Signal propagation and feature learning in ResNets

Emilie Zheng

September 1, 2024

Contents

1 Déroulement du stage

J'ai effectué mon stage de M1 du 5 février 2024 au 5 juillet 2024, pour une durée de 5 mois dans l'Institut de mathématiques de l'Ecole polytechnique fédérale de Lausanne (EPFL), une université en Suisse spécialisée en science. J'ai été accueillie dans l'équipe de recherche DOLA (Dynamics of learning algorithms) dirigée par le Pr. Lénaïc Chizat, qui était mon superviseur pendant ce stage. Pr. Chizat travaille sur des sujets tels que le tranport optimal et les réseaux de neurones, et pendant le stage, je travaillais sur la théorie du deep learning.

J'ai pu dans un premier temps me familiariser avec le thème de mon stage et dans un deuxième temps tenter d'avancer sur des sujets de recherche nouveaux avec l'aide de Pr. Chizat, qui m'orientait lors de nos nombreux rendez-vous, par exemple en indiquant les articles pertinents qui pouvaient m'aider. Les nombreuses conférences données dans l'EPFL ont été une très bonne opportunité de découvrir de nombreux domaines de recherche. De plus, je pouvais assister aux meetings hebdomadaires de l'équipe DOLA, ce qui m'a permis de comprendre un peu les différents sujets qui pouvaient intéresser les membres de l'équipe DOLA (des doctorants et post-docs principalement).

Je remercie grandement Pr. Chizat, qui m'a permis d'effectuer mon stage dans son équipe. Je remercie de même tous les membres de l'équipe DOLA, qui m'ont offert un très bon accueil et ont contribué à faire de ma première expérience de stage une bonne expérience.

2 Introduction

2.1 General context

Let us first introduce the structure studied throughout all this report. A neural network consists of neurons in several layers connected by connections. Each neuron gets the information given by the connections from the previous layers, and computes an output using this information and weights (that are given parameters and specific to the network) and an activation function (always non-linear). The first layer is the input, and the signal goes through the network in a forward pass, the last layer is the output layer. The number of layers in the network is called the depth of the network, while the number of neurons per layer is called the width of this layer, and in our case all the hidden layers have the same width. Typically, the general neural network can be written as a function that takes an input x, returns an output y, and the ℓ-th layer of neuron noted f_ℓ (called features) is obtained by $f_{\ell+1} = T_\ell(f_\ell, W_\ell)$, where T_ℓ is a function, and W_ℓ are the weights.

Neural networks are important structures that can learn information through training, where the goal is to minimize a loss function step after step by varying the parameters of the network, the weights. The training is done through different methods such as gradient descent, and often the weights are initialized randomly at the start of the training. There are different problems that can occur during the training, such as situations called lazy training where the parameters hardly vary whereas the training loss still converges to 0, or situations where the initial weights chosen make the training not progress in certain methods of training. To avoid these kind of situations, it is important for example to understand how the training goes after one step with different set of weights, and this is what will be studied in this report.

To be more specific, we are interested in the speed at which the features evolve after one step of gradient descent. To understand why some objets are studied, it is important to know that the study is mainly motivated by the theorem 2.4 of Chizat and Netrapalli [2024] (the first version).

In order to do this, we will first establish the stochastic differential equation of the forward pass for a specific type of residual networks that we study in the limit width $m \to \infty$ (Hayou [2023]), then try to study the spectrum of the Jacobian of the function $f_0 \to f_\ell$, an object

not entirely known yet and involved in the theorem that we are interested in Chizat and Netrapalli [2024].

2.2 Definitions

In the rest of this report, for any vector $x \in \mathbb{R}^n$, we will note $x^{(i)}$ the i-th coordinate of x. We note $(B_t)_{t\in\mathbb{R}_+}$ the standard Brownian motion in 1D. If non specified, the norms used are the euclidean norm for any vector $x \in \mathbb{R}^m$, and the frobenius norm for any matrix $\sigma \in \mathbb{R}^{m\times n}$.

The main object that will be studied in this report is the Residual neural network (Resnet). It is a type of neural network architecture. With input $x = f_0 \in \mathbb{R}^{m_0}$, the forward pass is given by $f_1 = W_1 x$ and for $\ell \in [2 : L-1]$,

$$\bar{f}_\ell = \phi(f_{\ell-1}), \quad f_\ell = \sqrt{1-\beta^2} f_{\ell-1} + \beta W_\ell \bar{f}_\ell, \quad \bar{f}_L = W_L f_{L-1}, \quad f_L = \mathcal{L} = \text{loss}(\bar{f}_L) \in \mathbb{R}$$

where the map $\phi : \mathbb{R} \to \mathbb{R}$ acts entrywise on vectors. Here d is the input width, m the hidden widths, k the output width and $\forall \ell \in [1 : L]$, $W_\ell \in \mathbb{R}^{m_\ell \times m_{\ell-1}}$. The factor $\beta \in [0, 1]$ which we refer to as the *branch strength*, allows to interpolate between ResNet architectures (for $\beta < 1$) and MLPs (Multilayer perceptron for $\beta = 1$, but it is not the structure that we will study here). The computational graph is represented on Figure 1.

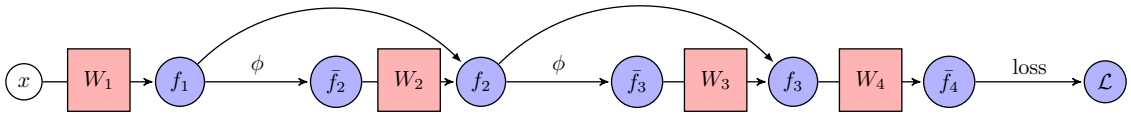

Figure 1: Computational graph of the forward pass for a simple ResNet.

In our case, let us consider the ResNet structure given above, with $\forall \ell \in [1 : L-1], m_\ell = m \in \mathbb{N}^*$, and $\beta = \frac{1}{\sqrt{L}}$ that will depend on the depth of the ResNet. The activation ϕ is ReLU, with $\phi(r) = max(r, 0)$ for all $r \in \mathbb{R}$.

