Rapport de stage

Bertrand Even

16th August 2022

1 Expériences de voyage

J'ai effectué un stage de recherche sous la supervision de Sebastian Goldt, à la SISSA (Trieste, Italie), dans le département de Data Science. Spatialement parlant, cela était fort agréable. En effet, cette ville donne sur un golfe de la mer Adriatique, sur lequel les fenêtres des bureaux donnaient. Je logeai dans le centre-ville, à côté d'un port, tandis que la SISSA se situe en dehors de la ville, dans des hauteurs qui surplombent la mer. Bien heureusement, un bus permet d'accéder facilement au site, moyennant un trajet de 20 minutes.

Je travaillai dans un bureau avec des étudiants en thèse et un stagiaire, ce qui était plutôt pratique au niveau social. Nous mangions dans le jardin tous les midis avec aussi les professeurs du département de Data Science, et nous enchainions bien souvent sur des parties de Tennis de Table endiablées. Je dois aussi avouer que j'ai porté haut les couleurs de l'ENS en remportant le tournoi de Tennis de Table de la SISSA sans perdre un seul set.

Quant au déroulement du stage, Sebastian m'a donné un sujet d'étude (développé dans ce rapport) ainsi que des articles à lire pour se donner de l'inspiration. Puis, nous nous retrouvions régulièrement pour parler des avancées et chercher les directions qui peuvent s'avérer intéressantes. A la fin du stage, j'ai pu présenter les résultats lors d'une Poster Session à une conférence à Trieste "Youth in High Dimension."

Ce stage m'a permis de découvrir le monde de la recherche, tant dans le cadre humain que dans le mode de réflexion.

Abstract

We analyse the implicit bias of feedback alignment in simple learning problems and compare them to the one of Gradient Descent, which have already been studied.

2 Introduction

Describe FA [1], DFA [2] : To learn neural networks, the most used method is Back-propagation, which uses the negative gradient of each layer to make updates on the weights. For Back-propagation, the layers are spread backward in order to calculate the gradients. This method achieves great performances. But, there are alternatives to this method. Instead of spreading the layers backwards, the method of Feedback Alignment (FA) uses Feedback matrices F_l to spread the updates backwards. Direct Feedback Alignment (DFA) directly calculates the update of a layer with the gradient of the output and the fixed Feedback matrix.

Consider a network with weights $w_l, l \in [1, L]$, a loss $\mathcal{L}(\hat{y})$ with $\hat{y}$ the output of the network. Denote by $a_l(t) = w_l^T h_{l-1}(t) + b_l$ and $h_l(t) = g(a_l(t))$ with $h_0 = x$ and $\hat{y} = w_L^T h_{L-1} + b_L$. Then, the update rule is $\delta w_l = -\eta h_{l-1} \delta a_l^T$ with:

1. Back-propagation: $\delta a_l = \frac{\partial \mathcal{L}}{\partial a_l} = [w_{l+1}^T \delta a_{l+1}] \odot g'(a_l)$ and $\delta a_L = \frac{\partial \mathcal{L}}{\partial \hat{y}}$
2. Feedback Alignment: $\delta a_L = \frac{\partial \mathcal{L}}{\partial \hat{y}}$ and $\delta a_l = (F_l \delta a_{l+1}) \odot g'(a_l)$

3. Direct Feedback alignment: $\delta a_L = \frac{\partial \mathcal{L}}{\partial \hat{y}}$ and $\delta a_l = (F_l \delta a_L) \odot g'(a_l)$.

These two alternative methods are more realistic when modelling the brain since we do not need to spread the layers in two directions.

Implicit bias Modern neural networks are strongly over-parameterised, which implies that the number of solutions to minimise a loss can be infinite. These different solutions of optimisation for a train dataset have different performances when looking at the generalisation. The algorithm of gradient descent tends to solutions that generalise well.

The goal of looking at the implicit bias of an algorithm and an architecture is to characterise which minima of the loss is selected. Studies on the implicit bias for gradient descent has been done previously. It has been shown [3] [3]that for fully-connected linear networks trained by gradient descent for a binary classification problem with an exponential tail, the solution selected is the direction of the one of norm one that maximises the margin (or equivalently the one of margin one that minimises the L_2 norm.) This classifier is called the max-margin classifier.

In the same article they also show that when using a convolutional architecture, the solution selected is different: it has the direction of the vector of margin one that the $L_{\frac{2}{L}}$ in the Fourier space, where L is the depth of the network.

More generally, for homogeneous networks, the solution selected to classify a dataset is the one that minimises the L_2 norm of the parameters [LyuLi and JiTelgarsky][4] [5].

We will try to compare this results with results on Direct Feedback alignment to understand if there is a theoretical difference on the solutions selected by these two different algorithms.

Our **main contributions** can be summarised as follows:

1. We show that for fully connected linear networks, Direct Feedback Alignment aligns the weights in such a way that the effective vector maximises the margin, as for back-propagation.

2. For convolutional neural networks, we show that the direction of the classifier is different than with back-propagation, as it minimises a different norm.

3. We finally show that for homogeneous networks, Direct Feedback alignment aligns the output to a direction which maximises the margin with respect to the hidden layers.

3 Definitions

We will study the implicit bias of Direct Feedback alignment on homogeneous architectures with an exponential loss.

Exponential loss: Given a dataset (x_n, y_n) with $x_n \in \mathrm{R}^D$ and $y_n \in [-1, 1]$, and a classification function f, the exponential loss is $\mathcal{L}(f) = \sum_n exp(-y_n f(x_n))$.

Margin of a homogeneous function. For a set of function ϕ_w homogeneous in w, ie there exists L such that for $\alpha > 0, \phi_{\alpha w}(x) = \alpha^L \phi w(x)$, and a dataset x_n, y_n, the margin of a parameter vector w is $min_n y_n \phi_w(x_n)$.

The margin is thus also homogeneous in w.

4 Linear networks, linearly separable data

We study the implicit bias of fully connected linear networks learned with Direct Feedback alignment.

Data model Describe linearly separable data model, define max-margin classifier.
We consider the case where we have a dataset $\{(x_n, y_n)\}_{[1,N]}$ where $x_n \in \mathrm{R}^D$ and y_n is a label in $\{-1, 1\}$, which is **linearly separable**, ie there exists a vector β such that $\forall n, y_n \langle x_n, \beta \rangle > 0$.
Amongst all classifiers with positive marge-in, the max marge-in classifier $\widetilde{\beta}$ is the one which direction maximises the marge-in. It is caracterized by $\hat{\beta} = argmin \|\beta\|$ with $\forall n, y_n \langle x_n, \beta \rangle \geq 1$.

KKT condition for the max-marge-in classifier $\hat{\beta}$ is the vector with marge-in equals to one that verifies:$\exists \alpha_n \geq 0, n \in \mathcal{S}$ such that $\hat{\beta} = \sum_{n \in \mathcal{S}} \alpha_n y_n x_n$ where $\mathcal{S} := \{n, x_n^T \hat{\beta} = 1\}$

4.1 Fully connected networks trained with DFA

Architecture Let us consider linear networks of depths L and w_l matrices that represent the l-th layer. For $l \in [1, L-1]$, $w_l \in \mathrm{R}^{D \times D}$ and $w_L \in \mathrm{R}^D$. Then, the network acts as $f_w(x) = (\prod_{l=1}^{L} w_l)^T x = \beta^T x$ with $\beta = \prod_{l=1}^{L} w_l$.

