Sigma^{-1}_{x,\tilde{N}-K})}{\sigma^2(\tilde{N}-K)}\right) + \frac{1}{2}log\,|\Sigma_{x,\tilde{N}-K}| \tag{4}$$

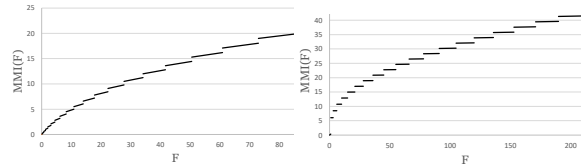

Figure 1. MMI(F) for a fully connected single layer architecture with 50 hidden units and $\sigma^2 = 1$ on a 100 dimensional input dataset whose eigenvalues are modeled via [with indices starting at 1] (Left) $\lambda_i = e^{-0.1(i-1)}$ (Right) $\lambda_i = \frac{1}{i}$.

Theorem 2 is a straightforward generalization of Theorem 1. The only additional insight is that the Frobenius norm constraint F acts to remove the smallest principal components from our maximum mutual information calculation. This role is similar to that of the hidden dimension. The MMI calculations in Theorems 1 and 2 will be seen to be very important - nearly every other case is a generalization of these two theorems. We've plotted some sample MMI curves as a function of the Frobenius norm F for this family of architectures in Figure 1.

5. Single Layer Relu Networks

We will now move on to studying what happens to the MMI values when we place relu activations on the hidden layers. The answer is quite nice: nothing changes at all. Thus we can take all of the insight from the previous sections and apply them to relu networks. Unfortunately, it takes quite a bit of setup and rigor to prove this fact.

We can provide some insight into the rigor that follows before diving into the proofs. Essentially, we will show that the mutual information between X and Z in a relu network is always bounded above by that of a corresponding linear network. However, we will be able to construct a sequence of relu networks whose marginal distributions on Z converge (weakly) to that of the maximum mutual information solution of the linear architectures. All we will need to do then is have some notion of continuity of maximum mutual information with respect to these marginals and we will have our proof. This penultimate step is taken care of primarily in Lemma 3.

Lemma 2. *Take all of the assumptions of either Theorem 1 or Theorem 2. Let W, b be fixed and define Z_{relu} through a new model given by $Z_{relu}|X,W,b \sim \mathcal{N}(relu(WX + b);\ \sigma^2 Id_{N_1})$ for the fully connected case. Let Σ_Z denote the covariance matrix of Z as defined in Theorem 1 and let $\Sigma_{Z_{relu}}$ denote the covariance matrix of Z_{relu}.*

Then:

$$H(Z_{relu}) \leq H(Z) \tag{5}$$

The next lemma relies heavily on the concept of a proba-

bilistic coupling. Thus we will first review this concept.

Definition 1. *Given two probability models $\mathbb{P}_{\tilde{S}}$ and $\mathbb{Q}_S$ on a list of variables S, a* ***coupling*** *of these models is a pair of random variables $(\tilde{S}, \hat{S})$ with joint distribution $\gamma_{\tilde{S},\hat{S}}$ such that the marginal distributions satisfy $\gamma_{\tilde{S}} = \mathbb{P}_{\tilde{S}}$ and $\gamma_{\hat{S}} = \mathbb{Q}_S$.*

In particular, we will rely on the concept of a *maximal coupling*, which is defined as follows:

Definition 2. *Given two probability measures $\mathbb{P}$ and $\hat{\mathbb{P}}$ on a Euclidean space $\mathcal{X}$ with probability density functions p and $\hat{p}$, the maximal coupling on this pair is defined as follows:*

First, define the function $m : \mathcal{X} \to [0, 1]$ through $m(a) := \int min\{p_{Y|X}(b|a), \hat{p}_{Y|X}(b|a)\}db$. Next, define a real number ρ as $\rho := \int m d\mathbb{P}$ and define J as a Bernoulli random variable with success probability ρ. Then define variables U, V and W with the following distributions:

$$p_U := \frac{m}{\rho},\ p_V := \frac{p-m}{1-\rho},\ p_W := \frac{\hat{p}-m}{1-\rho} \tag{6}$$

Next define $(\tilde{X}, \hat{X})$ as functions of the above random variables as follows:

$$\begin{cases} \tilde{X} = \hat{X} = U & \text{if } J = 1 \\ \tilde{X} = V, \quad \hat{X} = W, & \text{if } J = 0 \end{cases} \tag{7}$$

The pair ($\tilde{X}$, $\hat{X}$), and its distribution γ form the maximal coupling. One can show that $1 - \rho$ is the total variation distance between p and $\hat{p}$.

