Symmetry breaking for two-layer neural networks in the teacher-student setting

Mattia Mariantoni with the supervision of professor Lénaïc Chizat

January 2022

1 Introduction

Artificial neural networks are parametric functions that process their input through a sequence of nonlinear maps called activation functions, they are based on a collection of connected units which are called neurons, they are loosely inspired from the neurons found in a biological brain. Given a predictive or generative task, the parameters are adjusted by minimizing a loss function, we will focus on the mean squared error loss, which is estimated on training data via the gradient descent algorithm. Since Rosenblatt [1958], experimental research has gradually improved the performance of neural networks, to the point that they are now a method of choice in many machine learning tasks [Bengio et al., 2017], these successes have triggered considerable interest among theoreticians [Anthony and Bartlett, 2009], but the mechanism by which a neural network discovers the structure of high dimensional data given a limited number of training points is still largely an open question.

In this project we focus on shallow neural networks, only two-layers deep, differently than the common setting in which models are trained on datasets with predefined labels, we focused on the teacher-student framework and assume a teacher network model, also known as oracle, to provide the labels for given instances. The main objective is to utilize numerical experiments and prove theoretical results that can provide some insights on the "symmetry breaking" phase that happens in the training for this setting, which follows a first phase in which all neurons tend to maximally correlate with the signal. In this work we study the problem using the framework of statistical mechanics, leveraging the concept of order parameters, similarly to previous work of Saad and Solla [1995].

1.1 Details

We consider the problem of learning a teacher shallow neural network of width $m \in \mathbb{N}^*$ by training a student shallow neural network of width m via the

gradient flow of the expected square loss. Letting $\phi : \mathbb{R} \to \mathbb{R}$ be the non-linearity/activation function (in our case it will be $\phi : s \mapsto 1(s > 0)$ and $\rho \in \mathcal{P}\left(\mathbb{R}^d\right)$ be the probability distribution on inputs, the objective function $F : \left(\mathbb{R}^d\right)^m \to \mathbb{R}$ is

$$F(v) := \frac{1}{2} \int_{\mathbb{R}^d} \left(\sum_{i=1}^{m} \phi\left(x^\top v_j\right) - \sum_{j=1}^{m} \phi\left(x^\top u_i\right) \right)^2 \mathrm{d}\rho(x) \tag{1}$$

where $u = (u_1, \ldots, u_m) \in \left(\mathbb{R}^d\right)^m$ are the weights/parameters of the teacher (which are given and fixed). The gradient flow starting from $v(0) = (v_1(0), \ldots, v_m(0))$ is the unique solution to $v'(t) = -\nabla F(v(t))$.

By expanding the square, we can rewrite the objective as

$$F(v) = \frac{1}{2} \sum_{i,i'} k\left(v_i, v_{i'}\right) + \frac{1}{2} \sum_{j,j'} k\left(u_j, u_{j'}\right) - \sum_{i,j} k\left(v_i, u_j\right) \tag{2}$$

where for $u, v \in \mathbb{R}^d$,

$$k(u, v) := \int_{\mathbb{R}^d} \phi(u, x)\phi(v, x)\mathrm{d}\rho(x) \tag{3}$$

Eq. (2) can be interpreted as an interaction energy with a (symmetric semidefinite) kernel k. Denoting by ∇k its gradient in the first variable, the gradient flow of F satisfies, for $i = 1, \ldots, m$,

$$v_i'(t) = \sum_j \nabla k\left(v_i(t), u_j\right) - \sum_j \nabla k\left(v_i(t), v_j(t)\right).$$

For the sake of theoretical tractability, we make the following assumptions:
(**H1**) The distribution ρ is spherically symmetric;
(**H2**) The non-linearity ϕ is the step function $\phi(s) = 1(s > 0)$;
Distributions satisfying (**H1**) include the Gaussian distribution or the uniform distribution on the sphere. Under (**H1**) and (**H2**), we have that $k(u, v) = \psi(\widehat{u, v})$ where $\psi(\theta) = \frac{1}{2\pi}(\pi - \arccos(\theta))$ (here $\widehat{u, v} := u^\top v / (\|u\|\|v\|)$) is the angle between u and v) (proof in subsection **1.2**).

