كلية العلوم و التقنيات فاس
ⵜⴰⵖⵙⵓⵍⵜ ⵏ ⵜⵇⴰⵙⵙⵉⵡⵉⵏ ⴷ ⵜⵓⵙⵙⵔⵜⵉⵏ
Faculté des Sciences et Techniques de Fès

جامعة سيدي محمد بن عبد الله
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵙⵉⴷⵉ ⵎⵓⵃⴰⵎⴻⴷ ⴱⵏ ⵄⴱⴷⵍⵍⴰⵀ
Université Sidi Mohamed Ben Abdellah

DEPARTEMENT DES MATHEMATIQUES

Master Mathématiques et Applications au Calcul Scientifique (MACS)

MEMOIRE DE FIN D'ETUDES

Pour l'obtention du Diplôme de Master Sciences et Techniques (MST)

Théorie d'approximation des Réseaux de Neurones Artificiels, Application à la résolution des équations différentielles fractionnaires

Réalisé par: Salim El azami El idrissi

Encadré par: Pr.ETTAOUIL Mohamed

Soutenu le :20/06/2019

Devant le jury composé de:

- Pr. ETTAOUIL Mohamed	**FST Fès**
- Pr. EZZAKI Fatima	**FST Fès**
- Pr. YOUSSFI Ahmed	**ENSA Fès**
- Pr. EL KHAOULANI EL IDRISSI Rachid	**FST Fès**
- Pr. HADDOUCH Khalid	**ENSA AL-Hoceïma**
-Dr. RAMCHOUN Hassan	**Invité**

Année Universitaire : 2018 / 2019

Dédicaces

À mes parents :

Aucun mot ne pourra exprimer tous mes sentiments d'amour et de gratitude, je vous remercie beaucoup pour vos efforts le long de ces années, pour votre présence rassurante et pour tout l'amour que vous m'avez réservé, vous étiez et vous resterez pour moi les parents idéals qui m'ont toujours conseillé et orienté dans le bon sens, peut-être que j'ai encore des réserves sur des décisions que vous avez préféré de les prendre à ma place, mais soyez sûr que je n'ai jamais perdu ma croyance en votre sagesse que je la considère jusqu'à l'instant indiscutable.

chers parents, je vous remercie une autre fois pour vos sacrifices , que Dieu vous prête la longue vie et la bonne santé afin que je puisse vous rendre la pareille à mon tour,je vous aime très fort ♡.

À mes chers frères et sœurs :

Qui ont toujours été là pour moi, avec leur support leurs encouragements.
Qu'ils trouvent dans ce travail, toute ma reconnaissance et tout mon grand amour envers eux.
Je vous aime beaucoup

À tout les membres de ma famille :

En témoignage de profonds liens qui nous unissent, veuillez trouver à travers ce travail l'expression de mon grand amour, mon attachement et ma profonde reconnaissance.

Remerciements

Au terme de ce travail, je voudrais exprimer mes remerciements et ma profonde reconnaissance à tout ceux qui ont contribué de prêt ou de loin à sa réalisation.

Je voudrais, en tout premier lieu, exprimer ma profonde gratitude à mon encadrant, le professeur **ETTAOUIL MOHAMED**, d'avoir accepté de m'initier à la recherche et de diriger ce travail. Comme tous ceux qui ont eu la chance d'être sous sa direction, j'ai pu constater à quel point il est attachant, attentionné et profondément gentil, il restera pour moi un modèle, pour ses grandes compétences scientifiques, pour son exigence de clarté, sa franchise et ses qualités humaines. Je lui adresse mes remerciements les plus chaleureux, pour tout ce qu'il m'a appris durant la période pendant laquelle s'est déroulé ce travail.

j'aimerais bien adresser des chaleureux remerciements à tous les membres du jury :

— Madame **FATIMA EZZAKI** , Professeur et directrice du laboratoire LMCS à la Faculté des Sciences et Techniques de Fès.
— Monsieur **EL KHAOULANI EL IDRISSI RACHID** Professeur à la faculté des sciences et techniques de Fès.
— Monsieur **HADDOUCH RACHID** Professeur à l'Ecole Nationale des Sciences Appliquées d'Al-Hoceïma.
— et Monsieur **YOUSSFI AHMED** Professeur à l'Ecole Nationale des Sciences Appliquées de Fès.

qui m'ont fait l'honneur de bien vouloir juger ce travail et de l'enrichir par leurs remarques et leurs critiques.

Je tiens à remercier également mon cher professeur le docteur **HASSAN RAMCHOUN**, et qui n'a épargné ni temps ni effort pour m'aider à concrétiser ce modeste travail.

Je tiens également à remercier le corps professoral de la Faculté des Sciences et Techniques de Fès pour leur contribution à ma formation et pour leurs enseignements précieux et multidisciplinaires.

Enfin, un merci très spécial à mes deux collègues **BRAHIM BEN DABIHI** et **HAKIM HABRI** pour leurs aides précieuses, leur gentillesse et leurs encouragement, tout au long de la préparation de ce mémoire.

Sommaire

Résumé

Les équations différentielles fractionnaires (EDFs) apparaissent actuellement dans les différents domaines scientifiques, comme la physique, l'ingénierie, la médecine, l'électrochimie, la théorie du contrôle, … etc, et l'efficacité remarquable quelles présentent, réside essentiellement dans leur grande capacité de modéliser les différents phénomènes du monde réel, ce fait va motiver et encourager, beaucoup de chercheurs à les étudier attentivement, et dans les deux côtés théorique et applicatif.

L'objectif de ce mémoire est de proposer une approche numérique convenable pour la résolution des équations différentielles fractionnaires ordinaires, et selon laquelle on essayera de franchir l'invalidité des schémas numériques classiques pour ce types d'équations, à travers un recours au domaine de l'intelligence artificielle.

Nous étudions tout d'abord et dans un cadre général, l'aspect d'apprentissage automatique, et au cours duquel on présentera les réseaux de neurones artificiels et leurs algorithmes d'apprentissage, puis on passera à la justification de la grande aptitude de ces réseaux, et ceci en faisant appel à la théorie d'approximation, qui va s'appuyer sur plusieurs théorèmes fondamentaux d'analyse fonctionnelle pour fournir des résultats sur leur capacité à approcher une certaine fonction donnée, enfin ces résultats vont nous servir pour approcher la solution de notre équation différentielle, et que l'on fera pratiquement à l'aide des algorithmes d'apprentissage, ce qui rend par conséquent ces réseaux, des générateurs de schémas itératifs pour la résolution numérique de ce type de problèmes.

Mots clés : Apprentissage automatique**,** Réseaux de Neurones Artificiels**,** Perceptron**,** Algorithme de Rétropropagation**,** Théorie d'approximation**,** Dérivation fractionnaire**,** Équations différentielles fractionnaires.

Liste des abréviations

- **"ADALINE" :**ADAptive LInear NEuron.
- **"C.P-neurone" :**neurone de Mc.Culloch et Pitts.
- **"EDF" :** Equation Différentielle Fractionnaire
- **"LMS" :** Least Mean Squares.
- **"MADALINE" :** Many ADALINE.
- **"MV" :** Maximum de Vraisemblance.
- **"PMC" :** Perceptron Multi-Couches.
- **"RNA" :** Réseaux de Neurones Artificiels.
- **"SVM" :** Support Vector Machine.

Liste des figures

Liste des algorithmes

Liste des tableaux

Introduction générale

La classe des équations différentielles d'ordre fractionnaire s'est marquée par leur capacité de modéliser les différents problèmes et phénomènes du monde réelle, et aussi par le comportement de ses solutions qui sont très difficile à trouver, à décrire et à comprendre, pour cela on va proposer, et au cadre de ce mémoire, une approche qui se base sur des modèles d'apprentissage, pour estimer cette solution sur un domaine d'étude donné.

Dans ce contexte le modèle le plus classique qui peut être proposé, sera celui des réseaux de neurones artificiels, la particularité de ces derniers et qui occupent depuis environ soixante ans une place notable dans la recherche scientifique réside dans le fait que chaque structure régulière peut être considérée comme un approximateur universel, ce qui fait, que ce modèle est bien approprié au problème mentionné précédemment et qu'il peut estimer avec une précision approximative élevée la fonction inconnue sur le domaine dont on dispose.

Notre objectif dans la suite est de justifier d'abord le fait que ce modèle est un approximateur universel, et ensuite l'utiliser pour résoudre les problèmes d'équations différentielle fractionnaires .

Ce mémoire s'articule principalement autour de quatre chapitres :

Le premier chapitre est consacré à une étude détaillée de quelques théorèmes d'analyse fonctionnelle, comme celui d'Egorov, de Lusin, de Dini et de Stone-Weierstrass,$\cdots$etc, et qui vont nous servir par la suite pour faire l'étude de la théorie d'approximation dans le cadre des réseaux de neurones artificiels.

Le deuxième est une présentation générale de l'apprentissage automatique, et qui va englober dans un premier temps une introduction aux réseaux de neurones artificiels, et au cours de laquelle on traitera leur histoire, leurs définitions,leurs types, leurs différentes structures, ainsi que leurs procédures d'apprentissage, et dans un second nous ferons notre premier contact avec le modèle neuronal le plus célèbre, et qui sera nommé "perceptron multicouches", nous allons définir sa nature,sa topologie, ses variantes, et enfin son algorithme d'apprentissage et qui sera appelé l'algorithme de rétro-propagation de gradient.

Le troisième concerne la théorie d'approximation, et qui va établir que les réseaux de neurones artificiels feed forward à une seule couche cachée**,**sont des approximateurs universels qui peuvent approcher et **,**au degré souhaité de précision**,** plusieurs types de fonctions**,** et on verra en outre, que la capacité de ces réseaux **,** dépend essentiellement et d'un point de vue théorique **,** du type de la fonction d'activation choisie et du nombre des nœds adoptés**.**

Le dernier chapitre va se baser sur les résultats de la théorie d'approximation, pour appliquer les réseaux de neurones artificiels à la résolution des équations différentielles fractionnaires ordinaires.
Dans une première partie nous allons évoquer le cadre théorique de la dérivation fractionnaire au sens de Caputo , et dans une deuxième on présentera la forme de notre problème, la façon de sa discrétisation, et la procédure de sa résolution par un processus d'apprentissage , enfin, un exemple de test sera présenté pour mieux illustrer la méthodologie proposée.

Chapitre I

Quelques outils d'analyse fonctionnelle pour la théorie d'approximation :

Résumé :

Dans ce premier chapitre on va représenter et étudier d'une façon bien détaillée, ***quelques théorèmes généraux d'analyse fonctionnelle***, surtout ceux qui ***vont nous servir pour la suite*** dans le cadre des réseaux neuronaux artificiels (théorème de Lusin, de Stone-Weierstrass, de Dini... etc), et pour cela nous incitons le lecteur d'être un peu ***courageux***,et ***curieux*** lors de sa lecture de cette partie, car elle est très importante, et elle rendra la théorie d'approximation, qui sera traitée ultérieurement dans le chapitre III, beaucoup ***plus claire et compréhensible***.

Mots clés : Densité **,**approximation**,** la presque continuité**,** prolongeabilité.

Introduction :

Dans la pratique, on se pose souvent face à des problèmes où l'inconnue est une fonction appartenant à un certain espace fonctionnel déterminé, par exemple celui des fonction continues, dérivables, de classe C^∞...etc, et dans ce cas là l'analyse usuelle qui porte essentiellement sur l'étude des $\mathbb{R}$ ou $\mathbb{C}$ espaces vectoriels topologiques [b] de dimension finie ou dénombrable devient complètement inutile.

Cette incapacité face à la manipulation des fonctions et qui sont généralement des objets mathématiques d'une nature très complexe et indescriptible, va nous obliger de faire le recours au domaine d'analyse fonctionnelle, et qui peut être vu tout simplement comme une extension naturelle à la dimension infinie de la géométrie euclidienne usuelle qui est en dimension finie, ce passage dimensionnel du fini à l'infini n'est pas du tout évident car ça peut causer parfois la perte du validité d'une grande partie de propriétés topologiques [c], comme il peut créer des contradiction avec l'intuition géométrique [d].

Dans les problèmes d'équations aux dérivées partielles par exemple, ou juste d'équations différentielles ordinaires, ou même d'ordre fractionnaires comme on le verra après au cours du dernier chapitre, le calcul explicite des solutions est souvent hors de portée, ce qui va nous pousser d'aller ***chercher à décrire leur structure par des sous ensembles denses dans l'espace fonctionnel adap-***

b. Comme son nom l'indique c'est un espace vectoriel munit d'une topologie qui peut être définie abstraitement comme elle peut provenir d'une norme, d'une distance, d'un produit scalaire ou même d'un ordre.

c. La continuité de toutes les formes linéaires par exemple.

d. Par exemple dans le cas de dimension infinie, on peut trouver des sous espaces vectoriels non fermés.

tés au problème posé, et dont les éléments sont d'une forme descriptible.
afin d'arriver à ce but on présentera dans ce chapitre quelque résultats généraux qui vont nous aider après à justifier l'importance des fonctions réalisables par des réseaux de neurones artificiels, et qui formeront à priori par la suite cet ensemble dense dont on a parlé .

1 Passage de l'intégrabilité à la presque continuité :

1.1 Théorème d'Egorov :

Théorème 1.1.1 :(d'Egorov,[31] :)

Soit $(\Omega, \mathscr{F}, \mu)$ un espace mesuré ***de mesure finie*** , et $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions $\mathscr{F} - \mathfrak{B}_{\mathbb{R}}$ ***mesurables*** convergeant μ-***presque partout*** vers une fonction f réelle et mesurable à son tour sur Ω.
Alors, $(\forall \varepsilon > 0)\ (\exists A \in \mathscr{F})$ tel que :$\mu(A) < \varepsilon$ et f_n converge ***uniformément*** vers f ***sur*** $\mathbf{E} \setminus \mathbf{A}$.

Preuve :

Pour tout $(k,n) \in \mathbb{N}^*$ on considère la suite des ensembles $\left(B_k^n\right)_{(k,n)\in(\mathbb{N}^*)^2}$définie par :

$$B_k^n = \bigcap_{i\geq n} \left\{ x \in E / \left| f_i(x) - f(x) \right| \leq \frac{1}{k} \right\}.$$

Pour tout $k \geq 1$ ***fixé*** , la suite $\left(B_k^n\right)_{n\in\mathbb{N}^*}$ est croissante pour l'inclusion ,donc et grâce à ***la continuité croissante de la mesure*** :

$$\mu\left(\bigcup_{n\geq 1} B_k^n\right) = \lim_{n\to\infty} \mu(B_k^n)$$

or,et comme la suite des fonctions $(f_n)_{n\in\mathbb{N}^*}$ converge simplement μ-p.p. vers f, on aura :

$$\left(\forall k \in \mathbb{N}^*\right) \ :\mu\left(\bigcup_{n\geq 1} B_k^n\right) = \lim_{n\to\infty} \mu(B_k^n) = \mu(E). \quad \star$$

donc,et car :$\mu\left(E\right) < +\infty$ on peut formuler $\star$ ainsi :

$$\left(\forall k \in \mathbb{N}^*\right)\left(\forall \delta > 0\right)\left(\exists N_{(k,\delta)} \in \mathbb{N}^*\right) \text{ tel que : } \left(\forall n \geq N_{(k,\delta)}\right) : \quad \left| \underbrace{\mu(E) - \mu(B_k^n)}_{\text{positif}} \right| = \mu(E) - \mu(B_k^n) < \delta. \quad \star\star$$

On ***fixe*** $\varepsilon > 0$, grâce à $\star\star$,et en posant $\delta := \left(\frac{\varepsilon}{2}\right)^k$ on peut trouver pour chaque $k \geq 1$ un entier $n_k := N_{\left(k,\left(\frac{\varepsilon}{2}\right)^k\right)}$ tel que : $\mu(B_k^{n_k}) \geq \mu(E) - 2^{-k}\varepsilon.$

Alors, l'ensemble : $A = \bigcup_{k\geq 1}\left(E \setminus B_k^{n_k}\right)$ convient car :

$$\begin{aligned}
\mu\left(A\right) &= \mu\left(\bigcup_{k\geq 1}\left(E \setminus B_k^{n_k}\right)\right) \\
&\leq \sum_{k\geq 1} \mu\left(E \setminus B_k^{n_k}\right) \quad \text{(grâce à la sous } \sigma - \text{additivité de la mesure)} \\
&\leq \sum_{k\geq 1} \mu\left(E\right) - \mu\left(B_k^{n_k}\right) \\
&\leq \sum_{k\geq 1} 2^{-k}\varepsilon = \varepsilon.
\end{aligned}$$

et : $A = \bigcup_{k\geq 1}\left(E \setminus B_k^{n_k}\right) = E \setminus \left(\bigcap_{k\geq 1} B_k^{n_k}\right)$,donc :$E \setminus A = \left(\bigcap_{k\geq 1} B_k^{n_k}\right) = \bigcap_{k\geq 1}\bigcap_{i\geq n_k}\left\{x \in E / \left|f_i(x) - f(x)\right| \leq \frac{1}{k}\right\}$.
d'où ce qu'il faut prouver. ■

Remarque 1.1.1 :(l'importance de finitude de la mesure) :

1. Considérons les fonctions mesurables f_n définies sur l'ensemble des réels ,muni de la tribu borélienne et la mesure de Lebesgue, par : $(\forall n \in \mathbb{N}) : f_n = \mathbb{1}_{[n,+\infty[}$.
   $\left(\text{où} \mathbb{1}_{[n,+\infty[} \text{désigne la fonction indicatrice de l'ensemble} \left[n, +\infty\right[\right)$, alors,il est clair que la suite (f_n) converge simplement (donc μ-presque partout)vers la fonction nulle, ***mais*** il n'existe aucun borélien de mesure finie sur le complémentaire duquel cette convergence est uniforme.
   donc l'hypothèse que μ ***soit fini est essentiel***.
2. Ce résultat est d'une ***grande importance***,et ça se voit à travers son apparence dans plusieurs démonstrations de théorèmes axiaux de la théorie de mesure et d'intégration,voyez par exemples les corollaires ci dessous ,mais dans la cadre d'apprentissage :il va nous servir essentiellement pour prouver le fameux théorème de Lusin.

Corollaire 1.1.1 :

On suppose μ finie. La convergence $f_n \xrightarrow[n\to\infty]{\text{p.p.}} f$ implique la convergence $f_n \xrightarrow[n\to\infty]{\mu} f$.

Preuve :

Fixons $\alpha > 0$,grâce au théorème d'Egorov :
pour tout $\varepsilon > 0$ il existe A mesurable tel que : $\mu\left(E \setminus A\right) < \varepsilon$ et $f_n \underset{n\to\infty}{\longrightarrow} f$ ***uniformément*** sur A.
On aura donc et pour n assez grand de sorte que : $\sup_{x\in A}(|f_n - f| \leq \alpha)$:[a]

$$\mu\left(\left|f_n - f\right| > \alpha\right) \leq \mu\left(\complement_E^A\right) + \mu\left(\sup_{x\in A}|f_n - f| > \alpha\right) < \varepsilon.$$

ce qui achève la preuve. ■

Corollaire 1.1.2 :(théorème de convergence bornée :)

Soit $(f_n)_{n\in\mathbb{N}^*}$ une suite de fonctions mesurables $(f_n : \mathbb{R}^d \to \mathbb{R})$ satisfaisant :

— il existe une constante $M > 0$ avec $\left|f_n\right| \leq M$ pour tout $n \geq 1$.

— il existe $E \subset R^d$ mesurable avec $\mu(E) < \infty$ et $supp(f_n) \subset E$ pour tout $n \geq 1$.

— $f_n(x) \underset{n\to\infty}{\longrightarrow} f(x)$ pour presque tout $x \in E$.

alors , la fonction limite f est ***mesurable***, satisfait : $supp(f) \subset E$,et en plus :

$$\int_{\mathbb{R}^d}\left|f_n - f\right| \mathrm{d}\mu \underset{n\to+\infty}{\longrightarrow} 0$$

ce qui fait que : $$\lim_{n\to+\infty}\int_{\mathbb{R}^d} f_n \,\mathrm{d}\mu = \int_{\mathbb{R}^d}\lim_{n\to+\infty} f_n \,\mathrm{d}\mu = \int_{\mathbb{R}^d} f \,\mathrm{d}\mu\,.$$

a. un tel n existe bien évidemment car sur A : $f_n \underset{n\to\infty}{\longrightarrow} f$ ***uniformément***.

Preuve :

D'après les hypothèses, on voit immédiatement que la fonction limite f est bornée par la même constante M. On voit aussi que f s'annule hors de E, car : $supp(f) := \overline{\{x \in E / f(x) \neq 0\}} \subset E$. soit $\varepsilon > 0$ le ***théorème d'Egorov*** nous permet de trouver un sous-ensemble mesurable $E_\varepsilon \subset E$ avec : $\mu(E \setminus E_\varepsilon) \leqslant \varepsilon$ en restriction auquel on a la convergence uniforme : $f_n/E_\varepsilon \underset{n\to\infty}{\longrightarrow} f/E_\varepsilon$

Alors sur E_ε, nous pouvons trouver un entier N_ε assez grand pour lequel :

$$n \geqslant N_\varepsilon \implies (\forall x \in E_\varepsilon) : |f_n(x) - f(x)| \leqslant \varepsilon$$

donc,toujours pour $n \geqslant N_\varepsilon$:

$$\begin{aligned}\int_E |f_n - f| \, d\mu &= \int_{E_\varepsilon} |f_n - f| \, d\mu + \int_{E\setminus E_\varepsilon} |f_n - f| \, d\mu \\ &\leqslant \varepsilon\mu(E_\varepsilon) + 2M\mu(E \setminus E_\varepsilon) \\ &\leqslant \varepsilon\mu(E) + 2M.\varepsilon \\ &\leqslant \varepsilon(\mu(E) + 2M)\end{aligned}$$

ce qui achève la preuve car : $\varepsilon > 0$ était arbitraire. ■

1.2 Théorème de densité des fonctions continues dans les espaces L^p :

Définition 1.2.1 :(la distance d'un point à une partie :)

Soit (X,d) un espace métrique et Y une de ses sous-partie ,on définit la distance entre un élément $x \in X$ et Y par : $(\forall x \in X) : d(x,Y) := \begin{cases} \inf\{d(x,y)/y \in Y\} & \text{si : } Y \neq \emptyset \\ +\infty & \text{sinon} \end{cases}$

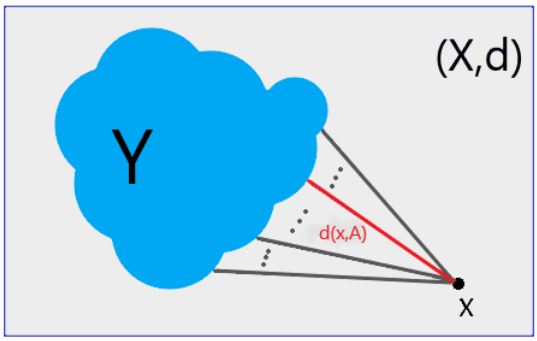


FIGURE I.1: Illustration de la distance de x à Y.

Lemme 1.2.1 :

Soit (X,d) un espace métrique, pour tout $Y \neq \emptyset$, un sous ensemble ***fixé*** de X, l'application : "$d(\cdot, Y)$" qui à tout élément : x dans X associe l'image $d(x,Y) = \inf\{d(x,y)/y \in A\}$ est bien définie, ***continue,et même 1-lipschitzienne.*** (voir :[31])

Preuve :

Soit x et t dans X,montrons que : $d(\cdot, Y)$ est 1-lipschitzienne.
prenons $\varepsilon > 0$,Par définition de $d(\cdot, Y)$ et qui est sous forme d'une borne inférieur,il existe $y_0 \in Y$ tel que :$d(x,y_0) \leqslant d(x,Y) + \varepsilon$,alors $d(t,Y) \leqslant d(t,x) + d(x,y_0) \leqslant d(t,x) + d(x,Y) + \varepsilon$, donc : $d(t,Y) \leqslant d(x,Y) + d(t,x) + \varepsilon$,et en faisant tendre ε vers 0 on obtient : $d(t,Y) \leq d(x,Y) + d(x,t)$.
D'une façon analogue on aura $d(x,Y) \leqslant d(t,Y) + d(x,y)$, donc $|d(x,Y) - d(t,Y)| \leqslant d(x,t)$
ce qui fait de $d(.,Y)$ une application 1-lipschitzienne et par suite on déduit immédiatement qu'elle est continue.

Remarque 1.2.1 :

Cette proposition servira un peu partout, par exemple pour l'approximation des ouverts par des compacts dans un espace de dimension fini, ou pour démontrer le théorème de projection sur un ***convexe complet*** [a] ,et il sera encore très utile pour approcher des fonctions intégrables par des fonctions C^∞ [b] ,et en plus, elle va apparaitre pour nous sortir vraiment d'un grand pétrin dans la preuve du théorème III.1.3.1, portant sur l'approximation des fonctions de décision par des réseaux de neurones à une seule couche cachée.

Lemme 1.2.2 :

Soit (X,d) un espace métrique, $\mathfrak{B}$ sa tribu ***borélienne*** et μ une ***mesure finie*** sur $(X,\mathfrak{B})$; pour tout borélien $B\in\mathfrak{B}$ et tout $\varepsilon>0$, il existe un fermé F et un ouvert U de X,tels que : $F\subset B\subset U$, $\underbrace{\mu(U\setminus F)=\mu(U)-\mu(F)}_{\text{car la mesure est finie.}}<\varepsilon$.

Preuve :

Notons $\mathfrak{A}$ la classe de parties de X formée de tous les $A\in\mathfrak{B}$ tel que :
pour tout $\varepsilon>0$ ***il existe*** un fermé F et un ouvert U , de sorte que $F\subset A\subset U$ et $\mu(U\setminus F)<\varepsilon$;
on va tout simplement essayer de montrer que la classe $\mathfrak{A}$ ***est une tribu qui contient tous les ouverts de*** X : on en déduira que $\mathfrak{A}=\mathfrak{B}$,ce qui est le résultat voulu.

► ***vérifions que*** $\mathfrak{A}$ ***est une tribu :***

- il est évident que $X\in\mathfrak{A}$ car : $X\subset X\subset X$ [c] et : $(\forall\varepsilon>0)$:$\mu(X\setminus X)=\mu(\emptyset)=0<\varepsilon$.
- Il est clair d'abord que $\mathfrak{A}$ ***est stable par complémentaire :***or, pour $A\in\mathfrak{A}$ le complémentaire A^c vérifie l'encadrement $U^c\subset A^c\subset F^c$,et car :$F^c\setminus U^c=U\setminus F$.,on déduit que : $\mu(U\setminus F)=\mu\left(F^c\setminus U^c\right)=\mu(U)-\mu(F)<\varepsilon$ ce qui fait :$A^c\in\mathfrak{A}$.
- ***Pour la stabilité de la réunion dénombrable :***on prend une suite $(A_n)_{n\in\mathbb{N}}\subset\mathfrak{A}$,d'éléments de $\mathfrak{A}$ on a :pour tout $\delta>0$ et $n\in\mathbb{N}$ ***il existe*** un fermé F_n et un ouvert U_n , de sorte que $F_n\subset A_n\subset U_n$ et $\mu(U\setminus F)<\varepsilon$ on prend $\varepsilon>0$ et on choisit $F_n\subset A_n\subset U_n$ pour tout $n\geqslant 0$ tels que $\mu\left(U_n\setminus F_n\right)<\varepsilon\left(\frac{1}{2}\right)^{n+2}$;on considère l'ouvert : $U=\bigcup\limits_{n\in\mathbb{N}}U_n$,on pose $A=\bigcup\limits_{n\in\mathbb{N}}A_n$ et $Y=\bigcup\limits_{n\in\mathbb{N}}F_n$; bien sur Y ***n'est pas fermé en général***, mais on a $Y\subset A\subset U$, et la relation :
$$U=\bigcup_{n\in\mathbb{N}}U_n\setminus\bigcup_{n\in\mathbb{N}}F_n\subset\bigcup_{n\in\mathbb{N}}\left(U_n\setminus F_n\right).$$
ce qui entraîne que : $\mu\left(U\setminus Y\right)\leqslant\sum\limits_{n=0}^{+\infty}\mu\left(U_n\setminus F_n\right)\leqslant\sum\limits_{n=0}^{+\infty}\varepsilon\left(\frac{1}{2}\right)^{n+2}=\frac{\varepsilon}{2}$.
Selon la continuité monotone de la mesure,on aura pour n assez grand :
$\mu(Y)-\mu(F_0,\cdots,F_n)<\frac{\varepsilon}{2}$ et le fermé $F=F_0\cup\cdots\cup F_n$ fournit alors l'encadrement $F\subset A\subset U$ avec :
$$\mu(U)-\mu(F)=\mu(U)-\mu(Y)+\mu(Y)-\mu(F)=\underbrace{\mu(U\setminus Y)}_{\text{car : }Y\subset F}+\underbrace{\mu(Y\subset U)}_{\text{car : }F\subset Y}<\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon.$$

► ***il reste à voir que*** $\mathfrak{A}$ ***contient tous les ouverts de*** **X :**

a. Attention :un convexe ***complet*** en général et pas juste fermé.
b. Ce qui est bien notre cas.
c. n'oubliez pas que ***pour la topologie induite de sa distance***,X est ***fermé*** et ***ouvert*** au même temps.

soit U un ouvert de X, on pose pour tout $n \geqslant 0$ $F_n = \left\{x \in X / d(x, U^c) \geqslant 2^{-n}\right\}$,Il est clair que F_n est fermé [a],et que $F_n \subset U$; de plus, comme U est ouvert, tout point $x \in U$ vérifie $d(x, U^c) > 0$ [b] donc appartient à F_n pour n assez grand. Il en résulte que :$U = \bigcup_{n\in\mathbb{N}} F_n$. Si on donne $\varepsilon > 0$, on aura $\mu(U) - \mu(F_n) < \varepsilon$ pour n assez grand. On choisit alors l'encadrement $F_n \subset U \subset U$ montre que $U \in \mathfrak{A}$.

donc et en guise de conclusion : $\mathfrak{A}$ est une tribu qui contient la topologie de X (c'est à dire l'ensemble de ses ouverts),alors il contiendra aussi la tribu engendrée par cette topologie et qui n'est autre que la tribu borélienne $\mathfrak{B}$ d'où $\mathfrak{B} \subset \mathfrak{A}$ et pour l'autre inclusion elle est triviale car $\mathfrak{A}$ ne contient que des ensembles mesurables,ce qui achève la preuve. ■

Théorème 1.2.1 :

De même que [60], on désignera par :$C_c^0(\mathbb{R}^d)$, l'espace des fonctions continues à support [c] compact définies sur $\mathbb{R}^d$.alors pour tout p tel que : $1 \leqslant p < +\infty$ l'espace $C_c^0(\mathbb{R}^d)$ est dense dans : $L^p\left(\mathbb{R}^d\right)$.