Lemma 3. *Take all of the assumptions of either Theorem 1 or Theorem 2 and all definitions from lemma 2. Denote the marginal probability laws of Z and Z_{relu} as $\mathbb{P}$ and $\hat{\mathbb{P}}$ with densities denoted $p(z)$ and $\hat{p}(z)$. Let $\mathcal{B}(Z)$ be the set of Borel measurable sets on $\mathcal{Z}$, and let δ denote the total variation distance between $\mathbb{P}$ and $\hat{\mathbb{P}}$. That is, $\delta = \sup_{A\in\mathcal{B}(Z)} |\mathbb{P}(A) - \hat{\mathbb{P}}(A)| = \frac{1}{2}\int |p(z) - \hat{p}(z)|dz$. Finally, let $h_2(\cdot)$ denote the binary entropy function. Let $\epsilon > 0$. Then for all W, b such that $\delta < \frac{1}{e}$, there exists a non-negative function $g(\delta(W, b))$ which is continuous in δ from the right at $\delta = 0$, has $g(0) = 0$, and*

$$|I_{W,b}(X; Z) - I_{W,b}(X; Z_{relu})| \leq g(\delta(W, b)) \tag{8}$$

Theorem 3. *Take all of the assumptions of either Theorem 1 or Theorem 2, and take all definitions from lemmas 2 and 3. Then the results of Theorem 1 and Theorem 2 hold for MMI$(X; Z_{relu})$. That is:*

$$MMI(X; Z_{relu}) = MMI(X; Z) \tag{9}$$

We see that all of the insights that we obtained for linear activated networks hold for relu activated networks as well.

6. Single Layer Fully Connected Networks with Bijective Activation Functions

We will now move on to deriving the MMI for one final family of architectures - single layer networks with bijective activation functions. This family includes sigmoidal activations, tanh activations, selu activations, and much more. Conveniently, we once again find ourselves looking back to the linear case for its calculation.

Theorem 4. *Let A denote the pre-activated representation variable of a single layer neural network. Let ϕ be a bijective activation whose log-derivative has finite expectation, and let the representation Z be given by $Z = \phi(A)$. Then:*

$$I(X; Z) = I(X; A) \tag{10}$$

From this theorem, we can immediately see that, if we take the linear case and place our noise injection on the pre-activated variable, and then use $Z = \phi(A)$, we will have the same MMI as we did in the linearly activated case.

7. Multilayer Fully Connected Linear, Relu, and Bijective Networks

We finally move on to the multi-layer case for linear and relu fully connected networks.

Theorem 5. *Take all assumptions and definitions from the previous theorems corresponding to a fully connected network (linear or relu), but assume that we are using a K layer neural network instead of a single layer network, with the noise placed on the Kth layer. Let $N_0, N_1, \cdots, N_K$ denote the number of hidden units in each layer. Redefine $\tilde{N}$ to $\tilde{N} \triangleq \min(N_0, N_1, \cdots, N_K)$. Then the results of those previous theorems hold.*

We see that all of our previous insights for these families of fully connected networks hold. However, a new MMI bottleneck can be identified: $\min\{N_0, N_1, \cdots, N_K\}$ - the dimension of the smallest layer of the network.

8. Conclusion

We have rigorously derived the Maximum Mutual Information for a large number of neural network architectures, and provided several insights along the way. Nearly every case generalizes from Theorems 1 and 2. All of the studied fully connected single layer architectures have the exact same MMI expressions as those studied in Theorems 1 and 2. Multi-layer networks generalize from these as well, but with the primary architectural parameter being given by the smallest hidden dimension in the network. Thus great care should be given to the design of this layer when attempting to control the network's mutual information.

References

Achille, A. and Soatto, S. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Anantharam, V., Gohari, A., Kamath, S., and Nair, C. On maximal correlation, hypercontractivity, and the data processing inequality studied by erkip and cover, 2013.

Bang, S., Xie, P., Lee, H., Wu, W., and Xing, E. Explaining a black-box using deep variational information bottleneck approach. *arXiv preprint arXiv:1902.06918*, 2019.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 531–540, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/belghazi18a.html.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pp. 2172–2180, 2016.

Cover, T. M. *Elements of information theory*. John Wiley & Sons, 1999.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 2012.

Foggo, B. and Yu, N. Analyzing data selection techniques with tools from the theory of information losses, 2019a.

Foggo, B. and Yu, N. Improving supervised phase identification through the theory of information losses. *IEEE Transactions on Smart Grid*, 2019b.

Foggo, B., Yu, N., Shi, J., and Gao, Y. Information losses in neural classifiers from sampling. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–11, 2019. ISSN 2162-2388. doi: 10.1109/tnnls.2019.2952029. URL http://dx.doi.org/10.1109/TNNLS.2019.2952029.

Friedman, N., Mosenzon, O., Slonim, N., and Tishby, N. Multivariate information bottleneck. *arXiv preprint arXiv:1301.2270*, 2013.

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

Gao, W., Kannan, S., Oh, S., and Viswanath, P. Estimating mutual information for discrete-continuous mixtures. In *Advances in neural information processing systems*, pp. 5986–5997, 2017.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Kolchinsky, A., Tracey, B. D., and Wolpert, D. H. Nonlinear information bottleneck. *Entropy*, 21(12):1181, 2019.

