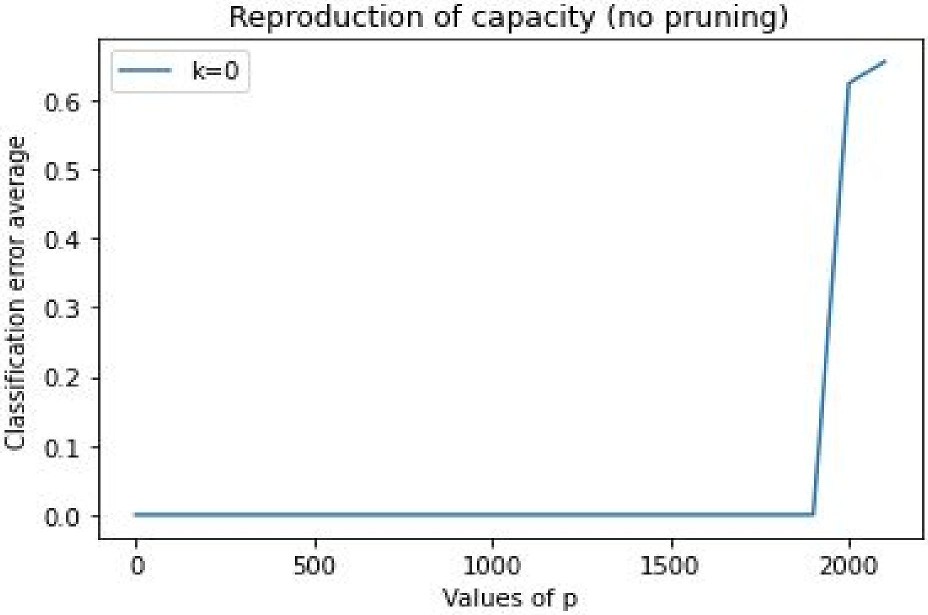
Mehdi Bouassami
14/03/21

Report

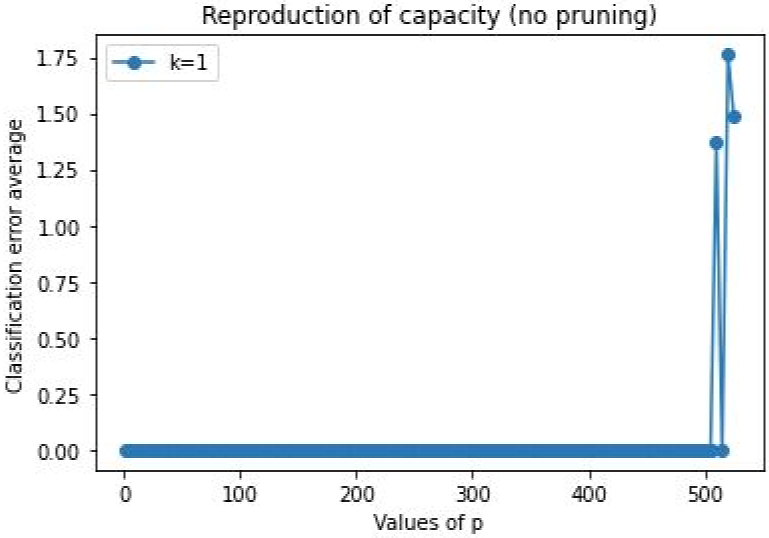
The strength of connections between neurons, called synaptic weights, enable the brain to function. It is known that even if some synapses turn over, the association between an input and an output can remain stable. However, the reason for this stability is still unknown. This is where the neuroscience motivation for pruning is from.
Our project tries to better our understanding of this mysterious stability by investigating a special perceptron model in which we reproduce synaptic turnover by setting half of the smallest synaptic weights in absolute value to zero.

