Learning to Compress: Local Rank and Information Compression in Deep Neural Networks

Niket Patel
Department of Mathematics
University of California, Los Angeles
niketpatel@ucla.edu

Ravid Shwartz Ziv
New York University, Wand.AI
ravid.shwartz.ziv@nyu.edu

Abstract

Deep neural networks tend to exhibit a bias toward low-rank solutions during training, implicitly learning low-dimensional feature representations. This paper investigates how deep multilayer perceptrons (MLPs) encode these feature manifolds and connects this behavior to the Information Bottleneck (IB) theory. We introduce the concept of *local rank* as a measure of feature manifold dimensionality and demonstrate, both theoretically and empirically, that this rank decreases during the final phase of training. We argue that networks that reduce the rank of their learned representations also compress mutual information between inputs and intermediate layers. This work bridges the gap between feature manifold rank and information compression, offering new insights into the interplay between information bottlenecks and representation learning.

1 Introduction

The **Data Manifold Hypothesis** (Olah, 2014) suggests that real-world datasets lie on a manifold whose intrinsic dimensionality is much lower than the ambient input space. While numerous models can fit training data, those that generalize well and exhibit robustness likely learn meaningful representations of this underlying manifold. Recent studies have observed that deep neural networks, particularly multilayer perceptrons (MLPs), exhibit an emergent bottleneck structure, where certain layers effectively compress input data into lower-dimensional feature manifolds Jacot (2023a).

Understanding how this emergent structure aligns with existing theories, such as the Information Bottleneck (IB) theory (Tishby et al., 2000; Shwartz-Ziv & Tishby, 2017), is crucial to advance our understanding of representation learning in deep networks. This paper aims to provide a theoretical and empirical exploration of this phenomenon by introducing the concept of *local rank* and analyzing its relationship with information compression during training.

This paper makes the following key contributions:

- **Definition and Analysis of Local Rank:** We define *local rank* as a metric to quantify the dimensionality of feature manifolds within a neural network. We provide theoretical insights into how local rank behaves during training and establish bounds based on implicit regularization.
- **Empirical Evidence of Rank Reduction:** We conduct experiments on synthetic and real-world datasets to demonstrate that local rank decreases during the terminal phase of training, indicating that networks compress the dimensionality of their learned representations.
- **Connection to Information Bottleneck Theory:** We explore the relationship between local rank and information compression, arguing that a reduction in local rank correlates with mutual information compression between inputs and intermediate representations.

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of related literature, situating our work within the broader context of implicit regularization, low-rank bias, and the Information Bottleneck theory. Section 3 introduces the notation used throughout the paper. In Section 4, we define the local rank and present theoretical and empirical analyses. Section 6 discusses the information-theoretic implications of local rank, connecting it to the Information Bottleneck theory. Finally, Section 7 concludes the paper and outlines future research directions.

2 Related Work

Implicit Regularization in Deep Learning. Implicit regularization refers to the phenomenon in which the training dynamics of neural networks, particularly under gradient descent, lead to solutions with desirable properties without explicit regularization terms. Neyshabur et al. (2014) and Zhang et al. (2021) investigated how overparameterized networks generalize despite being able to fit random labels. Gunasekar et al. (2017) and Arora et al. (2019) studied implicit bias in linear and deep matrix factorization, showing that gradient descent favors low-rank solutions.

Low-Rank Representations and Manifolds. The concept of neural networks learning low-dimensional manifolds has been explored in various contexts. Papyan et al. (2020) introduced the notion of *neural collapse*, where class means converge to a simplex Equiangular Tight Frame at the final layer. Ansuini et al. (2019) and Ben-Shaul et al. (2023) measured intrinsic dimensionality in deep networks, observing that representations become more compressed in deeper layers. Súkeník et al. (2024) explicitly analyzes the relation between low-rank bias and neural collapse in deep networks.

Rank of Jacobians and Feature Maps. Closely related to the local rank, Jacot (2023b) and Jacot (2023a) analyzed the rank of Jacobians in neural networks, connecting it to generalization properties. Humayun et al. (2024) studies a similar object in the context of assessing the geometry of diffusion models.Feng et al. (2022) studied the low-rank structure in the Jacobian matrices of neural networks, showing that rank deficiency can lead to better generalization.

Information Bottleneck Theory. The Information Bottleneck (IB) framework (Tishby et al., 2000) provides a theoretical lens to understand how neural networks balance compression and prediction. Shwartz-Ziv & Tishby (2017) applied IB to deep learning, proposing that networks first fit the data and then compress the representations. Subsequent works, such as Saxe et al. (2018) and Shwartz-Ziv et al. (2023), explored the universality of this phenomenon, leading to a richer understanding of information dynamics in training.