Algorithm Describe loss, finite step-size SGD, gradient flow.
We will use an exponential loss in order to find a classifier. $\mathcal{L}(w) = \sum_{n=1}^{N} exp(-y_n \beta^T x_n)$.
Let us denote $z(w) = -\nabla_\beta \mathcal{L} = \sum_{n=1}^{N} y_n exp(-y_n \beta^T x_n) x_n$. Then, $\forall l, \nabla_{w_l} \mathcal{L}(w) = -w_{1:l-1}^T z(t) w_{l+1:L}^T$
For DFA with a stepsize η, we initialize the weights at zero (the results would be the same with random initial weights but we suppose that to lighten the calculus) and we make the updates $\triangle w_l = \eta w_{1:l-1}^T z(t) b_l^T$ with b_l the feedback matrices taken randomly and then fixed during the all proceeding of the algorithm.
We also consider the flow corresponding to the case where η goes to zero. We study the behaviour of $\widetilde{w}(t)$ the solution of the differential equation $\frac{d}{dt} \widetilde{w}_l(t) = \widetilde{w_{1:l-1}}(t)^T \, z(t) b_l^T$

Results for backpropagation: [Gunasekar] Linear networks trained with gradient descent, when converging in direction, converge to a rescaling of the max-margin classifier.

4.1.1 Analysis of the continuous case.

We first study the continuous flow and then we will be able to generalise them to the discrete case when the stepsize is small enough.

In order to learn, we must have that the derivative of the weights is in a descent direction. That will be ensured by a alignment with the feedback matrices.

Proposition 1:*The layers $w_l, l \in [2, L-1]$ get aligned to $b_{l-1} b_l^T$ and w_L to b_{L-1}. Thus, $\mathcal{L}(w(t))$ is decreasing and tends to zero. Every layer has its norm going to infinity.*

Once, we now that the algorithm learns and achieves zero misclassification, we can study amongst all the directions that are perfect classifiers, which one it tends to.

Theorem 1:*The first layer also converges in direction and the direction of $\beta = w_1 ... w_L$ converges to a rescaling of the max-margin classifier.*

We will prove this in the annexe section. Let us just give us an intuition of why the classifier is biased towards the max-margin classifier. We have that $\beta(t)$ is aligned to $\int_0^t z(t')$ and then if its direction converges to a $\overline{\beta}$ it must converge to something that is in the cone created by $\{x_n y_n, n \in \mathcal{S}\}$ where $\mathcal{S} = argmin \overline{\beta}^{y_n x_n}$. That is exactly the KKT condition of the max-margin classifier.

Hence, the bias induced by Direct Feedback Alignment on fully connected networks is the same as for Gradient descent, at least for the continuous case. Now let us generalize this result to the discrete case.

4.1.2 Analysis of the discrete case

We will use the part on the continuous case in order to study the implicit bias on the finite-step algorithm.

Theorem 1[Discrete version]*For a finite $\eta > 0$ small enough, we have that the loss $\mathcal{L}(w(t))$ tends to zero. Moreover, we have that each layer converges in direction and that β converges in direction to a rescaling of the max-margin classifier.*

The proof, in the annex, follows this plan:

1. Proving that if the at one point, we have a positive margin for $\frac{\beta}{\|\beta\|}$ with the layers having a large enough norm, there exist $\delta > 0$ so that the margin of $\frac{\beta}{\|\beta\|}$ stays larger than δ.
2. Proving that if the margin stays larger than a δ, the loss goes to zero.
3. If the loss goes to zero, then β converges in direction to the max-margin classifier.

4.2 Convolutional neural network trained with DFA

Notation For a vector w in R^D, we will note by $\hat{w}$ its discrete Fourier transform given by $\hat{w} = \mathcal{F}w$ where $\mathcal{F}[d,p] = \frac{1}{\sqrt{d}} exp(\frac{2i\pi pd}{D})$

Properties of the Fourier matrix:

Architecture We consider convolutional network architectures where each non-output layer has exactly D units and the transformation from the layer l-1 to the layer l is a circular convolutional transformation parametrized by the vector $w_l \in \mathrm{R}^D$ as:

$$h_l(d) = \frac{1}{\sqrt{D}} \sum_{k=0}^{D-1} w_l(k) h_{l-1}(d + k[modD]) = (h_{l-1} \star w_l)(d)$$

The output layer is fully connected and parametrized by $w_L \in \mathrm{D}$. The network acts as:

$$f_w(x) = ((((x \star w_1) \star w_2)...) \star w_{L-1})^T w_L$$

The network can be represented by a vector $\beta(w)$ which is the unique vector that verifies that:$\forall x, f_w(x) = \langle \beta, x \rangle$. In this case we have that: $\beta(w) = (((w_L^{\downarrow} \star w_{L-1})...) \star w_1)^{\downarrow}$ where $w^{\downarrow}[d] = w[D-1-d]$ for $k \in [0, D-1]$.

Equivalence with a diagonal architecture This architecture is equivalent to having $f_w(x) = \hat{w}_L^* \odot \hat{w_{L-1}}^* \odot ... \odot \hat{w^1} \odot \hat{x}$ which is represented by $\hat{beta} = \mathcal{F}\beta = \hat{w_1} \odot ... \odot \hat{w_L}$. Moreover, we also have that $\nabla_{\hat{w}_l} \mathcal{L} = \mathcal{F} \nabla_{w_l} \mathcal{L}$

Algorithm Still with the exponential loss, we will study the behaviour of direct feedback alignment on this architecture. With $z(w) = \sum y_n exp(-y_n \beta^T x_n) x_n$, we have that $\nabla_{\hat{w}_l} \mathcal{L} = \hat{w}_1^* \odot \hat{w}_2^* \odot ... \odot \hat{w}_{l-1}^* \odot \hat{w}_{l+1}^* \odot ... \hat{w}_L^*$. Thus, for direct feedback alignment with finite stepsize η, we initialize the weights at 0 and we take $\forall l < L, \triangle \hat{w}_l = \eta \hat{w}_1^* \odot ... \odot \hat{w}_{l-1}^* \odot b_l^* \odot \hat{z}(t)$ and $\triangle \hat{w}_L = \eta \hat{w}_1^* \odot ... \odot \hat{w}_{L-1}^*$.

Results on Gradient Descent [Gunasekar et al] For gradient Descent on a covolutional linear network, it has been shown that when converging in direction the effective vector β converges in direction to the vector of margin one that minimises the $L_{\frac{2}{L}}$ norm in the Fourier space.

We will now show that the implicit bias for Direct Feedback Alignment is different.

4.2.1 Analysis of the continuous flow

Here, we analyse the continuous flow of direct feedback alignment and characterise its implicit bias.

Proposition:*Each layer except the first one, has its coefficient positively proportional to $b_{l-1} \hat{\odot} \hat{b}_l^\star$. Hence, the loss is decreasing and tends to zero.*

Theorem 2:*If we suppose that $(\frac{w_1}{\|w_1\|}, ..., \frac{w_L}{\|w_L\|})$ does not have an adherence value of margin 0, all the layers converge in direction and the direction of the effective vector at infinity is a rescaling of the vector of margin one which minimizes in the Fourier space:*

$$\psi(\hat{\beta}) = \left(\sum_{d=0}^{D-1} \frac{1}{\prod_{l=1}^{L-1} |\hat{b}_{L-l}[d]^{2^l}|^{\frac{1}{2^{L-1}}}} |\hat{\beta}[d]|^{\frac{1}{2^{L-1}}+1} \right)^{\frac{1}{\frac{1}{2^{L-1}}+1}}$$

.