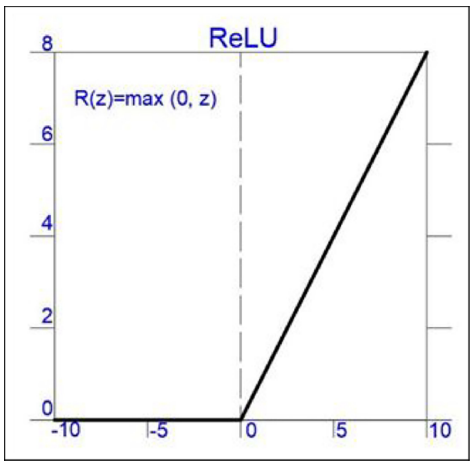

Figure 2: RELU graph

The weights are initialized with (before the training) :

$$\forall \ell \in [1 : L], \qquad \forall (i, j) \in [1 : m_{\ell-1}] \times [1 : m_\ell], \qquad W_\ell^{(i,j)} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\ell)^2$$

with $\sigma_1 = \sqrt{\frac{1}{m_0}}$ and for $\ell \in [2:L], \sigma_\ell = \sqrt{\frac{2}{m_{\ell-1}}}$, and all the variables in the weight matrices are independent. We will omit the dependence in L and m in every notation.

This is the basic Resnet architecture, but in practice we are more interested in Resnets where the forward pass is given by :

$$\begin{aligned} f_\ell &= W_\ell f_{\ell-1} && \text{if } \ell = 1 \text{ or } L \\ f_\ell &= \sqrt{1-\tfrac{1}{L^2}} f_\ell + \tfrac{1}{L} V_\ell \bar{f}_{\ell-1}, \ \bar{f}_\ell = \phi(U_\ell f_{\ell-1}) && \text{if } 1 < \ell < L \end{aligned}$$

Here, there are even more parameters since there are 2 matrices U_ℓ and V_ℓ per layer. This structure only appears in the section 3, we are only interested in the first type of Resnet in the other parts of this report.

3 Equations on the forward pass

In this section, we will study the networks in the limit depth $L \to \infty$. More precisely, we will show that the random process $(f_t^L)_{t\in[0,1]} = (f_{\lfloor tL \rfloor})_{t\in[0,1]})$, where $f_{\lfloor tL \rfloor}$ are features in a Resnet of depth L, will converge to the solution of a stochastic differential equation (SDE).

There are several ways to define the stochastic differential equations, and we refer to Pavliotis [2014] for example for a rigorous definition of it.

3.1 On the first type of Resnet

We look at the case where the depth L tends to infinity, which allows us to approximate the evolution of the features that are discrete in time by a continous time SDE, where the time $t \in [0,1]$, and the random process that we want to approximate is $f_t^L = f_{\lfloor tL \rfloor}$.

In fact, we can see this intuitively because rewriting our formula gives us the discretization with a step of time of $\frac{1}{L}$ of the limiting SDE :

$$\begin{aligned} f_{\ell+1} &= \sqrt{1-\frac{1}{L}} f_\ell + \sqrt{\frac{1}{L}} W_{\ell+1} \phi(f_\ell) \\ &= \sqrt{1-\frac{1}{L}} f_\ell + \sqrt{\frac{2}{m}} B_{1/L} \phi(f_\ell) \end{aligned}$$

The next proposition gives the limit of $(f_t^L)_{t\in[0,1]}$.

Proposition 3.1. *When the depth $L \to \infty$, the random process $f_t^L = f_{\lfloor tL \rfloor}$ converges in law towards the solution of the following stochastic differential equation :*

$$\begin{cases} df_t^{lim} = -\frac{1}{2} f_t^{lim} dt + \sqrt{\frac{2}{m}} dB_t \phi(f_t^{lim}) \quad \forall t \in [0,1] \\ f_0^{lim} \sim \mathcal{N}(0, \frac{\|x\|^2}{m_0} I_m) \end{cases} \tag{1}$$

where $x \in \mathbb{R}^{m_0}$ is the input (we consider that the SDE starts at $f_1 = W_1 x$).

Proof. A proof of this proposition can be found in Hayou [2023]. □

3.2 On the second type of Resnet

In this section, we consider the second Resnet architecture. Given an input $x = f_0$, the forward pass is :

$$f_1 = W_1 x, \quad f_\ell = \sqrt{1-\beta^2} f_{\ell-1} + \beta V_\ell \phi(U_\ell f_{\ell-1}), \quad f_L = W_L f_{L-1}, \quad \mathcal{L} = \text{loss}(\bar{f}_L) \in \mathbb{R}$$

As in the last section, $\beta = \frac{1}{\sqrt{L}}$, all the layers have the same width m and the weights are initialized with :

$$\forall \ell \in [1 : L-1], \quad U_\ell \sim \mathcal{N}(0, \tfrac{1}{m} I_m), V_\ell \sim \mathcal{N}(0, \tfrac{2}{m} I_m)$$

and W_1 and W_L are initialized like in the previous section.

Unlike the first Resnet, the equation that we have is not a discretization of any kind of SDE. This is why we need another approach if we want to study this architecture.

However, what we will do here in this section is not totally rigorous since we do not prove some key technical points, for example we use results on Itô diffusion without ever proving that the limiting process will be one. This section is thus only a sketch of what could be a way to study this type of Resnet.

We will try to deduce what could be potentially a limiting equation. For this, we use the infinitesimal generator of an Itô diffusion. This allows us to analyse the drift and diffusion coefficients of the SDE which solution could the process $(f_t^L)_{t\in[0,1]}$ converges towards. We start with a few definitions.