Relation to Our Work. Our paper builds on these foundational studies by leveraging a measurable quantity—the local rank—that captures the dimensionality of the learned feature manifolds. We theoretically link this to the implicit regularization of the ranks of weight matrices and empirically demonstrate its reduction during training. Moreover, we connect these geometric properties of neural representations to information-theoretic principles, offering new insights into how networks compress information.

3 Notation

We consider a neural network $\mathcal{N}$ parameterized by θ, mapping inputs to outputs as $\mathcal{N}_\theta : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$. The network has depth L, and each layer $l \in \{1, \ldots, L\}$ consists of an affine transformation followed by a nonlinearity:

$$h_l(x) = \phi(A_l(h_{l-1}(x))) = \phi(W_l h_{l-1}(x) + b_l), \tag{1}$$

where $h_0(x) = x$, and $h_l(x) = \mathcal{N}_\theta(x)$, $W_l \in \mathbb{R}^{n_l \times n_{l-1}}$ is the weight matrix, $b_l \in \mathbb{R}^{n_l}$ is the bias vector, and $\phi(\cdot)$ is an element-wise activation function (e.g., ReLU). The pre-activation at layer l is $p_l(x) = W_l h_{l-1}(x) + b_l$.

The Jacobian matrix of a function $f : \mathbb{R}^n \to \mathbb{R}^m$ at point x is denoted by $J_x f \in \mathbb{R}^{m \times n}$. The rank of a matrix A is denoted by $\text{rank}(A)$, and the ϵ-rank, $\text{rank}_\epsilon(A)$, counts the number of singular values of A greater than ϵ.

4 The Local Rank of Representations

In this section, we introduce the concept of *local rank* as a measure of the dimensionality of feature manifolds learned by neural networks. Consider the data distribution with support $\Omega \subseteq \mathbb{R}^{n_0}$. The feature manifold at layer l is defined as $\mathcal{M}_l = p_l(\Omega)$, where p_l maps input data to pre-activations at layer l.

Definition 1. *The **local rank** at layer l, denoted as* $\mathbf{LR}_l$*, is defined as the expected rank of the Jacobian of* p_l *with respect to the input:*

$$\mathbf{LR}_l = \mathop{\mathbb{E}}_{x \sim Data} \left[rank(J_x p_l) \right]. \tag{2}$$

Since the set of matrices without full rank is measure 0, we find it more convenient to look at an approximation for the local rank. For a threshold $\epsilon > 0$*, the **robust local rank** at layer l is:*

$$\mathbf{LR}_l^\epsilon = \mathop{\mathbb{E}}_{x \sim Data} \left[rank_\epsilon(J_x p_l) \right]. \tag{3}$$

This is a meaningful measure of the rank of the feature manifold since the Jacobian's null space identifies the input dimensions which vanish near x, so its rank captures the true number of feature dimensions.

4.1 Theoretical Analysis of Local Rank

We investigate how the local rank behaves under gradient flow dynamics and its connection to implicit regularization. Under suitable assumptions (Wang et al., 2021; Lyu & Li, 2020; Ji & Telgarsky, 2020), solutions to the gradient flow ODE with exponential tailed losses have been shown to converge in direction to a Karush-Kuhn-Tucker (KKT) point of the following optimization problem, where $\{(x_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^{n_0} \times \{-1, 1\}$ is the training dataset:

$$\min_\theta \frac{1}{2}\|\theta\|^2 \quad \text{subject to} \quad \forall i \in [n],\, y_i \mathcal{N}_\theta(x_i) \geq 1. \tag{4}$$

This convergence to a KKT point implies that the solution minimizes the norm of the weights while satisfying the classification constraints. The minimization of the norm leads to implicit regularization effects, including the potential reduction in the rank of weight matrices.

If an intermediate layer exists with a low local rank, it can be viewed as a bottleneck in terms of the dimensionality of the feature manifold. We can establish a connection between Equation (4) and the existence of a bottleneck layer with low local rank. Specifically, we can prove the existence of such a bottleneck layer under the global optimum of this problem as a consequence of the implicit regularization of the ranks of weight matrices.

Proposition 2. ***(Informal)*** *Let* $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^{n_0} \times \{-1, 1\}$ *be a binary classification dataset. Assume there exists a fully connected neural network with weight matrices uniformly bounded by* B *that correctly classifies every data point in* $\mathcal{D}$ *with margin 1. Then, for* $\theta = (W_1, \cdots W_L)$ *parameterizing a* L *layer neural network at the global optimum to 4, there exists a layer* l *and* $\epsilon_0 > 0$ *such that for all* $0 < \epsilon < \epsilon_0$*:*

$$\mathbf{LR}_l^\epsilon \leq \frac{2}{\epsilon^2} \left(\frac{B}{\sqrt{2}} \right)^{\frac{2K}{L}} \frac{L+1}{L} \|W_l\|_\sigma^2, \tag{5}$$

where $\|W_l\|_\sigma$ *denotes the operator norm of* W_l*.*

Proof. The proof leverages recent results from implicit regularization (Timor et al., 2023b) and properties of the Jacobian of ReLU networks. A formal statement with definitions for all constants and proof are provided in Appendix A.1. □

