Does the Geometry of the Data Control the Geometry of Neural Predictions? (Student Abstract)

Anirudh Cowlagi, Pratik Chaudhari
Department of Electrical and Systems Engineering,
General Robotics, Automation, Sensing and Perception (GRASP) Laboratory,
University of Pennsylvania,
3330 Walnut St, Philadelphia, PA 19104
acowlagi@seas.upenn.edu, pratikac@seas.upenn.edu

Abstract

This paper studies the over-parameterization of deep neural networks using the Fisher Information Matrix from information geometry. We identify several surprising trends in the structure of its eigenspectrum, and how this structure relates to the eigenspectrum of the data correlation matrix. We identify how the eigenspectrum relates to the topology of the predictions of the model and develop a "model reduction" method for deep networks. This ongoing investigation hypothesizes certain universal trends in the FIM of deep networks that may shed light on their effectiveness.

Introduction

Deep networks are mysterious. Classical statistics dictates that these massively over-parametrized models, i.e., the number of weights is much larger than the number data, should overfit to the training data. However, deep networks essentially do not overfit, this is a phenomenon known as "benign over-fitting" (Bartlett et al. 2020). This short paper presents our investigations into understanding over-parameterization. **Our contributions** are as follows:

1. Deep networks are known to have a large number of **redundant weights**; roughly about 95% of the weights can be pruned or heavily quantized to reduced precision after training without changing the predictions of the model on the test data. At the same time, results like the Lottery Ticket Hypothesis suggest that one cannot simply train the resultant smaller deep network with random initializations. We **study this phenomenon under the lens of "sloppy models"**, a rather universal phenomenon discovered across statistics, physics, biology, and chemistry (Transtrum, Machta, and Sethna 2011) suggesting that nonlinear parametric models fitted to naturally occurring data are bound to have such redundant weights. To study this, we formalize a "model manifold" which is the manifold of the network's predictions on n training data for different weight configurations.

2. We identify a **peculiar relationship between the eigenvalues of the data auto-correlation matrix and the eigenvalues of the Fisher Information Matrix (FIM)**. Roughly speaking, while the former tells us the how different samples in typical datasets used in machine learning are similar to each other, the latter tells us how the output of a deep network is also similar on different images in the training set. Surprisingly, both of these matrices are not only similar but their eigenspectra are extremely "sloppy" – very few eigenvalues are large (less than 5%) and capture the bulk of the eigenspectrum while the remainder are dramatically smaller.

3. We perform **model reduction**, i.e., project a deep network into a subspace of simpler deep networks using the eigenvectors of the FIM. We observe that such a projection preserves both the predictions of the network on the data and the topology of the model manifold.

Geometry of Model Manifolds

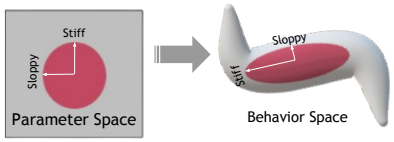

Figure 1: Manifold Structure of a Sloppy Model

Consider a nonlinear regression problem on dataset $D = \{(x_i, y_i)\}_{i=1}^n$ where a deep network with weights $w \in \mathbb{R}^p$ is fit to predict targets $\hat{y}_i = f(x_i; w)$. Predictions on the training set are a point $z(w) = [\hat{y}_1, \ldots, \hat{y}_n]$ in n-dimensional space parameterized by w (left panel, Fig. 1). The space of all predictions, $M = \{z(w) : w \in \mathbb{R}^p\}$ is called the model manifold (right panel Fig. 1). A Euclidean ball in weight (w) space skews and stretches when seen in the prediction ($z(w)$) space. The map between the two spaces is given by the Jacobian of the model, $J_{ij} = \partial z_i / \partial w_j$.

Like the Euclidean metric determines distance between weights $w, w' \in \mathbb{R}^p$, the Fisher Information Matrix (FIM) $g \propto J^\top J$ is the metric on the model manifold. The infinitesimal distance on the manifold turns out to be $\mathrm{d}s^2 = \sum_{i,j=1}^p g_{ij} \mathrm{d}w_i \mathrm{d}w_j$ (Amari 2016). Thus, the FIM allows us to understand when two deep networks w, w' predict similarly on inputs even if weights are far away in the Euclidean weight space. In particular, **large eigenvalues of the FIM** correspond to large changes in the predictions z despite small changes in the weights in the Euclidean

space. Such directions are **"stiff directions"**. **Small eigenvalues of the FIM** correspond to small changes in predictions z even if the weights undergo large changes in Euclidean space. Such directions are **"sloppy directions"**. This theory can also be applied to classification using $z = [\sqrt{p_w(y_1 \mid x_1)}, \ldots, \sqrt{p_w(y_n \mid x_n)}]$ where $p_w(y|x)$ is the softmax output. Finally, to ground intuition, note for linear regression, $g \propto XX^\top$, the data autocorrelation matrix.