Preuve :

On veut approcher $f \in L^p\left(\mathbb{R}^d\right)$ par une fonction φ continue à support compact, au sens de norme L^p, et pour celá on va procéder à travers plusieurs réductions :
puisque $f = f^+ - f^-$,il suffit de le faire pour $f \geqslant 0$,or dans ce cas ,f est une limite simple d'une certaine suite croissante $(f_n)_{n\in\mathbb{N}}$ de fonctions numériques étagées mesurables positives, $0 \leqslant f_n \leqslant f$.
La suite $(f - f_n)^p$ tend vers 0 en étant dominée. par la fonction intégrable f^p,donc $\|f - f_n\|_p \underset{n\to+\infty}{\longrightarrow} 0$ par convergence dominée [d]
Il suffit donc de pouvoir approcher toute fonction étagée g,Puisque g est sous la forme d'une combinaison linéaire de fonctions de la forme $\mathbb{1}_B$ avec $B \in \mathfrak{B}_{\mathbb{R}^d}$, il suffit d'approcher les fonctions indicatrices $\mathbb{1}_B$.
Enfin, si on pose $B_n = B \cap B(0, n[$, on montre comme avant que$\mathbb{1}_{B_n}$ tend vers $\mathbb{1}_B$ dans $f \in L^p\left(\mathbb{R}^d\right)$, par convergence dominée. Finalement on veut approcher une fonction indicatrice $\mathbb{1}_B$ d'un borélien borné B.
on a $B \subset B(0, R[\subset K = \overline{B(0, R[} = B(0, R]$, introduisons aussi l'ouvert $V = B(0, R+1[$;on a $B \subset K \subset V$,désignons par μ la mesure sur $\left(\mathbb{R}^d, \mathfrak{B}_{\mathbb{R}^d}\right)$, définie par : $\left(\forall A \in \mathfrak{B}_{\mathbb{R}^d}\right) : \mu(A) = \lambda\left(A \cap V\right)$.
C'est une mesure finie à laquelle le lemme I.1.2.2 s'applique.
On trouve ainsi,un fermé F et un ouvert U de $\mathbb{R}^d$ tels que : $F \subset B \subset U$ et : $\mu(U) - \mu(F) < \varepsilon$, c'est à dire : $\lambda(U \cap V) - \underbrace{\lambda(F \cap V)}_{=\lambda(F) \text{ car } :F\subset B\subset K\subset V} < \varepsilon$ le fermé F est borné. puisque $F \subset K$, donc $K_1 = F$ est compact et l'ouvert $U_1 = U \cap V$ contient B. On a donc $K_1 \subset B \subset U_1$ et $\lambda(U_1 \setminus K_1) < \varepsilon$.
Maintenant,il reste à construire une fonction continue $\varphi \in C_c^1(\mathbb{R}^d)$, qui vérifie :

$$\|\mathbb{1}_B - \varphi\|_p := \left(\int_\Omega \left|\mathbb{1}_B - \varphi\right|^p \, \mathrm{d}\mu\right)^{\frac{1}{p}} < \varepsilon$$

a. car et ***selon le lemme*** I.1.2.1 $d(x, U^c)$ est continue et l'image réciproque d'un fermé par une fonction continue reste toujours un fermé.

b. on rappelle que dans un espace métrique (X, d) on a :$\left(\forall A \in \mathscr{P}(X)\right)\left(\forall x \in X\right) : x \in \overline{A} \iff d(x, A) = 0$.

c. On appelle support de f, noté $supp(f)$, l'adhérence de l'ensemble des points en lesquels la fonction ne s'annule pas. $supp(f) = \overline{\{x \in X \mid f(x) \neq 0\}}$.

d. je pense dans ce cas qu'on peut utiliser juste le théorème de la convergence monotone .

et pour cela,on va la choisir d'une façon,telle que : $\mathbb{1}_{K_1} \leqslant \varphi \leqslant \mathbb{1}_{U_1}$.
on pose d'abord :

$$\rho = d\left(K_1, U_1^c\right) := \inf\left\{d(x, U_1^c) / x \in K_1\right\}$$

. on a : $\rho \geqslant 0$ puisque cette quantité est la borne inférieure d'un ensemble d'éléments positifs,et en plus ,on peut remarquer que :$\rho \neq 0$,car s'il n'est pas le cas,il existera forcement [a]une suite $\left(x_n\right)_{n\in\mathbb{N}}$ d'éléments de K_1 qui vérifie : $\lim\limits_{n\to+\infty} d(x_n, U_1^c) = 0$,or K_1 est compact, donc selon le théorème de ***Bolzano-Weierstrass***, il existe une sous suite $(x_{\psi(n)})_{n\in\mathbb{N}}$ qui converge vers un un élément :$\overline{x}$ de K_1.
une autre fois,on fera appel au ***lemme 1*** qui va nous justifier l'égalité :

$$\lim_{n\to+\infty} d(x_{\psi(n)}, U_1^c) = d\left(\lim_{n\to+\infty} x_{\psi(n)}, U_1^c\right) = d(\overline{x}, U_1^c) = 0$$

et cela implique que : $\overline{x} \in \overline{U_1^c} = U_1^c$ [b],ce qui est impossible car :$K_1 \subset U_1$ d'où $\rho > 0$;alors et par conséquent on peut aisément donner un sens à la fonction h définie par :

$$h(x) := 1 - \frac{d\left(x, K_1\right)}{\rho}.$$

de nouveau,on va se reposer sur le ***le lemme 1*** pour mettre en évidence la continuité de cette fonction. et par suite on peut déduire que la fonction φ définie par :

$$\left(\forall x \in X\right) : \varphi(x) := \begin{cases} h(x) & \text{si : } h(x) \geqslant 0 \\ 0 & \text{sinon} \end{cases}$$

est continue à son tour et vérifie :$0 \leqslant \varphi(x) \leqslant 1$.
or,on remarque que : si $x \in K_1$, on a $\mathrm{d}(x, K_1) = 0$ ce qui fait que : $\varphi(x) = 1$;et si $x \notin U_1$, on aura : $d(x, K_1) \geqslant \rho$ ce qui donne : $\varphi(x) = 0$; tout ceci montre que :$\mathbb{1}_{K_1} \leqslant \phi \leqslant \mathbb{1}_{U_1}$ [c],et comme on a aussi : $\mathbb{1}_{K_1} \leqslant \mathbb{1}_B \leqslant \mathbb{1}_{U_1}$ [d] , il en résulte que : $\left|1_B - \varphi\right| \leqslant \mathbb{1}_{U_1} - \mathbb{1}_{K_1} = \mathbb{1}_{U_1 \setminus K_1}$,donc et enfin de compte :

$$\left\|\mathbb{1}_B - \varphi\right\|_p := \left(\int_\Omega \left|\mathbb{1}_B - \varphi\right|^p \, \mathrm{d}\mu\right)^{\frac{1}{p}} = \mu\left(U_1 \setminus K_1\right) < \varepsilon$$

d'où ce qu'il faut prouver. ■

1.3 Premier théorème de Lusin :

Le théorème de Lusin a été cité pour la première fois en *1903* par Henri Lebesgue, qui a remarqué,dans un éclair de génie,que ***toute fonction intégrable est presque continue***, mais cet énoncé restait considéré comme une conjecture,jusqu'au *1905* quand Giusippe Vitali a réussi de l'établir,ce qui était redécouvert et complété en *1912* par Nicolas Lusin.
L'idée de base de ce théorème, est d'affirmer que toute fonction mesurable possède une restriction ***continue*** à une ***grande*** partie de son domaine de définition.

Lemme 1.3.1 :(La réciproque partielle de théorème de la convergence dominée :)

Soit $\left(\Omega, \mathscr{F}, \mu\right)$ un espace mesuré, on note :$\mathscr{L}^1\left(\Omega, \mathscr{F}, \mu\right)$ l'espace des fonctions ***réelles*** μ-intégrables.
Soit $(f_n)_{n\in\mathbb{N}}$ une suite de$\mathscr{L}^1\left(\Omega, \mathscr{F}, \mu\right)$ et $f \in \mathscr{L}^1\left(\Omega, \mathscr{F}, \mu\right)$ telles que : $\int_\Omega \left|f_n - f\right| \, \mathrm{d}\mu \underset{n\to\infty}{\longrightarrow} 0$,alors :

1. il existe une sous-suite $\left(f_{n_k}\right)$ qui converge $\mu-$p.p. vers la limite f.
2. il existe $h \in \mathscr{L}^1\left(\Omega, \mathscr{F}, \mu\right)$ telle que :$f_{n_k} \leqslant h$ $\mu-$p.p.

a. selon la définition de la borne inférieure.
b. U est ouvert,ce qui fait que U^c est un fermé,d'où $\overline{U_1^c} = U_1^c$.
c. On peut montrer ça à travers une simple double inégalité.
d. Car : $K_1 \subset B \subset U_1$.

Preuve :

On observe que :$(f_n)_{n\in\mathbb{N}^*}$ est une suite de Cauchy dans $\mathscr{L}^1(\Omega,\mathscr{F},\mu)$,donc on peut définir une sous suite : $(f_{n_k})_{k\in\mathbb{N}}$ par récurrence de telle sorte que :

$$(\forall k\geqslant 1):\left\|f_{n_k}-f_{n_{k+1}}\right\|_1\leqslant 2^{-k}\,.$$

on pose pour tout $x\in\Omega$ et $p\geqslant 1$:

$$g_p(x):=\sum_{k=1}^{p}\left|f_{n_{k+1}}(x)-f_{n_k}(x)\right|\ \text{ et }\ g(x):=\sum_{k=1}^{+\infty}\left|f_{n_{k+1}}(x)-f_{n_k}(x)\right|\in\overline{\mathbb{R}}^{+}\,.$$

il est clair que :$\left\|g_p\right\|_1\leqslant\sum_{k=1}^{p}2^{-k}\leqslant 1$, donc d'après le ***théorème de convergence dominée***, on en déduit que :$g\in\mathscr{L}^1(\Omega,\mathscr{F},\mu)$ et :$g_n\underset{n\to\infty}{\longrightarrow}g$ dans $\mathscr{L}^1(\Omega,\mathscr{F},\mu)$, On observe aussi que pour : $l>k$:

$$\left|f_{n_l}-f_{n_k}\right|\leqslant\sum_{i=k}^{l-1}\left|f_{n_{i+1}}(x)-f_{n_i}(x)\right|=g_l-g_k\leqslant g-g_k\,.$$

Il en résulte que (f_{n_k}) est (μ-p.p.) une suite de Cauchy donc (μ-p.p) une suite convergente [a],en notant $f^*=\lim_{k\to\infty}f_{n_k}$ sa limite (μ-p.p) et en faisant tendre $l\to+\infty$ dans l'inégalité précédente, on obtient :

$$\left|f^*-f_{n_k}\right|\leqslant g-g_k\leqslant g\,.$$

En utilisant le théorème de convergence dominée pour la suite :$f^*-f_{n_k}$, on déduit que : $f_{n_k}\underset{k\to+\infty}{\longrightarrow}f^*$, Par unicité de la limite, on a donc $f^*=f$ μ p.p. et on conclut en posant $h:=g+f^*$. ■

Théorème 1.3.1 :(de Lusin,version faible :)

Soit $f:\left([a,b],\mathfrak{B}_{\mathbb{R}}\sqcap[a,b]\right)\to\left(\mathbb{R},\mathfrak{B}_{\mathbb{R}}\right)$ une fonction ***intégrable,*** alors et selon [41]**,** on a : pour tout $\varepsilon>0$, il existe un ***compact*** $E_\varepsilon\subset[a,b]$ tel que : $\lambda\left([a,b]\setminus E_\varepsilon\right)\leqslant\varepsilon$ [b]et la restriction de f à E_ε ***est continue au sens de la topologie induite*** *sur E, et pour résumer ça, on dit dans ce cas que* f ***est*** $\varepsilon-$***presque continue sur*** $[\mathbf{a},\mathbf{b}]$.

Preuve :

Selon le théorème qu'on a traité précédemment portant sur la densité des fonctions continues à support compact dans les espaces L^p, on peut garantir dans ce cas, et en posant $p=1$, qu'il existe une suite $(g_n)_{n\in\mathbb{N}}$ de fonctions continues,telle que : $g_n\xrightarrow[n\to+\infty]{}f$ dans L^1, et via le ***lemme précédent*** qui a présenté une réciproque partielle de théorème de la convergence dominée, on peut de cette suite,extraire une sous-suite $(g_{n_k})_{k\in\mathbb{N}}$ telle que : $g_{n_k}\xrightarrow[k\to+\infty]{}f$ ***presque partout*** ensuite,et en utilisant le théorème d'Egorov, on trouvera que g_{n_k} converge ***uniformément*** sauf sur un ensemble de mesure inférieure à un certain ε choisi arbitrairement vers la fonction :f, et puisque la limite uniforme [c] d'une suite de fonctions continues reste encore continue,notre preuve sera bien finie. ■

Remarque 1.3.1 :

Ce résultat sera indispensable,pour montrer le théorème III.1.3.2,et qui va établir la capacité ou l'habileté des réseaux neuronaux, ***pour l'apprentissage des fonction de décision***.

a. car $\mathbb{R}$ est complet.
b. λ désigne la mesure de Lebesgue usuelle sur $\mathbb{R}$.
c. si elle existe bien sûr.

2 Aperçu sur la densité dans l'espace des fonctions continues :

2.1 Théorème de Dini :

Théorème 2.1.1 :(de Dini :)

Soit :$(f_n)_{n\in\mathbb{N}}$ et f des fonctions ***à valeurs réelles*** définies et ***continues*** sur un espace topologique ***compact*** (X,Θ),si :$(f_n(x))_{n\in\mathbb{N}}$ est croissante pour tout $x\in X$,et si f_n converge vers f ***simplement*** alors : cette convergence sera uniforme.

Preuve :

On va reprendre pour établir ce résultat les idées de la preuve proposée,dans [47] page :272 en introduisant quelques modifications pour s'adapter au cas général des espaces compacts :
on pose d'abord :$(\forall x\in X)(\forall n\in\mathbb{N}):\ \varphi_n(x)=f(x)-f_n(x)\geqslant 0$,on peut remarquer que la suite $(\varphi_n)_{n\in\mathbb{N}}$ est ***décroissante et qu'elle converge simplement vers la fonction nulle.*** fixons un nombre réel $\varepsilon>0$ et considérons les ensembles :$\Lambda_n(\varepsilon)=\left\{x\in X/\varphi_n(x)<\varepsilon\right\}$,puisque les φ_n sont continues alors les $(\Lambda_n(\varepsilon))_{n\in\mathbb{N},\varepsilon>0}$ sont des ouverts.
on a :

$$\begin{aligned}(\forall x\in X)(\forall n\in\mathbb{N}):f_n(x)\underset{n\to+\infty}{\longrightarrow}0 &\Leftrightarrow (\forall x\in X)(\forall\varepsilon>0)(\exists\eta_{x,\varepsilon}>0)\,\text{tq}:(\forall n>\eta_{x,\varepsilon}):\left|\varphi_n(x)\right|=\varphi_n(x)<\varepsilon\\ &\Rightarrow(\forall\varepsilon>0)(\forall x\in X)(\exists\eta_{x,\varepsilon}>0)\,\text{tq}:(\forall n>\eta_{x,\varepsilon}):x\in\Lambda_n(\varepsilon)\\ &\Rightarrow(\forall\varepsilon>0)(\forall x\in X)(\exists M=[\eta_{x,\varepsilon}]+1>0)\,\text{tq}:(\forall n>\eta_{x,\varepsilon}):x\in\Lambda_M(\varepsilon)\\ &\Rightarrow(\forall\varepsilon>0):X\subset\bigcup_{n\in\mathbb{N}}\Lambda_n(\varepsilon)\end{aligned}$$

comme X est compact, on peut extraire de : $(\Lambda_n(\varepsilon))_{n\in\mathbb{N}}$ un sous-recouvrement fini,alors pour tout $\varepsilon>0$ il existe un entier N_ε tel que :$X\subset\bigcup_{n\leq N_\varepsilon}\Lambda_n(\varepsilon)$.

maintenant soit : $\varepsilon>0$,il est facile de voir que : $\left(\forall(n,m)\in\mathbb{N}^2\right):n\leqslant m\implies\Lambda_n(\varepsilon)\subset\Lambda_m(\varepsilon)$.
or : $(\forall x\in X)(\forall n\leqslant m):x\in\Lambda_n(\varepsilon)\implies\varphi_n(x)<\varepsilon\implies\varphi_m(x)\leqslant\varphi_n(x)<\varepsilon\implies x\in\Lambda_m(\varepsilon)$.
donc :$X\subset\bigcup_{n\leq N_\varepsilon}\Lambda_n(\varepsilon)=\Lambda_{N_\varepsilon}(\varepsilon)\iff X=\Lambda_{N_\varepsilon}(\varepsilon)$.
en se basant une autre fois sur l'hypothèse de la monotonie nous obtenons :
$\forall n\geq N_\varepsilon,\ \varphi_n(x)\leq\varphi_{N_\varepsilon}(x)<\varepsilon$, on déduit donc que la convergence de (f_n) vers 0 est uniforme sur X ce qui implique que $f_n\xrightarrow[n\to+\infty]{\text{unifor}}f$. ■

Remarque 2.1.1 :

le but général de théorème de Dini est d'énoncer des conditions sous lesquelles ***la convergence simple d'une suite de fonctions implique la convergence uniforme***,ce théorème peut prendre d'autre versions plus généralisées, mais dans notre cas cet énoncé est suffisant et il va nous servir par la suite.

2.2 Théorème de Stone-Weierstrass :

Soit (X,d) un espace métrique ***compact***, et $C(X,\mathbb{R})$ l'espace des fonctions continues[a] de X dans $\mathbb{R}$,muni de la norme uniforme : $\|f\|_\infty=\sup_{x\in X}|f(x)|$.
Si $f,g\in C(X,\mathbb{R})$,on note : $f\times g\in C(X,\mathbb{R})$ la fonction $x\mapsto f(x)\cdot g(x)$, ***muni de cette loi de produit, l'espace vectoriel*** $(\mathbf{C(X},\mathbb{R}),+,\cdot)$ ***devient une algèbre.***[b]

a. On rappelle que $(C(X,\mathbb{R}),+,\cdot,\|.\|)$ est un espace de Banach.
b. Une algèbre sur un corps commutatif K n'est autre qu'un $K-$espace vectoriel $(E,+,\cdot)$ muni d'une loi de composition interne bilinéaire "$\times$".

Définition 2.2.1 :

1. On dit qu'une partie $\mathfrak{S} \subset C(X,\mathbb{R})$ est une sous-algèbre si et seulement si :
$$\Big(\forall (f,g) \in C(X,\mathbb{R})^2\Big)\big(\forall \lambda \in \mathbb{R}\big) \text{ on a : } f+g \in \mathfrak{S},\ f \times g \in \mathfrak{S}, \text{ et } \lambda \cdot f \in \mathfrak{S}.$$
2. On dit que $\mathfrak{S}$ est une sous-algèbre unitaire si la fonction constante $\mathbb{1}$ appartient à $\mathfrak{S}$.
3. On dit que $\mathfrak{S}$ sépare les points de X si : $\big(\forall x \neq y \in X\big)\big(\exists f \in \mathfrak{S}\big)$ tel que : $f(x) \neq f(y)$.

Exemple 2.2.1 :

L'ensemble des fonctions polynomiales sur un segment $[a,b] \subset \mathbb{R}$ (où $a < b$) est une sous-algèbre unitaire qui sépare les points de $[a,b]$.

Remarque 2.2.1 :

1. Si $\mathfrak{S} \subset C(X,\mathbb{R})$ est une sous-algèbre unitaire, alors pour tout $f \in \mathfrak{S}$ et tout polynôme $P \in \mathbb{R}[X]$ on a : $P(f) \in \mathfrak{S}$.
2. Si $\mathfrak{S} \subset C(X,\mathbb{R})$ est une sous-algèbre unitaire, alors l'adhérence $\overline{\mathfrak{S}} \subset C(X,\mathbb{R})$ est encore une sous-algèbre unitaire (démonstration facile).

Théorème 2.2.1 :(de Stone-Weierstrass algébrique,[58] :)

Soit (X,d) un ***espace métrique compact*** et $C(X,\mathbb{R})$ l'espace des fonctions continues de X dans $\mathbb{R}$, muni de la norme uniforme , si $\mathfrak{S} \subset C(X,\mathbb{R})$ est une sous-algèbre unitaire qui sépare les points de X, alors $\mathfrak{S}$ est dense dans $C(X,\mathbb{R})$.

Preuve :

On va suivre la même démarche que :[14],où la démonstration comportera 5 étapes, dont 3 sont préliminaires. Le but des deux premières étapes est de montrer que, si $f_1, f_2, \ldots, f_n \in \mathfrak{S}$, alors : $\max\limits_{1\leqslant i\leqslant n}(f_i) \in \overline{\mathfrak{S}}$, et $\min\limits_{1\leqslant i\leqslant n}(f_i) \in \overline{\mathfrak{S}}$.

►**Étape 1** : Il existe une suite de fonctions polynomiales (P_n) sur $[-1,1]$ convergeant uniformément vers la fonction : $t \mapsto |t|$.

définissons cette suite par récurrence en posant : $\begin{cases} \big(\forall t \in [-1,1]\big) : P_0(t) = 0 \\ \big(\forall t \in [-1,1]\big) : P_{n+1}(t) = P_n(t) + \dfrac{1}{2}\Big(t^2 - P_n(t)^2\Big) \end{cases}$

Il est facile de vérifier, par récurrence, que :

$$(\forall n \in \mathbb{N})\big(\forall t \in [-1,1]\big) :\ 0 \leqslant P_n(t) \leqslant |t| \qquad (*)$$

en effet, c'est vrai pour $n = 0$ et l'hypothèse de récurrence nous fournit déjà que : $P_{n+1}(t) \geqslant 0$. D'autre part :

$$|t| - P_{n+1}(t) = |t| - P_n(t) - \frac{1}{2}\big(|t| - P_n(t)\big)\big(|t| + P_n(t)\big)$$
$$\big(\text{et car : } |t| + P_n(t) \leqslant 2|t| \leqslant 2 \text{ sur } [-1,1]\big) \quad = \big(|t| - P_n(t)\big)\left(1 - \frac{1}{2}\big(|t| + P_n(t)\big)\right) \geqslant 0$$

l'inégalité $(*)$ montre aussi que $P_{n+1}(t) \geqslant P_n(t)$, donc pour tout $t \in [-1,1]$ la suite de nombres réels $n \mapsto P_n(t)$ est croissante,et car elle est majorée alors,alors elle sera convergente. notons : $u(t) = \lim_{n\to+\infty} P_n(t)$. La formule de récurrence pour P_n donne à la limite :

$$\big(\forall t \in [-1,1]\big) : u(t) = u(t) + \frac{1}{2}\Big(t^2 - u(t)^2\Big) \text{ ,donc : } u(t)^2 = t^2 .$$

et comme $u(t) \geqslant 0$, on obtient que : $u(t) = |t|$, le théorème de Dini I.2.1.1, nous permet de déduire la convergence uniforme de P_n vers la fonction $t \mapsto |t|$.

►**Étape 2** :Si $f, g \in \mathfrak{S}$, alors les fonctions $|f|$,$\max(f,g)$,et $\min(f,g)$ appartiennent à $\overline{\mathfrak{S}}$.
montrons d'abord que, si $f \in \mathfrak{S}$, alors $|f| \in \overline{\mathfrak{S}}$. c'est évident si $f = 0$. Dans le cas contraire, on note $h = \frac{f}{\|f\|_\infty}$ et on observe que $h \in \mathfrak{S}$ et $h(x) \in [-1,1]$ pour tout $x \in X$. Ainsi, d'après l'étape 1 :

$$\frac{|f|}{\|f\|_\infty} = |h| = \lim_{n\to\infty} P_n(h). \quad \text{(la convergence étant uniforme sur } X.)$$

Comme $P_n(h) \in \mathfrak{S}$ [a] pour tout $n \in \mathbb{N}$, on conclut que : $\frac{|f|}{\|f\|_\infty} \in \overline{\mathfrak{S}}$, donc $|f| \in \overline{\mathfrak{S}}$. soient maintenant $f, g \in \mathfrak{S}$.,alors :

$$\max(f,g) = \frac{f+g+|f-g|}{2} \in \overline{\mathfrak{S}} \quad \text{et :} \quad \min(f,g) = \frac{f+g-|f-g|}{2} \in \overline{\mathfrak{S}}$$

avec une simple itération de ce résultat on obtient que :

$$\left(\forall f_1, f_2, \ldots, f_n\right) \in \mathfrak{S} : \quad \max_{1\leqslant i\leqslant n}(f_i) \in \overline{\mathfrak{S}}, \text{ et } \min_{1\leqslant i\leqslant n}(f_i) \in \overline{\mathfrak{S}}.$$

►**Étape 3** :Pour tous les $x, y \in X$ tels que $x \neq y$ et tous les $\alpha, \beta \in \mathbb{R}$, il existe $g \in \mathfrak{S}$ tel que $g(x) = \alpha$ et $g(y) = \beta$.
En effet, comme $\mathfrak{S}$ sépare les points de X, il existe $h \in \mathfrak{S}$ tel que $h(x) \neq h(y)$. on prend alors g la fonction définie par :
$g : z \in X \mapsto \alpha + (\beta - \alpha)\frac{h(z)-h(x)}{h(y)-h(x)} = \lambda \cdot h(z) + \mu$, comme $\mathfrak{S}$ contient les constantes,on a bien $g \in \mathfrak{S}$.

►**Étape 4** : soit $f \in C(X,\mathbb{R})$ et $x_0 \in X$ on a :

$$\boxed{(\forall \varepsilon > 0)\left(\exists g \in \mathfrak{S}\right) \text{ tel que : } g(x_0) = f(x_0), \text{ et } \left(\forall x \in X\right) : \ g(x) < f(x) + \varepsilon.}$$

En effet, pour chaque $y \in X \setminus \{x_0\}$, l'étape 3 fournit l'existence d'une fonction $g_y \in C(X,\mathbb{R})$ telle que : $g_y(x_0) = f(x_0)$ et $g_y(y) = f(y)$. Comme g_y et f sont continues, l'ensemble :

$$\mathcal{U}_y = \left\{z \in X / / g_y(z) < f(z) + \varepsilon\right\}$$

est un ouvert de X contenant x_0 et y. il s'ensuit que la famille $\left(\mathcal{U}_y\right)_{y\in X\setminus\{x_0\}}$ est un recouvrement ouvert de l'espace X ***qui est compact***, donc :
il existe donc des points $y_1, \ldots, y_n \in X \setminus \{x_0\}$ tels que : $X = \bigcup_{i=1}^{n} \mathcal{U}_{y_i}$.
Soit à présent $g = \min\left(g_{y_1}, \cdots, g_{y_1}\right)$. alors : $g \in \overline{\mathfrak{S}}$ d'après la première étape,et $g(x_0) = f(x_0)$. en outre,pour tout $x \in X$, il existe $i \in 1, \cdots, n$ tel que $x \in U_{y_i}$, donc $g(x) \leqslant g_{y_i}(x) < f(x) + \varepsilon$.

►**Étape 5** : soit $f \in C(X,\mathbb{R})$ on a :

$$\boxed{(\forall \varepsilon > 0)\left(\exists h \in \mathfrak{S}\right) \text{ tel que : } \left(\forall x \in X\right) : \ f(x) - \varepsilon < h(x) < f(x) + \varepsilon.}$$

en effet, pour chaque $x \in X$, l'étape 4 fournit une fonction $h_x \in \overline{\mathfrak{S}}$ telle que $h_x(x) = f(x)$ et $h_x < f + \varepsilon$ sur tout X. l'ensemble : $V_x = \left\{z \in X / f(z) < h_x(z) + \varepsilon\right\}$ est un ouvert de X contenant x, et la famille

a. on a déjà signalé ça au premier point de la remarque :I.2.2.1

$(V_x)_{x\in X}$ constitue un recouvrement ouvert de l'espace X, ***qui est compact***. donc il existe des points $x_1,\cdots,x_m \in X$ tels que :

$$X=\bigcup_{i=1}^{m} \mathcal{V}_{x_i}.$$

Soit maintenant $h=\max(h_{x_1},\cdots,h_{x_m})$.***d'après l'étape 1***, $h\in\overline{\mathfrak{S}}$, et il est clair que :$h(x)<f(x)+\varepsilon$ pour tout $x\in X$, puisque : $\left(\forall i\in\{1,\cdots,m\}\right) : h_{x_i}<f+\varepsilon$. d'autre part, pour tout $z\in X$, il existe : $i\in\{1,\cdots,m\}$ tel que :$x\in V_{x_i}$,donc : $f(z)-\varepsilon<h_{x_i}(z)\leqslant h(z)$.
d'où : $\left(\forall x\in X\right) : f(z)-\varepsilon<h_{x_i}(z)<f(z)+\varepsilon$. ■

Corollaire 2.2.1 :(Théorème d'approximation de Weierstrass :)

Si $f:[a,b]\to\mathbb{R}$ est une fonction continue, alors pour tout $\varepsilon>0$ il existe une fonction polynomiale P telle que :

$$\left\|f-P\right\|_\infty=\sup_{x\in[a,b]}\left|f(x)-P(x)\right|\leqslant\varepsilon\,.$$

ce qui fait que :

$$\left(\forall f\in C^0([a,b])\right)\left(\exists(P_n)_{n\in\mathbb{N}}\in(\mathbb{R}[X])^n\right)\text{ tel que : }P_n\xrightarrow[n\to+\infty]{\text{uniforme}}f\text{ sur }[a,b]\,.$$

Preuve :

il suffit de remarques que l'espace des fonctions polynomiales sur $[a,b]$ est une sous algèbre unitaire qui sépare les points de $[a,b]$ (car la fonction identité est en particulier polynomiale) et d'appliquer le théorème :I.2.2.1· ■

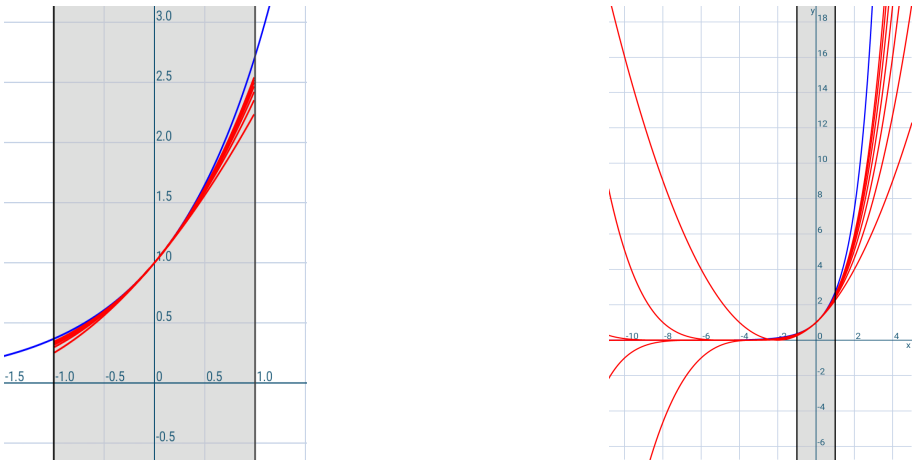


FIGURE I.2: Une suite de polynômes convergeant uniformément vers $x\mapsto\exp(x)$ sur $[-1,1]$.