We note that the right-hand side converges to $\frac{2\|W_l\|_\sigma^2}{\epsilon^2}$ as $L \to \infty$, so this bound is typically much better than the trivial bound on the local rank at layer l given by $\mathbf{LR}_l^\epsilon \leq ||W_l||_F^2/\epsilon^2$. This proposition implies that during training, certain layers in the network develop low-rank weight matrices, leading to a reduced local rank in the representations.

Next, we show an analogous result with even a tighter bound for minimum norm solutions to interpolating neural networks for training on **regression tasks**:

Proposition 3. ***(Informal)*** *Let $\{(x_i, y_i)\}_{i=1}^{n} \subset \mathbb{R}^{n_0} \times \mathbb{R}_+$ be a regression dataset. Assume there exists a fully connected neural network $\mathcal{N}$ with weight matrices uniformly bounded by B that interpolates all data points, $\mathcal{N}(x_i) = y_i$. Let $\mathcal{N}_\theta$ be a fully-connected neural network of depth L parameterized by $\theta = [W_1, \ldots, W_L]$, where θ is a global optimum of the following optimization problem:*

$$\min_{\theta} \|\theta\| \quad \textit{s.t.} \quad \forall i \in [n],\ \mathcal{N}_\theta(x_i) = y_i. \tag{6}$$

Then, there exist an $l \in \{1, \cdots, L\}$ and $\epsilon_0 > 0$ such that for $0 < \epsilon < \epsilon_0$ the following holds:

$$\boldsymbol{LR}_l^\epsilon \leq \frac{||W_l||_\sigma^2 \cdot B^{\frac{2K}{L}}}{\epsilon^2} \tag{7}$$

where $||W_l||_\sigma$ denotes the operator norm of W_l.

Proof. As before, we leverage results on implicit regularization from Timor et al. (2023b), and a full proof and formal statement can be found in Appendix A.2. □

5 Empirical Evidence of Local Rank Reduction

We empirically validate the theoretical insights by training MLPs on synthetic and real datasets and measuring the local rank during training.

Synthetic Data. We train a 3-layer MLP with an input dimension of 100, 200 neurons per hidden layer, and an output dimension of 2. The inputs and outputs are Gaussian with a random cross-covariance matrix, reflecting a scenario where the network learns to map between correlated Gaussian distributions. Here, we use Adam Optimizer with a learning rate of $1e^{-4}$, with the Mean Squared Error Loss function.

MNIST Data. We also train a 4-layer MLP on the MNIST dataset (Lecun et al., 1998). Each hidden layer has 200 neurons. We use the Adam optimizer with a learning rate of $1e^{-4}$, with the Cross Entropy Loss function.

As shown in Figure 1, we observe a significant drop in local rank during the final stages of training in both cases. This phenomenon occurs across all layers of the network, suggesting that neural networks inherently compress the dimensionality of their learned representations as they converge. This compression occurs in two phases, where the second phase corresponds to the compression phase identified by Shwartz-Ziv et al. (2019) in the IB theory.

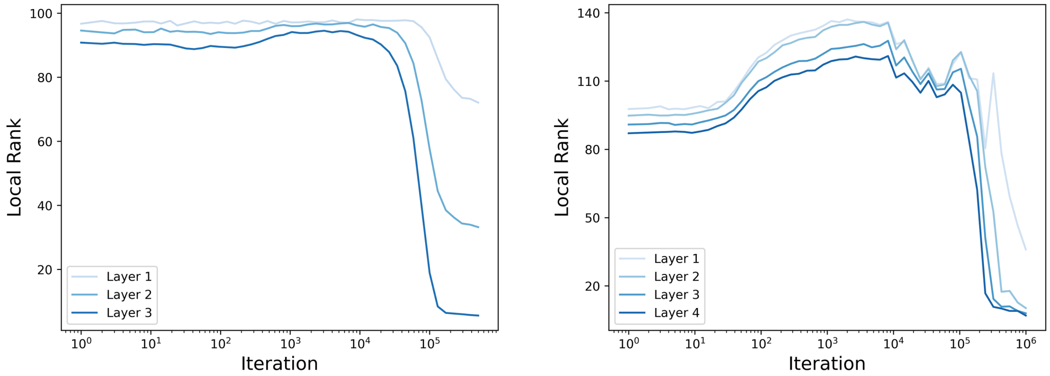

Figure 1: **Reduction in Local Rank During Training. Left:** A 3-layer MLP trained on synthetic Gaussian data. **Right:** A 4-layer MLP trained on MNIST. In both cases, the local rank decreases during the terminal phase of training, indicating compression of the feature manifold across all layers.