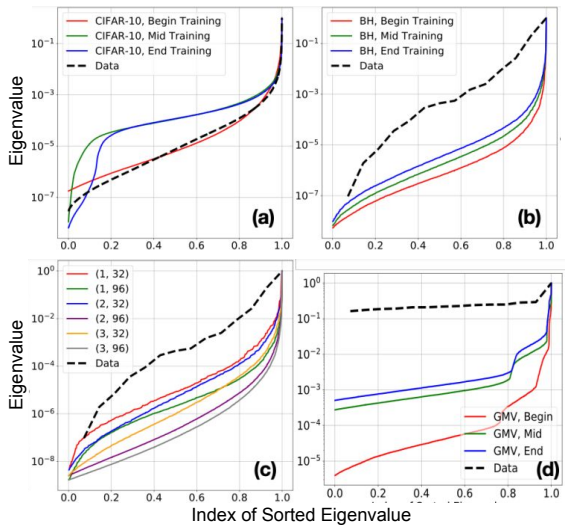

Figure 2: Eigenspectra of FIM at beginning (red), middle (green) and end (blue) of training compared to data covariance matrix $XX^\top$ (black dashed line) for different problems and architectures, **(a)** 9-layer CNN on CIFAR-10 image classification dataset, **(b)** 2 layer MLP on Boston Housing regression dataset, **(c)** different MLPs (#layers, #hidden neurons) on Boston Housing at end of training, and **(d)** perceptron from (b) on a random dataset with inputs as Gaussian random variables and outputs are sampled randomly.

The experiments in Fig. 2 conducted based on the observations above reveal surprising properties. Eigenspectra, of both data and FIM, in (a) and (b) span exponentially larger magnitudes: **low magnitude eigenvalues specify exactly the sloppy directions** described above. Eigenspectra for a random dataset (d) is not as sloppy as the others – only drops by about one to two orders after the initial drop (despite (b) having the same architecture), indicating that **input data sloppiness is necessary for model weight redundancy.** Eigenspectra in (c) of different networks trained on the same dataset are very similar: **slope of the linear (sloppy) range is similar for all**.

There is a peculiar sharp drop in the eigenspectrum at the right, most prominently for (a) and (b); most of these eigenspectra have less than 5% stiff eigenvalues; **we hypothesize that this property is related to the topology of the model manifold**. As training progresses, the trace of the FIM increases monotonically – sloppy eigenvalues increase in magnitude while stiff eigenvalues are essentially unchanged, suggesting the network may be learning representations correlated with sloppy eigenvectors of $XX^\top$ while **discarding extremely sloppy eigenvalues because there's less information** about these modes; this is not so for the random dataset in (c). This leads us to the hypothesis that the data's structure ultimately controls FIM structure (and network behavior).

Model Reduction

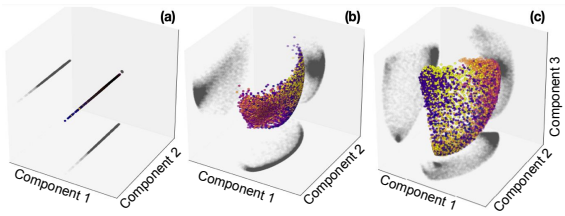

Figure 3: Each point is a datum colored by correct label on the CIFAR-10 dataset. (a) Highly reduced model retaining the top 1% eigenvectors of FIM makes poor predictions on all samples and is topologically a line. (b) Model that retains the top 5% eigenvectors has similar topology to the full model (c) and obtains similar accuracy on the validation set (75% vs 77 %).

The FIM eigenvectors determine the local coordinates of the model manifold. We can therefore project a trained model on the top few eigenvectors of its FIM to "reduce" it, by keeping the most important prediction directions of the model manifold and set sloppy directions to zero. We write the weight as $w = \sum_{i=1}^{p} c_i e_i = \sum_{i=1}^{p} \langle w, e_i \rangle e_i \approx \sum_{i=1}^{k} \langle w, e_i \rangle e_i$ where e_i are the FIM eigenvectors sorted by eigenvalue magnitude. We then use a dimensionality reduction method known as Intensive Principal Component Analysis designed for the isometric embedding of probabilistic models (like the predictions z) (Quinn et al. 2019) to compare the original deep network and reduced ones; this is shown in Fig. 3. Differences in topological complexity between models indicate there may exist a critical threshold of eigenvectors needed to ensure the topological (and predictive) accuracy of the reduced model. This threshold may depend on the ratio of stiff to sloppy eigenvalues in $XX^\top$.

Acknowledgments

The first author was funded by the Vagelos Integrated Program in Energy Research (VIPER) at Penn and conducted this research during the summer of his freshman year.

References

Amari, S.-i. 2016. *Information Geometry and Its Applications*. Springer Japan.

Bartlett, P. L.; Long, P. M.; Lugosi, G.; and Tsigler, A. 2020. Benign Overfitting in Linear Regression. *PNAS*.

Quinn, K. N.; Clement, C. B.; De Bernardis, F.; Niemack, M. D.; and Sethna, J. P. 2019. Visualizing Probabilistic Models: Intensive Principal Component Analysis. *PNAS*.

Transtrum, M. K.; Machta, B. B.; and Sethna, J. P. 2011. Geometry of nonlinear least squares with applications to sloppy models and optimization. *Physical Review E*.