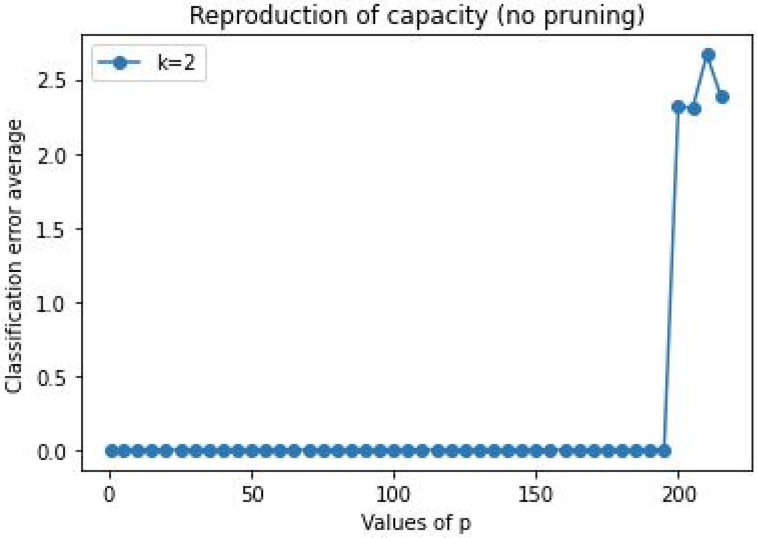
We started with a numerical investigation of learning input output associations using a simple perceptron model. The perceptron is a model of a neuron that adjusted synaptic weights in order to learn to classify the inputs it receives. And when a certain number of input is reached, they cannot be classified anymore. We say that we have reached capacity and that is what we started investigating. We kept increasing the number of input patterns (p) until we reached capacity (α) and we found out that there was a dependence on the number of associations that is given by the following equation: P = αN. We were able to show that numerically using python code:

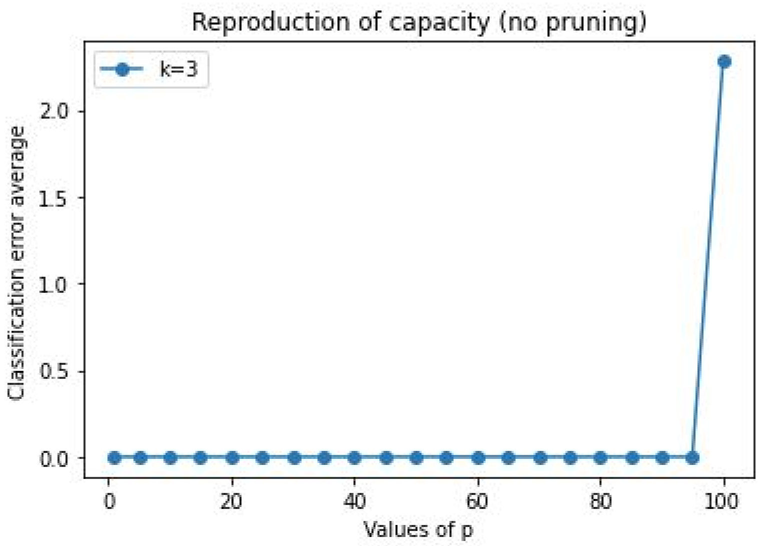


In this graph, N = 1000 so we see that the capacity is reached when p is approximately 2000 so α = 2.

In order to make our perceptron learn, we used the classification error which is equal to the actual value - the expected value of the correctly classified input. But we did not take into account how close to 0 $\sum_{i=1}^{N} w_i \xi_i^{\mu}$ (w being the weights and ξ being the matrix of input patterns) is. In order to take into account the margin of noise, we assumed that a classification is correct if $y^{\mu} * \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i \xi_i^{\mu}$ > k. If not, we would update the weights. By running different simulations, we understood that there was a dependence between learning input output associations and the value of k (the margin for noise).

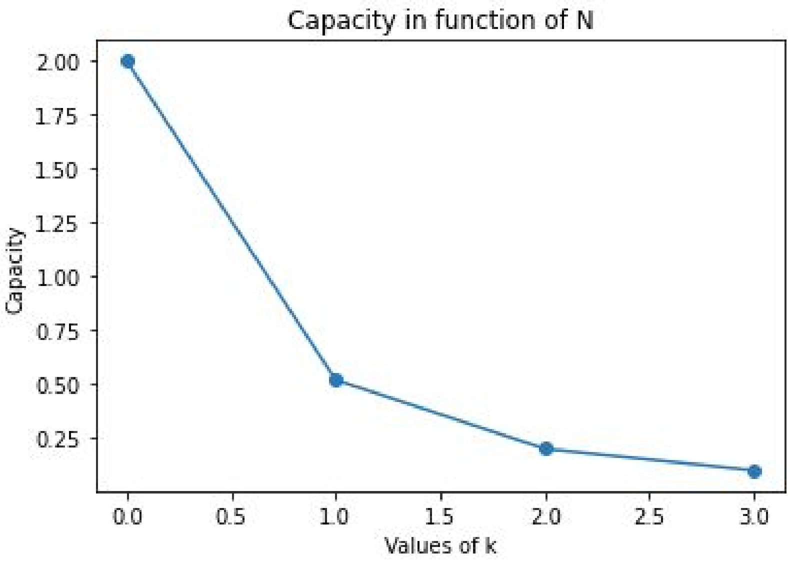






From the graphs above, we understand that the larger the value of k and the smaller the value of α. It is interesting to note that the values of k found in the graphs above follow E. Gardner's theory. Indeed, her formula for the capacity is:

$$\alpha_c = \left(\int_{-\kappa}^{\infty} \mathrm{D}t(t+\kappa)^2 \right)^{-1}$$