6 Information Theoretic Implications of Local Rank

A core challenge in developing an understanding of deep learning is to define what constitutes as good representation learning. The Information Bottleneck (IB) framework provides a theoretical foundation for this by proposing that optimal representations are those which are maximally informative about the targets while compressing redundant information. In this section, we explore the relationship between local rank and information compression, positioning our findings within the context of IB theory to provide insight into the structure of learned representations.

6.1 Information Bottleneck Framework

The Information Bottleneck (Tishby et al., 2000) and (Shwartz-Ziv, 2022) provides a principled approach to balance compression and relevance in representations. Given input X and output Y, the goal is to find a representation T that maximizes mutual information with Y while minimizing mutual information with X. The IB Lagrangian is:

$$\mathcal{L}_{\text{IB}} = I(T;X) - \beta I(T;Y), \tag{8}$$

where β controls the trade-off between compression and prediction.

6.2 Local Rank and the Gaussian Information Bottleneck

In general, finding analytical solutions to the IB problem is challenging. However, for jointly Gaussian variables, Chechik et al. (2003) found an explicit solution for the IB problem, which we can adjust:

Theorem 4. ***(Adapted from Chechik et al. (2003))*** *For jointly Gaussian variables X and Y, the solution to the IB optimization problem is a noisy linear transformation $T = A_\beta X + \eta$, where η is Gaussian noise. Moreover, there exist critical values β_n^c such that $0 \leq \beta_i^c \leq \beta_j^c$ whenever $i < j$, and*

$$\textit{rank}(A_\beta) = n \quad \textit{for} \quad \beta \in (\beta_n^c, \beta_{n+1}^c). \tag{9}$$

This theorem indicates that as we adjust the trade-off parameter β, there are bifurcation points where the rank of the optimal linear transformation A_β changes. This corresponds to changes in the dimensionality of the representation T, which is directly related to the local rank in the case of neural networks.

6.3 Empirical Evidence Connecting Local Rank and Information Compression

After establishing an analytical connection between local rank and the trade-off parameter β in the Gaussian case, we now empirically test this relationship on both synthetic and more complex datasets. We train Deep Variational Information Bottleneck (VIB) models (Alemi et al., 2016) and observe the effect of the IB trade-off parameter β and analytical phase transitions on the local rank of the encoder.

Gaussian Data. In the first experiment, we train Deep VIB models to map between two correlated Gaussian distributions in $\mathbb{R}^5$. In the **left** of Figure 2, we show that the KL divergence component of the loss varies with β, as expected. The points are colored according to the empirical local rank values, which correspond to the closest critical β values predicted by the theory. In the **right** of the figure, we plot the local rank as a function of β, observing that it increases with β and there is a distinct phase transition, aligning with the theoretical predictions. See Appendix B for more information.

MNIST and Fashion-MNIST Data. In the second experiment, we train Deep VIB models on the MNIST (Lecun et al., 1998) and Fashion-MNIST (Xiao et al., 2017) datasets. As we increase β, we observe that the local rank increases and the accuracy decreases.

In both experiments, changing β leads to a reduction in local rank, indicating increased information compression. This supports our hypothesis that local rank is indicative of the level of information compression in the network, even for complex datasets.

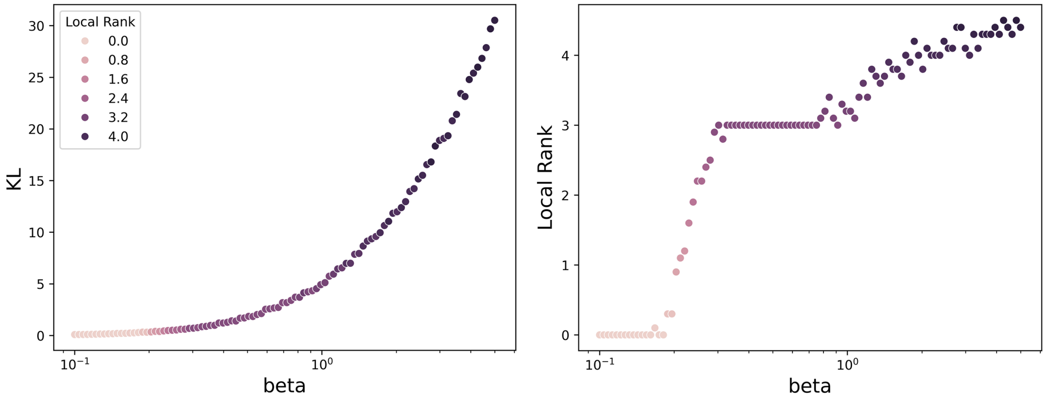

Figure 2: **Empirical Results on Gaussian Data using Deep-VIB. Left:** KL divergence component of the loss versus β, with points colored by empirical local rank corresponding to critical β values. **Right:** Local rank as a function of β, showing an increase with β and distinct phase transitions. We provide more information in Appendix B.