Remarque 2.2.2 :

Le théorème de Stone-Weierstrass, ne reste plus valable sans l'hypothèse de compacité, voir par exemple la fonction définie sur $\mathbb{R}^+$ par : $f(x)=e^{1-x}$, et qui ne peut jamais être approchée uniformément par une suite des polynômes car :

$$\left(\forall P\in\mathbb{R}\left[X\right]\right) : \lim_{x\to+\infty}\left|P(x)\right|=+\infty\Rightarrow\lim_{x\to+\infty}\left|P(x)-f(x)\right|\geqslant\lim_{x\to+\infty}\left|P(x)\right|-\overbrace{\lim_{x\to+\infty}\left|f(x)\right|}^{=0}=+\infty.$$

3 Passage de la mesurabilité à la presque continuité :

3.1 Mesures régulières et espaces de Radon :

Définitions 3.1.1 :

Soit (X,Θ) un ***espace topologique de Hausdorff*** (appelé aussi espace séparé) [a], et ν une mesure sur la tribus Borélienne :
$\mathfrak{B}_X := \tau(\Theta)$.

1. La mesure ν est appelé ***régulière interne*** ou serré si et seulement si :
   $\left(\forall B \in \mathfrak{B}_X\right) : \ \nu(B) = \sup\left\{\nu(K) \Big/ K^c \in \Theta \ \text{ et : } B \supset K\right\}$.
2. La mesure ν est appelée ***régulière externe*** si et seulement si :
   $\left(\forall B \in \mathfrak{B}_X\right) : \ \nu(B) = \inf\left\{\nu(U) \Big/ U \in \Theta \ \text{ et : } B \subset U\right\}$.
3. La mesure ν est dite ***régulière*** si et seulement si :elle vérifie les deux points précédents au même temps.
4. La mesure ν est appelée ***localement finie*** si et seulement si :
   $\left(\forall x \in X\right)\left(\exists U \in \Theta\right) : \ x \in U$ et $\nu(U) < +\infty$.
5. La mesure ν est appelée ***mesure du Radon*** si elle est régulière et localement finie à la fois,et dans ce cas le triplet$\left(X, \mathfrak{B}_X, \nu\right)$ sera appelé un ***espace mesuré de Radon*** .
   (voir pour ces définitions :[27]).

Proposition 3.1.1 :

Soit $\left(X,d\right)$ un espace métrique,on note $\mathfrak{B}_X$ sa tribu borélienne,et on considère une mesure μ définie sur cette tribu,si μ est ***finie***,alors elle sera forcement régulière.

Preuve :

On défini l'ensemble $\mathcal{R} \subset \mathfrak{B}_X$,par :

$$\left(\forall A \in \mathcal{P}\left(X\right)\right) : \ A \in \mathcal{R} \Longleftrightarrow \begin{cases} \mu(A) = \sup\left\{\mu(K) \big/ K \text{ est fermé et } A \supset K\right\} \\ \text{et :} \\ \mu(A) = \inf\left\{\mu(U) \big/ U \text{ est ouvert et } A \subset U\right\} \end{cases}$$

on veut établir que :$\mathcal{R} = \mathfrak{B}_X$,et pour cela ,on va montrer que :$\mathcal{R}$ est une tribu et que :$\Theta \subset \mathcal{R}$.

- ***vérifions que*** $\mathcal{R}$ ***est une tribu*** :
  - il est évident que $X \in \mathcal{R}$ car :
    $\left\{U \text{ est ouvert et } X \subset U\right\} = \left\{X\right\} \implies \mu(X) = \inf\left\{\mu(U) \big/ U \text{ est ouvert et } A \subset U\right\}$
    et :$X \in \left\{K \text{ est fermé et } A \supset K\right\} \implies \mu(X) = \inf\left\{\mu(U) \big/ U \text{ est ouvert et } A \subset U\right\}$ [b]
  - Il est clair d'abord que $\mathcal{R}$ ***est stable par complémentaire :***or, soit $A \in \mathcal{R}$ et $\varepsilon > 0$, prenons un ouvert U et un fermé C de sorte que :$C \subset A \subset U$ et $\mu(C) - \varepsilon < \mu(A) < \mu(U) + \varepsilon$.
    alors, le complémentaire A^c vérifie l'encadrement $U^c \subset A^c \subset F^c$,et de plus :
    $$\mu(U^c) - \varepsilon < \mu(A^c) < \mu(C^c) + \varepsilon$$
    ce qui fait :$A^c \in \mathcal{R}$.

a. c'est un espace topologique dans lequel deux éléments différents admettent toujours des voisinages ouverts disjoints.

b. n'oubliez pas que ***pour la topologie induite de sa distance***,X est ***fermé*** et ***ouvert*** à la fois.

- **Pour la stabilité de la réunion dénombrable :**on prend une suite $(A_n)_{n\in\mathbb{N}'} \subset \mathcal{R}$, d'éléments de $\mathcal{R}$ et un certain $\varepsilon > 0$,prenons pour tout : $i \in \mathbb{N}$ un ouvert U_i un fermé : C_i tel que :$\left(\forall i \in \mathbb{N}\right)$: $C_i \subset A_i \subset U_i$ et $\mu(U_i) - 2^i\varepsilon < \mu(A_i) < \mu(C_i) + 2^{i+1}\varepsilon$.
donc : $\bigcup\limits_{i\in\mathbb{N}} C_i \subset \bigcup\limits_{i\in\mathbb{N}} A_i \subset \bigcup\limits_{i\in\mathbb{N}} U_i$,et $\bigcup\limits_{i\in\mathbb{N}} U_i$ est ouvert,(la réunion quelconque des ouverts reste un ouvert) avec :

$$\begin{aligned}\mu\left(\bigcup_{i\in\mathbb{N}} U_i\right) - \mu\left(\bigcup_{i\in\mathbb{N}} A_i\right) &\leqslant \mu\left(\bigcup_{i\in\mathbb{N}} U_i \setminus \bigcup_{i\in\mathbb{N}} A_i\right)\\ &\leqslant \mu\left(\bigcup_{i\in\mathbb{N}} \left(U_i \setminus A_i\right)\right) \leqslant \sum_{i\in\mathbb{N}} \mu\left(U_i \setminus A_i\right)\\ &= \sum_{i\in\mathbb{N}} \left(\mu\left(U_i\right) - \mu\left(A_i\right)\right) < \sum_{i\in\mathbb{N}} 2^i\varepsilon = \varepsilon.\end{aligned}$$

en outre : et par la continuité croissante de la mesure :$\mu\left(\bigcup\limits_{i\in\mathbb{N}} C_i\right) = \lim\limits_{k\to+\infty} \mu\left(\bigcup\limits_{i=1}^{k} C_i\right)$
par conséquent :
il existe un certain k pour lequel :$\mu\left(\bigcup\limits_{i\in\mathbb{N}} C_i\right) - \mu\left(\bigcup\limits_{i=1}^{k} C_i\right) < \dfrac{\varepsilon}{2}$, donc : $C = \bigcup\limits_{i=1}^{k} C_i \subset \bigcup\limits_{i\in\mathbb{N}} A_i$ est fermé avec :

$$\begin{aligned}\mu\left(\bigcup_{i=1}^{+\infty} A_i\right) - \mu\left(C\right) &< \mu\left(\bigcup_{i=1}^{+\infty} A_i\right) - \mu\left(\bigcup_{i=1}^{+\infty} C_i\right) + \frac{\varepsilon}{2}\\ &\leqslant \mu\left(\bigcup_{i=1}^{+\infty} A_i\right) \setminus \bigcup_{i=1}^{+\infty} C_i + \frac{\varepsilon}{2}\\ &\leqslant \mu\left(\bigcup_{i=1}^{+\infty} \left(A_i \setminus C_i\right)\right) + \frac{\varepsilon}{2}\\ &\leqslant \sum_{i=1}^{+\infty} \mu\left(A_i \setminus C_i\right) + \frac{\varepsilon}{2}\\ &= \sum_{i=1}^{+\infty} \left(\mu\left(A_i\right) - \mu\left(C_i\right) + \frac{\varepsilon}{2} \leqslant \frac{\varepsilon}{2} + \frac{\varepsilon}{2}\right.\end{aligned}$$

donc on obtient : $\bigcup\limits_{i=1}^{+\infty} A_i \in \mathcal{R}$, et ceci justifie la stabilité de $\mathcal{R}$ par réunion dénombrable.

► ***il reste à vérifier que*** $\mathcal{R}$ ***contient tous les ouverts de*** $\mathbf{X}$:
Afin de montrer que : $\Theta \subset \mathcal{R}$ on va essayer d'établir que $\mathcal{R}$ contient tout les fermés de X,et on se basera sur la stabilité par passage au complémentaire pour déduire ce résultat pour les ouverts.
soit $A \subset X$ un fermé ,on note $U_n := \left\{x \in X \mid \mathrm{d}\left(x, A\right) < \frac{1}{n}\right\} = \left\{x \in X \mid \left(\exists a \in A\right) : \mathrm{d}\left(x, A\right) < \frac{1}{n}\right\}$
par le lemme I.1.2.1,la fonction $\mathrm{d}\left(\cdot, A\right)$ est continue, ce qui fait que les U_n sont des ouverts,et de plus :

$$\left(\forall i \in \mathbb{N}^*\right) \ : U_{i+1} \subset U_i \ \text{,et :} \ \bigcap_{i=1}^{+\infty} U_i = A \quad \left(\text{car } A \text{ est fermé.}\right) \ \text{[a] alors et}$$

car μ est finie on aura $\mu\left(U_1\right) < +\infty$ donc et selon ***la monotonie décroissante de la mesure*** :$\mu\left(A\right) = \lim\limits_{n\to+\infty} \mu\left(U_n\right) = \inf\limits_{n\in\mathbb{N}} \mu\left(U_n\right)$
donc :$\mu\left(A\right) \leqslant \inf\left\{\mu(U)/U \text{ est ouvert et } A \subset U\right\} \leqslant \inf\limits_{n\in\mathbb{N}} \mu\left(U_n\right) = \mu\left(A\right)$.

a. On rappelle que :$x \in \overline{A} \iff d\left(x, A\right) = 0$

donc et en guise de conclusion : $\mathcal{R}$ est une tribu qui contient la topologie de X (c'est à dire l'ensemble de ses ouverts),alors il contiendra aussi la tribu engendrée par cette topologie et qui n'est autre que la tribu borélienne $\mathfrak{B}$ d'où $\mathfrak{B} \subset \mathfrak{A}$ et pour l'autre inclusion elle est triviale car $\mathcal{R}$ ne contient que des ensembles mesurables,ce qui achève la preuve. ■

3.2 Deuxième théorème de Lusin :

Théorème 3.2.1 :(de Lusin,version généralisée :)

Soit $(E,\mathcal{E},\nu)$ un espace mesuré de Radon , (H,Θ) un espace topologique, à base [a] dénombrable muni de la tribus Borélienne et $f : E \to H$, une fonction **mesurable**, alors on a :
$(\forall \varepsilon > 0)\,(\forall A \in \mathcal{E}) : \nu(A) < +\infty \Rightarrow (\exists F_{(\varepsilon,A)} \subset A)$ tq : $F_{(\varepsilon,A)}$ est fermé,$\nu(A \setminus F_{(\varepsilon,A)}) < \varepsilon$ et $f_{/F_{(\varepsilon,A)}}$ est continue.
autrement dit,on peut restreindre toute les fonctions mesurables sur un espace de Radon à des autres qui sont continues sur une partie ***fermée*** ,aussi grande que voulue, de leur domaine.

Preuve :

À son tour,cette version va se baser,sur les résultats présentés précédemment pour la démonstration du premier théorème de Lusin,(surtout celui d'Egorov) ,mais,le fait de passer au cas général des espaces de Radon et de remplacer l'hypothèse d'intégrabilité par celui de mesurabilité , nécessite l'intervention d'une arsenal théorique très solide,surtout en topologie,pour cela,on l'acceptera pour le moment,et n'ayez pas peur,cette forme généralisée du théorème de Lusin, n'apparaîtra pas beaucoup dans la suite. ■

Remarque 3.2.1 :

Quand même ce théorème va nous servir pour justifier un petit passage du lemme III.2.2.3 et qui sera d'une grande utilité pour la preuve du théorème III.2.2.2.

4 C^0-prolongeabilité des fonctions continues sur un fermé :

4.1 Théorème de prolongement de Dungundji :

Définitions 4.1.1 :(Partition de l'unité :)

1. On appelle partition de l'unité d'un espace topologique X, toute famille $(\phi_i)_{i\in I}$ de fonctions ***continues***, définies de X à valeur dans $[0,1]$, telles que :
   i) pour tout point $x \in X$,il existe un voisinage de x tel que toutes les fonctions ϕ_i soient nulles sur ce voisinage à l'exception d'un nombre ***fini*** d'entre elles .
   ii) $(\forall x \in X) : \sum_{i\in I} \phi_i(x) = 1$.
2. On dit qu'une partition de l'unité : $(\phi_i)_{i\in I}$ est ***subordonnée à un recouvrement*** $(U_i)_{i\in I}$ de X si et seulement si elle est indexée par le même ensemble I que le recouvrement, avec : pour tout $i \in I$, le support de ϕ_i est inclus dans un U_j.

a. On désigne ici la base au sens topologique,et qui est une classe d'ouverts tel que tout élément de Θ soit une réunion d'ensembles de cette classe.
Pensez par exemple à la classe des intervalles ouverts dans l'ensemble des réels $\mathbb{R}$.

Définition 4.1.2 :(Espace topologique paracompact :)

Un espace topologique ***paracompact*** est un espace topologique ***séparé*** dans lequel tout recouvrement ouvert $\mathscr{U}$ admet un raffinement (un autre recouvrement dont chaque élément est inclus dans un autre de $\mathscr{U}$) ouvert ***localement fini***[a].

Attention ! : à ne pas confondre avec l'espace précompact dans lequel on peut extraire de toute suite une autre qui est de Cauchy.

Théorème 4.1.1 :(théorème de Dungundji :)

Soient (X,Θ) un espace topologique métrisable[b], $A\subset X$ un ***fermé*** de X et $(Y,+,.;\|\cdot\|)$ un espace vectoriel normé alors et selon [18] :
toute application continue f de A dans Y admet un prolongement ***continu*** $\tilde{f}$ de X dans Y, et on dit dans ce cas qu'elle est C^0-***prolongeable sur*** X.

Preuve :

Pour une distance **d** fixée sur X, considérons, dans l'ouvert $X\setminus A$, le recouvrement constitué des boules ouvertes $\left(B\left(x,\frac{\mathrm{d}(x,A)}{2}\right)\right)_{x\in X}$ de $X\setminus A$.
Selon le fameux ***critère de métrisabilité de Smirnov*** [c] tout espace métrique est paracompact, ce qui fait qu'il existe un recouvrement ouvert localement fini $(U_i)_{i\in I}$ de $X\setminus A$ dont chaque ouvert est inclus dans l'une de ces boules : $U_i\subset B\left(x,\frac{\mathrm{d}(x,A)}{2}\right)$.
On choisit alors une partition de l'unité $(\phi_i)_{i\in I}$ subordonnée à ce recouvrement et pour tout i, un point a_i de A tel que : $\mathrm{d}(x_i,a_i)\leqslant 2\cdot\mathrm{d}(x_i,A)$, et on prolonge f en posant :

$$\boxed{\tilde{f}(x)=\begin{cases} f(x) & \text{si } x\in A\\ \sum\limits_{i\in I}\phi_i(x)\cdot f(a_i) & \text{sinon.}\end{cases}}$$

L'application $\tilde{f}$ est clairement continue sur $X\setminus A$, il reste à montrer qu'elle l'est aussi sur A, c'est à dire :

$$(\forall a\in A)\,(\forall\varepsilon>0)\,(\exists\delta_\varepsilon>0),\text{tel que : } f\left(B\left(a,\delta_\varepsilon\right)\cap A\right)\subset B\left(a,\varepsilon\right).$$

Pour affirmer que pour tout $x\in B\left(a,\frac{\delta_\varepsilon}{6}\right)$, $f(x)\in B\left(a,\varepsilon\right)$, il suffit d'utiliser que pour tout U_i contenant x, $a_i\in B\left(a,\delta_\varepsilon\right)$, d'après les inégalités suivantes :
$\mathrm{d}(x,a_i)\leqslant\mathrm{d}(x,x_i)+\mathrm{d}(x_i,a_i)\leqslant\mathrm{d}(x,x_i)+2\cdot\mathrm{d}(x_i,A)\leqslant 5\cdot\mathrm{d}(x_i,A)-5\cdot\mathrm{d}(x,x_i)\leqslant 5\cdot\mathrm{d}(x,A)$,d'où :
$\mathrm{d}(a,a_i)\leqslant\mathrm{d}(a,x)+\mathrm{d}(x,a_i)\leqslant 6\cdot\mathrm{d}(a,x)$. ■

4.2 Théorème de l'extension de Tietze :

Théorème 4.2.1 :(de Tietze :)

Toute fonction $f:X\to\mathbb{R}$ continue à valeurs réelles définie sur un fermé d'un espace topologique métrisable (X,θ) peut être prolongée continument sur tout l'espace X.

a. Une famille de parties d'un espace topologique est dite localement finie lorsque chaque point possède un voisinage qui ne rencontre qu'un nombre fini d'éléments de la famille.

b. C'est à dire que sa topologie peut provenir d'une norme.

c. Il affirme qu'un espace topologique est métrisable $\Longleftrightarrow$ il est ***régulier***(dans lequel on peut séparer un point x et un fermé ne contenant pas x par deux ouverts disjoints) à base dénombrable localement finie.

Preuve :

Ce n'est autre qu'un cas particulier du théorème de Dungundji dans lequel l'espace Y est la droite réelle. ■

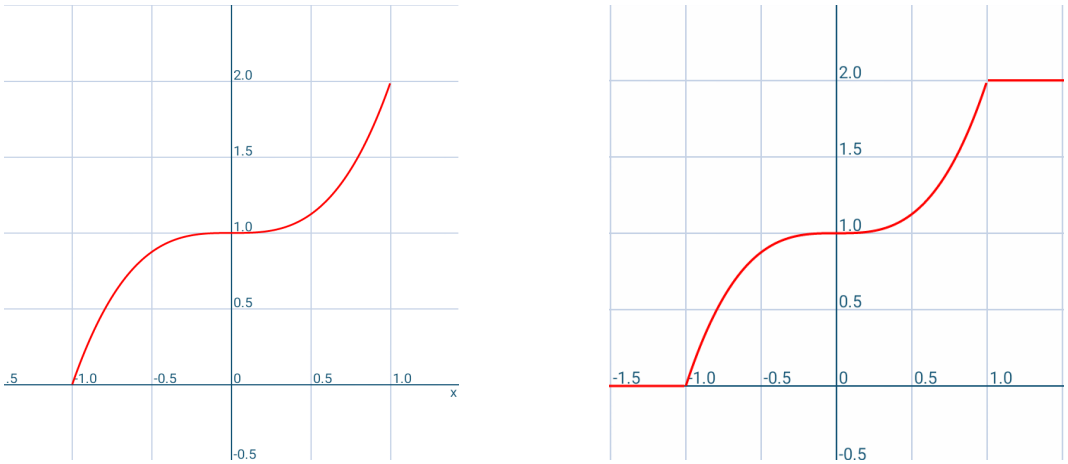


FIGURE I.3: Principe du théorème de Tietze pour $X = \mathbb{R}$.

Remarque 4.2.1 :

Ce résultat a été généralisée par le mathématicien russe Pavel Urysohn qui a remplaçé l'espace métrisable X par n'importe quel espace normal [a].

Conclusion :

Dans ce chapitre, et dans notre tentative de recouvrir une base théorique relativement solide et qui peut nous permettre de faire par la suite une lecture sans heurt des parties restantes, on a vu plusieurs résultats importants d'analyse fonctionnelle, comme les ***deux théorèmes de Lusin, le théorème de Stone-Weierstrass et le théorème de Tietze***, ces derniers sont exceptionnellement distingués par leurs énoncés qui vont justifier ultérieurement la grande aptitude d'une machine d'apprentissage appelée ***réseaux de neurones artificiels*****(RNA)**, et qui fera effectivement l'objet du chapitre suivant.

a. C'est un espace séparé dans lequel on peut trouver pour tous fermés disjoints F_1 et F_2,deux ouverts disjoints O_1 et O_2 tels que : $F_1 \subset U_1$ et $F_2 \subset U_2$.

Chapitre II

Apprentissage Automatique : Réseaux de Neurones Artificiels.

Résumé :

Dans ce chapitre, on va présenter une famille très intéressante de machines d'apprentissage et qui permet de faire le traitement et l'analyse de plusieurs classes de problèmes, que les outils ordinaires et les machines d'apprentissage classiques ne peuvent pas les réaliser ou juste ont des difficultés et du mal à résoudre.
le modèle de machines qu'on va proposer sera capable de reproduire et d'***imiter le plus fidèlement possible certains aspects de l'intelligence humaine,***et on verra que ceci revient à l'origine de l'inspiration de ce modèle du système nerveux, en outre l'introduction d'une telle machine qui portera le nom de ***"réseau de neurones artificiels"*** va causer l'émergence de plusieurs sous types remarquables, comme par exemple : le ***perceptron, les réseaux de Hopfield, réseaux à compétition...,etc*** , et même elle donnera naissance à des nouvelles gammes d'algorithmes appelés ***algorithmes d'apprentissages***.
dans la suite, nous allons citer les caractéristiques de ces réseaux, voir de proche leur fonctionnement et essayer d'exposer au lecteur leurs principaux algorithmes d'apprentissage, tout en faisant de notre mieux pour lever toute équivoque possible.

Mots clés : Réseaux de neurones, apprentissage, algorithme, intelligence artificielle.

Introduction :

Afin d'analyser et de comprendre le ***phénomène d'apprentissage*** ,l'Homme s'intéressa à l'étude du cerveau comme étant l'organe central de commande et de pensée ce qui va l'amener enfin de compte à faire ces études sur un système très compliqué ***qui est le système nerveux***, cette complexité au niveau de l'étude va le pousser à se restreindre aux élements de base qui le constituent et qui seront dans ce cas :***les neurones biologiques*** ,auxquels il proposera une représentation mathématique-informatique appelée : ***Le neurone formel ou artificiel.***

L'introduction du premier exemple de neurones formels a eu lieu en 1943, et c'était dû à Warren Mc.Culloch et Walter Pitts,qui se sont basés sur des observations neurophysiologiques,***récentes à l'époque***,pour nous proposer un modèle ***très simple*** du neurones formels, et qui s'agissait ,en ce temps là, d'une unité d'exécution binaire,retournant une sortie qui vaut 0 ou 1, le calcul de cette dernière,était par l'effectuation d'une somme pondérée des entrées, puis l'application d'une fonction de décision à seuil ,qui donne comme résultat 1 lorsque cette somme pondérée dépasseune

certaine valeur,et 0 au cas contraire.

après avoir constaté l'insuffisance du modèle qu'ils ont proposé et son incapacité de résoudre certains problèmes (le fameux problème "XOR" à titre d'exemple.) Mc.Culloch et Pitts ont décidé de faire l'analogie avec le cerveau humain en rendant leur neurone comme une unité élémentaire construisant une structure plus compliquée et puissante, nommée dans ce cadre : ***"réseaux de neurones artificiels"***.

Cette idée va marquer une étape charnière dans l'histoire de l'intelligence artificielle **,** et il développera par la suite tout le domaine de la bio-informatique , et surtout à travers l'introduction d'une famille d'algorithmes capables d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacun, et ceci fera des réseaux de neurones artificiels des systèmes habiles d'apprendre, et de mettre en œuvre le principe de l'induction, et spécialement avec l'émergence d'autre variantes de neurones qui ont été proposées,après les travaux de deux neurologues Mc.Culloch et Pitts.

1 De neurone biologique au neurone artificiel :

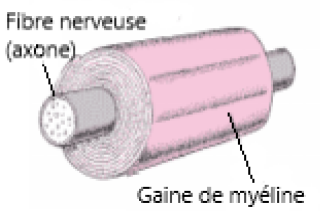
1.1 Neurone biologique :

1.1.1 Définitions générales et quelques statistiques :

Un neurone[a] ***biologique*** se considère comme une cellule nerveuse élémentaire qui constitue l'unité de base du système nerveux,son rôle principal est de recevoir des stimulations extérieures,les transformer en impulsions électriques pour les transmettre[b] enfin de compte à une autre cellule qui peut être nerveuse à son tour ou pas.

Le nombre total de ces neurones dans le corps humain est estimé à 100 milliards ,dont 89% se concentre dans le cerveau,et les 11% restantes se répartissent sur d'autres membres comme la moelle spinale qui compte presque 4 milliards, l'intestin qui compte plus de 450 millions et le cœur qui contient un nombre dans les environs de 60 mille …etc.

Les neurones sont les composantes principales les plus importantes du tissu nerveux, et elles possèdent des caractéristiques très remarquables et particulières, d'abord ***elles vivent très longtemps*** et ***elles nous accompagnent de la naissance à la mort,*** et en plus elles ***ne peuvent jamais être remplacées quand elles sont détruites*** **,** et pour cette raison, on trouve dans le système nerveux, un type spécial de cellules appelées ***les cellules gliales***, ou parfois ***les nervoglies***,ou tout simplement ***les glies*** ,qui s'occupent du soutien et de la nutrition des neurones ainsi que la production de ***la myéline***[c] qui protègent le tissu nerveux.



(a) Une gaine de myéline.

(b) Gaine de myéline normale face à une autre lésée.

a. Ce terme était introduit dans le lexique médical pour la première fois en 1879 par le médecin et le biologiste allemand Heinrich Wilhelm Waldeyer(1836-1921).

b. Ce mécanisme de transmission reflète deux propriétés physiologiques très importantes chez les neurones et qui sont :l'***excitabilité*** et la ***conductivité***.

c. C'est une sorte de substance qui sert à ***isoler*** et à protéger ***les fibres nerveuses***, exactement comme le fait du plastique autour des fils électriques ,et on peut trouver certains troubles qui provoquent une démyélinisation qui touche directement le système nerveux central,ou les nerfs d'autres parties du corps , par exemple : ***La polyneuropathie inflammatoire démyélinisante chronique*** **« PIDC »** .

Le nombre des nervoglies est beaucoup plus important que celui des neurones ,et selon les estimations des scientifiques, leur nombre dépasse trois fois celui des neurones, et pour cela on trouve qu'elles représentent à peu près 50% du volume cérébral et plus de 65% des cellules du cerveau.

1.1.2 La structure générale d'un neurone biologique :

Un neurone biologique peut être décrit tout simplement comme une cellule ordinaire prolongée par des ramifications,et sa structure complexe se caractérise essentiellement par les composantes principales suivante :

- ***les dendrites :*** d'une moyenne de 10 mille par neurone , elles font la réception de l'information sous forme des impulsions éléctriques, d'autres neurones auxquels elles sont attachées via des ***synapses*** [a].
- ***une membrane externe :*** elle s'occupe de la propagation de cette information le long du neurone, et on peut dire qu'elle est la seule responsable de la ***conductivité*** chez les cellules nerveuses.
- ***un corps cellulaire(ou Soma) :*** il fait le traitement de toutes les informations provenant des dendrites pour produire une sortie qui se propage via l'axone enfin de compte.
- ***un axone :*** d'un diamètre compris entre 10 et 15μm et d'une longueur qui varie d'un millimètre à plus d'un mètre**,** cette fibre nerveuse fait parvenir le potentiel d'action résultante à une autre cellule.

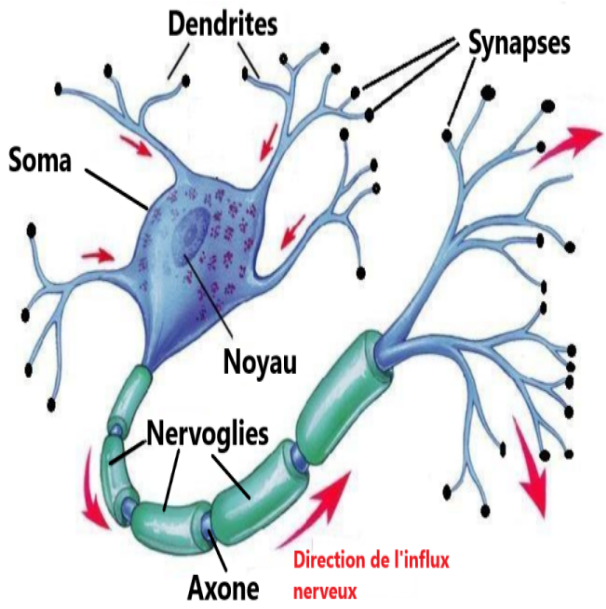


FIGURE II.2: Schéma d'un neurone biologique.

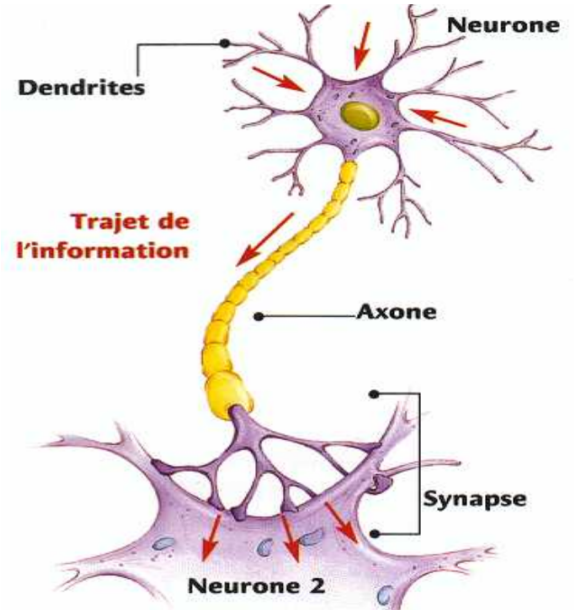


FIGURE II.3: Synapses entre deux neurones.