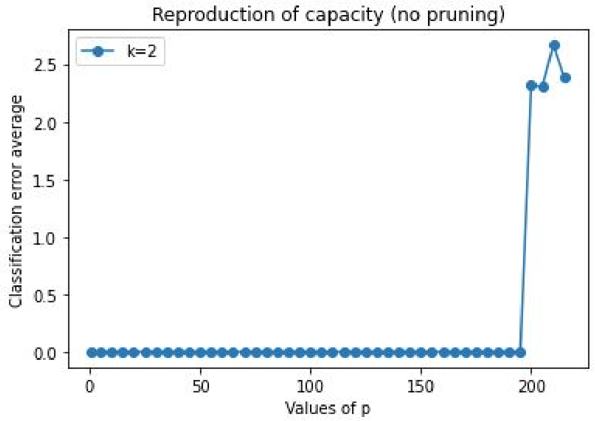


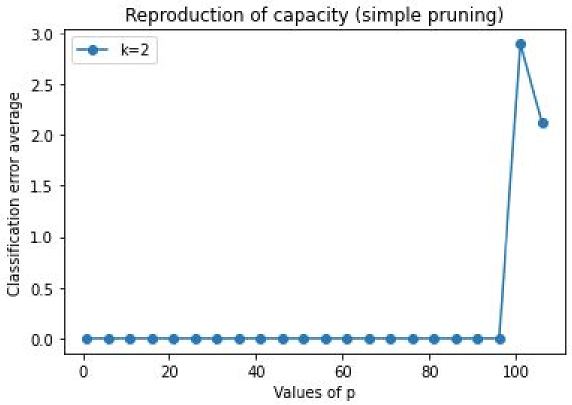
So when k = 0, we expect the graph to increase from 0 at N = 2000, for k = 1 at N = 500… and this is exactly what our graphs show.

Hence, from these simulations, we understood that learning input output associations is dependent on both the number of associations and the margin for noise.

After understanding these different calculations for capacity, we started to focus more on the actual idea of our project which is to try and understand what happens when some synapses turn over.

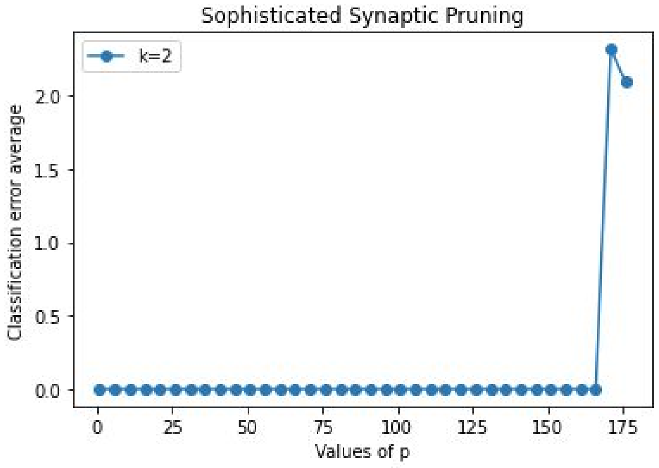
We started with simple pruning by setting half of the synaptic weights to 0 randomly. We made the assumption that the results would follow Gardner's theory as the capacity should only be what the capacity would have been without pruning divided by 2. Indeed, by setting half of the synaptic weights to 0 randomly, we just decrease the number of input patterns by 2.





Our assumption was true as shown with the graph representing our non-simply-pruned and pruned simulation.

We then decided to complexify the experience by setting half of the smallest synaptic weights in absolute values to 0. This is called sophisticated pruning. It is better than simple pruning because it is not a random process. Indeed, we set the synaptic weights which are the least important to 0 while in simple pruning, we might set some high value synaptic weights to 0. So we expect sophisticated pruning to be better than predicted by Gardner's theory. Hence, we did not expect the capacity to be as low as the one for simple pruning but at the same time, we did not expect it to be as high as the one with no pruning at all. And this is exactly what we found:



After getting these very exciting results, we started reading Gardner's paper to understand the theory of the capacity with no pruning. However, the paper was harder to understand than I thought because there were some mathematical concepts that were new and that required some reading to understand. The challenging mathematical concepts included Dirac delta function, Heaviside functions, Gaussian integrals, Fourrier representation of Dirac delta function and Heaviside functions. However, I was able to better my understanding of these concepts through the weeks by working on examples and proving some equations in Gardner's paper and was able to have a solid understanding of Gardner's paper in the end.
For example, I was able to prove the two identities below using Fourier expansion in the first equation and delta function properties in the second.