For convenience in the analysis, we introduce the so-called order parameters:

$$\alpha_{i,j} = u_i^\top u_j, \quad \beta_{i,j}(t) = u_i^\top v_j(t), \quad \gamma_{i,j}(t) = v_i(t)^\top v_j(t) \tag{4}$$

where only α does not depend on t (convention $\phi'(1) \cdot 0 = 0$). The dynamics is closed in the order parameter (i.e. they are solutions to an autonomous ODE),

indeed:

$$\beta'_{i,j} = u_i^\top v'_j = \sum_{k=1}^{m^*} \psi'\left(\beta_{k,j}\right)\left(\alpha_{i,k} - \beta_{k,j}\beta_{i,j}\right) - \sum_{k=1}^{m} \psi'\left(\gamma_{k,j}\right)\left(\beta_{i,k} - \gamma_{k,j}\beta_{i,j}\right),$$
$$\gamma'_{i,j} = v_i^\top v'_j + v_j^\top v'_i,$$
$$v_i^\top v'_j = \sum_{k=1}^{m^*} \psi'\left(\beta_{k,j}\right)\left(\beta_{k,i} - \beta_{k,j}\gamma_{i,j}\right) - \sum_{k=1}^{m} \psi'\left(\gamma_{k,j}\right)\left(\gamma_{i,k} - \gamma_{k,j}\gamma_{i,j}\right).$$

1.2 Angle formula

Assuming that the distribution ρ is symmetric and that $\phi(x) = 1(x > 0)$ with $u, v \in R^d$, our goal is to prove that

$$\int_{R^d} \phi(u^T x)\phi(v^T x) d\rho(x)$$

is equal to a function $\psi(t) = \frac{\pi - \arccos(t)}{2\pi}$ where $t = \frac{u^T v}{||u||||v||}$ which depends implicitly only on the angle between u, v.

Since $\phi(x)$ is zero homogeneous it's possible to reduce the original domain of integration to the sphere

$$\int_{S^{d-1}} \phi(u^T x)\phi(v^T x) d\rho(x)$$

substituting the value for the function $\phi(x)$

$$\int_{S^{d-1}} 1(u^T x > 0)1(v^T x > 0) d\rho(x)$$

which is equivalent to

$$\int_{S^{d-1}} 1(\frac{u^T x}{||u||||x||} > 0)1(\frac{v^T x}{||v||||x||} > 0) d\rho(x)$$

denoting the angle between vectors as $\widehat{u, x} = \arccos(\frac{u^T x}{||u||||x||})$

$$\int_{S^{d-1}} 1(\widehat{u, x} \in [-\frac{\pi}{2}, \frac{\pi}{2}])1(\widehat{v, x} \in [-\frac{\pi}{2}, \frac{\pi}{2}]) d\rho(x)$$

Now we consider the plane generated from the vectors u, v then we can apply a change of parametrisation for x in this u, v plane and the remaining dimensions, we note also that the integrand function's value is independent on those other dimensions. The integral is proportional to the angle $\theta = \widehat{u, v}$ between u, v in the 2 dimensional plane generated from those vectors, this proportionality can be expressed as a linear function of θ and given that ρ is spherically symmetric we have

$$\psi(\theta) = \frac{\pi - \arccos(\theta)}{2\pi}$$

2 Numerical experiments

The following hyper parameters are taken as input of the function: dimension of the input d, learning rate μ, number of iterations N, number of hidden neurons m. Two types of initialization have been tested, one randomly initialize the student weights $v \in \mathbb{R}^{d\times m}$ which will change and teacher weights $u \in \mathbb{R}^{d\times m}$ which will be fixed, each component of this two matrices is a sample from a standard normal random variable. The other case is the "orthogonal dynamics" one, in which the initialization matrix is an orthonormal matrix.