Definition 3.2. *Itô diffusion*
An Itô diffusion is a stochastic process $(X_t)_{t\in\mathbb{R}^+}$ satisfying a SDE of the form

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t, \qquad t \geq 0, \qquad X_0 = x_0$$

where B_t is m-dimensional Brownian motion and $b : \mathbb{R}^n \to \mathbb{R}^n$, $\sigma : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ satisfies the following condition :

$$|b(x) - b(y)| + |\sigma(x) - \sigma(y)| \leq D|x - y|, \qquad \forall x, y \in \mathbb{R}^n$$

where the norms are the euclidean norm for vectors and the frobenius norm for matrices.

Definition 3.3. *Infinitesimal generator*
For an Itô diffusion $(X_t)_{t\in\mathbb{R}^+}$, the infinitesimal generator A of X is defined by (if the limit exists) : $\forall x \in \mathbb{R}^m$, for $f : \mathbb{R}^m \to \mathbb{R}$ a function,

$$Af(x) = \lim_{t\downarrow 0} \frac{\mathbb{E}^x[f(X_t)] - f(x)}{t}$$

Here, the notation $\mathbb{E}^x$ is the expectation conditionned by the event $X_0 = x$.

This next theorem is the reason why we look at infinitesimal generators.

Theorem 3.4. *Let $(X_t)_{t\in\mathbb{R}^+}$ be an Itô diffusion defined as before. If $f \in \mathcal{C}_0^2(\mathbb{R}^m)$ then A exists and*

$$Af(x) = \sum_{i=1}^m b_i(x) \frac{\nabla f}{\nabla x_i} + \frac{1}{2} \sum_{i,j} (\sigma\sigma^T)_{i,j}(x) \frac{\partial^2 f}{\partial x_i \partial x_j} \tag{2}$$

Proof. The proof can be found in Øksendal [2010]. □

We now try to use this theorem. Let us assume that as in the previous case, there exists a limit f^{lim} when the depth $L \to \infty$, and that it is an Itô diffusion. Just as before, $f_t^L = f_{tL}$.

Proposition 3.5. *Considering that we can replace $f_{1/L}^{lim}$ by f_1 when we need to evaluate the limit for the generator, the generator A of f^{lim} is*

$$Ag(x) = -\frac{1}{2} \nabla g(x) \cdot x + \frac{1}{2} \sum_{i=1}^m \frac{\partial^2 g}{\partial^2 x_i} \frac{\|x\|^2}{m} \tag{3}$$

where g is a function. Furthermore, f^{lim} is the solution of the SDE :

$$df_t^{lim} = -\frac{1}{2} f_t^{lim} + \sqrt{\frac{\|f_t^{lim}\|^2}{m}} dB_t$$

Proof. Since we can replace $f_{1/L}^{lim}$ by f_1 when we need to evaluate the limit for the generator, we now need to evaluate the limit of $L(\mathbb{E}^x[g(f_1)] - g(x))$ when $L \to \infty$. We note $\delta = -\frac{1}{2L}x$ and $\Delta = \frac{1}{\sqrt{L}} V_1 \phi(U_1 x)$, we have $\delta, \Delta \to 0$ a. s. when $L \to \infty$. With these notations, we have :

$$\begin{aligned} f_1 &= \sqrt{1 - \frac{1}{L}} x + \frac{1}{\sqrt{L}} V_1 \phi(U_1 x) \\ &= x + \delta + \Delta + o(\frac{1}{L}) \end{aligned}$$

We then have : $g(f_1) = g(x) + \nabla g(x) \cdot (\delta + \Delta) + \frac{1}{2}(\delta + \Delta)^T \nabla^2 g(x)(\delta + \Delta) + o(\|\delta + \Delta\|^2) + o(\frac{1}{L})$. Since $\mathbb{E}^x[\Delta] = 0$, taking the expectation gives us

$$\begin{aligned} \mathbb{E}^x[g(f_1)] &= g(x) + \nabla g(x) \cdot \delta + \frac{1}{2}\mathbb{E}^x[\Delta^T \nabla^2 g(x) \Delta] + o(\frac{1}{L}) \\ &= g(x) + \nabla g(x) \cdot \delta + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{\partial^2 g}{\partial x_i \partial x_j} \mathbb{E}^x[\Delta^{(i)} \Delta^{(j)}] + o(\frac{1}{L}) \end{aligned}$$

Then, since $\mathbb{E}^x[\Delta^{(i)} \Delta^{(j)}] = \delta_{i,j} \frac{\|x\|^2}{m}$, we obtain :

$$L(\mathbb{E}^x[g(f_1)] - g(x)) = -\frac{1}{2} \nabla g(x) \cdot x + \frac{1}{2} \sum_{i=1}^{m} \frac{\partial^2 g}{\partial^2 x_i} \frac{\|x\|^2}{m}$$

Using the previous theorem, we obtain the first result of this proposition. We deduce the second result thanks to proposition 4.7 that is stated later. □

4 Spectrum of the Backward to feature Kernel (BFK)

4.1 Definitions and notations

Definition 4.1. *Spectral moments*
For a symmetric positive matrix $K \in \mathbb{R}^{m \times m}$, the p-th spectral moments of K is defined by

$$M_p(K) = \frac{1}{m} \sum_{i=1}^{m} \lambda_i^p$$

where $\lambda_1, \cdots, \lambda_m \geq 0$ are the eigenvalues of K.

We will denote for $v \leq \ell$, $J_{v \to \ell}$ the jacobian matrix of the function $f_v \to f_\ell$ and $J_\ell = J_{0 \to \ell}$.

The backward pass consists of the computation of all gradients $b_\ell := (\frac{\partial \mathcal{L}}{\partial f_\ell})^T$ and $\bar{b}_\ell = (\frac{\partial \mathcal{L}}{\partial \bar{f}_\ell})^T$. Using the chain rule, we can compute all these gradients one by one going backward from the loss (hence this name backward pass), they are given by $b_L = 1$ and for all $\ell \in [\![1, L-1]\!]$:

$$\bar{b}_L = \nabla \text{loss}(\bar{f}_L), \qquad b_\ell = \frac{1}{\sqrt{L}} D_\ell \bar{b}_{\ell+1} \qquad \bar{b}_{\ell+1} = \frac{1}{\sqrt{L}} W_{\ell+1}^T b_{\ell+1}$$

where D_ℓ is the diagonal matrix with $\phi'(f_\ell)$ on the diagonal.