4.2.2 Analysis of the finite step algorithm

Now, we study the discrete case.

Theorem 2:[Discrete case] *If we suppose that $\mathcal{L}(w)$ goes to zero, that all w_l converge in direction such that the direction of β converge to a vector $\overline{\beta}^\infty$ with positive marge-in, then we have that this direction is a rescaling of the vector of margin 1 which minimizes $\psi(\hat{\beta})$.*

We have a difference between the bias induced by direct feedback alignment and the bias induced by gradient descent. Indeed, the norm that is minimised in the Fourier space is not the same: it is sparsier for gradient descent. Here, we have a norm which is a weighted version of a L_α norm with $\alpha = \frac{1}{2^{L-1}+1}$ which tends to 1 with the depth of the network, whereas $\alpha = \frac{2}{L}$ for back-propagation.

5 Homogeneous networks.

Data model We consider the case where we have a dataset $\{(x_n, y_n)\}_{1,N}$ where $x_n \in^D$ and y_n is a label in $\{-1, 1\}$. The dataset is no longer supposed to be linearly separable.

5.1 Direction of the output layer in a homogeneous network.

Architecture We study a network composed of one homogeneous hidden part and a fully connected output. We denote by w the parameters of the hidden part and w_{out} the output layer. Then, we have that the network is of the form $f_{w,w_{out}}(w) = w_{out}^T \phi(w, x)$ with ϕ which verifies that for all w, x, α,$\alpha > 0$, $\phi(\alpha w, x) = \alpha^L \phi(w, x)$ for some L.

Theorem 3:[Continuous case]*Let $w(t), w_{out}(t), t \in \mathrm{R}_+$ such that:*

1. *$\mathcal{L}(w(t))$ tends to zero.*
2. *w converges in direction to a direction $\overline{w}$ such that $\forall n, \phi(\overline{w}, x_n) \neq 0$*
3. *$\frac{d}{dt} w_o ut \sim \nabla_{w_{out}} \mathcal{L}$*

Then, w_{out} also converges in direction and if its norm goes to infinity that direction is the max-margin classifier of the dataset $(\phi(\overline{w}, x_n), y_n)$.

Theorem 3[Discrete case]*Let $w(t), w_{out}(t), t \in \mathrm{N}$ such that:*

1. *$\mathcal{L}(w(t))$ tends to zero.*
2. *w converges in direction to a direction $\overline{w}$ such that $\forall n, \phi(\overline{w}, x_n) \neq 0$*
3. *$w_o ut(t+1) - w_o ut(t) \sim \eta \nabla_{w_{out}} \mathcal{L}$ for some $\eta > 0$.*

Then, w_{out} also converges in direction and if its norm goes to infinity that direction is the max-margin classifier of the dataset $(\phi(\overline{w}, x_n), y_n)$.

5.2 Analysis of Direct feedback alignment on 2-layer ReLU networks.

Architecture Let us consider a Relu network of depth 2 and w_l be the matrix that represents the l-th layer, with $w_1 \in \mathrm{R}^{D \times d_1}$ and $w_2 \in \mathrm{R}^{d_1}$, with d_1 the number of neurones in the first layer. Then, the network is represented by the function $f_w(x) = w_2^T g(w_1^T x$, with g being the ReLU function.

Algorithm To fit the dataset, we still use the exponential loss $\mathcal{L}(w) = \sum_{[1,N]} exp(-y_n f_w(x_n) = \sum \alpha_n(t)$. Then we have that $\nabla_{w_1} \mathcal{L}(w(t)) = -\sum \alpha_n(t) y_n x_n [w_2 \odot g'(w_1^T x_n)]^T$ and $\nabla_{w_2} \mathcal{L}(w) = -\sum \alpha_n(t) y_n g(w_1^T x_n)$. We study the flow corresponding to direct feedback alignment, which corresponds to having $\frac{d}{dt} w_1(t) = \sum \alpha_n(t) y_n x_n [b \odot g'(w_1^T x_n)]^T$ and $\frac{d}{dt} w_2(t) = \sum \alpha_n(t) y_n g(w_1^T x_n)$, with b being the feedback matrix, of the same dimension as w_2, and taking $w_2(0) = 0$ and $w_1 = 0$ (and taking $g'(0) = 1$ so that it does not stay to zero.

Proposition: *The coefficients of w_2 get of the same sign as the one of the feedback matrix b. That allows the parameters to learn: we have that $\mathcal{L}(w(t))$ decreases.*

Yet,that does not ensure that the loss goes to zero. We will now study the directional convergence of the parameters in the case where the flow ensures that the loss goes to zero.

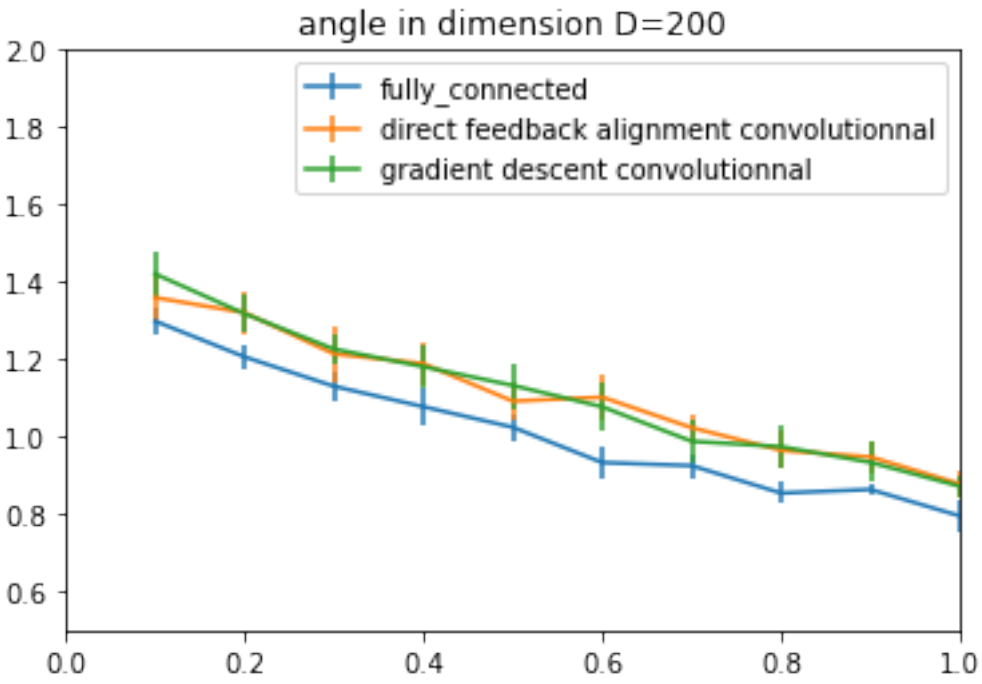
Theorem 4[Directional convergence for ReLU networks]*Let us suppose that the loss goes to zero and that $(\frac{w_1}{\|w_1\|}, \frac{w_2}{\|w_2\|}$ does not have an adherence value that has 0 as margin. Then, w_1 and w_2 both converge in direction and the direction $\overline{w_2}^\infty$ is the max-margin classifier of the dataset$(g(\overline{w_1}^\infty)^T x_n, y_n)$.*

Yet, we do not succeed in characterising the first layer as it is done with back-propagation where the couple (w_1, w_2) converge in direction to a KKT point of the optimisation problem $min \|(w_1, w_2)\|^2$ with the constraint of having the margin equal to 1. Here, the fact that the derivative of the first layer is not proportional to the gradient of the loss prevent us from using KKT conditions to characterise it. Hence, we are only able to characterise the direction of the second layer with respect to the direction of the first layer. Plus, the fact that the feedback matrix appears in the expression of its derivative make us believe that the direction can depend on this matrix, as it is the case for convolutional linear networks.