MacKay, D. J. and Mac Kay, D. J. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Paninski, L. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

Polyanskiy, Y. and Wu, Y. Dissipation of information in channels with input constraints. *IEEE Transactions on Information Theory*, 62(1):35–55, 2015.

Polyanskiy, Y. and Wu, Y. Strong data-processing inequalities for channels and bayesian networks. In *Convexity and Concentration*, pp. 211–249. Springer, 2017.

Prados, D. and Kak, S. Neural network capacity using delta rule. *Electronics Letters*, 25(3):197–199, 1989.

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

Slonim, N. and Tishby, N. Agglomerative information bottleneck. In *Advances in neural information processing systems*, pp. 617–623, 2000.

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

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method, 2000.

A. Proof of Theorem 1

Let Σ_x be a positive definite matrix and let $\sigma^2 > 0$. Let N_0 and N_1 be natural numbers representing the input and hidden dimensions. Let $\mathcal{N}(\mu; A)$ denote the Gaussian distribution with mean μ and covariance matrix A. Let $X \sim \mathcal{N}(0; \Sigma_x)$, $X \in \mathbb{R}^{N_0}$, $Z|X, W, b \sim \mathcal{N}(WX + b;\ \sigma^2 Id_{N_1})$, $Z \in \mathbb{R}^{N_1}$, where $W \in \mathbb{R}^{N_1 \times N_0}$, Id_{N_1} is the identity matrix in N_1 dimensions, and $b \in \mathbb{R}^{N_1}$ is the bias vector. Let $\tilde{N} = \min(N_0, N_1)$. Let $\Sigma_{x,\tilde{N}}$ denote $\tilde{N} \times \tilde{N}$ diagonal matrix containing the $\tilde{N}$ largest eigenvalues of Σ_x. Let $\lambda^x_{\tilde{N}}$ denote the smallest eigenvalue of $\Sigma_{x,\tilde{N}}$, and let $\rho_{\tilde{N}} \triangleq \sigma^2 \left(\frac{\tilde{N}}{\lambda^x_{\tilde{N}}} - Tr(\Sigma^{-1}_{x,\tilde{N}}) \right)$. Let $F \geq \rho_{\tilde{N}}$, $I_{W,b}(X; Z) \triangleq \mathbb{E}_{\sim p(x,z|W,b)} \left[log \frac{p(z|x,W,b)}{p(z|W,b)} \right]$, and define MMI$(X; Z) \triangleq \sup_{Tr(W^T W) \leq F} I_{W,b}(X; Z)$. Then:

$$\text{MMI}(X; Z) = \frac{\tilde{N}}{2} log \left(\frac{F + \sigma^2 Tr(\Sigma^{-1}_{x,\tilde{N}})}{\sigma^2 \tilde{N}} \right) + \frac{1}{2} log\, |\Sigma_{x,\tilde{N}}| \tag{11}$$

Proof. Since X is Gaussian and the network is linear, Z is Gaussian for all W, b. Thus, we can express $I_{W,b}(X; Z)$ as $\frac{1}{2} log\, \frac{|\Sigma_z|}{|\Sigma_{z|x}|} = \frac{1}{2} log\, \frac{|\sigma^2 Id_{N_1} + W \Sigma_x W^T|}{|\sigma^2 Id_{N_1}|} = \frac{1}{2} log\, |Id_{N_1} + \frac{1}{\sigma^2} W \Sigma_x W^T|$. Now, by the matrix determinant lemma, we have that $|Id_{N_1} + \frac{1}{\sigma^2} W \Sigma_x W^T| = |\sigma^2 \Sigma_x^{-1} + W^T W| \cdot |\frac{1}{\sigma^2} \Sigma_x|$, and so we can condense the dependence of the MMI optimization problem on W to obtain:

$$\text{MMI}(X; Z) = \frac{1}{2} log\, |\frac{1}{\sigma^2} \Sigma_x| + \sup_{Tr(W^T W) \leq F} \frac{1}{2} log\, |Q(W)| \tag{12}$$

where $Q(W) = \sigma^2 \Sigma_x^{-1} + W^T W$. Due to the positive definiteness of Q and Hadamard's inequality, we can cast this constrained maximization problem into the realm of eigenvalues since the optimal Q matrix will be diagonal, so $W^T W$ will have the same eigenvectors as Σ_x^{-1}:

$$\begin{aligned} &\sup_{\tilde{\lambda}_1, \cdots, \tilde{\lambda}_{N_0}} \sum_{i=1}^{N_0} log \left(\tilde{\lambda}_i + \frac{\sigma^2}{\lambda^x_i} \right), \\ &\text{s.t.} \sum_{i=1}^{N_0} \tilde{\lambda}_i \leq F,\ \tilde{\lambda}_i \geq 0,\ i = 1, 2, \cdots, N_0 \\ &\text{and } \tilde{\lambda}_i = 0, \text{ for at least } \max(0, N_0 - N_1) \text{ values of } i \end{aligned} \tag{13}$$