Donc,et après cette petite mise en situation ,on peut et sans problème parler d'une manière concrète ***d'un neurone artificiel***, ou encore ***neurone formel***.

1.2 Neurone artificiel :

1.2.1 Une description générale du neurone artificiel :

Un neurone artificiel ou formel n'est autre qu'une représentation mathématique informatique du neurone biologique et qui s'inspire exactement de ses composantes générales, et ***il se caracté-***

a. C'est une zone de contact fonctionnelle qui s'établit entre deux neurones, ou entre un neurone et une autre cellule.

rise par :

-***des entrées :*** qui correspondent aux dendrites.

-***des poids :*** qui représentent les actions excitatrices des synapses, ce qui justifie le fait de les appeler parfois ***les poids synaptiques***.

-***une fonction d'entrée totale :*** qui définie le pré-traitement effectué sur les entrées du neurone,et à la plupart des cas, cette fonction est choisie comme celle de la somme pondérée des entrées par les poids.

-***une fonction d'activation :*** elle agit sur l'entrée total produite par la fonction précédente pour nous donner ce qu'on appelle l'état du neurone.

-**une fonction de sortie :** tout simplement elle calcule la sortie du neurone en agissant sur son état,et habituellement elle est omise par une identification avec la fonction identité.

-**la sortie :** c'est le résultat final du traitement effectué par le neurone et il représente l'information qui se propage au long de l'axone.

-**et enfin les types d'entrées et de sortie :** il faut quand-même mentionner que ***ces types*** restent parmi les facteurs les plus importants du neurone artificiel ,et leur modification peut produire des changements radicaux sur sa capacité de résolution face à un problème précis ou à une classe de problèmes bien déterminée.

donc et d'une manière grossière, le neurone formel n'est autre que la composition des trois fonctions numériques définie précédemment, et de ce fait on pourra l'exprimer ainsi :

$$\underbrace{E \subset \mathbb{R}^n}_{\substack{x=(x_1,\cdots,x_n)^\top \\ \text{l'entrée : } x\in\mathbb{R}^n}} \overbrace{\xrightarrow{\Sigma(\cdot)}}^{\text{fonction d'entrée totale}} \underbrace{\mathscr{T} \subset \mathbb{R}}_{\substack{\omega^\top x-\theta \\ \text{l'entrée totale} \\ (\Sigma(\cdot)=\langle\omega,\cdot\rangle-\theta)}} \overbrace{\xrightarrow{\varphi(\cdot)}}^{\text{fonction d'activation}} \underbrace{\mathscr{E} \subset \mathbb{R}}_{\substack{\varphi(\omega^\top x-\theta) \\ \text{l'état du neurone}}} \overbrace{\xrightarrow{S(\cdot)}}^{\text{fonction de sortie}} \underbrace{\mathscr{S} \subset \mathbb{R}}_{\substack{\varphi(\omega^\top x-\theta) \\ \text{la sortie} \\ (S(\cdot)=\mathrm{Id}(\cdot))}}$$

(a) Neurone biologique.

(b) Neurone formel.

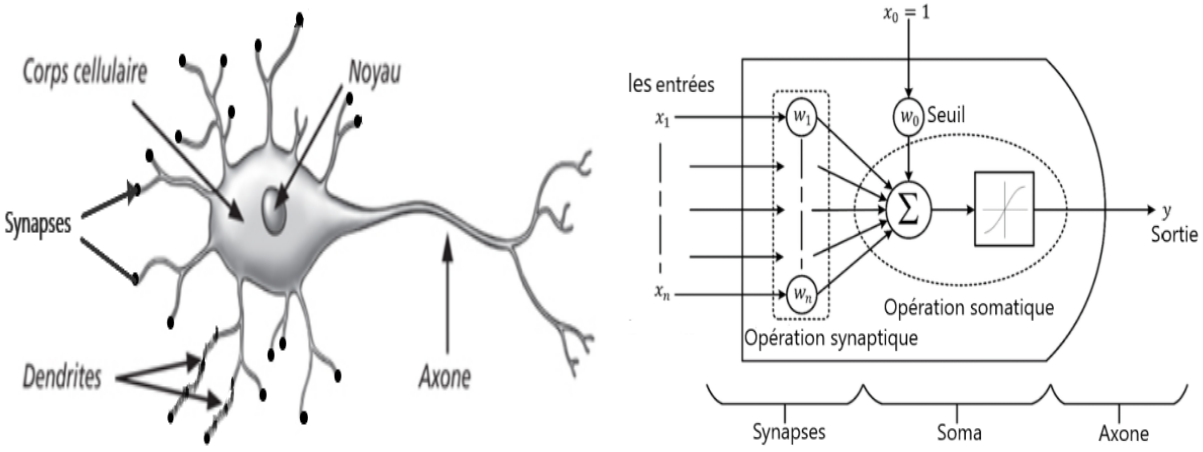


FIGURE II.4: Comparaison entre la structure d'un neurone biologique et d'un autre formel.

1.2.2 Le neurone artificiel d'un point de vue mathématique :

Définition : On appelle neurone formel ou artificiel, toute fonction numérique de n variables réelles $x_1, \cdots, x_n$ **,** et qui prend la forme suivante :

$$\begin{array}{rcccl} \Psi & : & E \subset \mathbb{R}^n & \longrightarrow & \mathscr{S} \subset \mathbb{R} \\ & & x & \longmapsto & \Psi\left(\omega_0 + \sum_{j=1}^{n} \omega_j x_j\right) \end{array}$$

avec :$n \in \mathbb{N}^*$ est ***un paramètre*** entier non nul indiquant ***la taille des entrées*** **,** $\omega = \left(\omega_1, \cdots, \omega_n\right)^\top \in \mathbb{R}^n$ **,** est un vecteur fixé à l'avance nommé : ***le vecteur des poids*** et $\omega_0 \in \mathbb{R}$ est une certaine constante appelée ***le biais*** **,** et dans ce cas là **,** l'image :$y = \Psi\left(\omega_0 + \sum_{j=1}^{n} \omega_j x_j\right) \in \mathbb{R}$**,** portera le nom de ***la sortie du neurone associée au vecteur :*** **x** et de plus les ensembles :$E \subset \mathbb{R}^n$ et $\mathscr{S} \subset \mathbb{R}$ **,** désignent respectivement ***l'espace d'entrée et de sortie*** du neurone

Remarque : Dans la définition précédente on a adopté les choix habituels pour les fonctions d'entrée [a] et de sortie , or la première était confondue avec celle des somme pondérées des poids avec les entrées,et la deuxième avec l'identité, mais ***faites attention!*** , ceci peut être modifié sans aucun problème dans certains cas , à savoir les $\Sigma\Pi$ neurones par exemple.

1.2.3 Comment on est arrivé à cette description du neurone formel ? :

On a bien vu dans ce qui précède que le neurone formel n'était qu'une simple présentation du neurone biologique, et même on a réussi de réduire et de simplifier sa description pour la rendre sous la forme d'une fonction mathématique possédant une expression un peu particulière et spécifique ,mais ***cette simple formulation n'était pas obtenue d'un seul coup,ou d'une façon miraculeuse***, cependant ***elle était le fruit d'une longue histoire de recherche et de travail*** ,dans la suite,on va présenter les grands repères historiques derrière l'évolution de cette description en exposant les différentes représentations adoptées avant son émergence.

★**Le neurone de Culloch-pitts** (1943) **:** Comme on a déjà dit, le premier neurone formel a été proposé en 1943 par les deux chercheurs américains **Mc.Culloch** et **Walter Pitts** (logicien) dans leur article [42],où il était présenté comme suit :

► **Description :**

-***les entrées*** **:** pour ce cas elles sont de type binaire **,**et elles peuvent être vues comme un vecteur : $x = \left(x_1, \cdots, x_n\right)^\top \in \{0,1\}^n$, en plus on pourra dans ce cadre faire la distinction entre deux types d'entrées :

- ***des entrées inhibitrices :*** sont celles qui ***ont un effet maximum sur la prise de décision*** quelles que soient les autres entrées.
- ***des entrées excitatrices :*** elles ne sont pas du tout celles qui vont déclencher le feu de neurone mais ***elles pourraient le déclencher quand elles sont combinées***.

×***les poids :*** en ce temps là cette notion n'était pas encore abordée ,et pour cela on va les négliger en leur attribuant tous la valeur 1.

a. Certaine sources appellent cette fonction ***la fonction d'agrégation***.

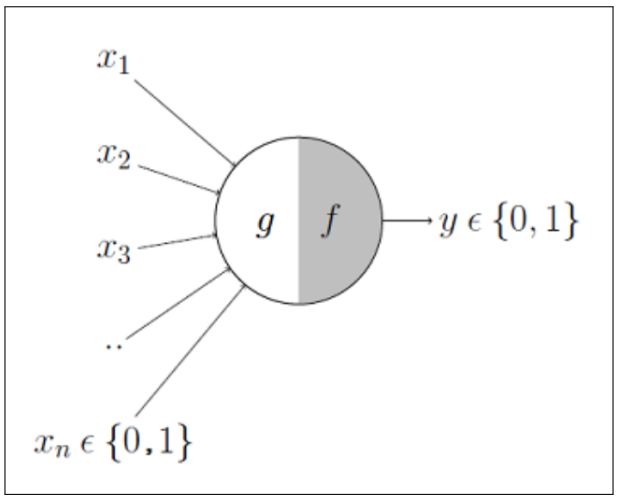


FIGURE II.5: La représentation simple d'un C.P-neurone.

-***la fonction d'entrée totale*** : à l'époque elle était juste la somme des entrées,c'est à dire :

$$\begin{array}{rccc} g & : & \{0,1\}^n & \longrightarrow & \mathbb{R} \\ & & x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} & \longmapsto & \sum_{j=1}^{n} x_j \end{array}$$

×***la fonction d'activation*** : selon la toute première définition proposée par Culloch et pitts dans leur article [42],c'était la fonction de décision définie par : $f(x) = \mathbb{1}_{[\theta,+\infty[}(x) := \begin{cases} 1 & \text{si : } x \geqslant \theta \\ 0 & \text{sinon.} \end{cases}$.
avec θ est un réel appelé ***paramètre de seuillage***.

×***la fonction de sortie*** : exactement comme les poids, cette notion n'avait pas de sens à l'époque,ce qui fait qu'on peut la confondre avec la fonction identité pour le moment.

- ***la sortie*** : clairement elle est de type binaire[a] avec :

$$y = \begin{cases} 1 & \text{si : } \sum_{j=1}^{n} x_j \geqslant \theta \\ 0 & \text{si : } \sum_{j=1}^{n} x_j < \theta \end{cases}$$

Remarque : cette expression de la sortie montre bien et justifie le fait d'appeler θ par paramètre de seuillage ,et on signale que pour améliorer la capacité de discrimination chez notre neurone et l'approcher d'un comportement désiré on doit régler et changer ***à la main*** la valeur prise par ce paramètre.

► **Mise en œuvre :** voyons maintenant comment ce modèle peut être utilisé pour représenter quelques fonctions booléennes.

1. La fonction logique "**AND**" :

a. c'est à dire qu'elle prend ses valeurs dans : $\{0,1\}$.

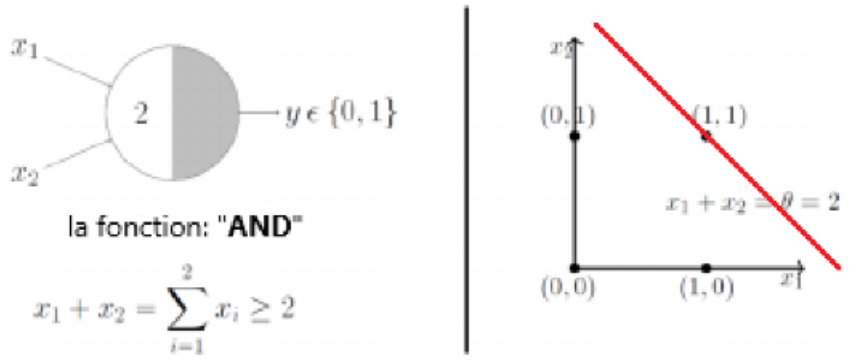


FIGURE II.6: Réalsation de la fonction "**AND**" par un C.P-neurone.

2. La fonction logique "**OR**" :

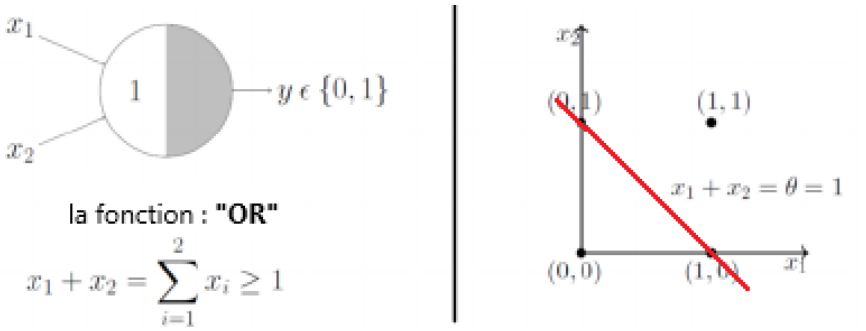


FIGURE II.7: Réalisation de la fonction "**OR**" par un C.P-neurone.

3. La fonction logique "**OR**" au cas de trois entrées : :

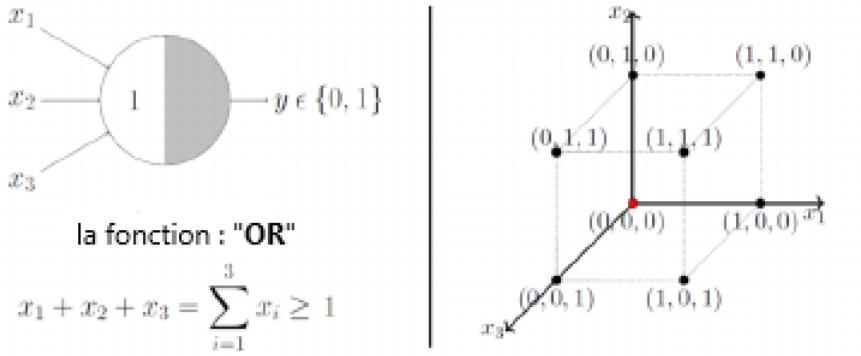


FIGURE II.8: Réalisation de la fonction "**OR**" à trois entrées par un C.P-neurone.

► **Les limites du modèle de Culloch-pitts :**

1. Que peut on faire au cas des entrées non booléennes (par exemple, réelles) ?·
2. Que faire si nous voulons attribuer plus d'importance à certaines entrées ?·
3. Avons-nous toujours besoin de coder ***manuellement*** le seuil ?·
4. Comment résoudre des fonctions qui ne sont pas linéairement séparables ? (la fonction "**XOR**" à titre d'exemple).

★**Le Perceptron simple de Rosenblatt** (1958) **:** Le psychologue américain **Frank Rosenblatt** va se baser sur les travaux de Culloch et Pitts pour introduire en 1958 une machine de ***classification linéaire*** très puissante appellée "***Perceptron***" et dans son article [50] , cette machine a été décrite ainsi :

► **Description :**

-***les entrées*** **:** dans ce cas sont des réels qui peuvent être vus comme un vecteur : $x = (1, x_1, \cdots, x_n)^\top \in \mathbb{R}^{n+1}$ où la première composante $x_0 = 1$ représente une ***entrée fictive*** qui revient à l'introduction du biais ω_0.

-***les poids*** **:** sont aussi des réels qui sont vus également comme un vecteur des pondérations : $\omega = (\omega_0, \omega_1, \cdots, \omega_n)^\top \in \mathbb{R}^{n+1}$,possédant une première composantes $\omega_0 = -\theta$ qui représente ***le biais*** et qui est tout simplement ***l'opposé du seuil***.

-***la fonction d'entrée totale*** **:** c'est tout simplement celle des sommes pondérées définie par.

$$\begin{array}{rccc} \Sigma & : & \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ & & x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} & \longmapsto & \left\langle \omega, \begin{pmatrix} 1 \\ x \end{pmatrix} \right\rangle = \omega_0 + \sum_{j=1}^{n} \omega_j x_j = \overbrace{\sum_{j=0}^{n} \omega_j x_j}^{\text{ici on a : } x_0 = 1.} \end{array}$$

-***la fonction d'activation*** **:** selon la toute première définition proposée par Culloch et pitts dans leur article [42],c'était la fonction Heavside définie par : $\varphi(x) = \mathbb{1}_{\mathbb{R}^+}(x) := \begin{cases} 1 & \text{si : } x \geqslant 0 \\ 0 & \text{sinon.} \end{cases}$.

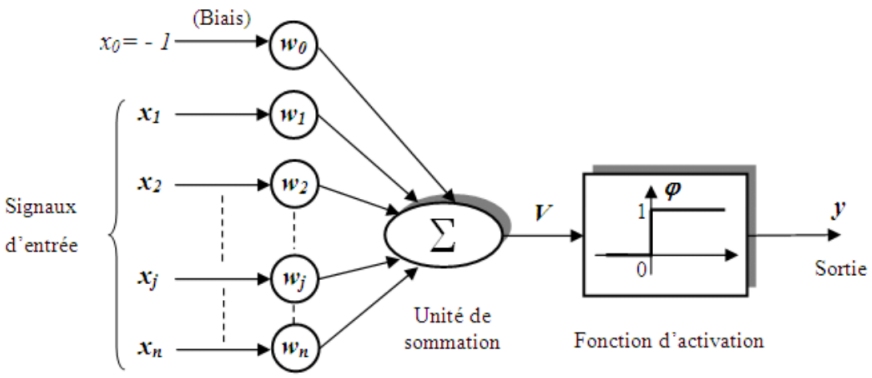


FIGURE II.9: Modèle du perceptron de Rosenblatt.

-***la fonction de sortie*** **:** comme d'habitude c'est la fonction identité.

- ***la sortie*** **:** clairement elle est de type binaire [a] avec :

$$S = \begin{cases} 1 & \text{si : } \sum_{j=1}^{n} \omega_j x_j \geqslant \theta \\ 0 & \text{si : } \sum_{j=1}^{n} \omega_j x_j < \theta \end{cases}$$

► **Mise en œuvre :** Une autre fois on va voir comment ce modèle peut être utilisé pour représenter quelques fonctions booléennes , mais cette fois ci on se contentera par l'étude de la simple fonction "**OR**" en laissant au lecteur le soin de traiter les autres exemples.

1. La fonction logique "**OR**" :

a. c'est à dire qu'elle prend ses valeurs dans : $\{0,1\}$.

x_1	x_2	OR	
0	0	0	$w_0 + \sum_{i=1}^{2} w_i x_i < 0$
1	0	1	$w_0 + \sum_{i=1}^{2} w_i x_i \geq 0$
0	1	1	$w_0 + \sum_{i=1}^{2} w_i x_i \geq 0$
1	1	1	$w_0 + \sum_{i=1}^{2} w_i x_i \geq 0$

$$w_0 + w_1 \cdot 0 + w_2 \cdot 0 < 0 \implies w_0 < 0$$
$$w_0 + w_1 \cdot 0 + w_2 \cdot 1 \geq 0 \implies w_2 > -w_0$$
$$w_0 + w_1 \cdot 1 + w_2 \cdot 0 \geq 0 \implies w_1 > -w_0$$
$$w_0 + w_1 \cdot 1 + w_2 \cdot 1 \geq 0 \implies w_1 + w_2 > -w_0$$

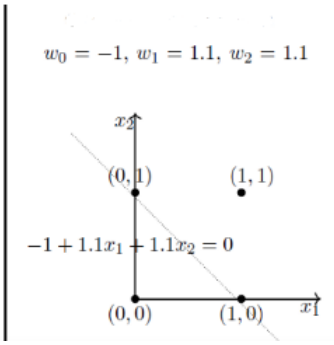


FIGURE II.10: Une solution possible de la fonction "**OR**" par un perceptron simple.

► **Les limites du Perceptron simple :** On ne peut pas nier que le perceptron est un modèle très puissant , qui a pu dépasser quelques limites du C.P-neurone en fournissant des solutions et des réponses à plusieurs questions , mais biensûr ***pas toutes les questions*** **,** et pour cette raison , on trouve qu'il y en a encore des interrogations qui se considèrent comme des bornes de capacité chez ce type de machines , et on propose à titre d'exemples les questions suivantes :

1. Avons-nous toujours besoin de coder ***manuellement*** le seuil?·
2. Comment résoudre des fonctions qui ne sont pas linéairement séparables? (la fonction "**XOR**" à titre d'exemple).

x_1	x_2	XOR	
0	0	0	$w_0 + \sum_{i=1}^{2} w_i x_i < 0$
1	0	1	$w_0 + \sum_{i=1}^{2} w_i x_i \geq 0$
0	1	1	$w_0 + \sum_{i=1}^{2} w_i x_i \geq 0$
1	1	0	$w_0 + \sum_{i=1}^{2} w_i x_i < 0$

$$w_0 + w_1 \cdot 0 + w_2 \cdot 0 < 0 \implies w_0 < 0$$
$$w_0 + w_1 \cdot 0 + w_2 \cdot 1 \geq 0 \implies w_2 > -w_0$$
$$w_0 + w_1 \cdot 1 + w_2 \cdot 0 \geq 0 \implies w_1 > -w_0$$
$$w_0 + w_1 \cdot 1 + w_2 \cdot 1 \geq 0 \implies w_1 + w_2 < -w_0$$

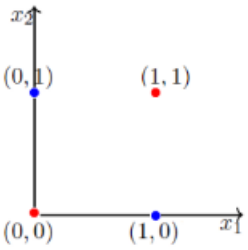


FIGURE II.11: L'incapacité du perceptron simple face au problème "**XOR**".

Remarque : en générale,un ***seul*** perceptron (simple) ne peut être utilisé que pour mettre en œuvre des fonctions linéairement séparables [a] ,et il reste totalement impuissant devant celles qui ne le sont pas, ce grand point faible sera à l'origine de l'apparence de l'idée de contribution entre les neurones, et de construction des ***réseaux neuronaux*** [b] par la suite.

★**l'Adaline de Widrow-Hoff (1960) :** le professeur américain **Bernard Widrow** et son étudiant, **Ted Hoff**,ont proposé en 1960 dans une partie de leur rapport technique [62], une machine d'apprentissage appelée "***neurone linéaire adaptatif***" en anglais : ***Adaptive Linear Neuron*** ou **« Adaline »** en abréviation.
en somme ce modèle n'était autre que le perceptron de Rosenblatt munit de la fonction d'activation linéaire pure ,c'est à dire "l'identité" , et la sortie dans ce cas sera sans surprise :

$$S = \sum_{i=0}^{n} \omega_i x_i \in \mathbb{R}.$$

a. C'est à dire qu'il est possible de tracer une ligne pour séparer les entrées positives des entrées négatives.
b. Nous reviendrons un peu plus tard à cette notion.

mais ce qui va marquer la différence complète de cette machine , c'est sa capacité ***de régler automatiquement ses paramètres en se basant sur des expériences déterminées et des données correctes*** , sans que ceci soit réalisé manuellement , ce fait va la distinguer de tous systèmes logique usuels à l'époque , et on verra par la suite que ce réglage se fait via un algorithme bien précis et qui était globalement inspiré d'une règle appelée la règle de Hebb.

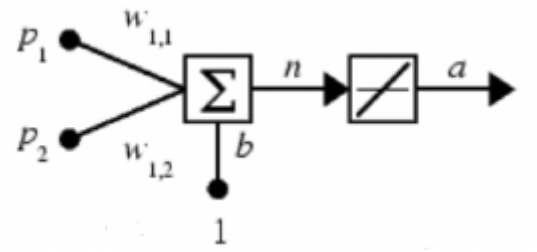


FIGURE II.12: Exemple d'un Adaline à deux entrées.

★ **Le perceptron de Minsky et Papert (1969) :**le modèle proposé par Rosenblatt va être analysé et amélioré par les deux neuropsycologues américains **Marvin Lee Minsky** et **Seymour Aubrey Papert** en 1969**,** dans leur article [43]**,** au sein duquel ils ont modifié légèrement quelques détails au niveau de la structure de cette machine en donnant plus d'opportunités à sa fonction d'activation (***la fonction signe à titre d'exemple***)**,** et ils se sont également inspirés des travaux de leurs collègues Widrow et Hoff**,** et qui ont été récentes à l'époque**,** pour attribuer à leur modèle un algorithme permettant de réaliser "le réglage" automatique des poids synaptiques**,** on va revenir par la suite à l'étude de cet algorithme d'une façon plus détaillée.

La plupart des neurones formels proposés après 1958 et jusqu'à maintenant ne sont que des variantes du perceptron de Rosenblatt dont la fonction d'activation est modifiée**,** par exemple :

★ **Les perceptrons sigmoïdes :** les neurones artificiels usuels qu'on a proposé jusqu'à cet instant disposent tous d'***une fonction d'activation à seuil***, qui ***produit un changement soudain*** au niveau de la décision de 0 à 1 lorsque la valeur de l'entrée totale : $S = \sum_{i=0}^{n} \omega_i x_i \in \mathbb{R}$ dépassera un certain seuil $\theta = -\omega_0$ **,** mais pour la plupart des applications réelles **,** ***cette logique de seuillage sévère est vraiment déconseillée*** **,** et ***surtout à cause des difficultés algorithmiques qui peuvent se produire*** à cause de la non différentiabilité de ces fonctions (l'invalidité des algorithmes basés sur la descente de gradient par exemple) et pour cela nous nous attendons à une fonction de décision plus fluide, ***qui passe progressivement*** de 0 à 1 , ce qui va nous pousser à introduire les neurones formels ou les perceptrons sigmoïdes où la fonction d'activation est beaucoup plus lisse que la fonction du pas ce qui semble être une chose logique et évidente à faire.
rappelez-vous qu'une fonction sigmoïde est une fonction mathématique d'une courbe en forme de "***S***" et qui passe de 0 à 1, Il existe de nombreuses fonctions qui ont cette caractérisation, par exemple : $\Psi : x \mapsto \frac{1}{2}\left(1+\tanh(x)\right)$.

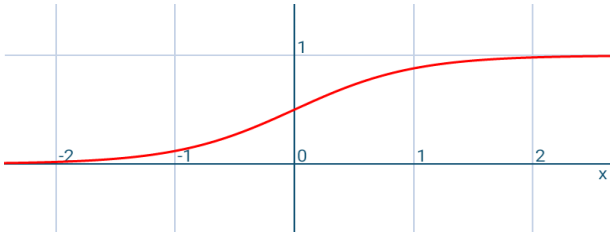


FIGURE II.13: La courbe de la fonction $: x \mapsto \frac{1}{2}\left(1+\tanh(x)\right)$.

(a) Les fonctions d'activation les plus courantes.

Nom de la fonction	Relation d'entrée/sortie	Icône
seuil	$a = 0$ si $n < 0$ $a = 1$ si $n \geq 0$	
seuil symétrique	$a = -1$ si $n < 0$ $a = 1$ si $n \geq 0$	
linéaire	$a = n$	
linéaire saturée	$a = 0$ si $n < 0$ $a = n$ si $0 \leq n \leq 1$ $a = 1$ si $n > 1$	
linéaire saturée symétrique	$a = -1$ si $n < -1$ $a = n$ si $-1 \leq n \leq 1$ $a = 1$ si $n > 1$	
linéaire positive	$a = 0$ si $n < 0$ $a = n$ si $n \geq 0$	
sigmoïde	$a = \frac{1}{1+\exp^{-n}}$	
tangente hyperbolique	$a = \frac{e^n - e^{-n}}{e^n + e^{-n}}$	
compétitive	$a = 1$ si n maximum $a = 0$ autrement	C

(b) Les fonctions d'activation les plus utilisées.

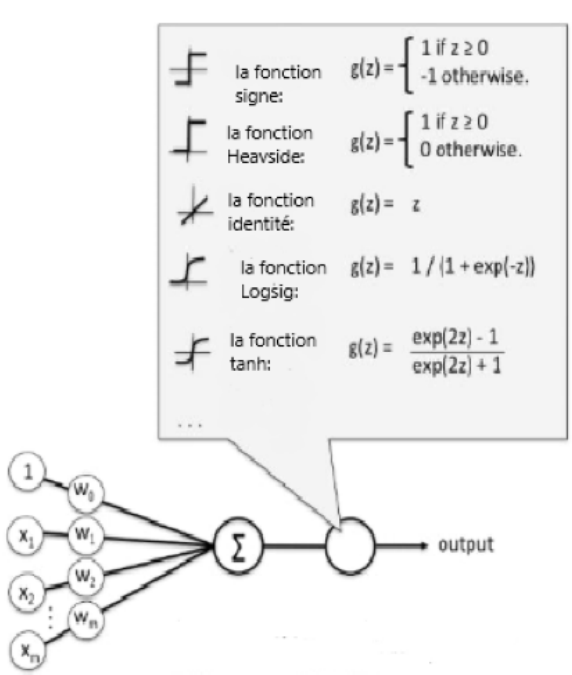


FIGURE II.14: Quelques choix possibles pour la fonction d'activation .

Remarque : on peut trouver d'autres applications numériques qui peuvent être proposées pour jouer le rôle de la fonction d'activation , ***mais faites attention***, les seules applications qui peuvent bénéficier de ce privilège sont celles qui garantissent la réalisation de la propriété ***d'approximation universelle*** chez les réseaux de neurones artificiels, ce qui rend la tâche du choix d'une fonction convenable un peu plus délicate.
pour le moment contentez vous des simples exemples qu'on a présenté, et ne vous inquiétez pas on reviendra dans le **chapitre III** pour voir en profondeur quelques modèles de fonctions qui peuvent bien nous servir dans ce sens.

2 Apprentissage d'un neurone artificiel :

2.1 Vue d'ensemble sur le concept d'apprentissage :

2.1.1 La notion d'apprentissage :

Dans un cadre général ***l'apprentissage*** désigne le fait d'***acquérir des informations*** ou des propriétés relativement stables ,et on peut le décrire dans le cadre humain comme étant un ***mélange*** complexe entre ***les contraintes innées*** et le ***hasard acquis***, mais dans le cadre de l'intelligence artificiels, l'apprentissage désigne une phase durant laquelle une machine ,comme le perceptron simple par exemple, ***se change*** et se développe,***en modifiant*** son comportement et ***ses paramètres pour arriver à une attitude désirée.***
le terme « ***apprentissage*** » a été introduit dans le cadre des machines pour la première fois en 1959 par l'informaticien américain **Arthur Samuel** dans son article [53] , et ceci à la suite de sa création d'un programme qui est capable de jouer au fameux Jeu de Dames et de s'améliorer en jouant.