6 From optimisation to generalisation

We have differences on the norms minimised when changing either the architecture of the network or the algorithm. We would like to see if that has consequences on the generalization error.

We now generate gaussian linearly separable data in dimension 200 with a teacher $w_{teacher}$ taken randomly and we plot the average of the angle (which is proportionnal to the generalisation error) as a function of n/D. We find that the angle is slightly lower for the Fully connected networks.



But, can we have theoretical results that link this difference of the norm optimized and the generalisation error?

A Details on the proofs

A.1 Proof of proposition 1

Let us prove that there exists $\phi_l(t) > 0$ such that each layer is of the form $w_l(t) = \phi_l(t) b_{l-1} b_l^T$. We have that $w_1(t) = \int_0^t z(t') b_1^T$ and so $w_2(t) = \int_{t'<t''<t} z(t')^T z(t'') b_1 b_2^T = \frac{1}{2} \| \int_0^t z(t') \|^2 b_1 b_2^T$ and so $\phi_1(t) = \frac{1}{2} \| \int_0^t z(t') \|^2 > 0$ since the dataset is linearly separable.

Let us suppose that there exists $\phi_1, ..., \phi_l(t) > 0$ such that $w_l(t) = \phi_l(t) b_{l-1} b_l^T$.

Then, $w_{l+1}(t) = \int_0^t \phi_1(t') ... \phi_{l-1}(t') \phi_l(t') dt' \|b_1\|^2 ... \|b_{l-1}\|^2 b_l b_{l+1}^T$. Thus we have $\phi_{l+1}(t) = \int_0^t \phi_1(t') ... \phi_{l-1}(t') \phi_l(t$
$\int_0^t \phi_l(t') \frac{d}{dt'} \phi_l(t') dt' > 0$. That implies the alignment of all the layers.

Then, we have that the derivative of each layer is aligned to the negative of the gradient. So the loss decreases. Thus, $\mathcal{L}(w(t))$ converges to a positive number l. But that aloso implies that its derivative is integrable and so $\|z(t)\|$ is integrable, as all the layers have there norm that tends to infinity. $\mathcal{L}$ is then in L_2 and then the limit is null.

Proving that all the layers have there norm that tends to infinity:

A.2 Proof of theorem 1 in the continuous setup

We only have to prove the convergence in direction of $\int_0^t z(t')$ to the max-margin classifier. Let us denote ϕ this quantity and $\overline{\phi} = \frac{\phi}{\|\phi\|}$. Note that $\|\phi\|$ goes to infinity.

First, let us prove that all the adherence values have the same marge-in. To do so, let us suppose that there exists $\theta_1 > \theta_2$ such that $\overline{\phi}$ has adherence values with marge-in θ_1 and θ_2 and let us find a contradiction. Denote by $\mathcal{F}_\theta = \{\beta, marge(\beta) = \theta\}$.
For $0 < \delta_1$ and $\delta_2 > 0$ small enough, we have that $\overline{\phi}(t)$ visits for arbitrary large t the set $\mathcal{D}(\delta_1) := \{\beta, \beta \in \mathcal{F}_\theta, \theta \leq \theta_1 - \delta_1\}$. For such a t, we take a $\overline{\beta}$ of margin θ_1 and norm 1 and such that $\mathcal{S}' = \{n, \overline{\phi}^T x_n \leq margin(\phi) + \delta_2\}$ is included in $\mathcal{S} = \{n, x_n^T \overline{\beta} = \theta_1\}$
Then, $\frac{d}{dt} \|\overline{\phi(t)} - \overline{\beta}\|^2 = \frac{1}{\|\phi(t)\|} \{\langle z(t), \overline{\phi(t)} - \overline{\beta} \rangle - \langle \overline{\phi}(t), \overline{\phi}(t) - \overline{\beta} \rangle \langle \overline{\phi}(t), z(t) \rangle\}$

Then we have that $z(t) = \sum_{n \in \mathcal{S}} exp(-x_n^T \widetilde{\beta}(t)) x_n + o(\|z(t)\|)$.
For n in $\mathcal{S}$, $\overline{\phi}(t)^T x_n \leq \theta_1 - \delta_1$.
For $n \in \mathcal{S}$:
$\langle x_n, \overline{\phi}(t) - \overline{\beta} \rangle \leq -\delta_1$
$\langle x_n, \overline{\phi}(t) \rangle \langle \overline{\phi}(t), \overline{\phi} - \overline{\beta} \rangle \geq 0$.
So,

$$\sum_{n \in \mathcal{S}} exp(-\beta^T x_n) \{\langle x_n, \overline{\phi(t)} - \overline{\beta} \rangle - \langle x_n, \overline{\phi}(t) \rangle \langle \overline{\phi}(t), \overline{\phi}(t) - \overline{\beta} \rangle\} \leq \sum_{n \in \mathcal{S}} exp(-^T x_n) - \delta_1$$

So, there exits T such that for $t \geq T$ and $\overline{\phi}$ has margin larger than $\theta_1 - \delta_1$, then for a vector $\overline{\beta}$ of margin θ_1 such that $\mathcal{S}'$ is included in $\mathcal{S}$, then $\frac{d}{dt} \|\overline{\beta} - \overline{\phi}\|^2 < 0$.

If we take δ_1 and δ_2 very small, δ_2 being also very small in front of δ_2 we can recover the zone of vectors of margin $\theta_1 - \delta_1$ and of norm 1 by circles of centers points of margin δ_1 and of ray tending to zero with δ_1 such that for a circle of center $\overline{\beta}$ with have that for all the points in this circle of margin lower than $\theta_1 - \delta_1$, $\mathcal{S}'$ is included in $\mathcal{S}$. That implies that $\overline{\phi}$ cannot reach for t big enough a margin equals to θ_2.

Hence, all the adherence values of $\overline{\phi}$ have the same margin. let us denote this margin by θ.