where λ^x_i is the ith largest eigenvalue of Σ_x, and $\tilde{\lambda}_i$ is the ith un-ordered eigenvalue of $W^T W$. The final constraint comes from the fact that $W^T W$ is only rank $\min(N_0, N_1)$. Furthermore, it must be the case that the mandatory 0-valued eigenvalues of $W^T W$, when they exist ($N_0 > N_1$), must be placed on the indices $i = N_1 + 1, \cdots, N_0$, as these correspond to the largest values of $\frac{\sigma^2}{\lambda^x_i}$. Indeed, suppose that we have placed a nonzero eigenvalue on one of these indices (without loss of generality, say index p, and that we set this eigenvalue to l) in such a way that all of the constraints are met. Then we must have placed a zero-valued eigenvalue on another index (which we will denote as q without loss of generality, $q < p$). Then the objective function can be increased without violating any constraints by taking l units of eigenvalue off of index p and placing it on index q, and so this cannot be a solution to our optimization problem. To see this, observe that:

$$\begin{aligned} &log \left(l + \frac{\sigma^2}{\lambda^x_p} \right) + log \left(\frac{\sigma^2}{\lambda^x_q} \right) = log \left(\frac{l\sigma^2}{\lambda^x_q} + \frac{\sigma^4}{\lambda^x_p \lambda^x_q} \right) \\ &\leq log \left(\frac{l\sigma^2}{\lambda^x_p} + \frac{\sigma^4}{\lambda^x_p \lambda^x_q} \right) = log \left(\frac{\sigma^2}{\lambda^x_p} \right) + log \left(l + \frac{\sigma^2}{\lambda^x_q} \right) \end{aligned} \tag{14}$$

We are thus left with the following optimization problem:

$$\begin{aligned} &\sup_{\tilde{\lambda}_1, \cdots, \tilde{\lambda}_{\tilde{N}}} \sum_{i=1}^{\tilde{N}} log \left(\tilde{\lambda}_i + \frac{\sigma^2}{\lambda^x_i} \right), \\ &\text{s.t.} \sum_{i=1}^{\tilde{N}} \tilde{\lambda}_i \leq F,\ \tilde{\lambda}_i \geq 0,\ i = 1, 2, \cdots, \tilde{N} \end{aligned} \tag{15}$$

This is a classic 'water-filling' problem with heights given by scaled versions of the inverses of the first $\tilde{N}$ eigenvalues of Σ_x. Thus, for a given 'water level', $\mu^*(F)$, a solution is readily available, being given by $\tilde{\lambda}_i = \max\left(0, \mu^*(F) - \frac{\sigma^2}{\lambda_i^x}\right)$ [optimality]. However, finding the relationship between $\mu^*(F)$ and F requires additional work, as $\mu^*(F)$ must satisfy $\sum_i \max\left(0, \mu^*(F) - \frac{\sigma^2}{\lambda_i^x}\right) = F$ [consistency]. We will show that our assumption, $F \geq \rho_{\tilde{N}}$, yields a consistent solution in which all maximums of the optimality equation are obtained in the second argument. To see this, note that under such a solution, the consistency equation yields $\mu^*(F) = \frac{F+\sigma^2 Tr(\Sigma_{x,\tilde{N}}^{-1})}{\tilde{N}}$, which coincides with the optimality equation because, for each i, $\frac{F+\sigma^2 Tr(\Sigma_{x,\tilde{N}}^{-1})}{\tilde{N}} - \frac{\sigma^2}{\lambda_i^x} \geq \frac{1}{\tilde{N}}\left(F - \rho_{\tilde{N}}\right) \geq 0$. Thus this solution holds and, in all, we have an MMI of:

$$
\begin{aligned}
&\frac{1}{2}log|\frac{\Sigma_x}{\sigma^2}| + \frac{\tilde{N}}{2}log\left(\frac{F+\sigma^2 Tr(\Sigma_{x,\tilde{N}}^{-1})}{\tilde{N}}\right) \\
&+\frac{1}{2}\sum_{i=\tilde{N}+1}^{N_0} log(\frac{\sigma^2}{\lambda_i^x}) \\
=&\frac{1}{2}\sum_{i=1}^{\tilde{N}} log(\frac{\lambda_i^x}{\sigma^2}) + \frac{\tilde{N}}{2}log\left(\frac{F+\sigma^2 Tr(\Sigma_{x,\tilde{N}}^{-1})}{\tilde{N}}\right) \\
=&\frac{1}{2}log|\Sigma_{x,\tilde{N}}| + \frac{\tilde{N}}{2}log\left(\frac{F+\sigma^2 Tr(\Sigma_{x,\tilde{N}}^{-1})}{\sigma^2\tilde{N}}\right)
\end{aligned} \tag{16}
$$

completing the proof. □

B. Proof of Lemma 1

Proof. First, ρ_1 is zero since:

$$
\rho_1 = \sigma^2\left(\frac{1}{\lambda_1^x} - \frac{1}{\lambda_1^x}\right) \tag{17}
$$