2.1.2 Les buts de l'apprentissage :

L'apprentissage possède plusieurs objectifs mais en somme on les résume comme suit :

1. l'***extraction des informations*** pertinentes à partir des données.
2. la ***classification*** des données ***et la prédiction*** si une certaine donnée appartient à une telle ou telle classe.
3. le ***regroupement*** des données en classes homogènes.
4. ***la mémorisation des informations*** afin d'***identifier*** toutes données bruitée ou incomplète à une autre déjà connue .
5. ***la généralisation*** qui nous permet d'***étendre nos résultats*** et de faire le passage ***à des nouvelles données***.

2.1.3 Les types principaux d'apprentissage :

Il existe beaucoup de types d'apprentissage, et dans la littérature on trouve souvent qu'ils sont au nombre de six (voir :[20]) mais principalement , on peux distinguer juste entre quatre grands types :

-l'apprentissage supervisé : il se fait sur une base de données particulière appelée ***base d'apprentissage*** et qui est constituée d'une collection finie de couples ayant la forme suivante :

$$(\textit{entrée,sortie désirée})$$

ces sorties désirées forment un ***étiquetage*** qui permet de subdiviser les entrées en classes, ce qui fait que nos exemples sont connus, nos classes sont prédéterminées,et le système n'a que d'***apprendre à les classer selon un modèle de classification bien précis*** ainsi ,la machine peut aussi apprendre à ***généraliser*** et à prédire la classification de futures données en s'appuyant sur les différentes techniques de ***la régression*** statistique .

-l'apprentissage non supervisé : comme son nom l'indique ,cet apprentissage se fait sur un autre type de base,appelé ***base de test*** et dans laquelle ***on ne dispose plus de sorties désirées***,juste des entrées ,et ce type d'apprentissage est voué à des utilisations particulières comme le ***clustering*** ou la répartition des données, or,on souhaite en l'occurrence de mettre des exemples hétérogènes sous forme de collections dont les éléments sont liés par des caractéristiques communes , et c'est à la machine de détecter les traits en communs sans aucune intervention extérieure , on mentionne que la similarité entre les données est souvent déterminée via une fonction de distance entre paires d'exemples.

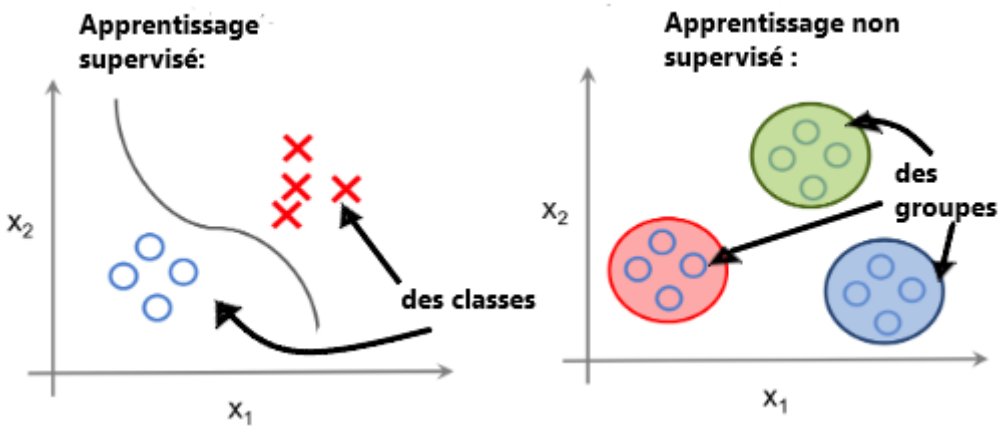


FIGURE II.15: La différence entre l'apprentissage supervisé et non supervisé.

-**l'apprentissage par renforcement :** ce type est basé sur trois notions ***l'agent***, ***l'environnement*** et ***le système de récompense***.
l'agent va réagir en fonction d'un état de l'***environnement***, pour renvoyer une action en fonction de celui-ci,le ***système de récompense*** va évaluer positivement ou négativement l'agent, en fonction de l'action qu'il a pris.
le but de l'agent est de ***collecter le maximum de points possible***, ce qui va l'obliger de comprendre la différence entre une bonne et une mauvaise action, et donc au fur et à mesure de favoriser les plus convenables.
cette façon d'apprentissage vise en fait de reproduire le mécanisme naturel d'acquisition des connaissances , et ce qui la rend extrêmement puissante , c'est le fait qu'elle n'a pas besoin de bases de données , contrairement aux deux types d'apprentissages précédents.

-**l'apprentissage par transfert :** comme son nom l'indique ce type d'apprentissage est basé sur la capacité d'un système à ***faire l'extension des connaissances*** et des compétences, acquises à partir des tâches antérieures ***sur des nouvelles tâches*** ou domaines ***partageant des propriétés communes***,(par exemple faire le passage de la résolution d'un cube de Rubik de dimensions $2 \times 2 \times 2$ à un autre de taille $3 \times 3 \times 3$) mais ce type d'apprentissage est rarement adopté en pratique, surtout à cause des difficultés qu'il présente au niveau de la détection des similitudes et le transfert des connaissances entre les tâches.

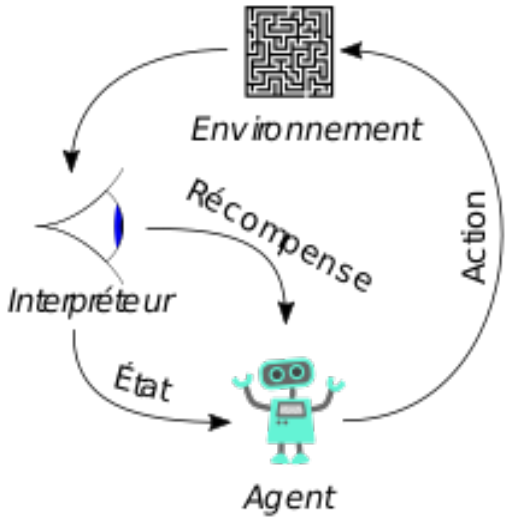


FIGURE II.16: Apprentissage par renforcement.

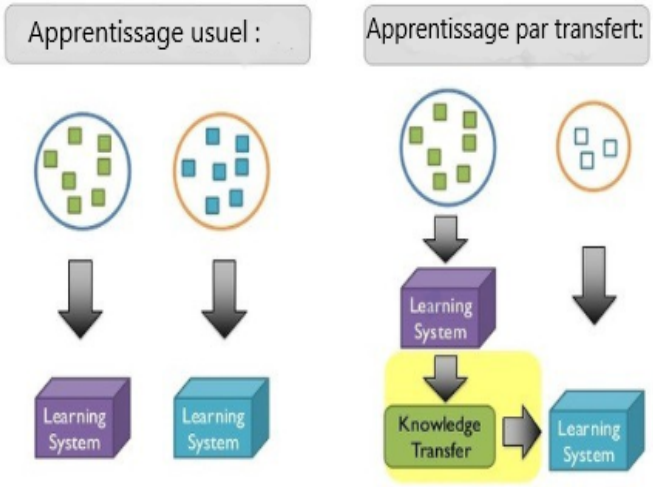


FIGURE II.17: Apprentissage par transfert face à l'apprentissage usuel.

2.1.4 Les facteurs qui influencent la qualité d'apprentissage :

La qualité de l'apprentissage dépend en générale de plusieurs facteurs,et qui sont essentiellement liés ***soit aux performances de la machine*** d'apprentissage utilisée **,** ***soit à la nature de la base de donnée*** fournie à cette machine (bien sûr lorsque son apprentissage nécessite la présence d'une telle base) :

► **les facteurs liés à la base des données :** et qui sont généralement les suivants :

- **la taille de la base de données :** lorsque le nombre d'exemplaires est ***acceptablement élevé ,*** la machine sera capable de se doter d'un nombre suffisant d'informations et de renseignements, ce qui va favoriser ***en principe*** ses chances de réaliser un apprentissage de bonne qualité.
- **la proportion des données incomplètes :** dans une base de taille importante la présence d'un tel type de données est fort probable **,** et ceci peut perturber notre machine

et par conséquent réduire sa qualité d'apprentissage , pour garder cette réduction dans le cadre de l'acceptable ,***on doit*** tout simplement ***se poser dans des cas où le pourcentage de cette catégorie de données est faible*** par rapport à la taille totale de la base .

- **la proportion des données bruité :** la présence des valeurs douteuses **,** par exemple des éléments erronés , mal classés ,ou non-conformes au mode de distribution générale des exemples,ou même ayant des valeurs bizarres et aberrantes, peut obscurcir la vision de notre machine , et par la suite affaiblir sa capacité d'apprendre , à l'instar des données incomplètes on va franchir ce problème en se mettant ***dans un cadre où la proportion de cette famille de données est négligeable*** par rapport à la taille totale de la base .
- **le nombre et le type des attributs :** les attributs qui décrivent nos exemples déterminent entièrement le nombre de leurs classes, ce qui montre qu'ils jouent quand même un rôle au niveau de l'amélioration de la qualité d'apprentissage, surtout au cas du supervisé, or, si ce nombre est très élevé la tâche d'apprendre sera plus compliquée, et spécialement pour une machine qui adopte un modèle de classification simple comme le modèle linaire par exemple**,** en plus le type de ces attributs et même des exemples peut apparaître comme un facteur dominant dans le processus d'apprentissage surtout pour le semi et le non-supervisé car dans lesquels on aura besoin de définir une ***distance*** entre deux exemples , une chose qui est simple dans le cas où ce type est quantitative (poids,âge,taille, ...) mais en revanche il ne l'est pas du tout pour le cas est qualitative (couleur,humeur, ...)**.**

► **les facteurs liés à la machine d'apprentissage :** la machine est un élément clé dans le processus d'apprentissage et sa capacité reste ***le paramètre le plus important*** qui peut parfaire et perfectionner la qualité de ce processus, et pour les facteurs principaux qui sont liés à cette machine, on trouve :

- **le modèle :** on distingue souvent entre trois grands modèles de machines d'apprentissage et dont chacune possède des propriétés et des particularités propres à elle :
  - ⋆ **les machines à vecteurs de support :** également appelées réseaux de vecteurs de support, ou les séparateurs à vaste marge **« SVM »**, ces machines constituent un ensemble de méthodes d'apprentissage supervisé ***consacré aux problèmes de classification et de régression***, or, et à partir d'une collection des ***exemples d'apprentissage***, dont chacun est identifié à une classe déterminée ,un algorithme d'apprentissage **SVM** fournit un modèle qui ***prédit*** si un nouvel exemple tombe dans une classe ou dans une autre.
    cette machine s'est développée grâce aux travaux de **Vladimir Vapnik** sur la théorie d'apprentissage statistique (voir :[61]), et elle reste jusqu'à présent ***un exemple flagrant d'une machine qui a été créée à base d'une étude théorique pure***.
  - ⋆ **les réseaux bayésiens :** ce sont des ***machines calculatoires des probabilités conditionnelles,*** et plus précisément des modèles graphique probabiliste qui représentent un ensemble de variables aléatoires et leur indépendance conditionnelle à l' aide d'***un graphe acyclique dirigé de causalité*** (voir : [44]).
    Un réseau bayésien peut nous servir par exemple, pour représenter les relations probabilistes entre les maladies et les symptômes, et compte tenu de ces derniers, le réseau peut calculer les probabilités de présence de diverses maladies.
    pour l'apprentissage de ces réseaux**,** il est basé sur la recherche d'une ***estimation des distributions de probabilités ou des paramètres des lois correspondantes,*** à partir de données disponibles.
    L'estimation de ces distributions est un sujet très vaste et complexe et dans ce sens

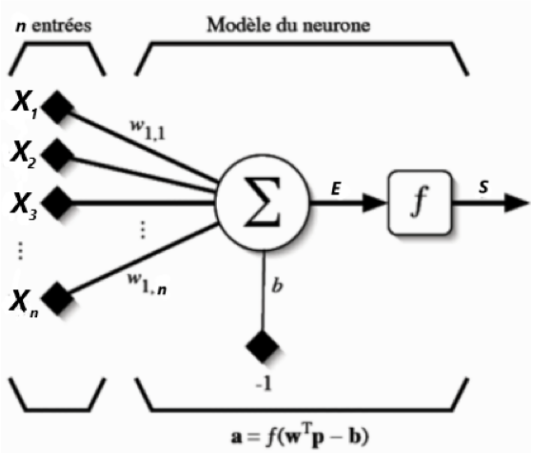
Il existe des méthodes et des algorithmes efficaces qui l'effectuent [a], et si vous êtes intéressés par ce sujet nous vous invitons d'aller voir : [39].

⋆ **les réseaux de neurones artificiels :** cette machine fera l'objet de la section suivante ,mais pour le moment, et sans entrer dans les détails,on peut la définir comme étant un groupe de neurones formels interconnectés entre eux, pour former un système semblable au réseau neuronal du cerveau.
Cette machine est très puissante et peut être utilisée pour résoudre plusieurs classes de problèmes par exemple les problèmes de ***classification***, de ***régression,***de ***prédiction,*** de ***détection d'anomalies,*** de ***clustering,*** ...***etc,*** et elle était notamment adoptée pour réaliser pas mal de tâches, comme, la reconnaissance de la parole, de l'image, la traduction automatique ,le diagnostic médical...etc.

- **le type :** si on a spécifié le modèle de la machine qu'on va utiliser ***pour résoudre un problème précis,*** alors, on doit ensuite déterminer le type de ce modèle**,** par exemple : si on a pris comme modèle ***les réseaux de neurones artificiels*** alors on doit signaler est ce qu'il est un perceptron multicouche, un madaline , un réseau de Hopfield, une carte auto adaptative de Kohonen ...etc**?.**
(on aura après l'opportunité de traiter quelques-uns de ces types.)
- **l'algorithme d'apprentissage :** à ce stade le modèle et le type de la machine sont déjà déterminés, mais la façon par laquelle elle apprend ne l'est pas encore, pour cela on doit toujours indiquer l'algorithme d'apprentissage adopté par notre machine car il peut vraiment marquer la différence au niveau de sa qualité.

2.2 Optimisation et apprentissage supervisé d'un neurone formel :

Une mise en situation : on considère un perceptron simple (de **Rosenblatt**) qui se présente comme suit :



avec :

— "$\Sigma(\cdot)$" est ***la fonction d'entrée totale*** qui fait la somme pondéré des x_i, et elle est définie par :

$$\Sigma\left(x_1,\cdots,x_n\right)=\sum_{i=1}^{n}\omega_i x_i+\overbrace{\omega_0}^{:=-b}=\sum_{i=0}^{n}\omega_i x_i=\left\langle\begin{pmatrix}1\\x\end{pmatrix},\omega\right\rangle.$$

a. Par exemple dans le cas où toutes les variables sont observées, la méthode la plus simple et la plus utilisée est l'***estimation statistique*** qui consiste à estimer la probabilité d'un événement par la fréquence son apparition dans la base de données, cette approche est appelée ***le maximum de vraisemblance*** (**MV**) et nous écrivons :

$$\widehat{P}\left(X_i=x_k/\text{parents}\left(X_i\right)=x_j\right)=\widehat{\theta}_{i,j,k}=\frac{N_{i,j,k}}{\sum_k N_{i,j,k}}$$

où $N_{i,j,k}$ est le nombre d'événements dans la base de données pour lesquels la variable X_i est dans l'état x_k et ses parents sont dans la configuration x_j **.**

— f est ***la fonction d'état ou d'activation*** qu'on peut la prendre comme : $\tanh(\cdot)$, $\text{sign}(\cdot)$, $\text{logsig}(\cdot)$,....

— $\omega = (\omega_0, \omega_1, \ldots, \omega_n)^\top \in \mathbb{R}^{n+1}$ est ***le vecteur des pondération*** associé à notre neurone.

— et enfin, $u = \sum_{i=1}^{n} \omega_i x_i + \omega_0$ et $S = f(u)$ sont respectivement l'entrée totale et la sortie fournie par le perceptron pour un vecteur d'entrée : $X = (x_1, \cdots, x_n) \in \mathbb{R}^n$.

D'abord, et au cours de cette section il faut savoir qu'***on se pose dans le cadre d'apprentissage supervisé,*** dans lequel on dispose d'une base d'apprentissage : $\mathcal{B} = \left(X^h, d^h\right)_{1 \le h \le \ell}$,avec :

⋆ ℓ : est la taille de la base $\mathcal{B}$.

⋆ et $\left(\forall h \in \{1, \cdots, \ell\}\right)$: $X^h = \left(x_1^h, \cdots, x_n^h\right)^\top \in \mathbb{R}^n$, et $\mathrm{d}^h \in \mathbb{R}$ sont respectivement les composantes du couple $\left(X^h, d^h\right)$ qui est un élément de notre base $\mathcal{B}$ qu'on l'appel **"*exemplaire*"**, le vecteur X^h est un ***vecteur exemple*** et d^h est tout simplement ***la sortie désirée associée à cet exemple***.

L'apprentissage de ce perceptron consiste en fait à trouver une façon ou une méthode pour le rendre "optimal" plus clairement on veut qu'il approche le mieux notre base d'apprentissage.
une manière traditionnelle pour attribuer un sens à ce terme de **"*mieux*",** est d'adopter le critère des moindres carrés **« L·M·S »**,(Least Mean Squares) ,et qui est inspiré essentiellement de la statistique.
ce critère est ***basé sur la minimisation d'une certaine erreur*** E comme étant ***une fonction des pondérations et de biais*** $(\omega_0, \omega_1, \cdots, \omega_n) \in \mathbb{R}^{n+1}$,et plus explicitement, pour trouver ***le perceptron optimal,*** il faut et il suffit de déterminer les poids :

$$(\omega_0^*, \cdots, \omega_n^*)^\top \in \mathbb{R}^{n+1}$$

qui minimisent l'erreur totale moyenne **:**

$$\boxed{C(\omega) = \frac{1}{\ell} \sum_{h=1}^{\ell} \frac{1}{2} \left(d^h - g(X^h, \omega)\right)^2 .}$$

Autrement dit, ***le problème d'apprentissage va se ramener enfin de compte au problème d'optimisation*** suivant :

$$\boxed{(\mathcal{P}) \begin{cases} \min\ C(\omega) = \dfrac{1}{2\ell} \displaystyle\sum_{h=1}^{\ell} \left(d^h - g(X^h, \omega)\right)^2 \\ \omega \in \mathbb{R}^{n+1} \end{cases}}$$

Remarques 2.2.1 :

1. Certain ouvrages ([16] par exemple) font la distinction entre trois types d'erreurs :

   -l'erreur locale : ou instantané c'est tout simplement l'erreur pour un seul exemple, et il est défini par : $e_h(\omega) = \frac{1}{2} \left(d^h - g(X^h, \omega)\right)^2$.

   -l'erreur totale : c'est l'erreur provenant de toute la base d'apprentissage et ce n'est autre que la somme de tout les erreurs locales de ses exemples, et on écrit donc :

   $$E(\omega) = \sum_{h=1}^{\ell} e_h(\omega) = \sum_{h=1}^{\ell} \frac{1}{2} \left(d^h - g(X^h, \omega)\right)^2 .$$

   -l'erreur moyenne : comme son nom l'indique c'est l'erreur en moyenne pour tout les exemples d'apprentissage, et on note : $C(\omega) = \frac{1}{\ell} E(\omega) = \frac{1}{\ell} \sum_{h=1}^{\ell} \frac{1}{2} \left(d^h - g(X^h, \omega)\right)^2$.

2. on peut voir facilement que : $E(\omega) = 0 \iff \left(\forall h \in \{1, \cdots, \ell\}\right)$: $e_h(\omega) = 0$.

3. pour le problème d'optimisation $(\mathcal{P})$, l'adjonction du coefficient $\frac{1}{\ell}$ dans l'expression de la fonction *"objectif"* est essentiellement due à l'étude du cas aléatoire, où sa présence était indispensable pour légitimer le passage de la minimisation de l'erreur théorique à la minimisation de l'erreur empirique en faisant appel au théorème central limite, mais dans notre cas ici on se pose dans un cadre déterministe et pour cela on pourra et sans aucun souci l'éliminer [a] de la syntaxe du problème $(\mathcal{P})$.

2.3 Quelques algorithmes d'apprentissage supervisé pour un neurone formel :

2.3.1 Algorithme de descente de gradient :

À ce stade, on a bien vu que le principe général d'apprentissage supervisé pour un neurone formel consiste en fait de trouver un élément neuronal optimal qui approche le mieux notre base de données,ce qui est revenu comme on a déjà dit précédemment, à la résolution du problème d'optimisation :

$$\boxed{(\mathcal{P})\begin{cases}\min\ E(\omega)=\dfrac{1}{2}\displaystyle\sum_{h=1}^{\ell}\left(d^h-g(X^h,\omega)\right)^2\\ \omega\in\mathbb{R}^{n+1}\end{cases}}$$

et pour le résoudre on va adopter des algorithmes ou des méthodes adaptatives ,qui vont se baser sur la génération d'une suite $(\omega^k)_{k\in\mathbb{N}}$ des poids qui convergent vers le minimum de l'erreur locale $e_h(\omega)$, c'est à dire que ces algorithmes vont minimiser l'erreur locale et non pas l'erreur total ou moyen.
la famille des algorithmes qui peut se présenter dans ce cas comme un candidat idéal est tout simplement la famille de ***descente de gradient***.
maintenant ***on suppose que*** f ***est différentiable*** ou plus simplement dérivable (car c'est une fonction d'une seule variable réelle).

$$\begin{aligned}\text{on a :}\quad \left(\forall i\in\{0,\cdots,n\}\right):\ \frac{\partial e_h(\omega)}{\partial\omega_i}&=\frac{\partial}{\partial\omega_i}\left(\frac{1}{2}\left(S^h-d^h\right)^2\right)=\frac{\partial}{\partial S^h}\left(\frac{1}{2}\left(S^h-d^h\right)^2\right)\times\frac{\partial S^h}{\partial\omega_i}\\ &=\left(S^h-d^h\right)\times\frac{\partial f(u)}{\partial u}\times\frac{\partial u}{\partial\omega_i}=\left(S^h-d^h\right)\times f'(u)\times\frac{\partial\sum_{k=0}^{n}\omega_k x_k^h}{\partial\omega_i}\\ &=\left(S^h-d^h\right)\times f'\left(\sum_{k=0}^{n}\omega_k x_k^h\right)\times x_i^h\\ &=-\underbrace{\left(d^h-S^h\right)\times f'\left(\sum_{k=0}^{n}\omega_k x_k^h\right)}_{:=\delta}\times x_i^h\\ &=-\delta.x_i^h\end{aligned}$$

alors le vecteur gradient de la fonction e_h s'écrira ainsi :

$$\nabla_\omega\ e_h(\omega):=\begin{pmatrix}\dfrac{\partial e_h(\omega)}{\partial\omega_0}\\ \vdots\\ \dfrac{\partial e_h(\omega)}{\partial\omega_n}\end{pmatrix}=\begin{pmatrix}-\delta.x_0^h\\ \vdots\\ -\delta.x_n^h\end{pmatrix}=-\delta X^h.$$

avec X^h est un exemple (un vecteur d'entrée) de notre classe.

a. bien sûr car le nombre ℓ est positif, en fait et pour la même raison on peut éliminer à fois le "2" de l'expression de la fonction à minimiser mais, il est préférable de la garder pour alléger l'expression du gradient par la suite.

donc dans ce cas l'algorithme d'apprentissage qu'on aura sera de la forme :

Algorithme 1 Algorithme d'apprentissage du perceptron simple par la descente de gradient.

Entrée: une base de données : $\left(X^h, d^h\right)_{1\leq h\leq \ell}$, une fonction d'activation dérivable f, des poids de départ :$\omega = \left(\omega_0, \cdots, \omega_n\right) \in \mathbb{R}^{n+1}$ et un nombre maximal d'itérations $N_{\max}$.
Sortie: les poids optimaux :$\omega = \left(\omega_0, \cdots, \omega_n\right)$

$k \leftarrow 1$;
tant que $k \leqslant N_{\max}$ **faire:**
  pour h allant de 1 à ℓ **faire:**
    pour i allant de 0 à n **faire:**
      $u^h \leftarrow \sum_{j=0}^{n} \omega_j^{(k)} x_j^h$;
      $S^h \leftarrow f\left(u^h\right)$;
      $e^h \leftarrow d^h - S^h$;
      $\delta \leftarrow e^h \times f'(u^h)$;
      ${\omega_i}^{(k+1)} \leftarrow {\omega_i}^{(k)} + \eta_k \times \delta \times x_i^h$;
    fin pour
  fin pour
fin tant que
retourner: $\omega = \left(\omega_0, \cdots, \omega_n\right)$

Remarques 2.3.1 :

1. Comme son nom l'indique cet algorithme est inspiré de celui de descente de gradient usuel (Cauchy 1847, Hadamard 1908), appliqué à la fonction d'erreur locale $e_h(\omega)$.
2. Le critère d'arrêt peut être choisi autrement, par exemple : $\left\|\nabla E(\omega)\right\| \leqslant \varepsilon$, $\left\|\omega^{(k+1)} - \omega^{(k)}\right\| \leqslant \varepsilon, \ldots$.
3. η_k est tout simplement le pas de descente qu'on l'appelle dans ce cas ***le taux d'apprentissage***, et dans la littérature il y en a plusieurs façons pour le choisir, par exemple le fixer dés le début (algorithme de gradient à pas fixé) ou avant chaque itération (à pas prédéterminé) ou le prendre par une procédure d'optimisation unidimensionnelle [a] (à pas optimal)... .
4. En général rien ne prouve que la diminution d'erreur pour un point va produire une diminution pour les autres aussi, mais cette façon de procéder est commune entre tous les algorithmes adaptatifs.

Théorème 2.3.1 :

1. Si on adopte pour l'algorithme précédent un choix prédéterminé du pas d'apprentissage et qui vérifie les deux conditions : $\lim\limits_{k\to+\infty} \eta_k = +\infty$ et $\sum\limits_{k=0}^{+\infty} \eta_k = +\infty$, alors il convergera globalement,c'est à dire pour n'importe quel point de départ,vers les pondérations optimales.
2. si la fonction d'erreur $\omega \mapsto E(\omega)$ est de classe $C^\infty(\mathbb{R})$, avec : $\lim\limits_{\|\omega\|\to+\infty} E(\omega) = +\infty$, et si notre actualisation du pas η est faite par une procédure d'optimisation unidimensionnelle, alors pour toute initialisation des poids, l'algorithme d'apprentissage par descente de gradient convergera mais vers des pondérations critiques et pas forcement optimales.

a. c'est à dire prendre :$\eta = \underset{\lambda>0}{\text{Argmin}}\, E\left(\omega^{old} - \lambda \cdot \nabla E\left(\omega^{old}\right)\right)$.

Preuve :

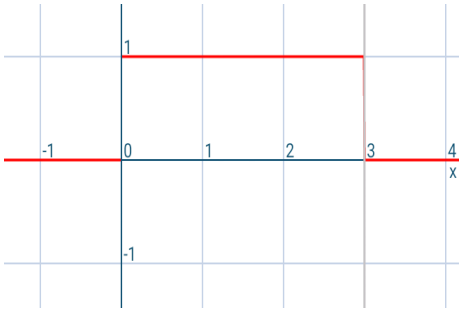
La preuve des deux critères précédents se découle nécessairement des théorèmes usuels de convergence pour l'algorithme de descente de gradient, voir pages :15, 17 dans [19] , et en plus on verra ultérieurement que cet algorithme n'est autre qu'un cas particulier de celui de la rétropropagation de gradient ce qui fait que ces critères vont être généralisés dans ce cas voir pour cela [10] page 215. ■

malgré la simplicité et l'efficacité de la procédure d'apprentissage par descente de gradient, ***la particularité du choix de la fonction d'activation entrave sa capacité***, or et dans plusieurs cas pratiques , par exemple les problèmes de décision, on peut se trouver face à des fonctions d'activation non dérivables, et dans ce contexte cet algorithme devient invalide, et pour cette raison on sera obligé d'introduire des modifications soit au niveau du problème, soit au niveau de l'algorithme.

▷ **modifier le problème :** le principe de cette approche est facile mais elle peut nous sortir de plusieurs pétrins, qui sont souvent rencontrés dans les problèmes de décision, et en somme il consiste à remplacer ou substituer la fonction d'activation f et qui est par hypothèse non dérivable par une autre $\widetilde{f}$ qui est dérivable et très proche de f.
cette idée va apparaitre dans la preuve du théorème : III.1.3.2 concernant l'approximation des fonctions de décision par des réseaux neuronaux à une seule couche cachée.

Exemple 2.3.1 :

(a) La courbe de : $\mathbb{1}_{[0,3]}(x)$.



(b) La courbe de : $\left(1+e^{200x-600}+e^{-200x}+e^{-600}\right)^{-1}$.

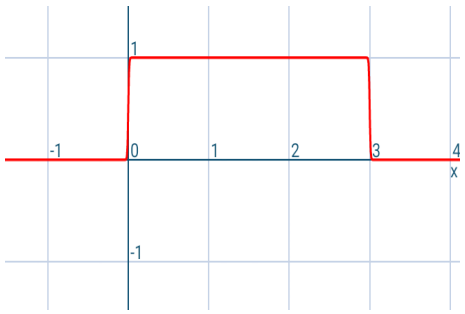


FIGURE II.18: La fonction (b) est régulière et elle peut bien jouer le rôle de la fonction (a).

Remarque 2.3.2 :

Pour l'approche de modification du problème, toutes les techniques permettant de faire l'approximation régulière [a], comme le développement limité, l'interpolation lagrangienne, l'interpolation Hermitienne [b] ... etc, peuvent être évoquées pour nous servir dans ce cas, mais malgré cela la réalisation d'une telle approximation est parfois très difficile et moins évidente à trouver (voir l'exemple précédent),ce qui ne nous laissera aucun choix que de modifier l'algorithme.

a. C'est à dire qui approchent notre fonction f par une autre régulière.(par exemple l'interpolation linéaire n'est pas acceptée dans ce cas).

b. C'est une extension de l'interpolation de Lagrange,qui consiste,pour une fonction dérivable en un nombre fini de points donnés à construire un polyôme interpolateur dont les valeurs de la dérivée à ces points coïncident avec celles de la dérivée de la fonction, l'avantage de cette méthode c'est qu'elle permet de donner une meilleure approximation lorsque le nombre des points augmentent, c'est à dire qu'elle peut franchir le phénomène de Runge (voir :[15]).