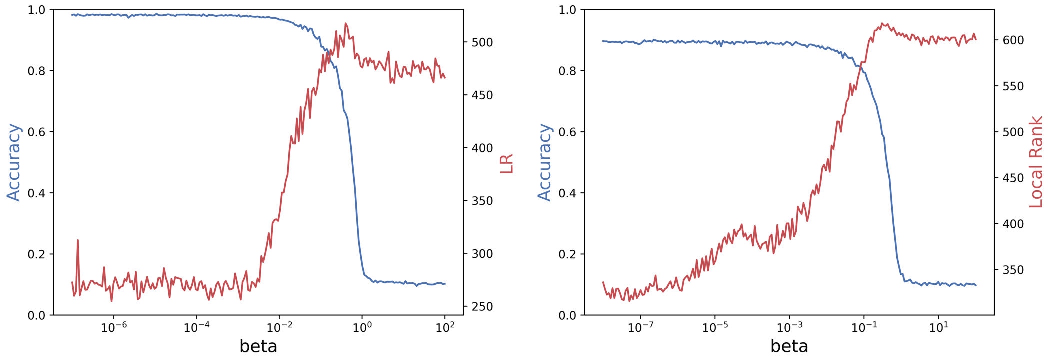

Figure 3: Empirical Results on MNIST and Fashion-MNIST. **Left:** MNIST dataset. **Right:** Fashion-MNIST dataset. As we increase β, the local rank increases, and accuracy decreases, indicating that higher β values correspond to less compressed representations and lower performance.

7 Discussion and Future Work

Our work introduces the concept of local rank as a meaningful measure of the dimensionality of feature manifolds in deep neural networks. We have demonstrated both theoretically and empirically that local rank decreases during the terminal phase of training, suggesting that networks compress the dimensionality of their representations.

By connecting local rank to the Information Bottleneck theory, we provide a new perspective on how neural networks manage the trade-off between compression and prediction. Our findings imply that networks naturally perform information compression by reducing the rank of their learned representations.

Understanding the behavior of local rank has implications for model compression, generalization, and the design of neural network architectures. Future research could formalize the relationship between local rank and mutual information in non-Gaussian settings, extend the analysis to other network architectures, and explore practical applications in compression techniques.

References

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

Ido Ben-Shaul, Ravid Shwartz-Ziv, Tomer Galanti, Shai Dekel, and Yann LeCun. Reverse engineering self-supervised learning. *Advances in Neural Information Processing Systems*, 36:58324–58345, 2023.

Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for gaussian variables. In S. Thrun, L. Saul, and B. Schölkopf (eds.), *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL `https://proceedings.neurips.cc/paper_files/paper/2003/file/7e05d6f828574fbc975a896b25bb011e-Paper.pdf`.

Ruili Feng, Kecheng Zheng, Yukun Huang, Deli Zhao, Michael Jordan, and Zheng-Jun Zha. Rank diminishing in deep neural networks. *Advances in Neural Information Processing Systems*, 35:33054–33065, 2022.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

Ahmed Imtiaz Humayun, Ibtihel Amara, Candice Schumann, Golnoosh Farnadi, Negar Rostamzadeh, and Mohammad Havaei. Understanding the local geometry of generative model manifolds. *arXiv preprint arXiv:2408.08307*, 2024.

Arthur Jacot. Bottleneck structure in learned features: Low-dimension vs regularity tradeoff. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 23607–23629. Curran Associates, Inc., 2023a. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/4a6695df88f2de0d49f875189ea181ef-Paper-Conference.pdf`.

Arthur Jacot. Implicit bias of large depth networks: a notion of rank for nonlinear functions. In *The Eleventh International Conference on Learning Representations*, 2023b.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17176–17186. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/c76e4b2fa54f8506719a5c0dc14c2eb9-Paper.pdf`.

Yann Lecun, Leon Bottou, Y. Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278 – 2324, 12 1998. doi: 10.1109/5.726791.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SJeLIgBKPS`.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Christopher Olah. Neural networks, manifolds, and topology. *Blog post*, 2014.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020. doi: 10.1073/pnas.2015509117. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2015509117`.

Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox. On the information bottleneck theory of deep learning. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=ry_WPG-A-`.

Ravid Shwartz-Ziv. Information flow in deep neural networks. *arXiv preprint arXiv:2202.06749*, 2022.

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

Ravid Shwartz-Ziv, Amichai Painsky, and Naftali Tishby. Representation compression and generalization in deep neural networks, 2019. In *URL https://openreview. net/forum*, 2019.