Next, we note that the difference $\rho_{\tilde{N}-K+1} - \rho_{\tilde{N}-K}$ is given by:

$$
\begin{aligned}
&\sigma^2\left(\frac{\tilde{N}-K+1}{\lambda_{\tilde{N}-K+1}^x} - \frac{\tilde{N}-K}{\lambda_{\tilde{N}-K}^x} - \frac{1}{\lambda_{\tilde{N}-K+1}^x}\right) \\
&= \sigma^2\left(\frac{\tilde{N}-K}{\lambda_{\tilde{N}-K+1}^x} - \frac{\tilde{N}-K}{\lambda_{\tilde{N}-K}^x}\right) \\
&\geq \sigma^2\left(\frac{\tilde{N}-K}{\lambda_{\tilde{N}-K}^x} - \frac{\tilde{N}-K}{\lambda_{\tilde{N}-K}^x}\right) = 0
\end{aligned} \tag{18}
$$

completing the proof. □

C. Proof of Theorem 2

Proof. We can follow the proof of Theorem 1 up until the water-filling optimization problem, whose solution will now be different because F has changed. Again, we need to find a solution consistent with $\tilde{\lambda}_i = \max\left(0, \mu^*(F) - \frac{\sigma^2}{\lambda_i^x}\right)$ [optimality] and $\sum_i \max\left(0, \mu^*(F) - \frac{\sigma^2}{\lambda_i^x}\right) = F$ [consistency]. We claim that our assumption, $\rho_{\tilde{N}-K+1} > F \geq \rho_{\tilde{N}-K}$ yields a consistent solution in which the $\tilde{\lambda}_i$ are zero for $i > \tilde{N} - K$ and nonzero otherwise. Under such a solution, the consistency equation yields $\mu^*(F) = \frac{F+\sigma^2 Tr(\Sigma_{x,\tilde{N}-K}^{-1})}{\tilde{N}-K}$. We will show that this coincides with the optimality equation. First,

define $l(i) \triangleq \frac{F+\sigma^2 Tr(\Sigma^{-1}_{x,\tilde{N}-K})}{\tilde{N}-K} - \frac{\sigma^2}{\lambda^x_i}$. Then, if $i \leq \tilde{N}-K$, we have:

$$\begin{aligned} l(i) &\geq \frac{F+\sigma^2 Tr(\Sigma^{-1}_{x,\tilde{N}-K})}{\tilde{N}-K} - \frac{\sigma^2}{\lambda^x_{\tilde{N}-K}} \\ &= \frac{1}{\tilde{N}-K}\left(F - \rho_{\tilde{N}-K}\right) \geq 0 \end{aligned} \tag{19}$$

On the other hand, if $i > \tilde{N}-K$, then:

$$\begin{aligned} l(i) &\leq \frac{F+\sigma^2 Tr(\Sigma^{-1}_{x,\tilde{N}-K})}{\tilde{N}-K} - \frac{\sigma^2}{\lambda^x_{\tilde{N}-K+1}} \\ &= \frac{F+\sigma^2 Tr(\Sigma^{-1}_{x,\tilde{N}-K+1}) - \frac{\sigma^2}{\lambda^x_{\tilde{N}-K+1}}}{\tilde{N}-K} - \frac{\sigma^2}{\lambda^x_{\tilde{N}-K+1}} \\ &= \frac{1}{\tilde{N}-K}\left(F - \rho_{\tilde{N}-K+1}\right) < 0 \end{aligned} \tag{20}$$

And so optimality is achieved. Under this solution, the objective function value is given by:

$$\sum_{i=\tilde{N}-K+1}^{\tilde{N}} log\left(\frac{\sigma^2}{\lambda^x_i}\right) + (\tilde{N}-K) log\left(\frac{F+\sigma^2 Tr(\Sigma^{-1}_{x,\tilde{N}-K})}{\tilde{N}-K}\right) \tag{21}$$

Thus, in all, we have that MMI$(X;Z)$ is given by:

$$\begin{aligned} &\frac{\tilde{N}-K}{2} log\left(\frac{F+\sigma^2 Tr(\Sigma^{-1}_{x,\tilde{N}-K})}{\sigma^2(\tilde{N}-K)}\right) \\ &\quad + \frac{1}{2} log\left(\frac{|\Sigma_{x,\tilde{N}}|}{\prod_{i=\tilde{N}-K+1}^{\tilde{N}} \lambda^x_i}\right) \end{aligned} \tag{22}$$