▷ **modifier l'algorithme :** contrairement au principe précédent, cette approche consiste à garder le problème tel qu'il est mais de modifier l'algorithme pour qu'il s'adapte à la non différentiabilité de la fonction d'activation, cette approche va donner naissance à deux fameux algorithmes d'apprentissage et qui sont ***l'algorithme d'apprentissage du perceptron*** et ***l'algorithme d'adaline de Widrow-Hoff*** .

2.3.2 L'algorithme d'apprentissage du perceptron :

La règle de Hebb : le neuropsychologue canadien **Donald Hebb** (1904 – 1985) s'est basé sur des conjectures biologiques pour affirmer que les neurones se lient entre eux s'ils s'excitent ensemble, c'est à dire que lorsque deux neurones sont excités conjointement, il se crée ou renforce un lien qui les unissent, cette idée sera la base de la notion d'apprentissage automatique, et elle fournira le tout premier algorithme de réajustement des poids, et qui était apparu en 1949 dans l'article de Hebb[29] sous la formulation moderne suivante :

Algorithme 2 Algorithme de réajustement des poids de Hebb.

Entrée: une collection d'exemples $(X^h)_{1\leqslant h\leqslant \ell}$,des poids de départ :$\omega = (\omega_0, \cdots, \omega_n) \in \mathbb{R}^{n+1}$, un facteur de correction η,et un nombre maximal d'itérations $N_{\max}$.
Sortie: des poids plus convenables :$\omega = (\omega_0, \cdots, \omega_n)$

$k \leftarrow 1$;
tant que $k \leqslant N_{\max}$ **faire:**
  pour h allant de 1 à ℓ **faire:**
    pour i allant de 0 à n **faire:**
      $u^h \leftarrow \sum_{j=0}^{n} \omega_j^{(k)} x_j^h$;
      $S^h \leftarrow \mathbb{1}_{\mathbb{R}^+}\left(u^h\right)$; (la fonction d'activation dans cas est la fonction Heavside.)
      ${\omega_i}^{(k+1)} \leftarrow {\omega_i}^{(k)} + \eta_k \times \left(S^h.x_i^h\right)$;
    fin pour
  fin pour
fin tant que
retourner: $\omega = (\omega_0, \cdots, \omega_n)$

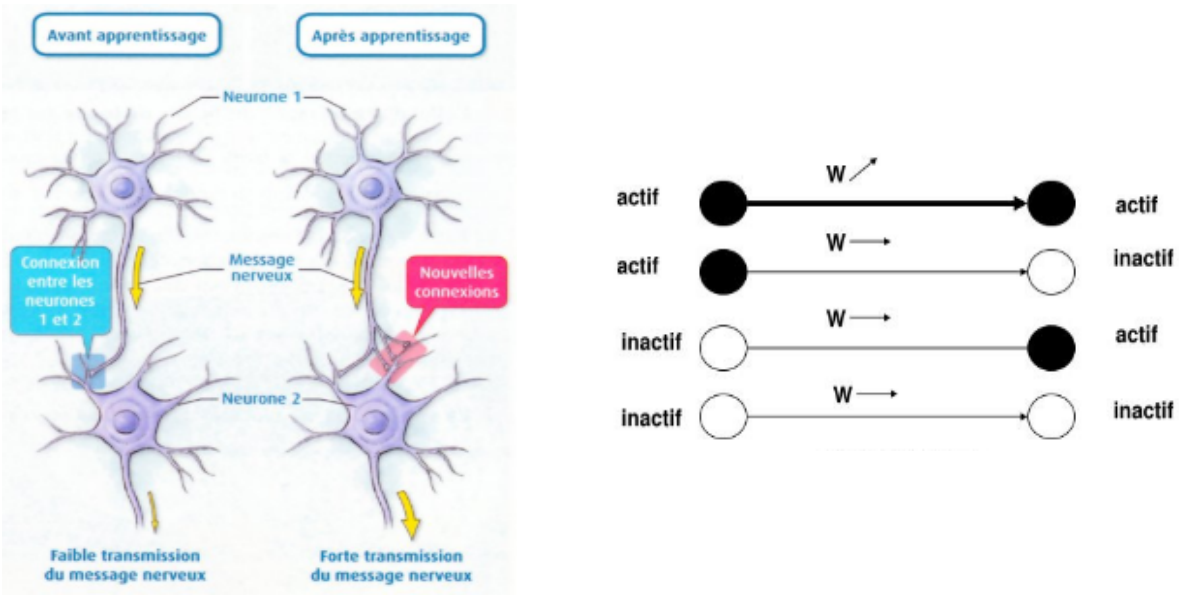


FIGURE II.19: Une illustration de la règle de Hebb.

Remarques 2.3.3 :

1. l'idée de cette règle était juste une conjecture et une vision théorique jusqu'à 2000, quand le médecin et le chercheur américain **Eric Kandel** a reçu le prix Nobel en fournissant une preuve expérimentale qui assure l'implication des mécanismes d'apprentissages de Hebb dans les synapses réels.
2. cette règle ne dispose pas des théorèmes de convergence, et reste inapplicable dans certains cas malgré l'existence de la solution, ce qui fait qu'elle n'est pas encore à la hauteur d'être considérée comme un algorithme d'apprentissage.

l'algorithme du perceptron cet algorithme est exactement analogue à la règle précédente, juste il a ajouté une idée un peu maline dans la correction des poids en s'inspirant de la descente de gradient, or, à la place de corriger comme suit : ${\omega_i}^{(k+1)} \leftarrow {\omega_i}^{(k)} + \eta_k \times \left(x_i^h.S^h\right) \times x_i^h$ on va le faire ainsi : ${\omega_i}^{(k+1)} \leftarrow {\omega_i}^{(k)} + \eta_k \times \left(d^h - S^h\right) \times x_i^h$, et de ce fait l'algorithme se présente ainsi :

Algorithme 3 Algorithme d'apprentissage du perceptron.

Entrée: une base de données : $\left(X^h, d^h\right)_{1\leq h\leq \ell}$, une fonction d'activation dérivable f, des poids de départ : $\omega = (\omega_0, \cdots, \omega_n) \in \mathbb{R}^{n+1}$ et un nombre maximal d'itérations $N_{\max}$.
Sortie: les poids optimaux :$\omega = (\omega_0, \cdots, \omega_n)$

$k \leftarrow 1$;
tant que $k \leqslant N_{\max}$ **faire:**
  pour h allant de 1 à ℓ **faire:**
    pour i allant de 0 à n **faire:**
      $u^h \leftarrow \sum_{j=0}^{n} \omega_j^{(k)} x_j^h$;
      $S^h \leftarrow \mathbb{1}_{\mathbb{R}^+}\left(u^h\right)$; (la fonction d'activation dans cas est la fonction Heavside.)
      $e^h \leftarrow d^h - S^h$;
      ${\omega_i}^{(k+1)} \leftarrow {\omega_i}^{k} + \eta_k \times e^h \times x_i^h$;
    fin pour
  fin pour
fin tant que
retourner: $\omega = (\omega_0, \cdots, \omega_n)$

Théorème 2.3.2 :(Novikov 1962 [1])

Si la base $\mathcal{B} = \left(X^h, d^h\right)_{1\leq h\leq \ell}$ ***est linéairement séparable*** par un hyperplan passant par l'origine, la procédure d'apprentissage précédente ***convergera globalement***, c'est à dire pour n'importe quelles pondérations initiales,***et en un nombre fini d'itérations*** vers un perceptron sans seuil résolvant le problème de décision déterminé par $\mathcal{B}$.

Preuve :

Soit $n+1$ le nombre d'entrées du perceptron, on suppose que chaque exemple possède une $n+1$-ème coordonnée égale à 1 (qui correspond au biais).
on note $\mathcal{E}_0$ et $\mathcal{E}_1$ respectivement l'ensemble ***des exemples*** de $\mathcal{B}$ pour lesquels le perceptron doit retourner la valeur 1 et la valeur 0.
par hypothèse, il existe un hyperplan de l'espace $\mathbb{R}^{n+1}$ qui sépare $\mathcal{E}_0$ de $\mathcal{E}_1$, ce qui fait qu'il existe

un vecteur $V=\left(v_0,\cdots,v_n\right)^{\intercal}$ tel que :

$$\begin{cases}\langle V,X\rangle<0 & \text{si : } X\in\mathcal{E}_0\\ \langle V,X\rangle\geqslant 0 & \text{si : } X\in\mathcal{E}_1\end{cases}$$

et puisque $\mathcal{B}$ est finie, on trouvera forcement un réel $\varepsilon>0$ strictement positif tel que :

$$\left(\forall X\in\mathcal{E}_0\right):\langle V,X\rangle\leqslant-\varepsilon \quad \text{et :} \quad \left(\forall X\in\mathcal{E}_1\right):\langle V,X\rangle\geqslant\varepsilon$$

alors pour tout X de $\mathcal{E}$, on a : $\left|\langle V,X\rangle\right|\geqslant\varepsilon$.
soit M un majorant [a] de $\left\{X^2=\langle X,X\rangle\,/\,X\in\mathcal{E}\right\}\subset\mathbb{R}^+$.

Soit $\omega^{(0)}=\left(\omega_0^{(0)},\omega_1^{(0)},\ldots,\omega_n^{(0)}\right)\in\mathbb{R}^{n+1}$ les valeurs d'initialisation des poids synaptiques. on procédera par l'absurde, en supposant que l'algorithme d'apprentissage ne s'arrête pas et en montrant par la suite que cette hypothèse va nous ramener à une contradiction.
on note $\left(\omega^{(k)}\right)_{k\in\mathbb{N}}\in\left(\mathbb{R}^{n+1}\right)^{\mathbb{N}}$ les états successifs différents [b] des coefficients synaptiques (c'est-à-dire que pour tout $k\in\mathbb{N}$, on a $\omega^{(k)}\neq\omega^{(k+1)}$), comme on a supposé que l'algorithme ne s'arrête pas, cela implique en particulier qu'il y a un nombre infini des vecteurs ω_i.
on va déterminer des bornes pour ; ω_i^2 et $\omega.V$.

1. ***Premier cas*** **:** on présente un exemple X de $\mathcal{E}_1$:
   - si $\left\langle\omega^{(k)},X\right\rangle\geqslant 0$, alors la réponse fournie par le neurone sera juste, ce qui fait que l'algorithme ne modifiera pas les poids.
   - si $\left\langle\omega^{(k)},X\right\rangle<0$, alors la réponse est fausse ce qui fait que les poids seront modifiés,donc : $\langle V,X\rangle>\varepsilon$ et : $\omega^{(k+1)}=\omega^{(k)}+\eta\left(d-S\right).X$, d'où :

$$\begin{aligned}\left(\omega^{(k+1)}\right)^2=\left(\omega^{(k)}+\eta\left(d-S\right).X\right)^2&=\left(\omega^{(k)}+\eta\left(1-0\right).X\right)^2\\&=\left(\omega^{(k)}\right)^2+2\eta.\left\langle\omega^{(k)},X\right\rangle+\left(\eta.X\right)^2\\&\leqslant\left(\omega^{(k)}\right)^2+\eta^2.M\end{aligned}$$

   d'une autre part : $\left\langle\omega^{(k+1)},V\right\rangle=\left\langle\omega^{(k)}+\eta.X,V\right\rangle=\left\langle\omega^{(k)},V\right\rangle+\eta.\langle X,V\rangle\geqslant\left\langle\omega^{(k)},V\right\rangle+\eta.\varepsilon$
2. ***deuxième cas*** **:** on présente maintenant un exemple X de $\mathcal{E}_0$:
   - si $\left\langle\omega^{(k)},X\right\rangle<0$, alors la réponse fournie par le neurone sera juste, ce qui fait que l'algorithme ne fera aucune modification sur les poids.
   - si $\left\langle\omega^{(k)},X\right\rangle\geqslant 0$, alors la réponse est fausse ce qui fait que les poids seront modifiés,donc : $\langle V,X\rangle<-\varepsilon$ et : $\omega^{(k+1)}=\omega^{(k)}+\eta\left(d-S\right).X$, d'où :

$$\begin{aligned}\left(\omega^{(k+1)}\right)^2=\left(\omega^{(k)}+\eta\left(d-S\right).X\right)^2&=\left(\omega^{(k)}+\eta\left(0-1\right).X\right)^2\\&=\left(\omega^{(k)}\right)^2-2\eta.\left\langle\omega^{(k)},X\right\rangle+\left(\eta.X\right)^2\\&\leqslant\left(\omega^{(k)}\right)^2+\eta^2.M\end{aligned}$$

   d'une autre part : $\left\langle\omega^{(k+1)},V\right\rangle=\left\langle\omega^{(k)}-\eta.X,V\right\rangle=\left\langle\omega^{(k)},V\right\rangle-\eta.\langle X,V\rangle\geqslant\left\langle\omega^{(k)},V\right\rangle+\eta.\varepsilon$

a. bien sûr cet ensemble est majoré car il est fini.
b. en fait pour cet algorithme si deux poids successifs sont identiques, alors il va aboutir son état de stabilité et il ne modifiera aucun poids par la suite.

finalement, on déduit que dans tout les cas, et pour n'importe quel $k \in \mathbb{N}$, on a :
$\left\langle \omega^{(k+1)}, V \right\rangle \geqslant \left\langle \omega^{(k)}, V \right\rangle + \eta.\varepsilon$ et : $\left(\omega^{(k+1)}\right)^2 \leqslant \left(\omega^{(k)}\right)^2 + \eta^2.M$.
on en déduit que pour tout : $k \geqslant 1$:

$$\left\langle \omega^{(k)}, V \right\rangle \geqslant \left\langle \omega^{(0)}, V \right\rangle + k \times \eta.\varepsilon \text{ et : } \left(\omega^{(k)}\right)^2 \leqslant \left(\omega^{(0)}\right)^2 + k \times \eta^2.M$$

comme : $\varepsilon > 0$, on trouvera forcement un $k_0 \in \mathbb{N}$**,** prenez par exemple : $k_0 = \left[\frac{-\left\langle \omega^{(0)}, V \right\rangle}{\varepsilon}\right] + 1$**,** tel que :$\left\langle \omega^{(0)}, V \right\rangle + k \times \eta.\varepsilon > 0$.
donc pour :$k \geqslant k_0$ on a :

$$\left(\langle \omega^{(0)}, V \rangle + k \times \eta.\varepsilon\right)^2 \leqslant \left(\langle \omega^{(k)}, V \rangle\right)^2 \leqslant V^2 \times \left(\omega^{(0)}\right)^2 + k \times \eta^2.M\textbf{.}$$

ce qui est impossible, car le terme de gauche est un polynôme de deuxième degré en k alors que le terme de droite est linéaire en k.
d'où ce qu'il faut prouver. ■

Remarque 2.3.4 :

On peut voir via la démonstration précédente que dans le cas d'un choix fixé du taux d'apprentissage, ce dernier n'a pas d'effet sur la convergence de l'algorithme, car il restera convergent en tout cas, mais malgré cela, un bon choix de ce paramètre peut réduire le nombre de corrections exécutées ,en outre cet algorithme nous ramène (dans le cas d'un choix fixé) vers des pondérations convenables mais qui ne sont pas forcement optimales.

2.3.3 L'algorithme d'apprentissage de Wirdrow-Hoff :

Cet algorithme a été proposé en 1960 par les deux chercheurs américains **Bernard Widrow** et **Ted Hoff** dans leur rapport technique [62], son idée générale est de franchir la non dérivabilité de f par une annulation totale de cette fonction.
autrement dit on va prendre exactement l'algorithme de descente de gradient mais en supposant que la fonction f n'existe plus c'est à dire qu'on va l'appliquer juste à la partie Adaline de notre perceptron ce qui justifie le fait de l'appeler parfois l'algorithme d'apprentissage d'adaline et on aura dans ce cas :

Algorithme 4 Algorithme d'apprentissage de Widrow-Hoff.

Entrée: une base de données : $\left(X^h, d^h\right)_{1 \leq h \leq \ell}$, une fonction d'activation non dérivable f, des poids de départ :$\omega = \left(\omega_0, \cdots, \omega_n\right) \in \mathbb{R}^{n+1}$ et un nombre maximal d'itérations $N_{\max}$.
Sortie: les poids optimaux :$\omega = \left(\omega_0, \cdots, \omega_n\right)$

```
k ← 1 ;
tant que k ⩽ N_max faire:
    pour h allant de 1 à ℓ faire:
        pour i allant de 0 à n faire:
            u^h ← Σ_{j=0}^{n} ω_j^(k) x_j^h ;
            ω_i^(k+1) ← ω_i^(k) + η_k × (d^h − u^h) × x_i^h ;
        fin pour
    fin pour
fin tant que
retourner: ω = (ω_0, ··· , ω_n)
```

On remarque que cet algorithme garde la même procédure de correction que l'algorithme d'apprentissage du perceptron, mais il modifie les pondérations à chaque itération et dans tous les cas, et pas seulement au cas de la présence d'une erreur [a] .
l'avantage de cet algorithme par rapport au précédent est le fait qu'il ne s'intéresse pas à la sortie calculée par le neurone, ce qui fait que même si l'information fournie est erronée, (c'est-à-dire que certains exemples de l'échantillon sont mal classés par le perceptron), l'algorithme restera encore en cours d'exécution et de plus si on choisit convenablement le paramètre η, il va converger vers une solution optimale (voir [16] page : 244).

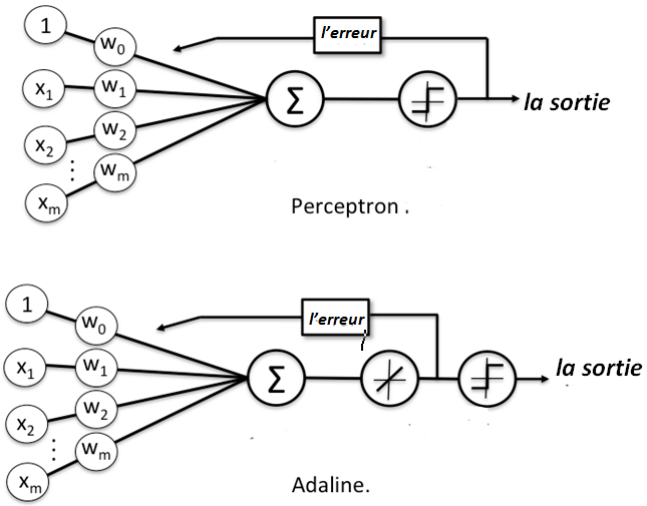


FIGURE II.20: La différence entre l'algorithme de Perceptron et celui de Widrow-Hoff.

Théorème 2.3.3 :

Si le problème déterminé par la base $\mathcal{B} = \left(X^h, d^h\right)_{1\leq h\leq \ell}$ admet une solution optimale ω^*, et si on affecte au pas d'apprentissage une valeur positive très petite alors les pondérations fournies par l'algorithme de Widrow-Hoff s'approchent [b], de cette solution à chaque itération.

Preuve :

Pour montrer la convergence de cet algorithme, nous allons calculer la distance entre les termes $\omega^k \in \mathbb{R}^{n+1}$ qu'il engendre et les pondérations optimales $\omega^* \in \mathbb{R}^{n+1}$, et dans ce cadre on a :

$$\begin{aligned}
\left\|\omega^{(k+1)}-\omega^*\right\|^2-\left\|\omega^{(k)}-\omega^*\right\|^2 &= \sum_{i=0}^{n}\left(\left(\omega_i^{(k+1)}-\omega_i^*\right)^2-\left(\omega_i^{(k)}-\omega_i^*\right)^2\right)\\
&= \sum_{i=0}^{n}\left(\left(\omega_i^{(k)}+\eta\times(d-u)\,x_i-\omega_i^*\right)^2-\left(\omega_i^{(k)}-\omega_i^*\right)^2\right)\\
&= \sum_{i=0}^{n}\left(2\times\eta\times(d-u)\left(\omega_i^{(k)}-\omega_i^*\right)x_i+\eta^2\times(d-u)^2x_i^2\right)\\
&= (d-u)\sum_{i=0}^{n}\left(2\eta\times\left(\omega_i^{(k)}-\omega_i^*\right)x_i+\eta^2(d-u)\,x_i^2\right)\\
\left(\text{car}: u=\sum_{i=0}^{n}\omega_i^{(k)}x_i \text{ et}: u^*=\sum_{i=0}^{n}\omega_i^*x_i\right) \quad &= 2\eta(d-u)\sum_{i=0}^{n}\left(\left(\omega_i^{(k)}-\omega_i^*\right)x_i\right)+\eta^2\times(d-u)^2\sum_{i=0}^{n}x_i^2\\
&= \eta^2(d-u)^2\|x\|^2-2\eta(d-u)(u^*-u)
\end{aligned}$$

a. Et pour ça on dit qu'il n'est pas un algorithme d'apprentissage par correction d'erreur
b. ***Attention!:*** s'approchent de cette solution et pas convergent vers elle.

si on prend maintenant un nombre M tel que : $\|X\| \leqslant M$ (par exemple : $M = \max_{h\in[\![1,\ell]\!]} \|x^h\|$), alors on aura :

$$\begin{aligned}
\eta^2 (d-u)^2 \|x\|^2 - 2\eta (d-u)(u^*-u) &\leqslant \eta^2 M^2 (d-u)^2 - 2\eta (d-u)(u^*-u) \\
&\leqslant \eta^2 M^2 (d-u)^2 - 2\eta (d-u)(u^*-d+d-u) \\
&\leqslant \eta^2 M^2 (d-u)^2 - 2\eta (d-u)^2 - 2\eta (d-u)(u^*-d) \\
&\leqslant \left(\eta M^2 - 2\right) \times \eta (d-u)^2 - 2\eta (d-u^*+u^*-u)(u^*-d) \\
&\leqslant \left(\eta M^2 - 2\right) \times \eta (d-u)^2 + 2\eta (d-u^*)^2 + 2\eta (d-u^*)(u^*-u) \\
\left(\text{car} : a\times b \leqslant \frac{1}{4}(a+b)^2\right) \quad &\leqslant \left(\eta M^2 - 2\right) \times \eta (d-u)^2 + 2\eta (d-u^*)^2 + \frac{\eta}{2}(d-u)^2 \\
&\leqslant \left(\eta M^2 - \frac{3}{2}\right) \times \eta (d-u)^2 + 2\eta (d-u^*)^2
\end{aligned}$$

donc on peut affirmer que :

$$\left\|\omega^{(k+1)} - \omega^*\right\|^2 - \left\|\omega^{(k)} - \omega^*\right\|^2 \leqslant \left(\eta M^2 - \frac{3}{2}\right) \times \eta (d-u)^2 + 2\eta (d-u^*)^2$$

de plus on a :

$$\begin{aligned}
\left(\eta M^2 - \frac{3}{2}\right) \times \eta (d-u)^2 + 2\eta (d-u^*)^2 \leqslant 0 &\iff 2\not{\eta} (d-u^*)^2 \leqslant \left(\frac{3}{2} - \eta M^2\right) \times \not{\eta} (d-u)^2 \\
&\iff 4(d-u^*)^2 \leqslant \left(3 - 2\times\eta M^2\right) \times (d-u)^2 \\
&\iff \underbrace{(d-u^*)^2}_{\text{erreur commise par } \omega^*} \leqslant \frac{\left(3-2\times\eta M^2\right)}{4} \times \underbrace{(d-u)^2}_{\text{erreur commise par } \omega}
\end{aligned}$$

donc si on prend η si petit, par exemple : $\eta \leqslant \dfrac{3}{2M^2}$, la suite $\left(\|\omega^{(k)} - \omega^*\|\right)_{k\in\mathbb{N}}$ sera décroissante, ce qui fait que les $\omega^{(k)}$ s'approchent à chaque itération de la solution optimale, d'où ce qu'il faut établir. ■

Remarque 2.3.5 :

Jusqu'à l'instant, il n'y a pas une règle qui peut nous garantir la convergence de l'algorithme d'apprentissage d'adaline vers la solution optimale,ou même vers un perceptron capable de classifier correctement tout les vecteurs d'entraînement (voir [40] pages 10 et 11).

3 Réseaux de neurones artificiels :

3.1 Nécessité d'introduire la structure de réseau :

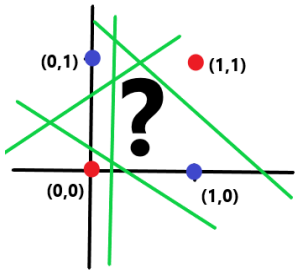
Parfois, on se pose dans des cas où un seul perceptron n'est plus capable de réaliser la tâche désirée ,et que la réalisation de cette dernière demande la contribution de plusieurs neurones ,à ce stade là on fera l'appel à ce qu'on appelle ***les réseaux neuronaux***.

Définition 3.1.1 :

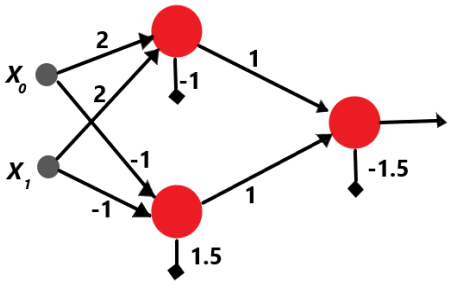
Un réseau neuronal est une collection de plusieurs neurones reliés et inter-connectés entre eux selon une architecture déterminée.
une architecture désigne en fait la topologie ,la façon ou la méthode selon laquelle les neurones

de notre réseau sont inter-connectés, on verra par la suite quelques exemples célèbres d'architectures.

Exemple 3.1.1 :



(a) l'incapacité d'un seul perceptron face au problème **"Xor"**.



(b) Un réseau résolvant le problème :**"Xor"**.

3.2 Architectures des réseaux de neurones artificiels :

On distingue entre deux grands types d'architectures de réseaux de neurones : Les réseaux feedforward non récurrents et les réseaux dynamiques récurrents.

3.2.1 Les réseaux feedforward :

Dans un réseaux feedforward, appelé aussi statique ou non récurrent, la sortie courante d'un nœud ne peut jamais apporter un effet sur ses sorties futures, c'est à-dire qu'elle ne peut pas être injectée à son entrée ni directement ni indirectement via d'autres neurones, ce qui fait que, l'information circule dans un seule sens, de l'entrée vers la sortie, et ceci justifie bien l'utilisation du terme "feedforward".
généralement ce type de réseaux réalisent des transformations de la forme : $O = \Psi(x)$ où $x \in \mathbb{R}^n$ et $O \in \mathbb{R}^m$,sont respectivement les vecteurs d'entrée et de sortie de notre réseau, en plus dans son architecture la plus générale, l'entrée de chaque neurone est connectée à toutes les sorties des neurones précédents, mais, la plupart des réseaux neuronaux statiques utilisés , sont organisés en couches , et pour ça ils sont appelés des réseaux multicouches ou perceptrons multicouches .
un réseau multicouche comporte généralement : une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie,et dans un tel réseaux, la sortie de chaque neurone d'une couche est connectée seulement aux entrées de chaque neurone de la couche suivante.

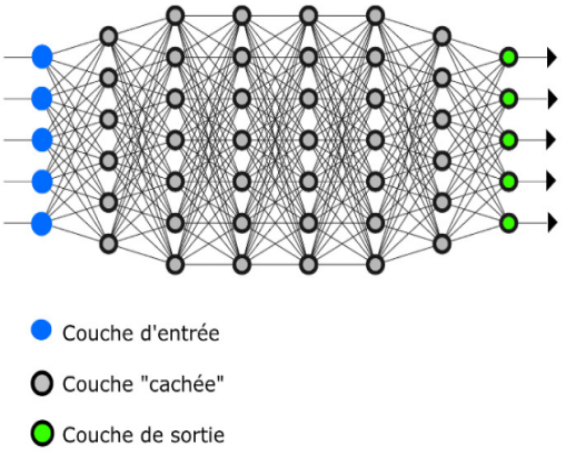


FIGURE II.22: La forme d'un réseau feedforward.

3.2.2 Les réseaux dynamiques :

Ces réseaux, qui sont appelés aussi réseaux récurrents, ou bouclés sont organisés de telle sorte que chaque neurone reçoit sur ses entrées en plus des informations externes, une partie ou la totalité de l'état du réseau c'est à dire des sorties d'autres neurones, ce qui s'interprète par le fait qu'il existe au moins un feedback d'une couche sur la précédente, autrement dit il existe une boucle dans l'ensemble de ses connexions.
dans ce cas l'état global du réseau dépend bien évidemment de ses états précédents, ce qui fait que l'équation du neurone, sera décrite par des suites récurrentes en temps discret, ou des équations différentielles en temps continu, et cela justifie bien l'emploi du terme "récurrent" .

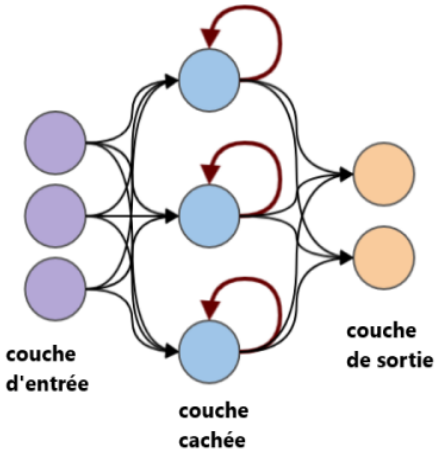


FIGURE II.23: La forme d'un réseau récurrent.

3.3 Quelques modèles de réseaux neuronaux artificiels :

Dans cette section on va essayer de présenter quelques modèles connexionnistes incontournables et qui décrient les différentes topologies de la grande majorité des réseaux neuronaux classiques.

3.3.1 Le modèle de Hopfield :

Le physicien **John Hopfield** $(1933,\cdots)$ a introduit en 1982 dans son article [32] un modèle connexionniste constitué de plusieurs neurones formels à sorties binaires, totalement connectés entre eux, et qui sont à la fois des unités d'entrée et de sortie, ce réseau récurrent, peut être décrit formellement comme un graphe complet non orienté valué $G = \left(N, [N]^2, f\right)$, doté d'une fonction de valuation symétrique f qui relie chaque paires (i, j) d'unités à la valeur de la pondération ω_{ij} correspondante.
les connexions dans un réseau Hopfield ont généralement les propriétés suivantes :

i) $\left(\forall (i,j) \in \{1,\cdots,n\}^2\right) : \omega_{ij} = \omega_{j,i}$.

ii) $\left(\forall (i,j) \in \{1,\cdots,n\}\right) : \omega_{ii} = 0$.

La mise à jour de la sortie d'une unité (un nœud du graphe qui simule un neurone artificiel) du réseau de Hopfield est effectuée à l'aide de la règle d'actualisation suivante :

$$\left(\forall i \in \{1,\cdots,n\}\right) : s_i^{(new)} = \begin{cases} +1 & \text{si} : \sum_{j=1}^{n} \omega_{ij} s_j^{(old)} \geq \theta_i, \\ 0 & \text{sinon.} \end{cases} .$$

cette procédure de mise à jour dans un réseau de Hopfield mis en valeur la notion du temps et montre qu'elle peut être effectuées de deux manières différentes :

Synchrone : toutes les unités sont mises à jour en même temps, ce qui nécessite une horloge centrale du système afin de maintenir la synchronisation, cette méthode est souvent considérée comme moins réaliste.