4.2 Motivation and general strategy

The main goal is to study how the features would evolve towards minimizing the loss after one step of gradient descent. One gradient descent step with a learning rate $\eta > 0$ updates every weight matrix W_ℓ by adding $\delta W_\ell = -\eta \nabla_\ell$, where ∇_ℓ is the matrix $(\frac{\partial \mathcal{L}}{\partial W_\ell^{(i,j)}})_{i,j} = \frac{1}{\sqrt{L}} b_\ell \phi(f_{\ell-1})^T$. The first order feature update on f_ℓ caused by an update of W_v for $v \leq \ell$ is :

$$\delta_v f_\ell = -\frac{\eta}{\sqrt{L}} J_{v\to\ell} \delta W_v \phi(f_{v-1}) \tag{4}$$

And the total feature update is given by

$$\delta f_\ell = \sum_{v \leq \ell} \delta_v f_\ell$$

If we could choose how the updates on the weights affect the features, we would like f_ℓ to not completely move in the orthogonal direction of b_ℓ, so that the features moves towards minimizing the loss. However, we do not really want them to move completely in the direction of b_ℓ either, since that would mean that we lose some of the complexity of the network. This argument is not that natural to understand, but basically the training on the weights should not be completely equivalent to directly training the features, since then we would lose the interest of having a lot of layers, and we could just train on one layer. That is why we look at the Backward Feature Angle (BFA) at feature f_ℓ, which is defined as followed.

Definition 4.2. *Backward to Feature Angle (BFA)*
The Backward to Feature Angle at feature f_ℓ is defined as

$$\widehat{b_\ell, \delta f_\ell} = arccos(\frac{\|\delta^a f_\ell\|_2}{\|\delta f_\ell\|_2}) \tag{5}$$

where $\delta^a f_\ell$ is the projection of δf_ℓ on span$\{b_\ell\}$.

We now introduce the main object of this study.

Definition 4.3. *Backward to feature Kernel (BFK)*
For $1 \leq v \leq \ell \leq L$, the partial BFK $K_{v\to\ell}$ is defined by

$$K_{v\to\ell} = \frac{\eta}{\sqrt{L}} \|\phi(f_{v-1})\|_2^2 J_{v\to\ell} J_{v\to\ell}^T \tag{6}$$

The (total) BFK is $K_\ell = \sum_{v\leq\ell} K_{v\to\ell}$

Combining this to (4) and the chain rule $b_v = J_{v\to\ell} b_\ell$, we obtain $\delta f_\ell = -K_\ell b_\ell$, hence its name of kernel.

With all these definitions, we can finally give the main theorem that motivated this study (Chizat and Netrapalli [2024]).

Theorem 4.4. *BFK spectrum and Alignment*
Let M_p be the p-th spectral moment of K_ℓ. If b_ℓ is Gaussian and independent of K_ℓ it holds, as $m_\ell \to \infty$,

$$cos(\widehat{b_\ell, \delta f_\ell}) = \frac{\|\delta^a f_\ell\|_2}{\|\delta f_\ell\|_2} \xrightarrow{pr.} \frac{M_1}{\sqrt{M_2}}$$

as soon as $\frac{\sqrt{M_2}}{M_1}$ and $\frac{\sqrt{M_4}}{M_2}$ are uniformly bounded (i. e. they are upper bounded by $C > 0$ with probability goinng to 1 as the width $m \to \infty$).

In our case, we do not study directly the spectrum of the BFK, but instead we look at the spectrum of $J_{v\to\ell}J_{v\to\ell}^T$, which is still indicative of the BFA thanks to the next proposition (Chizat and Netrapalli [2024]).

Proposition 4.5. *Let $M_p^{(v)}$ be the p-th spectral moment of $J_{v\to\ell}J_{v\to\ell}^T$ and suppose that $\frac{\sqrt{M_2^{(v)}}}{M_1^{(v)}}$ and $\frac{\sqrt{M_4^{(v)}}}{M_2^{(v)}}$ are uniformly bounded. Suppose that b_ℓ is Gaussian, independent from $K_{v\to\ell}$ for $v \le \ell$, and that $\|\delta^a f_v\|_2$ does not depend on v. Then as $m \to \infty$*

$$cos(\widehat{b_\ell, \delta f_\ell}) \ge \left(\frac{1}{\ell}\sum_{v\le\ell}\frac{\sqrt{M_2^{(v)}}}{M_1^{(v)}}\right)^{-1} + o(1)$$

In the rest of this section, we will try to utilize (1) to study the spectrum of $J_{v\to\ell}J_{v\to\ell}^T$ in the limit $m, L \to \infty$, mainly the first moment in the first type of Resnet that we introduced. More precisely, we will first derive a SDE on the Jacobian matrix. Then, we introduce a new vector $v_t = J_t z$ where $z \sim \mathcal{N}(0, Id)$, and use the proposition 4.7 in order to have a more convenient form. With this, we will try to study its convergence in the limit $m, L \to \infty$ and discuss how simplified situations would allow us to use some classical results of stochastic calculus, giving us information on the actual limit of the first moment.

Some technical points will not be discussed thoroughly, but proofs can be found in Vardhan [2007], Weinan E [2021], Pavliotis [2014] for example.

4.3 First moment of the Jacobian matrix

To simplify, we will only look at the case $v = 1$ when we study the jacobian matrix. Since we are interested in $J_{v\to\ell}$, we first derive a SDE on its limit J_t directly deduced from (1). From now on, we will only write f_t instead of f_t^{lim}.