Then, let us prove that $\overline{\phi}$ converges. We suppose that θ is not the margin of the max-margin classifier, otherwise the theorem is automatically verified. Let $\overline{\beta}$ be an adherence value of $\overline{\phi}$ that Let $\delta > 0$ and w of margin $\theta + \delta$ the closest point to $\overline{\beta}$ which verfifies that $\mathcal{S}(w) = \mathcal{S}(\overline{\beta})$(we can always find such a w otherwise that directly implies the convergence of $\overline{\phi}$ since there is a finite number of points where there is a dicrease of the possible points that reach a margin.) For t big enough and $\overline{\phi}$ close enough to $\overline{\beta}$, we have that $\frac{d}{dt}\|\overline{\phi} - w\|^2 < 0$. Let us denote by κ the distance between $\overline{\beta}$ and w. Then let $\mathcal{D} = \{\|\overline{\beta} - w\| \leq \frac{3}{2}\kappa\}$. For t big enough, every $\overline{\phi}$ that is in $\mathcal{D}$ verifies that $\mathcal{S}(\overline{\phi})$ is included in $\mathcal{S}(w)$ and thus for t big enough we have that if we are in $\mathcal{D}$, we stay in $\mathcal{D}$. Thus, the diameter of the adherence value of $\overline{\phi}$ can be arbitrary small. Thus it is a singleton.

Finally, let us prove that the limit of $\overline{\phi}$ is the max-margin classifier. Let us suppose that the limit of $\overline{\phi}$, that we write $\overline{\beta}$, has its margin equals to zero and find a contradiction. If its margin is equal to zero, then $z(t) \sim \sum_{x_n \in \overline{\beta}^T} \alpha(t) y_n x_n$ and so as the norm of ϕ tends to infinity, we have that $\overline{\beta}$ is orthogonal to itself, which is absurd.

So, the margin θ of $\overline{\beta}$ is positive. For $\mathcal{S}$ the set of n that verifies $\overline{\beta}^T x_n y_n = \theta$, we have that $z(t) \sim \sum_{n \in \mathcal{S}} \alpha_n(t) y_n x_n$ and so all its adherence values are in the cone created by the convex hall of $\mathcal{S}$. So, $\overline{\beta}$ is also in that cone. And, thus the rescaling of $\overline{\beta}$ of margin 1 verifies the KKT condition to minimise the euclidean norm with the constraint of having the margin equals to 1. That implies that $\overline{\beta}$ is the max-margin classifier.

A.3 Proof of theorem 1 in the discrete setup

Now, we will prove the same result for the finite step algorithm, provided the step size η is small enough. Let us denote $\beta(t) = w_1(t)...w_L(t)$ and $\phi(t) = \sum_{t' < t} z(t')$ and $\overline{\phi}$ its scaling of norm one.

First, let us prove that if at one point we have that a positive margin δ for $\frac{\beta}{\|\beta(t)\|}$ and such that the layers have a positive alignment to $b_{l-1} b_l^T$ superior to a $B(\delta)$ and such that for $\hat{w}$ the max-margin classifier,$\phi(t)^T \hat{w} > B$, then there exists δ' such that the margin of $\frac{\beta}{\|\beta\|}$ stays superior to δ' and such that the alignment of each layer is superior to B and such that $\phi(t)^T \hat{w}$ is still superior to B. Let $\delta' < \delta$. Then, if we have that the margin of $\frac{\beta}{\|\beta\|} = \overline{\phi}$ keeps being superior than δ', then the alignement between the layers and the feedback matrices increases and so does $\phi(t)^T \hat{w}$. In this case we have that $\overline{\phi} = \frac{\beta}{\|\beta\|}$;

We just have to prove that if we take B big enough, then we cannot escape from the zone of vectors of norm one that have there margin superior to δ'. Let $\delta_1 > 0$ small and κ that will also be taken very small. If we take B big enough, we have that between to steps, the margin of ϕ cannot increase or decrease by more than κ. Hence, to get lower than δ', the margin of $\overline{\phi}$ must be at one time between $\delta - \delta_1$ and $\delta - \delta_1 - \kappa$. Let us take $\delta_2 > 0$. For $\overline{\phi}$, let us denote $\mathcal{S}'(\overline{\phi}) = \{n, \overline{\phi}^T y_n x_n \leq margin(\overline{\phi} + \delta_2)\}$. We have that, with respect to B, $z(t) \sim \sum_{n \in \mathcal{S}'} \alpha_n(t) y_n x_n$. And, $\Delta\overline{\phi} \sim \frac{1}{\|\phi(t)\|} z(t) - \overline{\phi(t)} \langle \overline{\phi(t)}, z(t)$. And so, for B big enough and $\overline{\beta}$ of margin δ such that $\mathcal{S}'(\overline{\phi})$ is included in $\mathcal{S}$, we have that $\Delta\|\overline{\phi} - \overline{\beta}\| < 0$.

And, if we take δ_1,δ_2 and κ small enough, we can recover the zone of vector of margin between $\delta - \delta_1$ and $\delta - \delta_1 - \kappa$ by disks $\mathcal{D}(\overline{\beta})$ of centers points of margin δ such that for $\overline{\phi} \in \mathcal{D}(\overline{\beta})$, we have that $\mathcal{S}'(\overline{\phi}) \subset \mathcal{S}(\overline{\beta})$. That implies for η small enough, we have that the margin is larger than a $\delta > 0$ for t big enough.

Then, let us prove that if the margin stays larger than a δ, and that the norm of each layer is big enough, in the direction of $b_{l-1} b_l^T$, then the loss tends towards zero. Indeed, we have that the norm of the parameters will grow towards infinity (since z(t) has a positive alignment with the max-margin classifier.) So the norm of the effective vector will tend to infinity with a positive margin. That implies that the loss will tend to zero.

Finally, let us prove that if the loss goes to zero, and that every layer gets positively aligned to the feedback products then the first layer converges in direction in a way such that the classifier converges in direction to the max-margin classifier.(It is more or less the same as for the proof of the contonuous case.) We only need to prove that $\phi(t)$ converges in direction to the max-margin classifier. We will proceed in the same way than for the continuous case.

Let us prove that all the adherence of $\overline{\phi}$ have the same margin. Let us suppose that there exists $\theta_1 > \theta_2$ such that $\overline{\phi}$ has adherence values of margin θ_1 and θ_2 and let us find a contradiction.

For $0 < \delta_1, \delta_2, \alpha$ small enough, we have that $\overline{\phi}(t)$ visits for t arbitrary large the set of vector of margin between $\theta_1 - \delta_1$ and $\theta_1 - \delta_1 - \alpha$. For such a t, and $\overline{\beta}$ of margin θ_1, of norm 1, abd such that $\mathcal{S}'(\overline{\phi}) = \{n, \overline{\phi}^T x_n y_n \leq margin(\overline{\phi}) + \delta_2\}$ is included in $\mathcal{S}(\overline{\beta}) = \{n, y_n \overline{\beta}^T x_n = \theta_1\}$. Then, we have that $z(t) = \sum_{n \in \mathcal{S}'} exp(-y_n \beta(t)^T x_n) y_n x_n + o(\|z(t)\|)$ and $\delta\overline{\phi}(t) = \frac{1}{\|\phi(t)\|}\{\langle z(t), \overline{\phi}(t) - \overline{\beta}\rangle - \langle \overline{\phi}(t), \overline{\phi} - \overline{\beta}\rangle \langle \overline{\phi}(t) - z(t)\rangle + o(z(t))\}$. So, there exists t such that for T>t, and $\overline{\phi}$ with such a margin and such a $\overline{\beta}$, then $\delta\|\overline{\phi} - \overline{\beta}\| < 0$. If we take α very small in front of δ_1 and the same with δ_2 in front of δ_1 and with also δ_1 very small, we cannot escape from zone of vectors with margin larger than $\theta_1 - 2\delta_1$. That is absurd. So, all the adherence values of $\overline{\phi}$ have the same adherence value θ.

