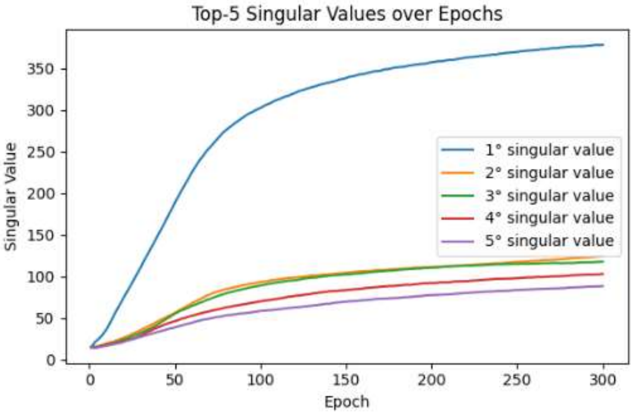
Spectral Analysis of weight matrix in learning for RBM

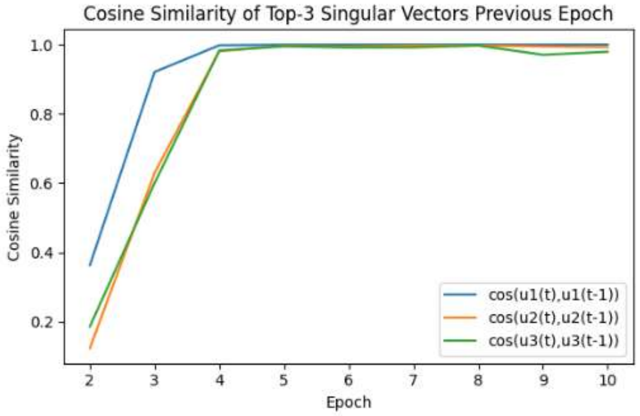
Daniele Vanzan 880054

Restricted Boltzmann machine (RBM) constitutes a common tool of machine learning. It is a generative model, in the sense that it defines a probability distribution, which can be learned to approximate any distribution of data points living in some N-dimensional space, with N potentially large. The RBMs are used for both unsupervised and supervised task, for the last one a softmax layer is added in the end to obtain probabilities of belonging to a class. In this project I investigate this kind of architecture on a hand writing classification tasks, based on MNIST dataset with images in black and white, that are encoded as 1 and 0, so the model of anlysis is a Bernoulli RBM for classification. The MNIST dataset is split in training (0.9) and test (0.1) to evaluate the performance of classifing hand writing numbers in decimal numbers.The standard learning procedure is called contrastive divergence from Hinton. At the same time an RBM can be regarded as a statistical physics model, being defined as a Boltzmann distribution with pairwise interactions on a bipartite graph. The energy of the states of RBM is $E(v,h)=-v^T W h-b^T v-c^T h$, so the weight matrix W is one of the components that defines the energy.

The objective of this project is to understand how the eigenvalues of the weight matrix evolve during training epochs and is divided in two parts depending on the tool used: PCA in the first part and Marchenko-Pastur distribution for the second. These informations can be used to reduce the dimension of the hidden neurons to the neurons where the learning process happens. The analysis of the spectrum are made for three RBM with hidden size respectively of 10, 100 and 1000 all trained for 300 epochs, and the input is a tensor of dimension of 784. The entries of the weight matrix are initialized as N(0, 0.5) . During the training epochs the weight matrix and the bias vectors are saved, then some spectral characteristics are plotted to understand the evolutions during learning epochs. The RBM with 100 hidden neurons achive the higher accuracy (about 0.91) after 20-30 epochs then the model won’t learn more complexity. After first 30 epochs the bias vectors are trained and will not change any more during the training, so after the bias are setted the evolution of weight matrix represent the further learning and the energy of the system.

The SVD of the weight matrix in this first epochs explain how the learning of RBM modify the weights to represent the distribution of the images in the hidden space, W is of shape hidden size × input size.The maximum singular value and the correspondent singular vector seems to explain the learning: for all the three order of magnitude of the hidden size, it appears that first singular value increase more and more than other singular value, meaning that RBM “learn in the direction of higher variance”. The direction is the first singular vector and, to understand how the vector changes during training, I plot the cosine similarity between the singular vector and the singular vector at the previous epoch. It is possible to see that after 4 epochs the first singular vector doesn’t change any more After these initial epochs the learning doesn’t consist in adjusting the singular vector, but only in increasing the first singular value, and less influent by increasing the other singular value. If the hidden size is sufficiently small and, so the explained variance of the first principal component is high, the learning of the RBM can be summarized as “first, adjust the direction of higher variance of weight matrix then learn more and more in this direction”, where the direction is a direction in the input space, that could be explained with less neurons.