Where the factors of $\frac{1}{2}$ and the term $log(\frac{1}{\sigma^2}|\Sigma_{x,\tilde{N}}|)$ have come from equation (12). Finally, the only factors remaining in post-cancellation of the second term are the eigenvalues of $\Sigma_{x,\tilde{N}}$ with indices smaller than or equal to $\tilde{N}-K$, transforming that term into what is presented in equation (4). This completes the proof. □

D. Proof of Lemma 2

Take all of the assumptions of either Theorem 1, Theorem 2, or Theorem **??**. Let W, b be fixed and define Z_{relu} through a new model given by $Z_{relu}|X, W, b \sim \mathcal{N}(\text{relu}(WX+b);\ \sigma^2 Id_{N_1})$ for the fully connected case, or $Z_{relu}|X, W, b \sim \mathcal{N}(\text{relu}(X \circledast W + b);\ \sigma^2 Id_{N_1})$ for the convolutional case. Let Σ_Z denote the covariance matrix of Z as defined in Theorem 1 and let $\Sigma_{Z_{relu}}$ denote the covariance matrix of Z_{relu}.

Then:

$$H(Z_{relu}) \leq H(Z) \tag{23}$$

Proof. Let η denote a multivariate Gaussian with covariance $\sigma^2 Id_{N_1}$. We denote $S \triangleq WX + b$. Then:

$$Z = S + \eta, \quad Z_{relu} = relu(S) + \eta \tag{24}$$

We can trivially write Z as:

$$\begin{aligned} Z &= 1_{S>0} Z + 1_{S\leq 0} Z \\ &= 1_{S>0} Z + 1_{S\leq 0}\eta + 1_{S\leq 0} S \end{aligned} \tag{25}$$

But the sum of the first two terms in this decomposition is nothing but Z_{relu}, so we have:

$$Z = Z_{relu} + 1_{S \leq 0} S \tag{26}$$

Furthermore:

$$\begin{aligned} H(Z, Z_{relu}) &= H(Z) + H(Z_{relu}|Z) \\ &= H(Z_{relu}) + H(Z|Z_{relu}) \end{aligned} \tag{27}$$

It follows that:

$$H(Z) = H(Z_{relu}) + H(Z|Z_{relu}) - H(Z_{relu}|Z) \tag{28}$$

We claim that $H(Z|Z_{relu}) > H(Z_{relu}|Z)$. To see this, note that:

$$\begin{aligned} H(Z|Z_{relu}) &= H(Z_{relu} + 1_{S \leq 0}|Z_{relu}) = H(1_{S \leq 0}|Z_{relu}) \\ &= H(1_{S \leq 0}|relu(S) + \eta) = H(1_{S \leq 0}|\eta) \\ &= H(1_{S \leq 0}) \end{aligned} \tag{29}$$

where the fourth equality holds because the integral in the entropy calculation only occurs over the region $S \leq 0$, in which $relu(S) = 0$. Meanwhile:

$$\begin{aligned} H(Z_{relu}|Z) &= H(Z - 1_{S \leq 0} S|Z) \\ &= H(-1_{S \leq 0} S|Z) \\ &= H(1_{S \leq 0} S|Z) \\ &\leq H(1_{S \leq 0} S) = H(Z|Z_r) \end{aligned} \tag{30}$$

Completing the proof. □

E. Proof of Lemma 3

Take all of the assumptions of either Theorem 1 or Theorem 2 and all definitions from lemma 2. Denote the marginal probability laws of Z and Z_{relu} as $\mathbb{P}$ and $\hat{\mathbb{P}}$ with densities denoted $p(z)$ and $\hat{p}(z)$. Let $\mathcal{B}(Z)$ be the set of Borel measurable sets on $\mathcal{Z}$, and let δ denote the total variation distance between $\mathbb{P}$ and $\hat{\mathbb{P}}$. That is, $\delta = \sup_{A \in \mathcal{B}(Z)} |\mathbb{P}(A) - \hat{\mathbb{P}}(A)| = \frac{1}{2} \int |p(z) - \hat{p}(z)| dz$. Finally, let $h_2(\cdot)$ denote the binary entropy function. Let $\epsilon > 0$. Then for all W, b such that $\delta < \frac{1}{e}$, there exists a non-negative function $g(\delta(W, b))$ which is continuous in δ from the right at $\delta = 0$, has $g(0) = 0$, and

$$|I_{W,b}(X; Z) - I_{W,b}(X; Z_{relu})| \leq g(\delta(W, b)) \tag{31}$$

Proof. Let $(\tilde{Z}, \hat{Z}, \gamma)$ denote the maximal coupling between $\mathbb{P}$ and $\hat{\mathbb{P}}$ on $(\tilde{Z}, \hat{Z})$. Then, as $H(\tilde{Z}|X) = H(\hat{Z}|X)$ (all of the uncertainty under these conditionals comes from the noise variable η), we have:

$$|I(X; Z) - I(X; \hat{Z})| = |H(\tilde{Z}) - H(\hat{Z})| \tag{32}$$

We can decompose these terms as:

$$\begin{aligned} H(\tilde{Z}) &= H(\tilde{Z}|J) + H(J) - H(J|\tilde{Z}),\ H(\hat{Z}) \\ &= H(\hat{Z}|J) + H(J) - H(J|\hat{Z}) \end{aligned} \tag{33}$$