Proposition 4.6. *We denote by J_t the jacobian matrix of the function $f_0 \to f_t$. Then J_t is the solution of the following SDE :*

$$\begin{cases} dJ_t = -\frac{1}{2}J_t dt + \sqrt{\frac{2}{m}}dB_t D_t J_t \\ J_0 = Id \end{cases} \tag{7}$$

where D_t is the diagonal matrix with $\phi'(f_t)$ on the diagonal.

Proof. See Vardhan [2007]. □

Actually, we are not that interested in J_t, but more in its spectrum. That is why we introduce a new variable : $v_t = J_t z$ where $z \sim \mathcal{N}(0, Id)$. Indeed, due to the choice of z, we have that

$$\mathbb{E}[\frac{\|v_\ell\|^2}{m} \mid J_\ell] = M_1(J_\ell^T J_\ell)$$

This means that knowing how v_t converges will give us some information on the first moment. In our case, we would like to show that v_t converges toward a deterministic limit in the limit $m \to \infty$, since that would mean that $M_1(J_\ell^T J_\ell)$ also converges to a deterministic limit. This new vector coupled with f_t gives us the following coupled SDE :

$$\begin{cases} df_t = -\frac{1}{2} f_t dt + \sqrt{\frac{2}{m}} dB_t \phi(f_t) \\ dv_t = -\frac{1}{2} v_t dt + \sqrt{\frac{2}{m}} dB_t D_t v_t \end{cases} \tag{8}$$
$$\text{and} \quad \begin{cases} f_0 \sim \mathcal{N}(0, \frac{\|x\|^2}{m_0} I_m) \\ v_0 \sim z \end{cases}$$

The next proposition gives us a more convenient form of (8).

Proposition 4.7. *Let $X_t \in \mathbb{R}^m$ (resp. Y_t) be a stochastic process solution of a diffusion SDE of the form $dX_t = b(t, X_t)dt + \sigma(t, X_t)dB_t$ (resp. $dY_t = b(t, Y_t)dt + \tilde{\sigma}(t, Y_t)d\tilde{B}_t$) with $b \in \mathbb{R}^m$, $\sigma \in \mathbb{R}^{m \times n_1}$ (resp. $\tilde{\sigma} \in \mathbb{R}^{m \times n_2}$) and $B_t \in \mathbb{R}^{n_1}$ (resp. $\tilde{B}_t \in \mathbb{R}^{n_2}$).*

We suppose that :

- $\sigma(t, x)\sigma(t, x)^T = \tilde{\sigma}(t, x)\tilde{\sigma}(t, x)^T$ *for all* $t \geq 0$, $x \in \mathbb{R}^m$
- $X_0 \sim Y_0$
- B_t *(resp.* $\tilde{B}_t$*) is independant from* X_0 *(resp.* Y_0*),*

then X_t and Y_t have the same distribution over the continuous path space.

Proof. The proof can be found in Øksendal [2010]. □

In our case, we define $X_t = (f_t^T, v_t^T)^T \in \mathbb{R}^{2m}$. We can rewrite the equation (8) in order to have a more standard form :

$$dX_t = -\frac{1}{2} X_t dt + \sqrt{2}\sigma(X_t) dB_t \quad \text{and} \quad X_0 \sim \begin{pmatrix} f_0 \\ v_0 \end{pmatrix} \tag{9}$$

where $B_t \in \mathbb{R}^{m^2}$ and $\sigma(X_t) = \frac{1}{\sqrt{m}} \begin{pmatrix} \phi(f_t)^T & 0 & 0 & \dots \\ 0 & \phi(f_t)^T & 0 & \dots \\ \dots & & & \\ (\phi'(f_t) \odot v_t)^T & 0 & 0 & \dots \\ 0 & (\phi'(f_t) \odot v_t)^T & 0 & \dots \\ \dots & & & \end{pmatrix} \in \mathbb{R}^{2m \times m^2}$

we have

$$2\sigma(X_t)\sigma(X_t)^T = \frac{2}{m} \begin{pmatrix} \|\phi(f_t)\|^2 & \langle \phi(f_t), v_t \rangle \\ \langle \phi(f_t), v_t \rangle & \|\phi'(f_t) \odot v_t\|^2 \end{pmatrix} \otimes I_m$$

All the coefficients in this matrix can be expressed as a function of the empirical measure $\mu_{X_t} = \frac{1}{m} \sum_i \delta_{X_t^{(i)}}$, and we want to exploit this structure. To do that, we need the following proposition that will allow us to recognize a McKean-Vlasov equation :

Proposition 4.8. *There exist coefficients $a(X_t^{(i)}, \mu_{X_t}), b(X_t^{(i)}, \mu_{X_t}), c(X_t^{(i)}, \mu_{X_t})$ and $\tilde{\sigma}(X_t) \in \mathbb{R}^{2m \times 2m}$ such that :*

$$2\sigma(X_t)\sigma(X_t)^T = 2\tilde{\sigma}(X_t)\sigma(\tilde{X}_t)^T$$

$$\text{where } \tilde{\sigma}(X_t) = \begin{pmatrix} a(X_t^{(1)}, \mu_{X_t}) & 0 & \dots & c(X_t^{(1)}, \mu_{X_t}) & 0 & \dots \\ 0 & a(X_t^{(2)}, \mu_{X_t}) & \dots & 0 & c(X_t^{(2)}, \mu_{X_t}) & \dots \\ \dots & & & & & \\ c(X_t^{(1)}, \mu_{X_t}) & 0 & \dots & b(X_t^{(1)}, \mu_{X_t}) & 0 & \dots \\ 0 & c(X_t^{(2)}, \mu_{X_t}) & \dots & 0 & b(X_t^{(2)}, \mu_{X_t}) & \dots \\ \dots & & & & & \end{pmatrix}$$