The second part of this analysis is inspired from the paper "Dyson Brownian motion and random matrix dynamics of weight matrices during learning", which uses the Marchenko–Pastur (MP) distribution to explain the learning dynamics of Gaussian RBMs and transformers. The central idea is that the weight matrix before training is initialized with entries drawn from a Gaussian distribution with zero mean and variance 0.5. This makes it a suitable object for analysis using the tools of Random Matrix Theory.

At initialization, the weight matrix behaves like a random matrix consisting purely of Gaussian noise. The MP distribution represent the distribution of the eigenvalues of the correlation matrix, defined as $\frac{1}{n}WW^T$ where W is a Random Matrix of dimension $p \times n$, with n,p ➔ ∞ and Q=p/n is higher than 0 and not divergent. This distribution is defined in an interval of eigenvalues (λ-,λ+) dependig on Q and the variance, this interval is called "bulk" and, outside this, the probability of finding an eigenvalue is zero, if W is a random matrix, meaning that the matrix represent only gaussian noise, if there are outliers from the "bulk" it means that the matrix represent a signal in the high dimensional space, and is no more a RM. Noticed that the eigenvalue of the correlation matrix are the square of the singular values of the SVD of W, of the previous analysis.

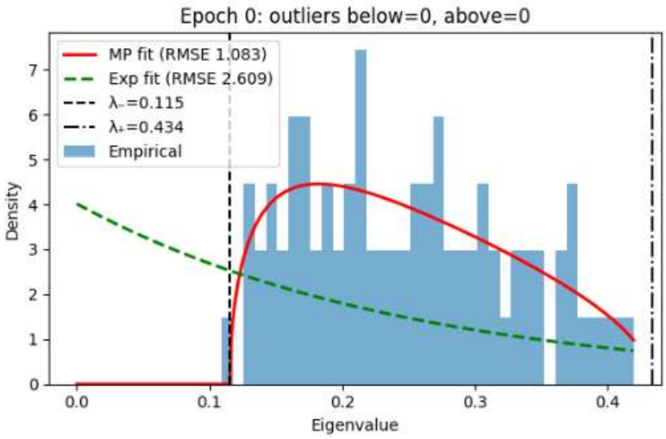
As training proceeds, the optimization process (contrastive divergence for RBMs) induces low-rank perturbations in the weight matrix. These perturbations correspond to the emergence of signal directions, which manifest as outlier eigenvalues that detach from the bulk of the MP spectrum. These outliers represent the signal learned by the RBM.

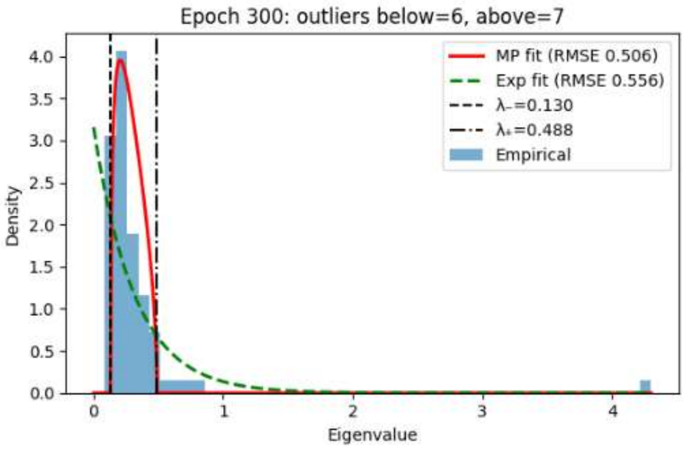
The outliers under the bulk are very close to zero, meaning that the weights could be represented in a hidden space smaller by removing the same dimensions and can be used as a rule to reduce the hidden neurons. The outliers above the bulk are the emergent signal components that capture meaningful structure learned during training. These correspond to directions in the weight space that encode relevant features or correlations and can be interpreted as the informative dimensions responsible for the model's predictive power.

This phenomenon can be understood in terms of Dyson Brownian motion: the eigenvalues of the weight matrix follow stochastic dynamics during training, driven by the optimization process. Initially, the spectrum is concentrated within the MP bulk. As training progresses, meaningful features begin to emerge in the form of eigenvalues that move beyond the MP upper edge.

The separation between noise and signal becomes detectable via the spectral density that track the learning progress, quantify the number of learned features, and even perform dimensionality reduction by removing components associated with noise (eigenvalues below λ−).

From these two spectal analysis it is possible to see that the first eigenvalue (the square of the first singular value) increase more and more than the other, meaning that the learning goes more and more in the direction of higher variance, after the direction is learned.





SOURCES:
https://ml4physicalsciences.github.io/2024/files/NeurIPS_ML4PS_2024_23.pdf

https://arxiv.org/pdf/1708.02917

MNIST dataset:
http://yann.lecun.com/exdb/mnist/