Then, let us prove that $\overline{\phi}$ converges. We suppose that θ is not the margin of the max-margin classifier otherwise the result is automatic. Let $\overline{\beta}$ be an adherence value of $\overline{\beta}$. Let $\delta > 0$ and w of margin $\theta + \delta$ the closest point to $\overline{\beta}$ that verifies $\mathcal{S}(w) = \mathcal{S}(\overline{\beta})$. For t big enough and $\overline{\phi}$ close enough to $\overline{\beta}$, we then have that $\delta\|\overline{\phi} - w\|^2 < 0$. Let us take κ the distance between $\overline{\beta}$ and w. Then if $\overline{\phi}$ is at distance less than $\frac{3}{2}\kappa$ (if δ small enough) we will have that $\delta\|\overline{\phi} - w\|^2 < 0$.. That means that the diameter of the adherence values of $\overline{\phi}$ is arbitrary small. Hence the convergence of $\overline{\phi}$.

The convergence of $\overline{\phi}$ to the max-margin classifier follows from that in the same way than for the continuous case.

A.4 Proof of proposition 2

Let us prove that there exists $\phi_l(t)$ with positive coefficient such that $\hat{w_l}(t) = \phi_l(t) \odot \hat{b}_{l-1} \odot \hat{b}_l^\star$. We have that $\hat{w_1}(t) = \int_0^t \hat{z}(t) \odot b_1^\star$ and so $\hat{w_2}(t) = \int_{0<t'<t} \hat{z}(t'') \odot \hat{z}(t') \odot \hat{b}_1 \odot \hat{b}_2^\star$. Moreover, $\int_{0<t'<t} \hat{z}(t'') \odot \hat{z}(t')$ is the Fourier space of real valued vector and its real part is equal to $\frac{1}{2}|\int_0^t z(t')|^{\odot 2}$ which is also the Fourier transform of a real valued vector. So, since the real part of the Fourier matrix is invertible, we have that $\int_{0<t'<t} \hat{z}(t'') \odot \hat{z}(t') = \frac{1}{2}|\int_0^t z(t')|^{\odot 2}$. That concludes the result for the second layer.

For the l-th layer, we have that, the thrid equality following the same argument as before of the injectivity of the real part of the Fourier matrix:

$$\begin{aligned}
\hat{w_l}(t) &= \int_0^t \hat{w}_{1:l-1}^\star(t) \odot \hat{z}(t) \odot \hat{b}_l^\star \\
&= \int_0^t \hat{w}_{l-1}^\star(t') \odot \frac{d}{dt'} \hat{w}_{l-1}(t') \odot b_l^\star \odot (b_{l-1}^{\odot -1})^\star \\
&= \frac{1}{2}|\hat{w}_{l-1}(t)|^{\odot 2} \odot \hat{b}_l^\star \odot \hat{b_{l-1}}^{\star \odot -1} \\
&= \frac{1}{2}|\phi_{l-1}(t) \odot \hat{b}_{l-2}|^{\odot 2} \odot \hat{b_{l-1}} \odot \hat{b}_l^\star
\end{aligned}$$

That concludes the proof for the positive alignment with the feedback matrices.

Hence, the loss decreases. The demonstration also shows that each layer has its norm that does not have 0 as an adherence value. So, $z(t)$ converges to zero. So, $\mathcal{L}(w(t))$ also converges to zero. That also implies that the norm of each layer tends to infinity.

A.5 Proof of Theorem 2 in the continuous setup

Proving that all the layers converge in direction provided that we have that the margin of the direction that does not have 0 as an adherence value. Each layer has its norm of the order of the

square of the norm of the previous layer. So, the norm of the effective vector $\hat{\beta}(t) = \odot_0^L \hat{w}_l(t)$ is of the order of $\|w_1\|^{2^{L+1}-1}$(since the margin of the rescaled network stays larger than a δ for t big enough). Note $\lambda(t) = \frac{min\beta^T x_n y_n}{\|w_1\|^{2^{L+1}-1}}$. Then, with $y := \|w_1\|$:
- $\frac{d}{dt}y(t) \geq C_1 exp(-\lambda(t)y(t)^{2^{L+1}-1})$
- $\frac{d}{dt}\frac{w_1}{\|w_1(t)\|} \leq C_2 \frac{1}{y(t)exp(\lambda(t)y(t)^{2^{L+1}-1})}$

Then ,let us study $\psi(t) = \frac{1}{\lambda(t)^{2^{L+1}-1}} ln(t)^{\frac{1}{2^{L+1}-1}} + ln(ln(t))^\alpha$. We have that $\frac{d}{dt}\psi(t) = \frac{1}{2^{L+1}-1\lambda(t)^{2^{L+1}-1}} ln(t)^{\frac{1}{2^{L+1}-1}-} \alpha \frac{1}{tln(t)} ln(ln(t))^{\alpha-1} - (2^{L+1}-1)(\frac{d}{dt}\lambda(t))\frac{1}{\lambda^{2^{L+1}}}$. Hence, $\frac{d}{dt}\psi(t) = o(exp(-\lambda(t)\psi(t)^{2^{L+1}-1}))$.

So, $\psi(t) = o(y(t))$ and $\frac{d}{dt}\frac{w_1}{\|w_1(t)\|} \leq C_3 \frac{1}{\psi(t)exp(\lambda(t)\psi(t)^{2^{L+1}-1})}$ that being true for all α. We take α such that this is integrable and that concludes the demonstration for w_1.

Ensuite, $\frac{d}{dt}\frac{w_2}{\|w_2\|} = O(\frac{\frac{d}{dt}w_2}{\|w_2\|}) = O(\|w_1\|\frac{d}{dt}w_1\frac{1}{\|w_2\|}) = O(\frac{\frac{d}{dt}w_1}{\|w_1\|})$ is also integrable.

Proving that the limit in direction of β minimises $\Psi(\hat{\beta})$. We have that every layer converges in direction. The first layer $\hat{w}_1 = \int \hat{z} \odot \hat{b}_1^\star$ converges in direction and so we have that $\int \hat{z}$ converges in direction (we suppose that the feedback matrices have no coefficients that are null). Since $\hat{z}$ has all its adherence value in the cone generated by $\hat{x}_n, n \in \mathcal{S}$, then we have that $\int \hat{z}$ has its limit in direction in this cone. Hence, there exists $\alpha_n \geq 0, n \in \mathcal{S}$ such that the limit in direction of $\int \hat{z}$ is equal to $\hat{z}^\infty := \sum_{n\in\mathcal{S}} \alpha_n \hat{x}_n$.

Moreover, we know that $\hat{w}_l(t) = \frac{1}{2}|\hat{w}_{l-1}(t)|^{\odot 2} \odot \hat{b}_l^\star \odot (\hat{b}_{l-1}^\star)^{\odot -1}$. Thus we have that the limit in direction of $\hat{w}_l$ is proportional to $|\hat{z}^\infty|^{\odot 2^{l-1}} \odot |\hat{b}_1|^{\odot 2^{l-2}} \odot ... \odot |\hat{b}_k|^{\odot 2^{l-1-k}} \odot \hat{b}_{l-1} \odot \hat{b}_l^\star$. Thus, we have that $\hat{\beta}$ has its direction which converges to something proportional to $\hat{\beta}^\infty = |\hat{z}^\infty|^{\odot 2^L - 2} \odot \hat{z}^\infty \odot |\hat{b}_1|^{\odot 2^{L-1}} \odot ... \odot |\hat{b_{L-1}}|^2$.