Proof. The existence of $\tilde{\sigma}$ implies that for all $i \in [1:m]$:

$$\begin{cases} a(X_t^{(i)}, \mu_{X_t})^2 + c(X_t^{(i)}, \mu_{X_t})^2 = \dfrac{\|\phi(f_t)\|^2}{m} \\ b(X_t^{(i)}, \mu_{X_t})^2 + c(X_t^{(i)}, \mu_{X_t})^2 = \dfrac{\|\phi'(f_t) \odot v_t\|^2}{m} \\ c(X_t^{(i)}, \mu_{X_t}) \cdot (a(X_t^{(i)}, \mu_{X_t}) + b(X_t^{(i)}, \mu_{X_t})) = \dfrac{\langle \phi(f_t), v_t \rangle}{m} \end{cases}$$

We actually see that a, b, c do not even have to depend on $X_t^{(i)}$, and that this is just a system of 3 equations with 3 variables that always has a solution. We do not detail the solution, but we will assume that the solution chosen will always have coefficients not depending on $X_t^{(i)}$. □

By proposition 4.8, we can rewrite (9) as :

$$dX_t = -\frac{1}{2} X_t dt + \sqrt{\frac{2}{m}} \tilde{\sigma}(X_t) d\tilde{B}_t \quad \text{and} \quad X_0 \sim \begin{pmatrix} f_0 \\ v_0 \end{pmatrix}$$

where $\tilde{B}_t \in \mathbb{R}^{2m}$.

We have for each coordinate :

$$dX_t^{(i)} = -\frac{1}{2} X_t^{(i)} dt + \sqrt{\frac{2}{m}} \begin{pmatrix} a(X_t^{(i)}, \mu_{X_t}) & c(X_t^{(i)}, \mu_{X_t}) \\ c(X_t^{(i)}, \mu_{X_t}) & b(X_t^{(i)}, \mu_{X_t}) \end{pmatrix} \begin{pmatrix} d\tilde{B}_t^{(i)} \\ d\tilde{B}_t^{(i+m)} \end{pmatrix} \tag{10}$$

We recognize a discrete system of stochastic differential equations which limit when $m \to \infty$ will be a 2-dimensional McKean-Vlasov process. From here, we first suppose that we can find proper coefficients a, b, c such that these coefficients are lipschitz, which means that we have for 2 vectors X, Y and 2 measures μ, ν on $\mathbb{R}^m$:

$$\|a(X, \mu) - a(Y, \nu)\| \leq \|X - Y\| + \|\mu - \nu\|$$

The distance on the measure space is the Wasserstein 2-distance defined by :

Definition 4.9. *Wasserstein distance*
For $p \in]1, +\infty[$, we define the Wasserstein p-distance by

$$W_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^p \mathrm{d}\pi(x, y) \right)^{1/p}$$

where $\Pi(\mu, \nu)$ is the set of measures on $\mathbb{R}^m \times \mathbb{R}^m$ which marginals are μ and ν.

With this supposition, we can use some standard theorem on McKean-Vlasov process (see Chaintron and Diez [2022] for example) to deduce the next proposition.

Proposition 4.10. *In the limit $m \to \infty$, the measure μ_{X_t} converges in distribution towards μ_t, where μ_t is the probability measure of X_t^{lim} solution of the following SDE :*

$$dX_t^{lim} = -\frac{1}{2} X_t^{lim} dt + \sqrt{2} \Sigma(X_t^{lim}, \mu_t) dB_t \quad \text{and} \quad X_0^{lim} \sim \mathcal{N}(0, \Sigma_0) \tag{11}$$

where $X_t^{lim}, B_t \in \mathbb{R}^2$, and
$\Sigma(X_t^{lim}, \mu_t) = \begin{pmatrix} a(X_t^{lim}, \mu_t) & c(X_t^{lim}, \mu_t) \\ c(X_t^{lim}, \mu_t) & b(X_t^{lim}, \mu_t) \end{pmatrix}$ *and* $\Sigma_0 = \begin{pmatrix} \frac{\|x\|^2}{m_0} & 0 \\ 0 & 1 \end{pmatrix}$.
Furthermore, the equations satisfied by $a(X_t^{lim}, \mu_t), b(X_t^{lim}, \mu_t), c(X_t^{lim}, \mu_t)$ become :

$$\begin{cases} a(X_t^{lim},\mu_t)^2 + c(X_t^{lim},\mu_t)^2 = \mathbb{E}[\phi(f_t^{lim})^2] \\ b(X_t^{lim},\mu_t)^2 + c(X_t^{lim},\mu_t)^2 = \mathbb{E}[(\phi'(f_t^{lim})\cdot v_t^{lim})^2] \\ c(X_t^{lim},\mu_t)\cdot(a(X_t^{lim},\mu_t)+b(X_t^{lim},\mu_t)) = \mathbb{E}[\phi(f_t^{lim})\cdot v_t^{lim}] \end{cases} \tag{12}$$

Here, the convergence is the convergence in law of μ_{X_t} (which is a random variable taking values in the space of probability measures) towards μ_t (which is deterministic). As a corollary, we have :

Theorem 4.11. *We have* $\frac{\|v_t\|^2}{m} \xrightarrow{\mathcal{L}} \mathbb{E}[(v_t^{lim})^2]$.

As mentionned earlier, this would mean that the first moment also converges towards a deterministic variable. We will discuss in the next section what that variable could be.

4.4 Potential improvements

In the last section, we supposed that our coefficients were regular enough. So a natural next step would be to look for cases where we can find such regular coefficients. Unfortunately, the system of equation that the coefficients q, b, c satisfy is a bit too complex and I could not find reasonable simplifications (on ϕ for example) that could help finding the right regularity.