Ravid Shwartz-Ziv, Randall Balestriero, Kenji Kawaguchi, Tim G. J. Rudner, and Yann LeCun. An information theory perspective on variance-invariance-covariance regularization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 33965–33998. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6b1d4c03391b0aa6ddde0b807a78c950-Paper-Conference.pdf.

Peter Súkeník, Marco Mondelli, and Christoph Lampert. Neural collapse versus low-rank bias: Is deep neural collapse really optimal?, 2024. URL https://arxiv.org/abs/2405.14468.

Nadav Timor, Gal Vardi, and Ohad Shamir. Implicit regularization towards rank minimization in relu networks. In Shipra Agrawal and Francesco Orabona (eds.), *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pp. 1429–1459. PMLR, 20 Feb–23 Feb 2023a. URL https://proceedings.mlr.press/v201/timor23a.html.

Nadav Timor, Gal Vardi, and Ohad Shamir. Implicit regularization towards rank minimization in relu networks. In *International Conference on Algorithmic Learning Theory*, pp. 1429–1459. PMLR, 2023b.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Bohan Wang, Qi Meng, Wei Chen, and Tie-Yan Liu. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In *International Conference on Machine Learning*, pp. 10849–10858. PMLR, 2021.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

A Proofs

A.1 For Binary Classification

Lemma 5. *There exists $\epsilon_0 > 0$ such that for all $\epsilon < \epsilon_0$:*

$$\mathit{rank}_\epsilon(J_x p_l(x)) \leq \mathit{rank}_\epsilon(W_l) \tag{10}$$

Proof. Notice first that a calculation gives us the following, where W_i are the weight matrices and D_i are diagonal matrices with 0 or 1 on the diagonal. These diagonal matrices correspond to the activation pattern of the ReLU functions at a specific layer.

$$J_x p_l(x) = W_l D_{l-1} W_{l-1} D_{l-2} \cdots W_0$$

Notice now it is clear that:

$$\text{rank}(J_x p_l(x)) \leq \text{rank}(W_l)$$

Notice also that for any matrix A: $\lim_{\epsilon \to 0} \text{rank}_\epsilon(A) = \text{rank}(A)$. Since $\text{rank}_\epsilon(\cdot)$ takes on values in a discrete set, its clear that there exists some $\epsilon_0 > 0$ such that $\epsilon < \epsilon_0$ implies that:

$$\text{rank}_\epsilon(J_x p_l(x)) \leq \text{rank}_\epsilon(W_l)$$

□

We now recall a theorem courtesy of Timor et al. (2023a):

Theorem 6. ***(Quoted from Timor et al. (2023a))*** *Let $\{(x_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^{n_0} \times \{-1, 1\}$ be a binary classification dataset, and assume that there is $i \in [n]$ with $\|x_i\| \leq 1$. Assume that there is a fully-connected neural network N of width $m \geq 2$ and depth $k \geq 2$, such that for all $i \in [n]$ we have $y_i N(x_i) \geq 1$, and the weight matrices $W_1, \ldots, W_k$ of N satisfy $\|W_i\|_F \leq B$ for some $B > 0$. Let N_θ be a fully-connected neural network of width $m' \geq m$ and depth $k' > k$ parameterized by θ. Let $\theta^* = [W_1^*, \ldots, W_L^*]$ be a global optimum of the above* *optimization problem 4. Namely, θ^* parameterizes a minimum-norm fully-connected network of width n_l and depth L that labels the dataset correctly with margin 1. Then, we have*

$$\frac{1}{L}\sum_{i=1}^{L}\frac{\|W_i^*\|_\sigma}{\|W_i^*\|_F} \geq \frac{1}{\sqrt{2}}\cdot\left(\frac{\sqrt{2}}{B}\right)^{\frac{k}{L}}\cdot\sqrt{\frac{L}{L+1}}. \tag{11}$$

Equivalently, we have the following upper bound on the harmonic mean of the ratios $\frac{\|W_i^*\|_F}{\|W_i^*\|_\sigma}$:

$$\frac{L}{\sum_{i=1}^{L}\left(\frac{\|W_i^*\|_F}{\|W_i^*\|_\sigma}\right)^{-1}} \leq \sqrt{2}\cdot\left(\frac{B}{\sqrt{2}}\right)^{\frac{k}{L}}\cdot\sqrt{\frac{L+1}{L}}. \tag{12}$$

For convenience we restate our proposition in full formality:

Proposition 7. *Let $\{(x_i,y_i)\}_{i=1}^n \subseteq \mathbb{R}^{n_0}\times\{-1,1\}$ be a binary classification dataset, and assume that there is $i\in[n]$ with $\|x_i\|\leq 1$. Assume that there is a fully-connected neural network N of width $m\geq 2$ and depth $k\geq 2$, such that for all $i\in[n]$ we have $y_iN(x_i)\geq 1$, and the weight matrices $W_1,\ldots,W_k$ of N satisfy $\|W_i\|_F\leq B$ for some $B>0$. Let N_θ be a fully-connected neural network of width $m'\geq m$ and depth $k'>k$ parameterized by θ. Let $\theta^*=[W_1^*,\ldots,W_L^*]$ be a global optimum of the above optimization problem 4. Namely, θ^* parameterizes a minimum-norm fully-connected network of width n_l and depth L that labels the dataset correctly with margin 1. Then, there exists an $l\in\{1,\cdots,L\}$ and $\epsilon_0>0$ such that for $0<\epsilon<\epsilon_0$ the following holds:*

$$\frac{\boldsymbol{LR}_l^\epsilon}{||W_l^*||_\sigma^2} \leq \frac{2}{\epsilon^2}\cdot(\frac{B}{\sqrt{2}})^{\frac{2k}{L}}\cdot\frac{L+1}{L} \tag{13}$$

Proof. Notice first that for any arbitrary matrix A, we have that,

$$||A||_F = \sqrt{\sum_{i=1}^{n}\sigma_i^2} \geq \sqrt{\text{rank}_\epsilon(A)\epsilon^2} = \epsilon\sqrt{\text{rank}_\epsilon(A)}$$

So then,

$$\frac{||A||_F}{||A||_\sigma} \geq \frac{\epsilon}{||A||_\sigma}\sqrt{\text{rank}_\epsilon(A)} \tag{14}$$

Notice now that from application of the theorem in Timor et al. (2023a), we can get that harmonic mean of the quantities $\frac{||W_i^*||_F}{||W_i^*||_\sigma}$ is bounded. In particular this means that there exists at least one $l\in\{1,\cdots,L\}$ such that $\frac{||W_l^*||_F}{||W_l^*||_\sigma}$ lies below this harmonic mean. So then,

$$\sqrt{2}\cdot\left(\frac{B}{\sqrt{2}}\right)^{\frac{k}{L}}\cdot\sqrt{\frac{L+1}{L}} \geq \frac{||W_l^*||_F}{||W_l^*||_\sigma} \tag{15}$$

$$\geq \frac{\epsilon}{||W_l^*||_\sigma}\sqrt{\text{rank}_\epsilon(W_l^*)} \tag{16}$$

Re-arranging terms and squaring both sides gives us:

$$\frac{\text{rank}_\epsilon(W_l^*)}{||W_l^*||_\sigma^2} \leq \frac{2}{\epsilon^2}\cdot(\frac{B}{\sqrt{2}})^{\frac{2k}{L}}\cdot\frac{L+1}{L}$$

Now if we can apply lemma 1 and we get that there exists $\epsilon'>0$ such that for any $0<\epsilon<\epsilon'$:

$$\text{rank}_\epsilon(J_xp_l(x)) \leq \text{rank}_\epsilon(W_l^*)$$

So then, we get:

$$\frac{\text{rank}_\epsilon(J_xp_l(x))}{||W_l^*||_\sigma^2} \leq \frac{2}{\epsilon^2}\cdot(\frac{B}{\sqrt{2}})^{\frac{2k}{L}}\cdot\frac{L+1}{L} \tag{17}$$

Since this holds for any x, we can then take expectation with respect to the data distribution on the left hand side and we get that:

$$\frac{\mathbf{LR}_l^\epsilon}{||W_l^*||_\sigma^2} \leq \frac{2}{\epsilon^2} \cdot (\frac{B}{\sqrt{2}})^{\frac{2k}{L}} \cdot \frac{L+1}{L} \tag{18}$$

□

A.2 For Interpolating Neural Networks

As before we recall the analogous theorem from Timor et al. (2023b) for interpolating neural networks.

Theorem 8. ***(Quoted from Timor et al. (2023a))*** *Let $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^{n_0} \times \mathbb{R}_+$ be a dataset, and assume that there is $i \in [n]$ with $\|x_i\| \leq 1$ and $y_i \geq 1$. Assume that there is a fully-connected neural network N of width $m \geq 2$ and depth $k \geq 2$, such that for all $i \in [n]$ we have $N(x_i) = y_i$, and the weight matrices $W_1, \ldots, W_k$ of N satisfy $\|W_i\|_F \leq B$ for some $B > 0$. Let N_θ be a fully-connected neural network of width $m' \geq m$ and depth $L > k$ parameterized by θ. Let $\theta^* = [W_1^*, \ldots, W_L^*]$ be a global optimum of the following problem:*

$$\min_\theta \|\theta\| \quad \textit{s.t.} \quad \forall i \in [n] \; N_\theta(x_i) = y_i. \tag{19}$$

Then,

$$\frac{1}{L}\sum_{i=1}^{L} \frac{\|W_i^*\|_\sigma}{\|W_i^*\|_F} \geq \left(\frac{1}{B}\right)^{\frac{k}{L}}. \tag{20}$$

Equivalently, we have the following upper bound on the harmonic mean of the ratios $\frac{\|W_i^*\|_F}{\|W_i^*\|_\sigma}$:

$$\frac{L}{\sum_{i=1}^{L}\left(\frac{\|W_i^*\|_F}{\|W_i^*\|_\sigma}\right)^{-1}} \leq B^{\frac{k}{L}}. \tag{21}$$

We can now restate our proposition with a proof.