Asynchrone : chaque unité est mise à jour à la fois, et elle peut être choisie au hasard, ou à un ordre prédéfini, comme elle peut être imposé dès le début.

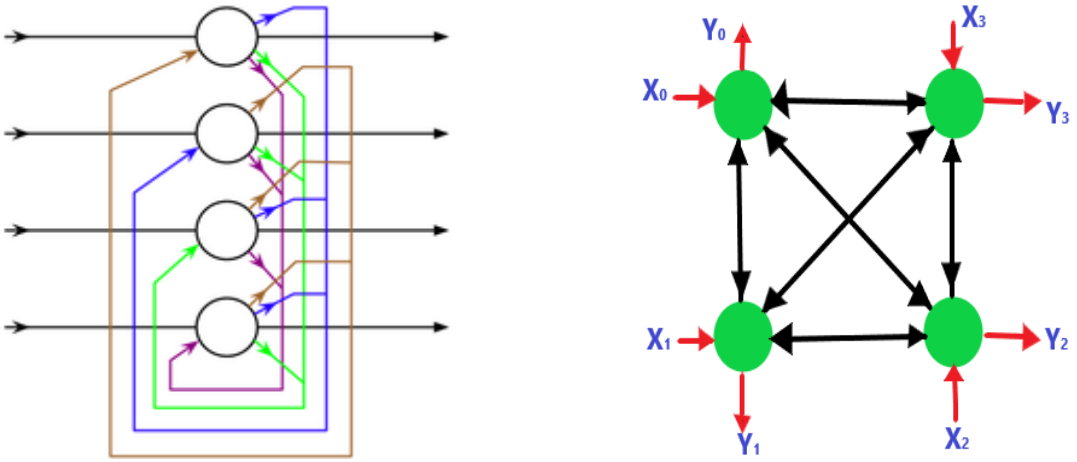


FIGURE II.24: Deux représentations différentes d'un réseau de Hopfield simple.

Remarque 3.3.1 :

Les modèles de Hopfield représentent une importance plus historique que pratique, et qui revient au fait de leur émergence pendant un tournant de l'histoire du connexionnisme, où ils ont été à la base de son redémarrage, en revanche ils ne sont plus utilisés maintenant dans leur version de base en raison de leur coût en terme de temps de calculs.

3.3.2 Le modèle de Kohonen :(la carte auto adaptative)

Le chercheur finlandais **Teuvo Kohonen** $(1934,\cdots)$ a proposé en 1982 dans son article [38] de projeter l'espace des données $\mathscr{D}$, sur un autre espace de faible dimension, en général 1, 2 ou $3D$ au maximum, cet espace de projection sera appelé une , qu'on la note $\mathscr{C}$ pour la suite .
cette carte est constitué d'un ensemble de neurones interconnectés selon une structure de graphe non orienté et qui peut prendre plusieurs formes,plan, cylindre,tore ...**etc** , mais en général on la réalise à partir d'un réseau neuronal à deux couches, une en entrée et une en sortie.
les neurones de la couche d'entrée sont entièrement connectés à ceux de la couche de sortie et qui sont placés dans un espace d'une ou de deux dimensions en général, chacun de ces neurones possède des voisins dans cet espace, et donc des connexions latérales récurrentes dans sa couche, ce qui va lui permettre de procéder en inhibant les neurones éloignés et laissant agir ceux qui sont voisins [a].
en somme le réseau de Kohonen est un réseau de neurones dont la particularité est d'agir en tant que compresseur de données, qui se base sur la quantification vectorielle pour conserver uniquement les informations caractérisantes.

a. la notion de voisinage nécessite l'introduction d'une distance, et souvent on prend celle qui est définie comme étant la longueur du plus court chemin entre deux sommets du graphe $\mathscr{C}$.

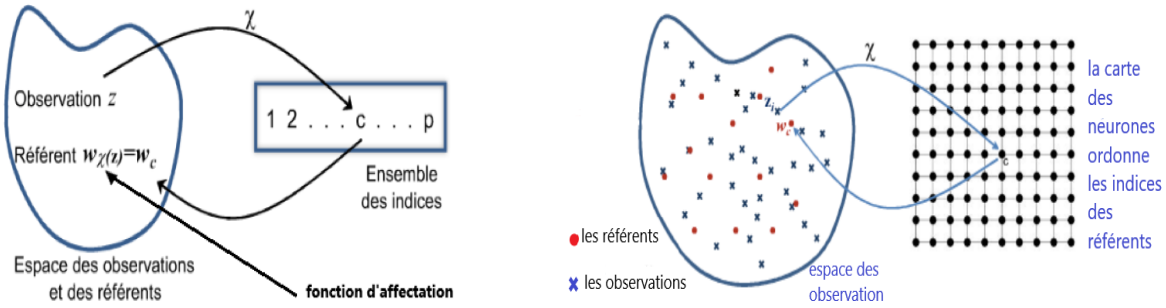


FIGURE II.25: Principe général de la modélisation par une carte de Kohonen.

3.3.3 Le Perceptron multicouche :

Le perceptron multicouche, et comme son nom l'indique est un réseau neuronal multicouche à propagation avant constitué de trois couches dont chacune a un rôle particulier.
La première couche s'appelle ***la rétine*** et son rôle est la réception de l'information de l'extérieur.
La deuxième s'appelle ***une couche cachée*** de pré-traitement : cette couche se situe entre la rétine et la couche de sortie ,elle est composée d'un nombre de neurones appelés neurones d'association et son rôle est de faire le pré traitement des informations provenantes de la rétine .
Et ***la dernière couche*** s'appelle une ***couche de sortie :*** tout simplement elle fait le traitement de ce qui provient de la couche cachée pour produire la sortie de notre réseau et il faut signaler qu'on plus de cet architecture particulière un perceptron est entièrement déterminé par ses pondérations .

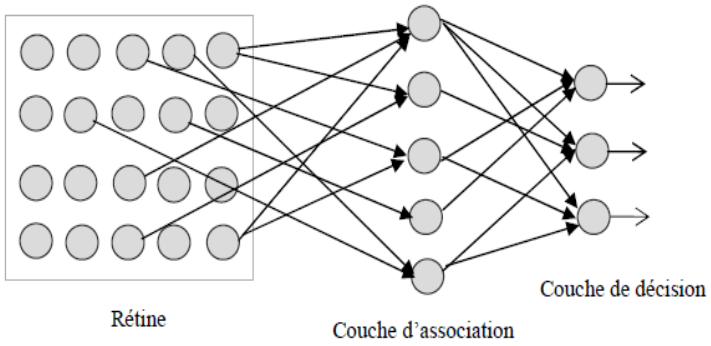


FIGURE II.26: Schéma d'un perceptron multicouches.

3.4 Apprentissage supervisé du perceptron multicouches :

Pour la suite, et afin d'éviter toute amboguïté, on signale qu'on se pose toujours dans le même cadre usuel d'apprentissage supervisé, et qui était déjà présenté pour le perceptron simple dans la page 35.

3.4.1 L'algorithme de Rétro-propagation du gradient :

Le problème de l'apprentissage pour les perceptrons multi-couches consiste à reconnaitre l'influence de chaque poids sur l'erreur globale du réseau, et l'algorithme de rétro-propagation du gradient permet de faire ceci en trois étapes :

1. la propagation de l'information de l'entrée jusqu'à la sortie.
2. le calcul de l'erreur en sortie.

3. la rétro-propagation de l'erreur de la sortie jusqu'aux entrées.

nous décrirons dans la suite les différentes équations qui permettent de faire la mise à jour des pondérations de notre réseau.

⋆ ***Pour la couche de sortie*** :
prenons un neurone j de la couche de sortie, lorsque le $n^{\text{ème}}$ ***exemple est lui présenté***, il produit une sortie $y_j(n)$ tandis que c'est la valeur $d_j(n)$ qui est attendue,ce qui fait qu'il commet une erreur :

$$e_j(n) = d_j(n) - y_j(n)$$

et de ce fait, l'erreur globale du réseau pour l'exemple n est :

$$\varepsilon(n) = \frac{1}{2}\sum_{j=1}^{m} e_j^2(n) = \left\| d(n) - y(n) \right\|_2 \quad (*)$$

où m est le nombre de neurones dans la couche de sortie.
comme d'habitude, le but de l'apprentissage est de minimiser l'erreur moyenne correspondante aux N exemples d'apprentissage, c'est à dire minimiser :

$$\varepsilon_{\text{moy}} = \frac{1}{N}\sum_{n=1}^{N} \varepsilon(n) \quad (**)$$

on note :

$$v_j(n) = \left(\sum_i \omega_{ij}(n) \times y_i(n)\right) - b_j(n) \quad (\blacklozenge)$$

l'entrée totale d'un neurone dont la fonction d'activation est f_j, sa sortie s'écrit alors sous la forme de

$$y_j(n) = f_j\left(v_j(n)\right)$$

Le mécanisme de rétro-propagation est basé sur l'effectuation d'une correction $\Delta\omega_{ji}(n)$ et $\Delta b_j(n)$ aux poids et aux biais.
en ce qui concerne les , on utilisera la règle dite du delta (delta rule), c'est à dire que la modification sera proportionnelle au gradient suivant :

$$\frac{\partial\varepsilon(n)}{\partial\omega_{ji}(n)} = \frac{\partial\varepsilon(n)}{\partial e_j(n)} \times \frac{\partial e_j(n)}{\partial y_j(n)} \times \frac{\partial y_j(n)}{\partial v_j(n)} \times \frac{\partial v_j(n)}{\partial\omega_{ji}(n)}$$

qui détermine la direction dans laquelle on recherche les valeurs de $\omega_{ij}(n)$,nous appellerons taux d'apprentissage le facteur de proportionnalité et nous le noterons η par la suite.
d'après l'équation $(**)$, on peut écrire :

$$\frac{\partial\varepsilon(n)}{\partial e_j(n)} = e_j(n)$$

de plus avec $*$:

$$\frac{\partial e_j(n)}{\partial y_j(n)} = f_j'\left(v_j(n)\right)$$

finalement et à partir de $(\blacklozenge)$:

$$\frac{\partial v_j(n)}{\partial\omega_{ji}(n)} = y_i(n)$$

soit $\delta_j(n)$ le gradient local défini par :

$$\delta_j(n) = -\frac{\partial \varepsilon(n)}{\partial e_j(n)} \times \frac{\partial e_j(n)}{\partial y_j(n)} \times \frac{\partial y_j(n)}{\partial v_j(n)}$$

d'après ce qui précède ceci est égal à :

$$\delta_j(n) = e_j(n) \times f_j'\Big(v_j(n)\Big) \quad (\blacklozenge\blacklozenge)$$

La correction appliquée au poids $\omega_{ji}(n)$ sera donc la suivante (règle du delta) :

$$\Delta\omega_{ji}(n) = -\eta \times \frac{\partial \varepsilon(n)}{\partial \omega_{ji}(n)}$$

soit encore :

$$\Delta\omega_{ji}(n) = \eta \times \delta_j(n) y_i(n)$$

du même, la correction apportée au biais s'écrira sous la forme

$$\Delta b_j(n) = \eta \times \frac{\partial \varepsilon(n)}{\partial b_j(n)}$$

en suivant un raisonnement analogue à ce qui précède on obtient

$$\begin{aligned}\frac{\partial \varepsilon(n)}{\partial b_j(n)} &= \frac{\partial \varepsilon(n)}{\partial e_j(n)} \times \frac{\partial e_j(n)}{\partial y_j(n)} \times \frac{\partial y_j(n)}{\partial v_j(n)} \times \frac{\partial v_j(n)}{\partial b_j(n)} \\ &= -\delta_j(n) \times \frac{\partial v_j(n)}{\partial b_j(n)} \\ &= \delta_j(n)\end{aligned}$$

d'où :

$$\Delta b_j(n) = -\eta.\delta_j(n)$$

ces deux équations de mise à jour ne sont valables que pour la couche de sortie, car on a utilisé l'erreur de sortie $e_j(n) = d_j(n) - y_j(n)$.

⋆ ***Pour les couches cachées*** :
En ce qui concerne les autres couches, on ne peut plus utiliser cette formule pour l'erreur, alors nous serons forcement obligés de déterminer les nouvelles équations de mise à jour.
▷ considérons un neurone j sur la dernière couche cachée, (celle qui se place juste avant celle de sortie.), nous pouvons et en utilisant une formule analogue à $(\blacklozenge\blacklozenge)$ définir le gradient local par :

$$\begin{aligned}\delta_j(n) &= \frac{\partial \varepsilon(n)}{\partial y_j(n)} \times \frac{\partial y_j(n)}{\partial v_j(n)} \\ &= \frac{\partial \varepsilon(n)}{\partial y_j(n)} \times f_j'\Big(v_j(n)\Big)\end{aligned}$$

et d'après $(**)$ on peut écrire :

$$\begin{aligned}\frac{\partial \varepsilon(n)}{\partial y_j(n)} &= \sum_k e_k(n) \times \frac{\partial e_k(n)}{\partial y_j(n)} \\ &= \sum_k e_k(n) \times \frac{\partial e_k(n)}{\partial v_k(n)} \times \frac{\partial v_k(n)}{\partial y_j(n)}\end{aligned}$$

or, le neurone k est sur la couche de sortie, d'où

$$\begin{aligned}\frac{\partial e_k(n)}{\partial v_k(n)} &= \frac{\partial(d_k(n)-y_k(n))}{\partial v_k(n))} \\ &= \frac{\partial(d_k(n)-f_k(v_k(n))}{\partial v_k(n)} \\ &= f_k'\left(v_k(n)\right)\end{aligned}$$

en outre, et d'après (♦) :

$$\frac{\partial e_k(n)}{\partial y_j(n)} = \omega_{kj}(n)$$

donc nous obtenons :

$$\begin{aligned}\frac{\partial \varepsilon(n)}{\partial y_j(n)} &= -\sum_k e_k(n) \times f_k'\left(v_k(n)\right) \times \omega_{kj}(n) \\ &= -\sum_k \delta_k(n) \times \omega_{kj}(n)\end{aligned}$$

le gradient local pour un neurone de la dernière couche cachée est alors donné par :

$$\delta_j(n) = \left(\sum_k \delta_k(n) \times \omega_{kj}(n)\right) \times f_k'\left(v_j(n)\right)$$

ceci peut se généraliser et sans aucun souci aux autres couches internes, nous obtenons ainsi les règles de mise à jour des poids et des biais avec la méthode de rétro-propagation du gradient.
pour cette raison, les poids et les biais doivent être actualisés en utilisant :

$$\begin{aligned}\omega_{ji} &\longleftarrow \omega_{ji} + \eta \times \delta_j(n) \times y_j(n) \\ b_j &\longleftarrow b_j - \eta \times \delta_j(n)\end{aligned}$$

avec :

$$\delta_j(n) = e_j(n) \times f_j'(v_j(n))$$

et :

$$e_j(n) = \begin{cases} d_j(n) - y_j(n) & \text{si } j \text{ est un neurone de sortie.} \\ \sum_k \delta_k(n) \times \omega_{kj}(n) & \text{si } j \text{ est un neurone interne.} \end{cases}$$

Ainsi, l'erreur de sortie se trouve propagée vers l'entrée (c'est à dire dans le sens inverse de celui qui est utilisé pour propager un signal à travers le réseau) afin de corriger de couche en couche les valeurs des poids et des biais.

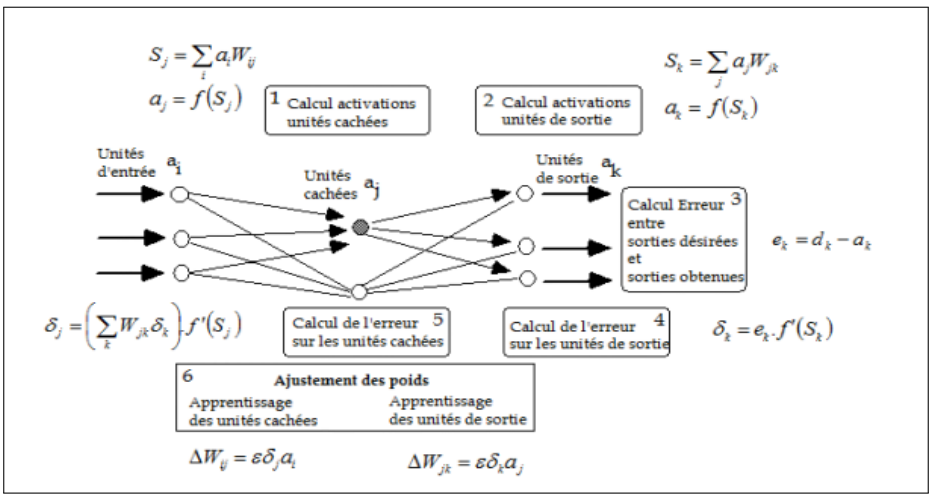


FIGURE II.27: Les étapes de l'algorithme de rétropropagation du gradient.

Algorithme 5 Algorithme de rétropropagation du gradient.

Entrée: une base de données : $\left(X^h, d^h\right)_{1\leq h\leq \ell}$, une fonction d'activation non dérivable f.
Sortie: les poids optimaux : ω_{ij}.

1. **Initialisation** :
   initialiser tous les poids et les biais du réseau de neurones.
2. **Présentation d'un exemple en entrée :**
   présenter un exemple en entrée $X^n = \left(x_1(n), x_2(n), \cdots, x_p(n)\right)$ ainsi que la sortie désirée correspondante $d^n = \left(d_1(n), d_2(n), \cdots, d_m(n)\right)$.
3. **Calculer les sorties :**
   calcules successivement les sorties des différentes couches pour l'entrée x.
4. La mise à jour des poids et des biais :
   modifier récursivement les poids et les biais du réseau suivant les règles détaillées avant :

$$\omega_{ji} \longleftarrow \omega_{ji} + \eta \times \delta_j(n) \times y_j(n)$$

$$b_j \longleftarrow b_j - \eta \times \delta_j(n)$$

   avec

$$\delta_j(n) = e_j(n) \times f_j'(v_j(n))$$

   et

$$e_j(n) = \begin{cases} d_j(n) - y_j(n) & \text{si } j \text{ est un neurone de sortie} \\ \sum\limits_k \delta_k(n) \times \omega_{kj}(n) & \text{si } j \text{ est un neurone interne} \end{cases}$$

5. **Critère d'arrêt :**
   répéter les étapes 2, 3 et 4 jusqu'au nombre maximum d'itérations ou jusqu'à ce que l'erreur quadratique moyenne soit inférieure à un certain seuil.

retourner: ω_{ij}

Conclusion :

Á partir du cerveau humain et du comportement des éléments neuronaux qui le constituent, on a pu construire des modèles artificiels très complexes appelés : "***réseaux de neurones artificiels***", dans ce chapitre on a bien traité ces réseaux, et on spécifié leurs définitions, leurs types, leurs concepts de base et aussi leurs mécanisme d'apprentissage, surtout pour le perceptron multicouches.

ces réseaux ont formé un outil très puissant, capable de traiter plusieurs familles de problèmes, que ça soit pour des problèmes de classification, de régression, de clustring, de prédiction …etc, et cette capacité revient essentiellement à leurs propriétés d'approximations universelles, et qui feront effectivement l'objet du chapitre suivant.

Chapitre III

Aptitude des Réseaux de Neurones Artificiels. (Théorie d'approximation) :

Résumé :

Dans ce chapitre, nous démontrons que les combinaisons linéaires ***finies*** du compositions d'une fonction ***fixe*** uni-variée (ou d'une seule variable réelle) et d'un ensemble de fonctionnelles affines peuvent approximer,***uniformément***, toute fonction continue de n variables réelles ***avec un support inclus dans l'hypercube unité*** ;notons que,seules des conditions faibles sont imposées à notre fonction uni-variée.

les résultats de ce chapitre vont répondre à une question très importante [b] et qui porte sur la représentabilité de la classe des réseaux de neurones ***à une seule couche cachée***.

En particulier, nous montrons que des régions de décision arbitraires peuvent être ***bien approchées*** [c] par des réseaux de neurones à propagation en avant (feedforward neural network) et avec une seule couche interne dont les nœuds ont tous une fonction d'activation ***sigmoïdale continue*** et ***commune***.

on va discuter aussi des propriétés d'approximation d'autres types possibles de fonctions,et qui peuvent être mises en œuvre par des réseaux de neurones artificiels.

Mots clés : Réseaux de neurones, approximation, complétude,densité.

Introduction :

Un certain nombre de domaines d'application,sont basés sur l'utilisation des fonctions de vecteurs réelles à n dimensions,$\left(x \in \mathbb{R}^n\right)$,et qui s'expriment par une combinaison linéaire ***finie*** de la forme :

$$\boxed{f(x) = \sum_{j=1}^{N} \alpha_j \, \sigma\left(\omega_j^{\top} x + \theta_j\right)} \tag{III.1}$$

où $\omega_j \in \mathbb{R}^n$, $\theta_j \in \mathbb{R}$, $\alpha_j \in \mathbb{R}$ et $\sigma \in \mathscr{F}(\mathbb{R},\mathbb{R})$ sont ***fixés*** [d] .($\omega^{\top}$ désigne ici la transposée de vecteur ω donc,et sans surprise $\omega^{\top} x$ ne va être que le produit scalaire usuel de x et ω).

La fonction $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$ dépend fortement du contexte de l'application,mais,notre principale

b. en fait elle était ouverte jusqu'au 1988,avant les travaux de **A.R.Gallant et H.White**.
c. on verra après dans quel sens.
d. c'est à dire qu'ils sont considérés comme des paramètres.

préoccupation concerne souvent ce qu'on appelle ***les fonctions sigmoïdales*** [a] **,**car elles apparaissent naturellement dans la théorie des réseaux neuronaux en tant que fonctions d'activation d'un nœud neuronal (ou d'une unité, qui devient plutôt le terme le plus courant).
Le but principal de ce chapitre est de démontrer le fait que les sommes de la forme III.1, avec σ est sigmoïdale continue,(une condition qui peut être modifiée ou plus affaiblie par la suite.) sont ***denses*** dans l'espace des fonctions continues sur l'hypercube unitaire de dimention n **.**

Les utilisations possibles d'un réseau de neurones artificiels et surtout dans le traitement et le contrôle du signal,a généré une attention considérable depuis les années quatre-vingt-dix ,et on peut constater en adoptant un point de vu un peu "mathématique", que sa description reste très facile,malgré sa puissance et ses différentes applications,or,un tel réseau est formé à partir de compositions et de superpositions d'***une*** simple fonction d'activation [b],en conséquence,la sortie qu'il fournit n'est autre que la valeur de la fonction qui résulte de cette composition.

En particulier, la plus simple classe non triviale de réseaux,qu'on peut imaginer,sont ceux d'une seule couche interne et ils ***implémentent*** exactement ***la classe de fonctions donnée par (1)***.

Donc,vous pouvez dire que ce chapitre vise de mettre en évidence le fait qu'un réseau de neurones mono-couche,***et d'un point de vu théorique***,possède une grande aptitude,et il est presque ***capable de tout faire***.

1 Théorèmes d'approximation au cas d'une fonction d'activation discriminatoire :

Pour une bonne lecture de cette section, il est indispensable de se référer à l'annexe A page : 124, qui va ***actualiser et enrichir*** le pré-acquis de notre lecteur, par quelques théorèmes d'analyse fonctionnelle et d'intégration, surtout ceux qui ***vont apparaitre*** pour légitimer et justifier quelques passages dans ***le cadre de l'approximation des fonctions par des réseaux de neurones artificiels***.
pour cette raison,nous vous invitons d'***aller voir cette annexe avant de traiter ce qui suit.***

1.1 Vocabulaire et quelques notions de base :

Définition 1.1.1 :(mesure extérieure :)

On appelle une ***mesure extérieure*** ou ***mesure de Crathéodory*** sur un ensemble Ω toute application :

$$\begin{array}{rccc} \upsilon & : & \mathcal{P}(\Omega) & \longrightarrow & \overline{\mathbb{R}}^{+} = [0,+\infty] \\ & & A & \longmapsto & \upsilon(A) \end{array}$$

qui vérifie les trois propriétés suivantes :

1. $\upsilon(\varnothing) = 0$.
2. $\left(\forall (A,B) \in \mathcal{P}(\Omega)\right) : \quad A \subset B \implies \upsilon(A) \leqslant \upsilon(B).$ (monotonie.)
3. $\left(\forall (A_i)_{i\in\mathbb{N}} \in \mathcal{P}(\Omega)^{\mathbb{N}}\right) : \quad \upsilon\left(\bigcup_{i\in\mathbb{N}} A_i\right) \leqslant \sum_{i\in\mathbb{N}} \upsilon(A_i).$ (sous additivité dénombrable.)

Exemple 1.1.1 :

Il est clair que toute mesure positive sur $\left(\Omega, \mathcal{P}(\Omega)\right)$ est en particulier une mesure extérieure.

a. tout simplement,ce sont des fonctions $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$ qui vérifient les propriétés suivantes :
$$\lim_{x\to-\infty} \sigma(x) = 0 \quad \text{et} \quad \lim_{x\to+\infty} \sigma(x) = 1.$$

b. ***Attention!***, ***on a supposé*** ici que les unités constituant notre réseaux sont ***identiques***.

Définition 1.1.2 :(mesurabilité au sens de Carathéodory :)

Soit Ω un ensemble quelconque(même vide),et υ une mesure extérieure sur cet ensemble,une sous partie $E \subset \Omega$ est dite υ-***mesurable*** ou mesurable au sens de Carathéodory par υ si et seulement si :

$$\left(\forall A \in \mathcal{P}(\Omega)\right) : \quad \upsilon(A) = \upsilon\left(E^c \cap A\right) + \upsilon(E \cap A).$$

autrement dit,un sous ensemble est υ mesurable si et seulement si elle divise toute partie A de Ω d'une manière additive par rapport à la mesure extérieure υ.

Remarque 1.1.1 :

Notons qu'en vertu de la sous $\sigma-$additivité, l'inégalité :$\upsilon(A) \leqslant \upsilon\left(E^c \cap A\right) + \upsilon(E \cap A)$ aura toujours lieu ;donc pour vérifier qu'une partie E est υ-mesurable, il suffit de vérifier l'inégalité inverse.

Définition 1.1.3 :(mesure de Borel régulière :)

Une mesure extérieure υ sur $\mathbb{R}^n$ est dite :mesure régulière de Borel si et seulement si : $\left(\forall B \in \mathfrak{B}_{\mathbb{R}^n}\right)$:B est υ-mesurable au sens de Carathéodorie, et $\left(\forall A \subset \mathbb{R}^n\right)\left(\exists B \in \mathfrak{B}_{\mathbb{R}^n}\right)$:$A \subset B$ et $\upsilon\left(A\right) = \upsilon\left(B\right)$.

Définition 1.1.4 :(mesure signée :)

Dans un espace mesurable, $\left(\Omega, \mathscr{F}\right)$ une mesure signée est tout simplement une fonction $\mu : \mathscr{F} \to \overline{\mathbb{R}}$ (à valeurs dans :$\overline{\mathbb{R}}$), qui vérifie :

1. $\left(\exists A \in \mathscr{F}\right)$:$\mu(A) = +\infty$ (resp $-\infty$) $\Rightarrow \left(\forall A \in \mathscr{F}\right)$:$\mu(A) \neq -\infty$(resp $+\infty$) .
2. $\mu(\varnothing) = 0$.
3. pour toute suite d'éléments deux à deux disjointes $\left(A_i\right)_{i \in \mathbb{N}} \in \mathscr{F}^{\mathbb{N}}$ on a : $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$.

c'est à dire qu'elle garde exactement la même définition qu'une mesure ordinaire sauf qu'on change l'espace d'arrivée, et qu'on ajoute la condition *"1"* pour qu'elle soit bien définie.

Exemples 1.1.2 :

- Considérons ***une mesure positive*** μ sur un espace mesuré $\left(\Omega, \mathscr{F}\right)$,et prenons une fonction $\left(\Omega, \mathfrak{B}_{\mathbb{R}}\right)$ mesurable $f : \Omega \to \mathbb{R}$ telle que :$f \in L^1\left(\Omega, \mathscr{F}\right)$,c'est à dire que : $\int_{\Omega} \left|f(x)\right| \, \mathrm{d}\mu(x) < +\infty$ alors :
  l'application $\upsilon : \mathscr{F} \to \mathbb{R}$ donnée par : $\upsilon(A) = \int_A f(x) \, \mathrm{d}\mu(x)$ pour tout $A \in \mathscr{F}$, est bien définie, et en plus, ***c'est une mesure signée*** à valeurs dans $\mathbb{R}$,et pour lui permettre de prendre $+\infty$ (ou $-\infty$) comme valeur,il suffit d'affaiblir l'hypothèse selon laquelle f est absolument intégrable et ceci en la remplaçant par :

  $$\int_{\Omega} f^-(x) \, \mathrm{d}\mu(x) < +\infty, \left(\text{ou} \int_{\Omega} f^+(x) \, \mathrm{d}\mu(x) < +\infty\right)$$

  avec :$f^+ = \max\left(f, 0\right)$ et $f^- = \max\left(-f, 0\right)$.
- Considérons cette fois ci ***deux mesures positives*** μ et ν sur un espace mesuré $\left(\Omega, \mathscr{F}\right)$,et ***supposons que l'une d'elles est finie*** ,alors, et sans surprise :l'application :$\upsilon = \mu - \nu$ sera bien ***une mesure signée*** à valeurs dans $\overline{\mathbb{R}}$,mais en revanche,et ce qui est un peu surprenant est le fait que ***n'importe qu'elle mesure signée est forcement de cette forme***,c'est un résultat très important,mais un peu compliqué à démontrer surtout qu'il se base essentiellement sur ***le***

théorème de décomposition de Hahn [a], donc ne vous laissez pas tromper par la sympathie de l'énoncé.

Remarques 1.1.2 :

1. Malgré cette définition qui est simple d'apparence,***l'étude des mesures signées est très compliquée***, et même elle présente un champ de recherche riche dont pas mal de questions sont encore ouvertes [b],surtout que plusieurs propositions, comme le lemme des classes monotones ou le théorème de Dynkin-Sierpinski [c] ou même celui de convergence monotones de Beppo-Levi ne restent plus valables,ce qui va poser ,et même dans notre cadre de travail, beaucoup de problèmes [d].
2. Pour n'importe qu'elle notion analogue,comme mesure extérieure,mesure régulière,mesure de Borel...etc,l'attribution du terme ***"signée"*** signifie la considération de notre mesure comme une différence de deux mesures positives de même type,avec ,bien sûr, une réalisation d'un changement de l'espace d'arrivée de $\overline{\mathbb{R}}^+$ à $\overline{\mathbb{R}}$ et l'ajout de la condition *"1"* pour éviter la forme indéterminée "$(+\infty)+(-\infty)$".