We could also try to know what is actually the deterministic limit in the theorem 4.10. We can actually calculate it by supposing that X^{lim} is actually a gaussian variable. In order to simplify (11) and delete the drift term, we introduce a new process $\tilde{X}_t^{lim} = \exp(\frac{t}{2})X_t^{lim}$. Then $\tilde{X}_t^{lim}$ is the solution of the new SDE :

$$\begin{aligned} d\tilde{X}_t^{lim} &= X_t^{lim} d(\exp(\frac{t}{2})) + \exp(\frac{t}{2}) dX_t^{lim} \\ &= \frac{1}{2}\exp(\frac{t}{2})X_t^{lim}dt - \frac{1}{2}\exp(\frac{t}{2})X_t^{lim}dt + \sqrt{2}\exp(\frac{t}{2})\Sigma(X_t^{lim},\mu_t)dB_t \\ &= \sqrt{2}\exp(\frac{t}{2})\Sigma(X_t^{lim},\mu_t)dB_t \end{aligned} \tag{13}$$

If we integrate (13) and take the expected value of the $(\tilde{v}_t^{lim})^2 = \exp(t)(v_t^{lim})^2$, we obtain :

$$\tilde{X}_t^{lim} = \tilde{X}_0^{lim} + \sqrt{2}\int_0^t \exp\left(\frac{s}{2}\right)\left(c(X_s^{lim},\mu_s)\,\mathrm{d}B_s^{(1)} + b(X_s^{lim},\mu_s)\,\mathrm{d}B_s^{(2)}\right)$$

$$\begin{aligned} \mathbb{E}[(\tilde{v}_t^{lim})^2] &= \mathbb{E}[(\tilde{v}_0^{lim})^2] + 2\mathbb{E}\left[\left(\int_0^t \exp\left(\frac{s}{2}\right)\left(c(X_s^{lim},\mu_s)\,\mathrm{d}B_s^{(1)} + b(X_s^{lim},\mu_s)\,\mathrm{d}B_s^{(2)}\right)\right)^2\right] \\ &\quad + 2\sqrt{2}\mathbb{E}[\tilde{v}_0^{lim}\int_0^t \exp\left(\frac{s}{2}\right)\left(c(X_s^{lim},\mu_s)\,\mathrm{d}B_s^{(1)} + b(X_s^{lim},\mu_s)\,\mathrm{d}B_s^{(2)}\right)] \qquad (14a) \\ &= 1 + 2\int_0^t \exp(s)\mathbb{E}[(\phi'(f_s^{lim})\cdot v_s^{lim})^2]\,\mathrm{d}s \qquad (14b) \\ &= 1 + 2\int_0^t \mathbb{E}[(\phi'(\tilde{f}_s^{lim})\cdot \tilde{v}_s^{lim})^2]\,\mathrm{d}s \qquad (14c) \end{aligned}$$

We obtain (14b) from (14a) by remarking that $\tilde{X}_0^{lim}$ is a centered variable and independent from B_t (that's why the last term of equality (14a) is 0) and using Itô isometry.

Now, $\tilde{X}_t^{lim}$ is a gaussian vector of mean 0, so uniquely characterized by its matrix of covariance.

That's why we can replace $(\tilde{f}_t^{lim}, \tilde{v}_t^{lim})$ by $(\sqrt{\mathbb{E}[(\tilde{f}_t^{lim})^2]}Z_1, \sqrt{\mathbb{E}[(\tilde{v}_t^{lim})^2]}(c_t Z_1 + \sqrt{1-c_t^2}Z_2))$, where $Z_1, Z_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and c_t is the correlation between $\tilde{v}_t^{lim}$ and $\tilde{f}_t^{lim}$. This allows us to calculate :

$$\begin{aligned}\mathbb{E}[(\phi'(\tilde{f}_s^{lim}) \cdot \tilde{v}_s^{lim})^2] &= \mathbb{E}[\phi'(\sqrt{\mathbb{E}[(\tilde{f}_s^{lim})^2]}Z_1) \cdot \mathbb{E}[(\tilde{v}_s^{lim})^2](c_s Z_1 + \sqrt{1-c_s^2}Z_2)^2] \\ &= \frac{1}{2}\mathbb{E}[(\tilde{v}_s^{lim})^2]\end{aligned}$$

Substituting this result in (14c), we obtain an ODE on $y(t) = \int_0^t \mathbb{E}[(\tilde{v}_s^{lim})^2]\,\mathrm{d}s$:

$$\begin{cases} y'(t) = 1 + y(t) \\ y(0) = 0 \end{cases}$$

which gives us $\mathbb{E}[(\tilde{v}_t^{lim})^2] = y'(t) = \exp(t)$. By doing a similar reasoning, we also have : $\mathbb{E}[\phi'(\tilde{f}_t^{lim})^2] = \frac{\exp(t)}{2}$, $\mathbb{E}[\phi(\tilde{f}_t^{lim})\tilde{v}_t^{lim}] = \frac{\exp(t)}{2}$. By plugging these values into the system (12), we have two solutions : $a = b = c = \frac{1}{2}$ or $a = b = c = -\frac{1}{2}$ (note : it will depend on how we choose the solutions for the system of equations we had before taking the limit $m \to \infty$. By the proposition 4.7, having $\frac{1}{2}$ or $-\frac{1}{2}$ does not change the law of the solution of the SDE.)

So the limit SDE should just be :

$$dX_t^{lim} = -\frac{1}{2}X_t^{lim}dt + \frac{\sqrt{2}}{2}\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}dB_t \quad \text{and} \quad X_0^{lim} \sim \mathcal{N}(0, \Sigma_0)$$

which is a well-defined process for which we know admits a strong solution. Furthermore, we can also check that the diffusion coefficients of this equation satisfy the system (12). Though the value 1 was found using a lot of assumptions, we will see in the next section that it is not a random value that we came across.

5 Numerical experiments

5.1 On the first moment

During this internship, a lot of our work were backed by numerical experiments that would either infirm or confirm our hypothesis.