Consequently, $\hat{\beta}^\infty$ has its coefficient of the same phase as $\hat{z}^\infty$ and we have that $|\hat{\beta}^\infty| \propto |\hat{z}^\infty|^{\odot 2^L - 1} \odot \odot_{l=1}^{L-1} |\hat{b}_{L-l}|^{2^l}$ which means that $|\hat{z}^\infty| \propto |\hat{\beta}^\infty|^{\odot \frac{1}{2^L-1}} \odot \odot_{l=1}^{L-1} |\hat{b}_{L-l}|^{2^{-l}}$. Then, if we take the rescaling of $\hat{\beta}^\infty$ with margin one, it verifies the KKT condition of the minimisation of $\Psi(\hat{\beta})$ with the constraint of having the margin larger than one. That concludes the proof.

KKT condition for the minimisation of $\Psi(\hat{\beta})$ with the margin larger than one. A vector $\hat{\beta}$ in the complex space verifies the KKT condition for minimising $\Psi(\hat{\beta})$ with the constraint $\forall n, \langle \hat{\beta}, \hat{x}_n \rangle y_n \geq 1$ if and only if there exists $\alpha_n \geq 0, n \in \mathcal{S}$ such that for all d, $\frac{d}{d\hat{\beta}[d]}\Psi(\hat{\beta}) = \sum_{n\in\mathcal{S}} \alpha_n \hat{x}_n$. And, we have that $\frac{d}{d\hat{\beta}[d]}\Psi(\hat{\beta}) = \frac{1}{\prod_{l=1}^{L-1} |\hat{b}_{L-l}[d]^{2^l}|^{\frac{1}{2^L-1}}} |\hat{\beta}[d]|^{\frac{1}{2^L-1}}) exp(i\phi_{\hat{\beta}[d]}) \Psi(\hat{\beta})$.

A.6 Proof of Theorem 2 in the discrete setup.

We suppose that every layer converges in direction (that means the Fourier transform too.) to $\overline{\hat{w}_l}^\infty$ We have that $\hat{w}_1 = \sum_{t'<t} \hat{z}(t') \odot \hat{b}_1^\star$. Thus, there exists $\alpha_n \geq 0, n \in \mathcal{S}$ such that the limit in direction of $\sum_{t'<t} \hat{z}(t') = \sum_{n\in\mathcal{S}} {}_n y_n \hat{x}_n := \overline{\hat{z}}^\infty$.

The second layer can be written $\hat{w}_2(t) = \sum_{t'<t} \hat{w}_1(t')^\star \odot \hat{z}(t') \odot \hat{b}_2^\star = \hat{b}_2^\star \odot \hat{b}_1^{\star -1 \odot} \odot \sum_{t'<t} \sum_{t''<t'} \Delta\hat{w}_1(t'')^\star \odot \delta\hat{w}_1(t')$. The quantity $\sum_{t'<t} \sum_{t''<t'} \Delta\hat{w}_1(t'')^\star \odot \delta\hat{w}_1(t')$ is equivalent to $\sum_{t''\leq t'\leq t} \Delta\hat{w}_1(t'')^\star \odot \delta\hat{w}_1(t')$ which is equal, using the injectivity of the real part of the Fourier matrix, to $\frac{1}{2}|\hat{w}_1|^{\odot 2}$ and hence, $\overline{\hat{w}_2}^\infty \propto |\hat{z}^\infty|^{\odot 2} \odot \hat{b}_1 \odot \hat{b}_2^\star$.

With the same method, we have that $\overline{\hat{w}_l}^\infty \propto |\hat{z}^\infty|^{\odot 2^{l-1}} \odot |b_1|^{\odot 2^{l-2}} ... \odot |b_{l-2}|^{\odot 2} \odot b_{l-1} \odot b_l^\star$. That concludes the proof.

A.7 Proof of Theorem 3 in the continuous setup.

We have that for all n, $w_{out}(t)^T\phi(w(t),x_n)y_n$ goes to infinity. So, we have that for t big enough $\overline{w}_out(t)^T\phi(\frac{w}{\|w\|},x_n)y_n>0$, with $\overline{w}_{out}(t)=\frac{w_{out}(t)}{\|w_{out}(t)\|}$. So, all the adherence value of $\overline{w}_{out}(t)$ verify that $\overline{w}_{out}^T\phi(\overline{w},x_n)y_n\geq 0$. Let us study the adherence values.

Let us show that all the adherence values of $\overline{w}_{out}$ have the same margin. Suppose that there exists $\theta_1>\theta_2$ such that $\overline{w}_{out}$ has adherence values with margin θ_1 and θ_2 and find a contradiction.(Here, we considier the marge with respect to the linearly separable dataset $\phi(\overline{w},x_n)$)

For $0<\delta_1$ and $\delta_2>0$ small enough, we have that $\overline{w}_{out}$ visits for t arbitrary large the set of vectors of margin lower than $\theta_1-\delta_1$. Let us denote $\mathcal{S}'(\overline{w}_{out})=\{n,\overline{w}_{out}^T phi(\overline{w},x_n)y_n\leq marge(\overline{w}_{out})+\delta_2\}$. We have that for t going to infinity, uniformly on the direction of $\overline{w}_{out}$, that $\nabla w_{out}(t)\mathcal{L}\sim\sum_{n\in\mathcal{S}'}y_n exp(-y_n w_{out}^5\phi(w,x_n))\phi(w,x_n)\sim\|w\|\sum_{n\in\mathcal{S}'}y_n exp(-y_n w_{out}^5\phi(w,x_n))\phi(\overline{w},x_n)$ in the case where the dataset is strictly linearly separable (otherwise there is only one vector that classify the dataset and $\overline{w}_{out}$ automaticly converges in that direction). So, we have that for t big enough and $\overline{\beta}$ of margin θ_1 and such that $\mathcal{S}'(\overline{w}_{out})\subset\mathcal{S}(\overline{\beta})$, we have that $\frac{d}{dt}\|\overline{w}_{out}(t)-\overline{\beta}\|^2<0$.

Then, with the same argument than in theorem 1, we have that all the adherence values have the same margin.

Let us show that $\overline{w}_{out}$ converges If the margin θ is not the one of the max-margin classifier (otherwise we have the convergence to the max-margin classifier), denote by $\overline{w}_{out}^\infty$ an adherence value and for $\delta>0$ small enough . Take $\overline{\beta}$ of margin $\theta+\delta$ such that $\mathcal{S}(\overline{w}_{out}^\infty)=\mathcal{S}(\overline{\beta})$ (we can always find such a $\overline{\beta}$ otherwise that directly implies the convergence of $\overline{w}_{out}$ since there is a finite number of points where there is a decrease of the possible points that reach a margin.) For t big enough and $\overline{w}_{out}$ close enough to $\overline{w}_{out}^\infty$, we have that $\frac{d}{dt}\|\overline{w}_{out}(t)-\overline{\beta}\|^{2\leq 0}$. Let us denote by κ the distance between $\overline{w}_{out}^\infty$ and $\overline{\beta}$. Let us denote $\mathcal{D}=\{\|\overline{w}_{out}(t)-\overline{\beta}\|\leq\frac{3}{2}\kappa\}$. If we take δ small enough, then δ will be arbitrary small and then every $\overline{w}_{out}(t)$ in $\mathcal{D}$ will verify that $\mathcal{S}'(w_{out})\subset\mathcal{S}(\overline{\lfloor\rceil\sqcup\dashv})$ and so, if we are in $\mathcal{D}$, we stay in $\mathcal{D}$.