Given the following equations: $\theta(u,v) = \frac{u^T v}{||u||||v||}$ and $k(u,v) = \psi(\theta)$ where $\psi(\theta) = \frac{\pi - \arccos(\theta)}{2\pi}$ and $\psi'(\theta) = \frac{1}{2\pi\sqrt{1-\theta^2}}$ and $k'(u,v) = \psi'(\theta(u,v))\frac{v - u^T v \frac{u}{||u||^2}}{||u||||v||}$

The loss function for the current student and teacher weights is

$$F(v) = \frac{1}{2}\sum_{i,i'} k\left(v_i, v_{i'}\right) + \frac{1}{2}\sum_{j,j'} k\left(u_j, u_{j'}\right) - \sum_{i,j} k\left(v_i, u_j\right)$$

The gradient for the student's weights is

$$v_i'(t) = \sum_j k'\left(v_i(t), u_j\right) - \sum_j k'\left(v_i(t), v_j(t)\right)$$

and the update for the student's weights is

$$v_i(t+1) = v_i(t) - \mu v_i'(t)$$

2.1 Experiment's results

The following graphs have the expected loss reported on the y-axis and the iteration number on the x-axis, both having with a logarithmic scale.
Each graph represents a certain hyperparameters combination noted on the title of the graph, with the blue color are represented experiments with orthogonal initialization, all of them converge to a sub optimal point. Meanwhile with the yellow color are reported the experiments with the normal random variable initialization, depending on the hyperparameters we have convergence to a sub optimal point or optimal convergence, up to machine error on the order of 10^{-9}.

Code is available on github: Repository Link

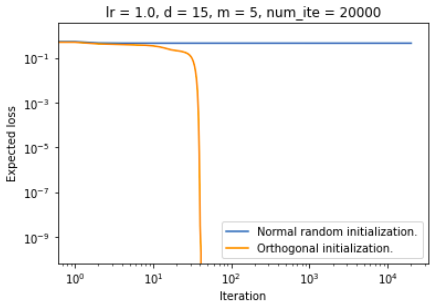

Figure 1: Sub optimal for orthogonal initialization, optimal for normal random variable initialization.

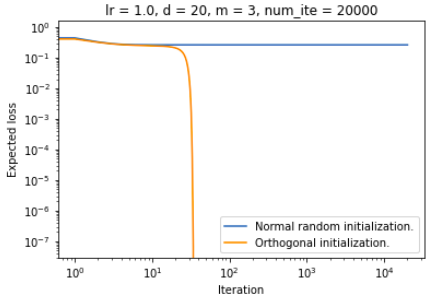

Figure 2: Sub optimal for orthogonal initialization, optimal for normal random variable initialization.

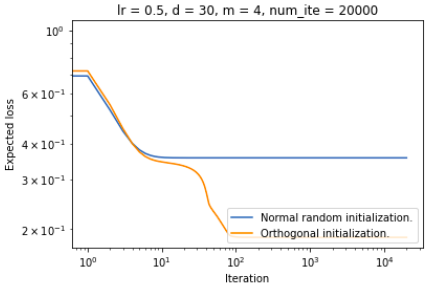

Figure 3: Sub optimal for orthogonal initialization, sub optimal for normal random variable initialization.

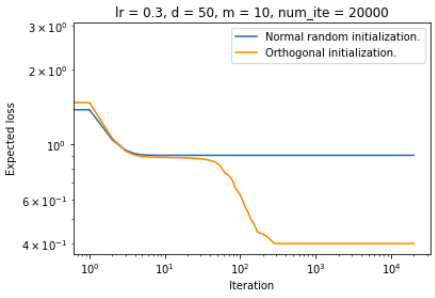

Figure 4: Sub optimal for orthogonal initialization, sub optimal for normal random variable initialization.

3 Theoretical results

The matrix $\theta(t) \in \mathbb{R}^{2m \times 2m}$ represents the teacher and student weights at iteration t with a standard normal random variable initialization, meanwhile $\theta_O(t) \in \mathbb{R}^{d \times 2m}$ represents the teacher and student weights at iteration t with orthogonal initialization.