$$\begin{aligned} \delta\left(y-z\right) &= \int \frac{dx}{2\pi} e^{ix(y-z)} \\ \Theta\left(z-\kappa\right) &= \int_{\kappa}^{\infty} dy \delta\left(y-z\right) = \int_{\kappa}^{\infty} dy \int \frac{dx}{2\pi} e^{ix(y-z)} \end{aligned}$$

I also had a hard time understanding the volume's calculation by Gardner.

$$V = \frac{\int \left(\prod_i dw_i\right) \prod_{\mu} \Theta\left(Z_{\mu} \frac{1}{\sqrt{N}} \sum_i w_i \xi_{i\mu} - \kappa\right) \delta\left(\sum_i w_i^2 - N\right)}{\int \left(\prod_i dw_i\right) \delta\left(\sum_i w_i^2 - N\right)}$$

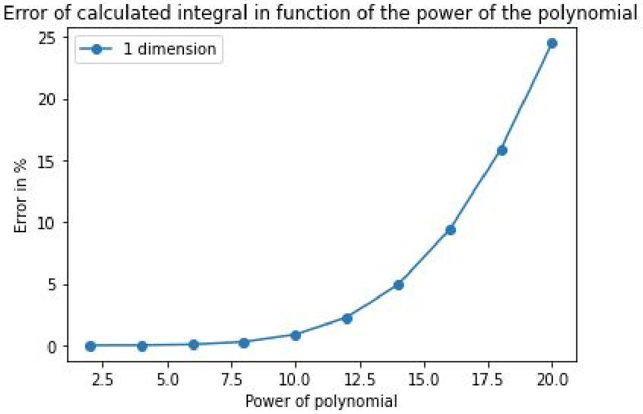
Indeed, the equation for the volume is very scary at first and I didn't understand what the volume meant in terms of synaptic weights. But I was amazed to understand how the volume's concept would help get a formula for the capacity. Every time we add an input output association, we add a constraint that can be thought of as being a plane. The volume between the planes is the solution. However, the more input output associations and the more planes are added and thus, the more constraints are accumulated. And at a certain point, the volume cannot be calculated because there are too many constraints which means that we have reached capacity.

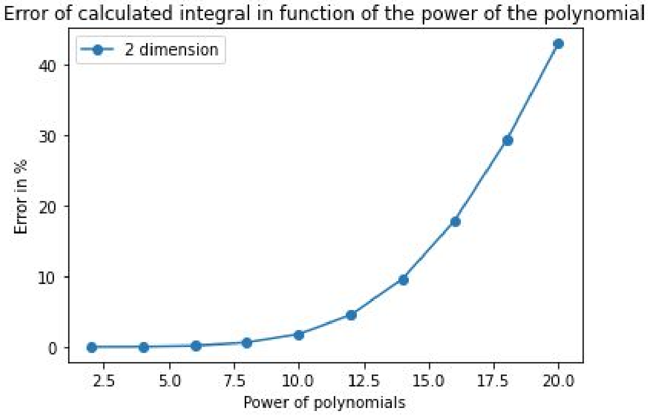
Now, in order to compute the volume (V^n), we had to introduce the concept of replicas. We introduce new replica symbols in order to facilitate the calculation of the volume, however, we pay a price with the correlation of solution between different copies of the weight space.

Similarly, pruning results in some movement in weight spaces. And the direction in which a synaptic weight goes varies depending on the copy. So we need additional order parameters in order to describe pruning in one copy of weight space compared to another.

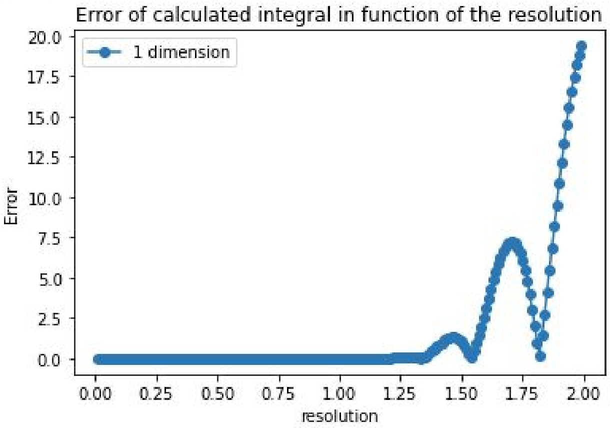
In the extended Gardner theory, we could not find one equation that computes the capacity. We came up with an explicit formula for α containing 3 equations that can only be solved by starting with a guess for the order parameter and adjusting our guesses iteratively. In order to solve these complex equations containing 2 and 3 dimensional Gaussian integrals, we started by writing some code that solves multi-dimensional Gaussian integrals. However, it was not easy at all because of the time it takes to compute 3 dimensional Gaussian integrals.