Hence, the diameter of the adherence values of $\overline{w}_{out}$ has an arbitrary small diameter. Thus, there is only one adherence value and that concludes the proof.

Let us show that the limit of $\overline{w}_{out}$ is the max-margin classifier of the dataset $(\phi(\overline{w},x_n),y_n)$ when its norm is infinite. The derivative of w_{out} is arbitrary close to the cone genrated by $\phi(\overline{w},x_n)y_n, n\in\mathcal{S}(\overline{w}_{out}^\infty)$ when t gets large and so it is the same with the direction of w_{out}. So, it verifies the KKT condition.

A.8 Proof of Theorem 3 in the discrete setup.

A.9 Proof of proposition 3

We have that $w_1(t)=\sum_m\int_0^t\alpha_m(t')y_m x_m[b\odot g'(w_1(t')^T x_m)]^T$ and thus, since we have $w_2(t)=\int_0^t\sum_n\alpha_n(t')y_n[w_1(t')^T x_n]\odot g'(w_1(t')^T x_n)$, then $w_2(t)=\sum_{n,m}\int_0^t\int_0^{t'}\alpha_n(t')\alpha_m(t'')y_n y_m x_m^T x_n[b\odot g'(w_1(t'')^T x_m)\odot g'(w_1(t')^T x_n)]$.
Hence:

$$\begin{aligned}w_2(t)&=b\odot\int_0^t\int_0^{t'}\sum_{n,m}\alpha_n(t')\alpha_m(t'')y_n y_m g'(w_1(t')^T x_n)\odot g'(w_1(t'')^T x_m)\\&=\frac{1}{2}b\odot\int_0^t\int_0^t\sum_{n,m}\alpha_n(t')\alpha_m(t'')y_n y_m g'(w_1(t')^T x_n)\odot g'(w_1(t'')^T x_m)\end{aligned}$$

Thus, we have that for all d, $w_2[d]$ is of the same sign than $b[d]$.

Hence, the sign of the coefficient of the derivative of w_1 are of the same sign of the negative of the gradient of the loss with respect to this same layer. That implies that the loss decreases.

A.10 Proof of Theorem 4

To prove this result, we only have to prove that the first layer converges in direction.

First, let us prove that there exists α and β such that $\alpha\|w_1\|^2 \leq \|w_2\| \leq \beta\|w_1\|^2$. We have that $\frac{d}{dt}\|w_2\|^2 = 2\sum_n \alpha_n(t) w_2^T g(w_1^T x_n)$ with $\alpha_n(t) = y_n exp(-y_n w_2^T g(w_1^T x_n))$. So there exists C so that for t big enough, since the margin is superior to a δ, $\frac{d}{dt}\|w_2\|^2 \geq \|w_2\|\|w_1\|\|\frac{d}{dt}w_1\|$. That proves the first inequality.

For the second inequality, we have that $\frac{d}{dt}Tr(w_1^T w_1) = 2Tr(\sum \alpha_n(t) w_1^T x_n[b \odot g'(w_1^T x_n)])$. And the vector made of the diagonal of this matrix is equal to $2b \odot \frac{d}{dt}w_2$. That proves the second point, due to the positive alignment of the coefficient of w_2 and b.

Let us prove that $\frac{d}{dt}\frac{w_1}{\|w_1\|}$ is integrable. Since $\frac{d}{dt}\frac{w_1}{\|w_1\|} = O(\frac{\frac{d}{dt}w_1}{\|w_1\|})$ we just have to prove that this last quantity is integrable. Let us denote $\lambda(t) := \frac{min(w_2^T g(w_1^T x_n))y_n}{\|w_1\|^3}$ and $y(t) = \|w_1\|$.

Then:

1. $\frac{d}{dt}y(t) \geq C_1 exp(-\lambda(t)y(t)^3)$
2. $\frac{\frac{d}{dt}w_1}{\|w_1\|} \leq C_2 \frac{exp(-\lambda(t)y(t)^3)}{y(t)}$

Indeed, for the first point, we have that:

$$\begin{aligned}\frac{d}{dt}y(t)^2 &= Tr(2b \odot \frac{d}{dt}w_2) \\ &\geq C\|w_1\| exp(-\lambda(t)y(t)^3) \\ &\geq Cy(t)exp(-\lambda(t)y(t)^3)\end{aligned}$$

For the second point, we have that: $\frac{d}{dt}w_1 = O(max(\alpha_n(t))) = O(exp(-(t)y(t)^3))$.

Then, let us denote, for $c > 0$, $\Psi(t) := \lambda(t)^{-3} ln(t)^{\frac{1}{3}} + ln(ln(t))^c$. Then, we have that $\frac{d}{dt}\Psi(t) = \frac{1}{3\lambda(t)^3} ln(t)^{-\frac{2}{3}} \frac{1}{t} + c\frac{1}{tln(t)} ln(ln(t))^{c-1} - 3\frac{\frac{d}{dt}\lambda(t)}{\lambda(t)^3} ln(t)^{\frac{1}{3}}$. The two first terms are negligible in front of $exp(-\lambda(t)\Psi(t)^3)$.

Let us now deal with the last term. We have that $\frac{d}{dt}\lambda(t) = O(\frac{exp(-\lambda(t)y(t)^3)}{y(t)})$. And $ln(t)^{\frac{1}{3}} = o(y(t))$. Indeed, $\frac{d}{dt}\gamma ln(t)^{\frac{1}{3}} = \frac{1}{3}\gamma\frac{1}{t}ln(t)^{-\frac{2}{3}} = o(exp(-\lambda(t)\gamma^3 ln(t)))$ for γ small enough.

Hence, the last term is negligible in front of $exp(-\lambda(t)y(t)^3)$. Therefore, we have that $\Psi(t)$ is negligible in front of $y(t)$.

Therefore, $\frac{\frac{d}{dt}w_1}{\|w_1\|} \leq \frac{C}{ln(t)^{\frac{1}{3}} exp(ln(t))ln(t)^{3c\lambda(t)}} \leq \frac{C}{tln(t)^2}$ for c big enough. That concludes the proof.

References

1. Lillicrap, T., Cownden, D., Tweed, D. & Akerman, C. Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications* **7,** 1–10 (2016).
2. Nøkland, A. *Direct Feedback Alignment Provides Learning in Deep Neural Networks* in *Advances in Neural Information Processing Systems 29* (2016).
3. Soudry, D., Hoffer, E. & Srebro, N. *The Implicit Bias of Gradient Descent on Separable Data* in *International Conference on Learning Representations* (2018).
4. Lyu, K. & Li, J. *Gradient Descent Maximizes the Margin of Homogeneous Neural Networks* in *International Conference on Learning Representations* (2020).
5. Ji, Z. & Telgarsky, M. *Directional convergence and alignment in deep learning* in *Advances in Neural Information Processing Systems* **33** (2020), 17176–17186.