The $H(J)$ terms cancel in the subtraction. Furthermore, both $H(J|\tilde{Z})$ and $H(J|\hat{Z})$ are bounded by $H(J) = h_2(\delta)$. Thus, by an application of the triangle inequality, we have:

$$|I(X; Z) - I(X; \hat{Z})| \leq |H(\tilde{Z}|J) - H(\hat{Z}|J)| + 2h_2(\delta) \tag{34}$$

We can further decompose the remaining terms as:

$$H(\tilde{Z}|J) = (1-\delta)H(U) + \delta H(V),$$
$$H(\hat{Z}|J) = (1-\delta)H(U) + \delta H(W) \tag{35}$$

leaving us with:

$$|I(X;Z) - I(X;\hat{Z})| \leq \delta|H(V) - H(W)| + 2h_2(\delta) \tag{36}$$

Now, observe that the absolute entropy difference, $|H(V) - H(W)|$, expands to $\frac{1}{\delta}$ multiplied by the absolute value of the following expression:

$$\int (p(z) - m(z)) log \left(\frac{p(z) - m(z)}{\delta}\right) dz$$
$$- \int (\hat{p}(z) - m(z)) log \left(\frac{\hat{p}(z) - m(z)}{\delta}\right) dz$$

Let A denote the set of points in $Range(Z)$ such that $p(z) \geq \hat{p}(z)$. Then the second term in the expression becomes zero over A and the first term becomes zero over A^c. We are then left with $\frac{1}{\delta}$ multiplied by the absolute value of the next expression:

$$\int_A (p(z) - \hat{p}(z)) log \left(\frac{p(z) - \hat{p}(z)}{\delta}\right) dz$$
$$- \int_{A^c} (\hat{p}(z) - p(z)) log \left(\frac{\hat{p}(z) - p(z)}{\delta}\right) dz$$

which can equivalently be written as:

$$\frac{1}{\delta}\left|\int (p(z) - \hat{p}(z)) log\, |p(z) - \hat{p}(z)| dz\right|$$
$$\leq \frac{1}{\delta}\int |p(z) - \hat{p}(z)||log\, |p(z) - \hat{p}(z)||dz \tag{37}$$

Now, split the domain of this last integral into three disjoint subregions Q_1, Q_2, and Q_3 where Q_1 is defined by the condition that $|p(z) - \hat{p}(z)| \leq 2\delta$, Q_2 is given by the condition that $2\delta < |p(z) - \hat{p}(z)| \leq 1$, and Q_3 is given by the condition that $|p(z) - \hat{p}(z)| > 1$. We then have by the definition of total variation that $\frac{1}{2}\int |p(z) - \hat{p}(z)| = \delta$, so it must be the case that:

$$\int_{Q_2} |p(z) - \hat{p}(z)| \leq 2\delta - \int_{Q_1 \cup Q_3} |p(z) - \hat{p}(z)| \leq 2\delta \tag{38}$$

Futhermore, in Q_2, $|log|p(z) - \hat{p}(z)||$ is decreasing in $|p(z) - \hat{p}(z)|$, so we can bound it above by $|log(2\delta)|$. We then have that:

$$\int_{Q_2} |p(z) - \hat{p}(z)||log\, |p(z) - \hat{p}(z)||dz$$
$$\leq |log\, 2\delta| \int_{Q_2} |p(z) - \hat{p}(z)|dz \leq 2\delta|log(2\delta)| \tag{39}$$

Also, in Q_1, if $\delta < \frac{1}{e}$, as was assumed in this Lemma's hypothesis, $|p(z) - \hat{p}(z)||log\, |p(z) - \hat{p}(z)||$ is decreasing in $|p(z) - \hat{p}(z)|$. It can thus be bounded above by $2\delta|log(2\delta)|$. Finally, the addition of the noise term η in both models guarantees a uniform upper bound via Young's Convolution Inequality given by $|p(z) - \hat{p}(z)| < \frac{1}{(2\pi\sigma^2)^{N_1/2}}$ (independent of both z and δ). A derivation of this bound is provided at the end of this proof. Given that we are in Q_3, we then have an upper bound given by $M = max\{1, \frac{1}{(2\pi\sigma^2)^{N_1/2}}\}$. Thus the Q_3 integral can be bounded as:

$$\int_{Q_3} |p(z) - \hat{p}(z)||log\, |p(z) - \hat{p}(z)||dz$$
$$\leq |log(M)| \int_{Q_3} |p(z) - \hat{p}(z)| \leq 2\delta|log(M)| \tag{40}$$

Thus the full entropy difference expression can be bounded above by $\frac{4\delta|log(2\delta)|+2\delta|log(M)|}{\delta}$. Plugging this back into equation (36) gives us:

$$|I(X;Z) - I(X;\hat{Z})| \leq 4\delta|log(2\delta)| + 2\delta|log(M)| + 2h_2(\delta) \tag{41}$$

The term on the right hand side is the desired continuous function $g(\delta)$.