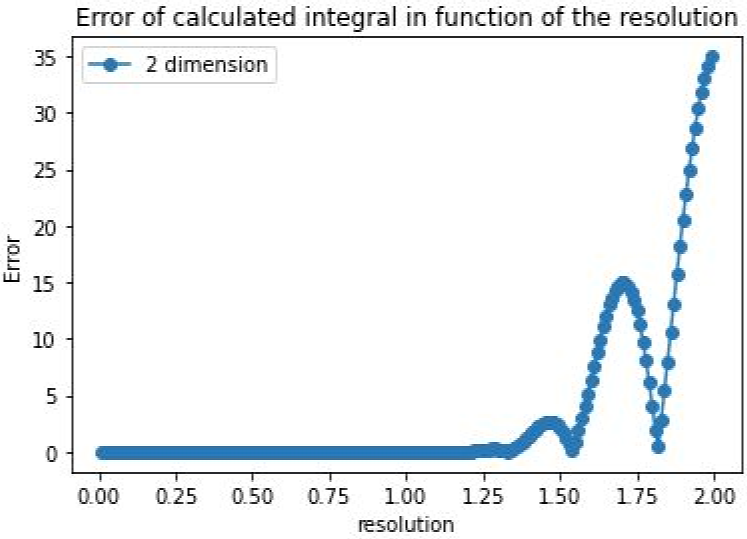
I also noticed that the higher the power of the polynomial function and the larger the resolution, the less accurate our computation becomes.





Note: The result for the 3 dimension gaussian integral is omitted because of the time it took to compute.





Note: The polynomials were x^4, x^4y^4, $x^4y^4z^4$ for Gaussian integrals of dimension 1, 2 and 3 respectively

From the graphs above, we understand that the higher the power of the polynomial and the larger the resolution and the less accurate our results for computing the Gaussian integrals. Indeed, we see that the error increases for both the graphs with the power of the polynomial and the resolution.

If we look at the equations that will solve the advanced Gardner theory below:

$$
\begin{aligned}
\hat{q} &= \frac{q}{(1-q)^2} \\
\frac{q^2}{1-q} &= \alpha \int Dt\mathcal{R}^2\left(\frac{\sqrt{q}}{\sqrt{1-q}}t\right) \\
\hat{p} &= x + \frac{\alpha}{\sqrt{(1-q)(1-r)}} \int D\omega \int Dt\mathcal{R}\left(\tilde{t}\right)\mathcal{R}\left(h\left(\omega,t,y=\tilde{t}\right)\right) \\
p &= \frac{\hat{p}\left(1+2\hat{q}\right)-\hat{q}\hat{p}'}{\left(1+\hat{r}\right)\left(1+\hat{q}\right)^2}
\end{aligned}
$$

$$
\begin{aligned}
\hat{p}' &= \frac{\alpha}{\sqrt{(1-q)(1-r)}} \int D\omega \int DtH^{-1}\left(\tilde{t}\right)\mathcal{R}\left(\tilde{t}\right)\int_{\tilde{t}}^{\infty} Dy\mathcal{R}\left(h\right) \\
p' &= \frac{\hat{p}'+\hat{q}\hat{p}}{\left(1+\hat{r}\right)\left(1+\hat{q}\right)^2} \\
\hat{r} &= \frac{1}{1-r}\int DtH^{-1}\left(\tilde{t}\right)\int D\omega \int_{\tilde{t}}^{\infty} Dy\mathcal{R}^2\left(h\right) \\
r &= 1+\frac{1}{\left(1+\hat{r}\right)^2}\left[1+2\hat{r}+\frac{\left(\hat{p}-\hat{p}'\right)\left(\hat{p}+\hat{p}'+2\hat{q}\hat{p}\right)}{\left(1+\hat{q}\right)^2}\right]
\end{aligned}
$$

we understand that q does not depend on other order parameters, thus, that the equation for q is standalone. It makes sense because additional order parameters are essential to describe the pruning but q is the correlation with the non-pruned original vector. So is there a way to find q as a function of α? The answer is yes because we know that as α tends to 2, q tends to 1. However, I do not really know how to solve this problem.