P is defined such that is equal to $argmin_P ||\theta(0) - P\theta_O(0)||_F^2$ where P needs to be have orthonormal columns, then given that $\theta_O(0)$ is the idenitity matrix of dimension $2m \times 2m$ the problem reduces to $P = argmin_P ||\theta(0) - P||_F^2$.

Given that the matrices are reals, we have the following equivalence which follows from the definition of the Frobenius norm

$$\|\theta(0) - P\|_{\mathrm{F}}^2 = \mathrm{trace}((\theta(0) - P)^T(\theta(0) - P)) = \mathrm{trace}(\theta(0)^T\theta(0) - P^T\theta(0) - \theta(0)^T P + I)$$

therefore the optimization problem reduces to $P = argmax(\mathrm{trace}(P^T\theta(0)))$
Given that the singular value decomposition always exists, we can rewrite
$\theta(0) = USV^T$ and define $A = \sqrt{S}$
Then

$$\mathrm{Tr}(P\theta(0)) = \mathrm{Tr}\left(PUA^2V^T\right) = \mathrm{Tr}\left((PUA)(VA)^T\right) = \langle PUA, VA\rangle$$

By the Cauchy-Schwarz inequality, we get

$$\mathrm{Tr}(P\theta(0)) \leq \|PUA\|_F \|VA\|_F = \|A\|_F \|A\|_F = \mathrm{Tr}\left(AA^T\right) = \mathrm{Tr}(S)$$

where we used the invariance of the $\|\cdot\|_F$ under orthogonal transformations. Therefore the maximum is attained taking $P = VU^T$.

3.1 Lemma 1

Assuming $||\theta(0) - P\theta_O(0)||_F < \epsilon$ then $||\theta(t) - P\theta_O(t)||_F < f(t, \epsilon)$ where f is an exponentially decaying function.

Proof. The elements of the matrix $\theta(t)$ which correspond to the teacher weights will not vary depending on t, focusing on one student weight we get $(\theta(0) - P\theta_O(0))_{i,j} < \epsilon' \leq \epsilon$ from our assumption on initial condition and the definition of the frobenius norm.
The aim would be to show that $(\theta(t) - P\theta_O(t))_{i,j} < g(t, \epsilon)$ for any element of the matrix, then we could use the definition of the Frobenius norm

$$||\theta||_F = \sqrt{\sum_{i=1}^{2m}\sum_{j=1}^{2m}\theta_{i,j}^2}$$

to claim that $f(t, \epsilon) = 4m^2 g(t, \epsilon)^2$.

If x and y are two different solutions of some ODE $y' = f(t, y)$ with initial conditions $y(0) = a$ and $x(0) = b$, then the difference in the Picard integral equation is

$$x(t) - y(t) = b - a + \int_0^t (f(s, x(s)) - f(s, y(s)))\, ds$$

With Lipschitz continuity there is some constant $L > 0$ such that:

$$\|f(s, x) - f(s, y)\| \leq L\|x - y\|$$

Using that, the difference of two solutions y, x of the ODE satisfies the following integral inequality as a consequence of the Picard integral equation

$$\|x(t) - y(t)\| \leq \|x(0) - y(0)\| + L \int_0^t \|x(s) - y(s)\|\, ds$$

By the Gronwall lemma this implies

$$\|x(t) - y(t)\| \leq \|x(0) - y(0)\| e^{L|t|}$$

Given the Lipschitz constant L we can apply the above and obtain:

$$g(t, \epsilon) = e^{Lt}\epsilon$$

and therefore:

$$f(t, \epsilon) = 4m^2\epsilon^2 e^{2Lt}$$

A tighter bound is obtainable by considering directly the matrix form of the problem.
If we are able to prove that $G : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz then:

$$\begin{cases} X'(t) = G(X(t)) \\ \tilde{X}'(t) = G(\tilde{X}(t)) \end{cases} \tag{5}$$