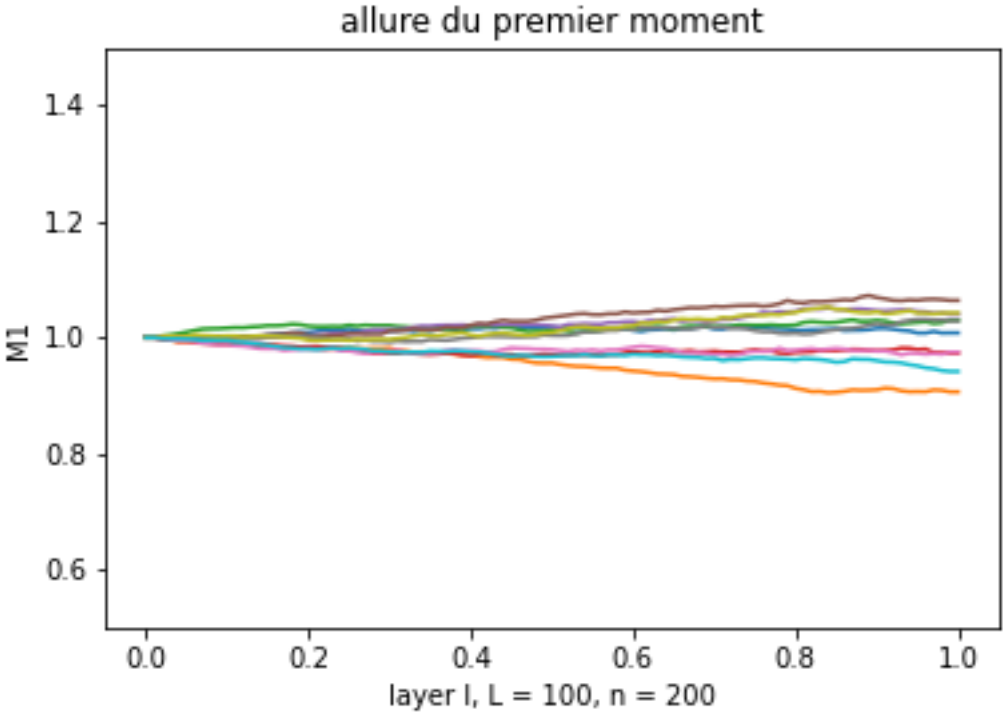

Figure 3: Different plots of the first moment

Also, at the very end of the internship, we stumbled upon Carrillo et al. [2021] that seemingly treats the exact same problem that us, but with a very different approach that uses coupling methods to prove the convergence. I actually did read a lot of litterature in order to see if I could apply any of the classical methods to prove the propagation of chaos (see Chaintron and Diez [2022] for example for an overview of those methods), but the contexts always seemed too different to be applied for me.

5.2 On the second moment

Using numerical experiments, we also get tried to get an overview of what the second moment would look like, and how the value $\left(\frac{1}{\ell}\sum_{v\leq\ell}\frac{\sqrt{M_2^{(v)}}}{M_1^{(v)}}\right)^{-1}$ would fluctuate. I actually plotted a lot of graphs in order to try to get a hunch of how the second moment would look like, but I did not find anything concluant. As a matter of example, a typical plot of the second moment would look like this :

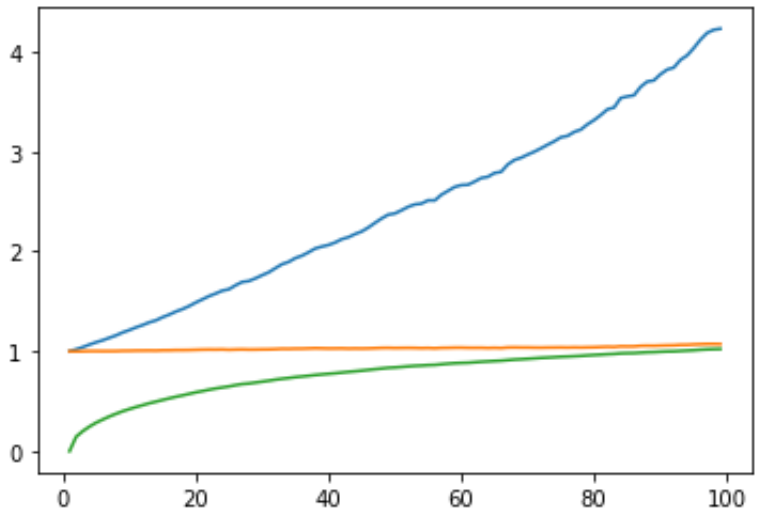

Figure 4: Blue : second moment
Orange : first moment
Green : $\frac{\sqrt{M_2}}{M_1}$

Unfortunately, the study of the second moment did not go any further than some really superficial calculations due to lack of time and the complexity of even the first moment.

References

J. A. Carrillo, F. Hoffmann, A. M. Stuart, and U. Vaes. Consensus based sampling, 2021. URL https://arxiv.org/abs/2106.02519.

Louis-Pierre Chaintron and Antoine Diez. Propagation of chaos: A review of models, methods and applications. i. models and methods. *Kinetic and Related Models*, 15(6):895, 2022. ISSN 1937-5077. doi: 10.3934/krm.2022017. URL http://dx.doi.org/10.3934/krm.2022017.

Lénaïc Chizat and Praneeth Netrapalli. The feature speed formula: a flexible approach to scale hyper-parameters of deep neural networks, 2024. URL https://arxiv.org/abs/2311.18718.

Soufiane Hayou. On the infinite-depth limit of finite-width neural networks. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=RbLsYz1Az9.

Grigorios A. Pavliotis. *Stochastic Processes and Applications*. Springer Berlin, Heidelberg, Oct. 2014.

S. R. S. Vardhan. *Stochastic processes*. American Mathematical Society, Oct. 2007. ISBN 978-08218408567.

Eric Vanden-Eijnden Weinan E, Tiejun Li. *Applied Stochastic analysis*. American Mathematical Society, Oct. 2021. ISBN 978-08218408567.

Bernt Øksendal. *Stochastic Differential Equations*. Springer Berlin, Heidelberg, Nov. 2010. ISBN 978-3-642-14394-6.