The bound on $|p(z) - \hat{p}(z)|$ is derived from Young's Convolution Inequality. Letting p_r be the density function of the random variable $Relu(WX + b)$ and p_η that of η. Then $\hat{p}$ is the convolution of p_r and p_η. We then have that $\|\hat{p}\|_\infty \leq \int p_r(z)dz \cdot \|p_\eta\|_\infty$ ($p = 1, q = r = \infty$ in Young's inequality). The integral evaluates to one, and we are left with a bound of the supremum of p_η, which is given by its multivariate normal pdf evaluated at its mean. Since p is bounded in exactly the same way, we have the desired result. □

F. Proof of Theorem 3

Proof. First, we note that, for all W, b, we have:

$$\begin{aligned} I_{W,b}(X;Z_{relu}) &= H_{W,b}(Z_{relu}) - H(Z_{relu}|X) \\ &= H_{W,b}(Z_{relu}) - H(Z|X) \\ &\leq H_{W,b}(Z) - H(Z|X) = I_{W,b}(X,Z) \end{aligned} \tag{42}$$

where the inequality follows from lemma 2. It follows immediately that MMI$(X;Z_{relu}) \leq$ MMI$(X;Z)$. We now show that MMI$(X;Z)$ is an achievable value for $I_{W,b}(X;Z_{relu})$ given the constraint $Tr(W^TW) \leq F$. Fix $\frac{1}{e} > \epsilon > 0$. Then given any value of F, we can set b large enough in each dimension such that $\hat{\mathbb{P}}(\cup_{i=1}^{N_1}\{z|z_i < 0\})$ is less than ϵ for all W satisfying $Tr(W^TW) \leq F$. When this is the case, the total variation between $\mathbb{P}$ and $\hat{\mathbb{P}}$ is also bounded above by ϵ. Then by lemma 3, there exists $b^* \in \mathbb{R}^{N_1}$ such that:

$$|I_{W,b^*}(X;Z) - I_{W,b^*}(X;Z_{relu})| \leq g(\epsilon) \tag{43}$$

for all W satisfying this constraint, and where the right hand side of this inequality is a continuous function of ϵ and satisfies $\lim_{\epsilon\to 0^+} g(\epsilon) = g(0) = 0$. Thus we can achieve $I_{W,b^*}(X;Z_{relu}) \geq I_{W,b^*}(X;Z) - g(\epsilon)$ for all W satisfying the constraint. Since $g \geq 0$ can be made arbitrary small via continuity, we can achieve:

$$I_{W,b^*}(X;Z_{relu}) \geq I_{W,b^*}(X;Z)\ \forall\ W \text{ s.t. } Tr(W^TW) \leq F \tag{44}$$

Inputting the MMI achieving matrix from Theorem 1 and Theorem 2 into (44) yields the result. □

G. Proof of Theorem 4

Proof. This follows immediately from the following set of equalities:

$$\begin{aligned} I(X;Z) &= H(Z) - H(Z|X), \\ H(Z) &= H(A) + \mathbb{E}_{\mathbb{P}_A}\left[log\left|\frac{d\phi}{da}\right|\right] \\ H(Z|X) &= H(A|X) + \mathbb{E}_{\mathbb{P}_{X,A}}\left[log\left|\frac{d\phi}{da}\right|\right],\ \mathbb{E}_{\mathbb{P}_{X,A}}\left[log\left|\frac{d\phi}{da}\right|\right] \\ &= \mathbb{E}_{\mathbb{P}_A}\left[log\left|\frac{d\phi}{da}\right|\right] \\ I(X;A) &= H(A) - H(A|X) \end{aligned} \tag{45}$$

where $\left|\frac{d\phi}{da}\right|$ is the determinant of the Jacobian matrix of ϕ. □

H. Proof of Theorem 5

Proof. In the linear case, we can take the proof of theorem 1 by replacing W with $W_K \cdots W_2W_1$ (the biases have no effect on the mutual information). We will only need to note that the corresponding inner-product matrix,

$W_1^T W_2^T \cdots W_K^T W_K \cdots W_2 W_1$ has rank $\tilde{N}$ (as redefined in this theorem's hypothesis). In the relu case we can follow the exact sequence of steps that were performed in the single layer case, noting that we can get the marginal total variation $\delta < \epsilon$ (for any fixed $\epsilon > 0$) by fixing each bias to be large enough such that sufficiently small amounts of marginal probability are contained in the saturated regions of each layer. $\square$