Now we can define $h(t) = ||X(t) - \tilde{X}(t)||^2$ and thanks to the L-Lipschitz of G obtain that:

$$\frac{d}{dt}\frac{h(t)}{2} = (G(X(t)) - G(\tilde{X}(t)))^T (X(t) - \tilde{X}(t)) \leq L||X(t) - \tilde{X}(t)||^2$$

and therefore:

$$h'(t) \leq 2Lh(t)$$

$$\frac{h'(t)}{h(t)} = (\log h(t))' \leq 2L$$

then integrating and considering the initial condition $h(0) = ||X(0) - \tilde{X}(0)||^2$ we obtain:

$$h(t) \leq h(0) \exp 2Lt$$

□

3.2 Lemma 2

$\lim_{d\to\infty} min_P ||\theta(0) - P\theta_O(0)||_F \to 0$

Proof. As shown previously we have that:

$$\|\theta(0) - P\|_{\mathrm{F}} = \mathrm{trace}(\theta(0)^T\theta(0) - P^T\theta(0) - \theta(0)^T P + I)$$

since the trace operator is linear and $\mathrm{trace}(A) = \mathrm{trace}(A^T)$

$$\mathrm{trace}(\theta(0)^T\theta(0)) - 2\,\mathrm{trace}(P^T\theta(0)) + \mathrm{trace}(I)$$

Using that the trace is invariant under cyclic permutations and $\theta(0) = USV^T$

$$\mathrm{trace}(S^2) - 2\,\mathrm{trace}(S) + 2m$$

now using that $S = diag[(s_i)_{i=1,\dots,2m}]$ the expression above becomes:

$$\sum_{i=1}^{2m}(s_i^2 - 2s_i - 1) = \sum_{i=1}^{2m}(s_i - 1)^2 \tag{6}$$

Given the bounds on singular values of matrix with elements from a standard normal random variable:

Theorem 1 (Feng Wei, 2018)**.** *Let A be an $2m \times d$ matrix whose entries are independent standard normal random variables. Then for every $t \geq 0$, with probability at least $1 - 2\exp\left(t^2/2\right)$ one has:*

$$1 - \frac{\sqrt{2m}}{\sqrt{d}} - \frac{t}{\sqrt{d}} \leq s_{\min}(A) \leq s_{\max}(A) \leq 1 + \frac{\sqrt{2m}}{\sqrt{d}} + \frac{t}{\sqrt{d}}$$

the theorem above implies that the expression 6 for d that grows to infinity becomes:

$$\sum_{i=1}^{2m}(s_i - 1)^2 = \sum_{i=1}^{2m}|\sqrt{\frac{2m}{d}} + \frac{t}{2d}|^2 = 2m|\sqrt{\frac{2m}{d}} + \frac{t}{2d}|^2 \to 0$$

□

4 References

Martin Anthony and Peter L. Bartlett. *Neural network learning: Theoretical foundations.* Cambridge University Press, 2009.

Yoshua Bengio, Ian Goodfellow, and Aaron Courville. *Deep learning*, volume 1. MIT press Massachusetts, USA:, 2017.

Andreas Engel and Christian Van den Broeck. *Statistical mechanics of learning.* Cambridge University Press, 2001.

Sebastian Goldt, Madhu S. Advani, Andrew M. Saxe, Florent Krzakala, and Lenka Zdeborová. *Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup.* Journal of Statistical Mechanics: Theory and Experiment, 2020(12) : 124010, 2020.

Frank Rosenblatt. *The perceptron: a probabilistic model for information storage and organization in the brain.* Psychological Review, 65(6) : 386, 1958.

David Saad and Sara A. Solla. *On-line learning in soft committee machines.* Physical Review E, 52(4) : 4225, 1995.

Yuandong Tian. *Symmetry-breaking convergence analysis of certain two-layered neural networks with ReLU nonlinearity.* 2017.

Feng Wei. *Measure Concentration and Non-asymptotic Singular Values Distributions of Random Matrices.* 2018