Proposition 9. *Let $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^{n_0} \times \mathbb{R}_+$ be a dataset, and assume that there is $i \in [n]$ with $\|x_i\| \leq 1$ and $y_i \geq 1$. Assume that there is a fully-connected neural network N of width $m \geq 2$ and depth $k \geq 2$, such that for all $i \in [n]$ we have $N(x_i) = y_i$, and the weight matrices $W_1, \ldots, W_k$ of N satisfy $\|W_i\|_F \leq B$ for some $B > 0$. Let N_θ be a fully-connected neural network of width $m' \geq m$ and depth $L > k$ parameterized by θ. Let $\theta^* = [W_1^*, \ldots, W_L^*]$ be a global optimum of the following problem:*

$$\min_\theta \|\theta\| \quad \textit{s.t.} \quad \forall i \in [n], \; \mathcal{N}_\theta(x_i) = y_i. \tag{22}$$

Then, there exist an $l \in \{1, \cdots, L\}$ and $\epsilon_0 > 0$ such that for $0 < \epsilon < \epsilon_0$ the following holds:

$$\frac{\boldsymbol{LR}_l^\epsilon}{||W_l||_\sigma^2} \leq \frac{B^{\frac{2k}{L}}}{\epsilon^2} \tag{23}$$

Proof. We first apply the prior theorem to get that the harmonic mean of the ratios of the Frobenius norm to the operator norm of the weight matrices are bounded like:

$$\frac{L}{\sum_{i=1}^{L}\left(\frac{\|W_i^*\|_F}{\|W_i^*\|_\sigma}\right)^{-1}} \leq B^{\frac{k}{L}}. \tag{24}$$

In particular, there exists some layer l such that its ratio falls below the Harmonic mean, so then:

$$\frac{||W_i^*\|_F}{||W_i^*\|_\sigma} \leq B^{\frac{k}{L}}. \tag{25}$$

Now recall that for any matrix A we have:

$$\frac{||A||_F}{||A||_\sigma} \geq \frac{\epsilon}{||A||_\sigma}\sqrt{\text{rank}_\epsilon(A)}. \tag{26}$$

Now apply this to W_l and we get that:

$$\frac{||W_l||_F}{||W_l||_\sigma} \geq \frac{\epsilon}{||W_l||_\sigma}\sqrt{\text{rank}_\epsilon(W_l)}. \tag{27}$$

We can now apply the lemma and we get that there exists $\epsilon_0 > 0$ such that for any $0 < \epsilon < \epsilon_0$:

$$\frac{||W_l||_F}{||W_l||_\sigma} \geq \frac{\epsilon}{||W_l||_\sigma}\sqrt{\text{rank}_\epsilon(W_l)} \geq \frac{\epsilon}{||W_l||_\sigma}\sqrt{\text{rank}_\epsilon(J_x p_l(x))}. \tag{28}$$

So then it follows that:

$$\frac{\epsilon}{||W_l||_\sigma}\sqrt{\text{rank}_\epsilon(J_x p_l(x))} \leq B^{\frac{k}{L}}. \tag{29}$$

Or equivalently,

$$\frac{\epsilon^2}{||W_l||_\sigma^2}\text{rank}_\epsilon(J_x p_l(x)) \leq B^{\frac{2k}{L}}. \tag{30}$$

Taking expectation over $x \sim$ **Data** now completes the proof as:

$$\frac{\mathbf{LR}_l^\epsilon}{||W_l||_\sigma^2} \leq \frac{B^{\frac{2k}{L}}}{\epsilon^2}. \tag{31}$$

□

B On the Gaussian Deep VIB

For figure 2, our Deep VIB model is trained to map X to Y using a Deep Linear network as the encoder. Here we take both of these to be isotropic Gaussians in $\mathbb{R}^5$. We use the following cross-covariance matrix:

$$\Sigma_{XY} = \begin{pmatrix} 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.5 \end{pmatrix}. \tag{32}$$

For an intuition, this means that the last 3 variables are highly correlated between X and Y, whereas the first two variables are only somewhat correlated. The theory would then suggest a phase transition, and a distinct jump from a rank to a rank of 3 as we lower β. We note that we can observe this structure in figure 2 (right).