Définition 1.1.5 :(la transformée de Fourier d'une mesure finie :)

Soit μ une mesure signée ***finie*** sur $\left(\mathbb{R}^d, \mathfrak{B}_{\mathbb{R}^d}\right)$, on appelle ***transformée de Fourier*** de μ la fonction $\widehat{\mu}:\mathbb{R}^d \to \mathbb{C}$, définie par :

$$\boxed{\left(\forall\, t \in \mathbb{R}^d\right) : \quad \widehat{\mu}(t) := \int_{\mathbb{R}^d} \exp\left(i\langle t, x\rangle\right)\, \mathrm{d}\mu(x).}$$

Remarques 1.1.3 :

- On notera que c'est l'hypothèse de la finitude de la mesure $\left(\mu(\mathbb{R}^d) < +\infty\right)$ qui nous a permis de donner un sens à la transformée de Fourier,car elle rend automatiquement la fonction $\Psi_t : x \mapsto \exp\left(i\langle t, x\rangle\right)$ intégrable,sur $\mathbb{R}^d$ pour tout $t \in \mathbb{R}^d$,or :

$$\int_{\mathbb{R}^d} \exp\left(i\langle t, x\rangle\right)\, \mathrm{d}\mu(x) \leqslant \int_{\mathbb{R}^d} \left|\exp\left(i\langle t, x\rangle\right)\right|\, \mathrm{d}\mu(x) = \int_{\mathbb{R}^d} 1\, \mathrm{d}\mu(x) = \mu\left(\mathbb{R}^d\right) < +\infty\,.$$

  mais faites attention,s'il n'est pas le cas, on ne parle plus d'une telle notion,par exemple :Ψ n'est pas λ_d-intégrable sur $\mathbb{R}^d$ ce qui fait qu'on ne peut pas définir $\widehat{\lambda_d}$.
- Lorsque μ est la loi P_X d'un vecteur aléatoire X,$\widehat{\mu}$ n'est autre que ***la fonction caractéristique*** de X c'est à dire :

$$\mu = P_X \implies \widehat{\mu}(t) = \varphi_X(t) = \mathbb{E}\left(e^{i\langle t, X\rangle}\right).$$

a. ce théorème affirme que : pour tout espace mesurable $(\Omega, \mathscr{F})$ et ***toute mesure signée*** μ définie sur $\mathscr{F}$, il existe deux ensembles mesurables :$P, N \in \mathscr{F}$ tel que :

$$\bullet\, P \cup N = \Omega\,,\, P \cap N = \varnothing \qquad \bullet\, (\forall E \in \mathscr{F}) : E \subseteq P \Rightarrow \mu(E) \geq 0 \qquad \bullet\, (\forall E \in \mathscr{F}) : E \subseteq N \Rightarrow \mu(E) \leq 0\,.$$

de plus, ***cette décomposition est unique presque surement***.

b. Par exemple quelles conditions on doit ajouter dans le cas des mesures signées , pour garder la validité du théorème de Levy, qui relie la convergence en loi d'une suite de variables aléatoires réelles, avec la convergence ponctuelle de leurs fonctions caractéristiques?.

c. Il énonce que deux mesures σ-finies confondues sur un π-système, seront aussi confondus sur la tribu qu'il engendre.

d. Par exemple au niveau de la démonstration que toute fonction sigmoïdal mesurable bornée est en particulier discriminatoire (voir la preuve du lemme 2.2.1 page 63).

- si la mesure μ est à densité par rapport à celle de Lebesgue, c'est à dire :

$$\left(\exists f \in L^1(\mathbb{R}^d)\right) \text{ tel que : } \left(\forall A \in \mathfrak{B}_{\mathbb{R}}\right) : \quad \mu(A) = \int_A f d\lambda$$

alors, la transformée de Fourier de la mesure μ sera exactement la transformée de Fourier usuelle de la fonction f.

Proposition 1.1.1 :

Soit μ une mesure signée finie sur $\left(\mathbb{R}^d, \mathfrak{B}_{\mathbb{R}^d}\right)$,on note $\widehat{\mu}$ sa transformée de Fourier associée,alors on a :

1. $\widehat{\mu}$ est toujours continue et bornée sur $\mathbb{R}^d$.
2. la ***transformation de Fourier***[a] $\mathcal{F}$ définie sur l'ensemble $M(\mathbb{R}^d)$ des mesures signées finies sur $\mathbb{R}^d$ par :

$$\begin{array}{rccc} \mathcal{F} & : & M(\mathbb{R}^d) & \longrightarrow & C^0(\mathbb{R},\mathbb{C}) \\ & & \mu & \longmapsto & \mathcal{F}(\mu) = \widehat{\mu} \end{array}$$

est une application ***bien définie***, ***linéaire***,et en plus **injective**, et c'est ce qui est très important (voir :[37]).

Preuve :

1. ⋆***Pour la continuité :*** on a d'une part : $\left(\forall t \in \mathbb{R}^d\right) : \quad \widehat{\mu}(t) := \int_{\mathbb{R}^d} \exp\left(i\langle t,x\rangle\right) \mathrm{d}\mu(x)$ et d'une autre :
   - pour tout $(x \in \mathbb{R}^d)$, $t \mapsto e^{i\langle t,x\rangle}$ est continue sur $\mathbb{R}^d$.
   - pour tout $(t \in \mathbb{R}^d)$, $x \mapsto e^{i\langle t,x\rangle}$ est continue sur $\mathbb{R}^d$.
   - pour tout : $t, x \in \mathbb{R}^d \left|e^{i\langle t,x\rangle}\right| \leqslant 1$, et car μ ***est fini***,on obtiendra immédiatement l'intégrabilité de la fonction constante $\mathbb{1}$,or :$\mu\left(\mathbb{R}^d\right) = \int_{\mathbb{R}^d} 1 \, \mathrm{d}\mu(x) < +\infty$.

   donc , et selon ***le théorème de continuité des intégrales paramétriques***[b] on peut déduire que : $\widehat{\mu}(t) := \int_{\mathbb{R}^d} e^{i\langle t,x\rangle} \, \mathrm{d}\mu(x)$ est continue.
   ⋆ ***pour la bornitude*** :,il suffit de remarquer que :
   $\left(\forall t \in \mathbb{R}^d\right) : \quad \int_{\mathbb{R}^d} e^{i\langle t,x\rangle} \, \mathrm{d}\mu(x) \leqslant \int_{\mathbb{R}^d} \left|e^{i\langle t,x\rangle}\right| \, \mathrm{d}\mu(x) = \int_{\mathbb{R}^d} 1 \, \mathrm{d}\mu(x) = \mu\left(\mathbb{R}^d\right) \Longrightarrow \widehat{\mu} \leqslant \mu\left(\mathbb{R}^d\right)$.
2. ⋆ ***montrons que l'application*** $\mathcal{F}$ ***est bien définie :*** d'abord, l'existence de l'image est un fait qui provient directement du premier point(la transformée de Fourier d'une mesure est toujours continue), et l'unicité est triviale,et on peut dire même qu'elle est vérifiée par définition.
   ⋆ ***maintenant pour la linéarité :***
   soit μ et ν deux mesures signées ***bornées*** sur $\mathbb{R}^d$,on a :

$$\left(\forall A \in \mathfrak{B}_{\mathbb{R}^d}\right) : \quad \int_{\mathbb{R}^d} \mathbb{1}_A \, \mathrm{d}(\mu+\nu) := [\mu+\nu](A) := \mu(A) + \nu(A) = \int_{\mathbb{R}^d} \mathbb{1}_A \, \mathrm{d}\mu + \int_{\mathbb{R}^d} \mathbb{1}_A \, \mathrm{d}\nu.$$

a. Attention,c'est la transformation de Fourier et pas la transformée,soyez prudents et faites la différence entre ces deux notions.

b. ***un théorème classique de l'analyse*** qui s'énonce ainsi :soit $(\Omega, \mathscr{F}, \mu)$ un espace mesuré et $f : (t,x) \mapsto f(t,x)$ une fonction de $I \times \Omega$ dans $\mathbb{C}$ (où I est un intervalle de R) on suppose que :
♣$\left(\forall t \in I\right)$: $x \mapsto f(t,x)$ est mesurable. ♣$\left(\forall x \in \Omega\right)$: $x \mapsto f(t,x)$ est continue sur I.
♣$(\exists \phi \in L\left((\Omega, \mathscr{F}, \mu); \mathbb{R}^+\right)$)tel que : $\left(\forall t \in I\right) \left|f(t,x)\right| \leq \phi(x) \, \mu \, p.p$ sur E .
alors la fonction $F : t \mapsto F(t) = \int_E f(t,x) \, \mathrm{d}\mu$ est bien définie et continue.

donc pour toute fonction numérique étagée mesurables positive (appelée également fonction simple) φ de la forme :$\varphi=\sum_{i=1}^{n} a_i \mathbb{1}_{A_i}$ où $\left(A_i\right)_{i\in\{1\cdots n\}} \subset \mathfrak{B}_{\mathbb{R}^d}$sont deux-à-deux disjoints et $(a_i)_{i\in\mathbb{N}} \subset \mathbb{R}^+$, on a :

$$\int_A \varphi \,\mathrm{d}(\mu+\nu) := \int_{\mathbb{R}^d} \sum_{i=1}^{n} a_i \mathbb{1}_{A_i} \,\mathrm{d}(\mu+\nu) = \sum_{i=1}^{n} a_i \int_{\mathbb{R}^d} \mathbb{1}_{A_i} \,\mathrm{d}(\mu+\nu) = \sum_{i=1}^{n} a_i \int_{\mathbb{R}^d} \mathbb{1} \,\mathrm{d}\mu + \sum_{i=1}^{n} a_i \int_{\mathbb{R}^d} \mathbb{1} \,\mathrm{d}\nu$$
$$= \int_{\mathbb{R}^d} \varphi \,\mathrm{d}\mu + \int_{\mathbb{R}^d} \varphi \,\mathrm{d}\mu. \ *$$

Soit maintenant f une fonction mesurable positive ***bornée***,d'une part et selon le théorème d'approximation, f est la limite simple d'une suite ***croissante*** de fonctions étagées mesurables positives $(\varphi_n)_{n\in\mathbb{N}}$,et d'une autre ,et comme on a déjà signalé au deuxième point de la remarque III.1.1.2, que n'importe quelle mesure signée peut s'écrire comme une différence de deux mesures positives,donc et en appliquant ce fait aux deux mesures μ et ν :on pourra bien écrire :$\mu=\mu^+-\mu^-$ et $\nu=\nu^+-\nu^-$ avec :μ^+,μ^- et ν^+,ν^- sont tous des mesures positives [a].

en outre, on a selon le résultat $*$:

$$\boxed{(\forall n\in\mathbb{N}) : \quad \int_{\mathbb{R}^d} \varphi_n d(\mu+\nu) = \int_{\mathbb{R}^d} \varphi_n \,\mathrm{d}\mu + \int_{\mathbb{R}^d} \varphi_n d\nu} \ \blacklozenge$$

et encore par le même résultat il est évident que :

$$(\forall n\in\mathbb{N}) : \int_{\mathbb{R}^d} \varphi_n \,\mathrm{d}\mu = \int_{\mathbb{R}^d} \varphi_n d(\mu^+ + \mu^-) = \int_{\mathbb{R}^d} \varphi_n \,\mathrm{d}\mu^+ + \int_{\mathbb{R}^d} \varphi_n \,\mathrm{d}\mu^- = \int_{\mathbb{R}^d} \varphi_n \,\mathrm{d}\mu^+ \overbrace{- \int_{\mathbb{R}^d} \varphi_n d(-\mu^-)}^{\text{par définition.}}.$$

or, et selon le théorème de la convergence monotone on a :

$$\lim_{n\to+\infty} \int_{\mathbb{R}^d} \varphi_n \,\mathrm{d}\mu^+ = \int_{\mathbb{R}^d} \lim_{n\to+\infty} \varphi_n \,\mathrm{d}\mu^+ = \int_{\mathbb{R}^d} f \,\mathrm{d}\mu^+ < +\infty$$
$$\lim_{n\to+\infty} \int_{\mathbb{R}^d} \varphi_n d(-\mu^-) = \int_{\mathbb{R}^d} \lim_{n\to+\infty} \varphi_n d(-\mu^-) = \int_{\mathbb{R}^d} f d(-\mu^-) < +\infty$$

donc et en faisant tendre $n \longrightarrow +\infty$ on peut déduire que :

$$\lim_{n\to+\infty} \int_{\mathbb{R}^d} \varphi_n \,\mathrm{d}\mu = \lim_{n\to+\infty} \int_{\mathbb{R}^d} \varphi_n \,\mathrm{d}\mu^+ - \lim_{n\to+\infty} \int_{\mathbb{R}^d} \varphi_n d(-\mu^-) = \int_{\mathbb{R}^d} \lim_{n\to+\infty} \varphi_n \,\mathrm{d}\mu^+ - \int_{\mathbb{R}^d} \lim_{n\to+\infty} \varphi_n d(-\mu^-)$$
$$= \int_{\mathbb{R}^d} f \,\mathrm{d}\mu^+ - \int_{\mathbb{R}^d} f d(-\mu^-) = \int_{\mathbb{R}^d} f \,\mathrm{d}\mu \lim_{n\to+\infty} \int_{\mathbb{R}^d} \varphi_n \,\mathrm{d}\mu \Rightarrow \boxed{\lim_{n\to+\infty} \int_{\mathbb{R}^d} \varphi_n \,\mathrm{d}\mu = \int_{\mathbb{R}^d} f \,\mathrm{d}\mu \lim_{n\to+\infty} \int_{\mathbb{R}^d} \varphi_n \,\mathrm{d}\mu}$$

ce qui veut dire que le théorème de convergence monotone est valable pour les mesures signées ,***mais à condition que la fonction limite soit bornée***.

en faisant de même pour les mesures ν et $\nu+\mu$ on obtiendra :

$$\boxed{\lim_{n\to+\infty} \int_{\mathbb{R}^d} \varphi_n \,\mathrm{d}\mu = \int_{\mathbb{R}^d} \lim_{n\to+\infty} \varphi_n d\nu = \int_{\mathbb{R}^d} f d\nu}$$

et :

$$\boxed{\lim_{n\to+\infty} \int_{\mathbb{R}^d} \varphi_n d(\mu+\nu) = \int_{\mathbb{R}^d} \lim_{n\to+\infty} \varphi_n d(\mu+\nu) = \int_{\mathbb{R}^d} f d(\mu+\nu)}$$

donc on obtiendra à travers les résultats obtenus précédemment,et en faisant tendre $n \to +\infty$ dans $\blacklozenge$ que :

$$\boxed{\int_{\mathbb{R}^d} f d(\mu+\nu) = \int_{\mathbb{R}^d} f \,\mathrm{d}\mu + \int_{\mathbb{R}^d} f d\nu.} \ **$$

prenons à ce stade f une fonction mesurable ***bornée*** quelconque (pas forcement positive),on a :$f=f^+-f^-$ avec :$f^+=\max(f,0)$ et $f^-=\max(-f,0)$.

a. Attention :cette écriture n'est pas unique,car on peut tout simplement ajouter et retrancher une certaine mesure positive ***finie***.

alors il suffit d'utiliser le résultat$**$ pour les fonctions f^+ et f^- ,pour déduire que :

pour toute fonction mesurable ***bornée*** f on a : $\boxed{\int_{\mathbb{R}^d} f\,d(\mu+\nu)=\int_{\mathbb{R}^d} f\ \mathrm{d}\mu+\int_{\mathbb{R}^d} f\,d\nu}$ ♦·

enfin,considérons une fonction complexe quelconque f,avec ***des parties réelles et imaginaires bornées***,on a :

$\int_{\mathbb{R}^d} f\,d(\mu+\nu)=\int_{\mathbb{R}^d} \mathcal{R}e(f)+\mathcal{I}m(f)i\ \mathrm{d}(\mu+\nu)=\int_{\mathbb{R}^d} \mathcal{R}e(f)\ d(\mu+\nu)+i\int_{\mathbb{R}^d} \mathcal{I}m(f)\ d(\mu+\nu)$

$=\int_{\mathbb{R}^d} \mathcal{R}e(f)\ \mathrm{d}\mu+\int_{\mathbb{R}^d} \mathcal{R}e(f)\ d\nu+i\int_{\mathbb{R}^d} \mathcal{I}m(f)\ \mathrm{d}\mu+i\int_{\mathbb{R}^d} \mathcal{I}m(f)\ d\nu$ (selon ♦)

$=\int_{\mathbb{R}^d} f\ \mathrm{d}\mu+\int_{\mathbb{R}^d} f\,d\nu$

donc et en appliquant ce résultat à la fonction $x\mapsto e^{i\langle t,x\rangle}=\underbrace{\cos(\langle t,x\rangle)}_{\text{fonction bornée}}+i\ \underbrace{\sin(\langle t,x\rangle)}_{\text{fonction bornée}}$ pour n'importe quel $t\in\mathbb{R}$ on déduit que :

$$\left(\forall \nu,\mu\in M(\mathbb{R}^d)\right):\quad \widehat{\mu+\nu}=\widehat{\mu}+\widehat{\nu}$$

en plus, et sans aucune difficulté :

$\left(\forall\lambda\in\mathbb{R}\right)\left(\forall\mu\in M(\mathbb{R}^d)\right)\left(\forall\, t\in\mathbb{R}^d\right):\quad \widehat{\lambda\mu}(t)=\int_{\mathbb{R}^d} e^{i\langle t,x\rangle}d(\lambda\mu(x))=\int_{\mathbb{R}^d}\lambda e^{i\langle t,x\rangle}\ \mathrm{d}\mu(x)$

$=\lambda\int_{\mathbb{R}^d} e^{i\langle t,x\rangle}\ \mathrm{d}\mu(x)=\lambda\widehat{\mu}(t).$

donc la transformation de Fourier est bien linéaire.

⋆ ***finalement l'injectivité de la transformation de Fourier :***

la preuve est vraiment difficile, et elle se base sur plusieurs lemmes qui nécessitent un prés-requis riche en ***analyse harmonique***[a],mais elle est classique et très renommée,on ne rentrera pas complètement dans ses détails,néanmoins on va vous présentez quand même les grandes lignes :

— pour $\sigma>0$ on définit la fonction suivante sur $\mathbb{R}^d$,en adoptant la notation $x=(x_1,x_2,\cdots,x_d)$:

$$g_{d,\sigma}(x)=\prod_{j=1}^{d} g_\sigma(x_j)=\prod_{j=1}^{d}\frac{1}{\sqrt{2\pi}}e^{-x_j^2/2\sigma^2}=\frac{1}{\sqrt{2\pi}}e^{-|x|^2/2\sigma^2}\ .$$

— on peut facilement ,et par un simple calcul, établir que la transformée de Fourier de la fonction g est :

$$\hat{g}_{d,\sigma}(x)=\prod_{j=1}^{d}\hat{g}_\sigma(x_j)=e^{-2\pi^2\sigma^2|x|^2}\ .$$

— d'une autre part on peut vérifier en utilisant le théorème de Fubini que :

$$(g_{d,\sigma}\star\mu)(x):=\int_{\mathbb{R}^d} g_{d,\sigma}(x-y)\mu\ \mathrm{d}y=\int_{\mathbb{R}^d}\hat{\mu}(t)e^{2i\pi\langle t,x\rangle-2\pi^2\sigma^2|t|^2}\ \mathrm{d}t\ .$$

et en plus,et c'est ce qui est très intéressant mais un peu dur à établir :toute fonction bornée h sur $\mathbb{R}^d$,l'intégrale : $\int_{\mathbb{R}^d} h\ \mathrm{d}\mu$ est la limite de : $\int_{\mathbb{R}^d}(g_{d,\sigma}\star\mu)h(x)\ \mathrm{d}x$ lorsque : $\sigma\to 0$.

— enfin il suffit de relier les trois points précédents pour arriver à l'injectivité de la trans-

a. C'est une branche des mathématiques qui étudie la représentation des fonctions ou des signaux comme superposition d'ondes de base qu'on les appellent les harmoniques, d'où le nom de la discipline.

formation de Fourier,or :

$$\begin{aligned}(\forall \nu,\mu\in M(\mathbb{R}^d)):\quad \widehat{\mu}=\widehat{\nu} &\Longrightarrow \int_{\mathbb{R}^d}\hat{\mu}(t)e^{2i\pi\langle t,x\rangle-2\pi^2\sigma^2|t|^2}\,\mathrm{d}t=\int_{\mathbb{R}^d}\hat{\nu}(t)e^{2i\pi\langle t,x\rangle-2\pi^2\sigma^2|t|^2}\,\mathrm{d}t.\\ &\Longrightarrow (g_{d,\sigma}\star\mu)(x)=(g_{d,\sigma}\star\nu)(x).\\ &\Longrightarrow \int_{\mathbb{R}^d}(g_{d,\sigma}\star\mu)h(x)\,\mathrm{d}x=\int_{\mathbb{R}^d}(g_{d,\sigma}\star\nu)h(x)\,\mathrm{d}x\quad(\forall h\in L^\infty(\mathbb{R}^d))\\ \text{en faisant tendre } \sigma\to 0 &\Longrightarrow \int_{\mathbb{R}^d}h\,\mathrm{d}\mu=\int_{\mathbb{R}^d}h\,\mathrm{d}\nu\quad(\forall h\in L^\infty(\mathbb{R}^d))\end{aligned}$$

donc il suffit de prendre : $h=\mathbb{1}_{[a,+\infty[}$ avec $a\in\mathbb{R}$ pour voir que μ et ν se coïncident sur le π-système générant la tribus borélienne $\mathfrak{B}_{\mathbb{R}^d}$ et par suite déduire selon le lemme des classes monotone (appelé également théorème de Dynkin-Sierpinski) l'égalité de ces deux mesures. ■

Exemple 1.1.3 :

Un exemple pratique de la transformée de Fourier d'une mesure est la fameuse ***série de Fourier***[a] et qui est une série de terme général $a_n e^{2in\pi x}$ indexée par : $n\in\mathbb{Z}$ lorsque les $a_n\in\mathbb{R}$ et $\sum_{n\in\mathbb{Z}}a_n<+\infty$.
la somme d'une telle série n'est autre que la transformée de Fourier de la mesure définie sur $\mathbb{R}$ par :$\mu=\sum_{n\in\mathbb{Z}}a_n\delta_{-n}$ où δ_{-n} désigne ***la masse de Dirac*** centrée,en $-n$[b].

Notations :

Pour toute la suite ,sauf mention contraire, on va adopter les notations suivantes :

— I_n désigne l'hypercube unité de dimension n :$[0,1]^n$.

— $C(I_n)$ est l'ensemble des fonctions ***numériques*** continues sur I_n.

— $(\forall f\in C(I_n))$ on note :$\|f\|=\sup_{x\in I_n}|f(x)|$.

— $M(I_n)$ est l'ensemble des mesures régulières de Borel sur I_n et qui sont ***signées et finies***.

Définition 1.1.6 :(fonction discriminatoire :)

On dit qu'une fonction : $\sigma:I_n\to\mathbb{R}$ est ***discriminatoire***,dans $M(I_n)$, si et seulement si :

$$\left(\forall\mu\in M(I_n)\right):\quad\left[\left(\forall\omega\in\mathbb{R}^n\right)\left(\forall\theta\in\mathbb{R}\right):\ \int_{I_n}\sigma\left(\omega^{\mathsf{T}}x+\theta\right)\,\mathrm{d}\mu=0\right]\Longrightarrow\mu=0\,.$$

Remarque 1.1.4 :

— Afin d'éviter toute confusion, nous proposons le diagramme suivant, qui représente les fonctions qu'on utilise,en montrant les différents paramètres et variables qui apparaissent dans leurs expressions.

a. ***les séries de Fourier*** qui ont été introduites par ***Joseph Fourier*** au 19$^{\text{éme}}$ siècle,présentent un outil puissant pour ***l'étude des fonctions périodiques*** et jusqu'à maintenant elles font encore un objet de recherches actives.
ces séries ont donné naissance à plusieurs branches nouvelles : analyse harmonique, théorie du signal, ..., etc, ***le but de leur introduction est de pouvoir écrire une fonction T-périodique comme la somme de fonctions sinusoïdales*** : $f(x)=\sum_{n=-\infty}^{+\infty}\lambda_n(f)e^{i2\pi\frac{n}{T}x}$ avec $\lambda_n(f)$, appelés coefficients de Fourier de f, définis par :

$$\lambda_n(f)=\frac{1}{T}\int_{-T/2}^{T/2}f(t)e^{-i2\pi\frac{n}{T}t}\mathrm{d}t.$$

b. la masse de Dirac centrée en un point $x_0\in\mathbb{R}^d$ (fixé à l'avance) est une mesure positive définie sur la tribu Borélienne de $\mathbb{R}^d$ comme suit : $(\forall A\in\mathfrak{B}_{\mathbb{R}^d}):\delta_{x_0}(A)=1_A(x_0)$

$$\begin{array}{ccccc} I_n & \xrightarrow{h} & \mathbb{R} & \xrightarrow{\sigma} & \mathbb{R} \\ x & \longmapsto & \omega^{\intercal}x+\theta & \longmapsto & \sigma(\omega^{\intercal}x+\theta) \end{array}$$

— pour la définition de la fonction discriminatoire [a] et qui était introduite pour la première fois dans l'article [13], le choix de l'espace $M(I_n)$ n'est pas obligatoire,et on peut aisément le changer par un autre espace de mesures.

Exemple 1.1.4 :

1. la fonction ***identité*** $\sigma_o : x \mapsto x$ est discriminatoire,dans l'ensemble $M^+(I_n)$ des mesures régulières positives de Borel sur I_n car :

$$\begin{aligned} \left(\forall \mu \in M^+(I_n)\right) : \quad \left[\left(\forall \omega \in \mathbb{R}^n\right)\left(\forall \theta \in \mathbb{R}\right) : \int_{I_n} \sigma_o\left(\omega^{\intercal}x+\theta\right)\,\mathrm{d}\mu = 0\right] &\Longrightarrow \int_{I_n} \sigma_o\left(0^{\intercal}_{\mathbb{R}^n}x+1\right)\,\mathrm{d}\mu = 0 \\ &\Longrightarrow \int_{I_n} 1\,\mathrm{d}\mu = \mu\left(I_n\right) = 0 \\ \left(\text{et puisque : } \left(\forall A \in \mathcal{P}(I_n)\right) : 0 \leqslant \mu(A) \leqslant \mu(I_n)\right) &\Longrightarrow \mu = 0. \end{aligned}$$

faites attention,cette fonction n'est pas du tout discriminatoire dans $M(I_n)$.

Définition 1.1.7 :(fonction sigmoïdale :)

On dit qu'une fonction : $\sigma : \mathbb{R} \to \mathbb{R}$ est sigmoïdale si et seulement si : $\lim\limits_{x\to-\infty} \sigma(x) = 0$ et $\lim\limits_{x\to+\infty} \sigma(x) = 1$.

Exemples 1.1.5 :

1. la fonction : $\varphi(x) : \mathbb{R} \to \mathbb{R}$ définie comme suit : $\varphi(x) = \frac{1}{2} \times \left(1 + \tanh\left(x + \sin\left(x^2\right)\right)\right)$. est une fonction sigmoïdale.

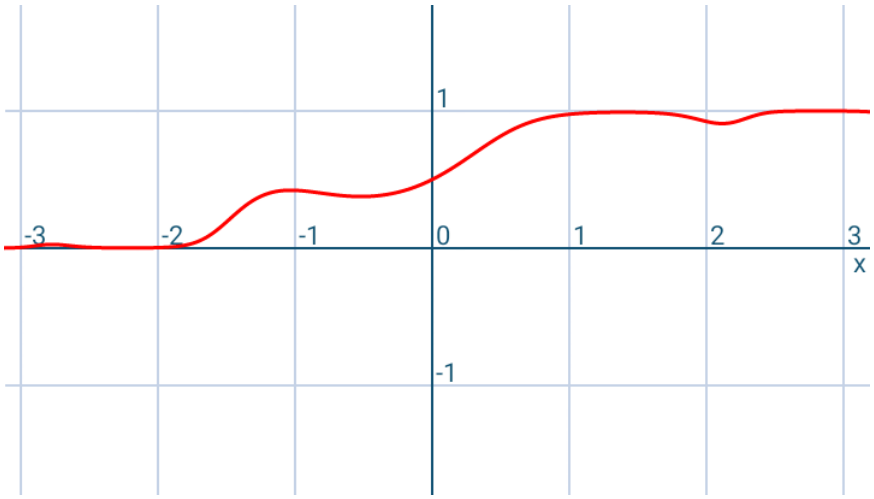


FIGURE III.1: La courbe de la fonction φ.

2. il est facile de remarquer que toute fonction de répartition de probabilité est en particulier sigmoïdale.

a. on peut dire également fonction discriminante,discriminative ou discriminatrice.

1.2 Résultats principaux :

Théorème 1.2.1 :

Soit σ une fonction ***continue discriminatoire*** dans $M(I_n)$,alors,l'ensemble :

$$\mathcal{N}_\sigma = \left\{ x \mapsto G(x) = \sum_{j=1}^{N} \alpha_j \, \sigma(\omega_j^\intercal x + \theta_j) \Big/ N \in \mathbb{N} \, , (\alpha_1, \cdots, \alpha_N) \in \mathbb{R}^N \, , (\theta_1, \cdots, \theta_N) \in \mathbb{R}^N \, , (\omega_1, \cdots, \omega_N) \in (\mathbb{R}^n)^N \right\}$$

est dense dans $C(I_n)$ ***pour la norme de la convergence uniforme***[a].
et comme il était énoncé dans [13] ,ceci veut dire que : pour toute fonction $f \in C(I_n)$ on a :

$$(\forall \varepsilon > 0) \left(\exists N_\varepsilon \in \mathbb{N}\right) \left(\exists (\alpha^\varepsilon, \theta^\varepsilon) \in \left(\mathbb{R}^{N_\varepsilon}\right)^2\right) \left(\exists (\omega_1^\varepsilon, \cdots, \omega_{N_\varepsilon}^\varepsilon) \in (\mathbb{R}^n)^{N_\varepsilon}\right) \text{tq} : \quad \left\| \sum_{j=1}^{N} \alpha_j^\varepsilon \, \sigma\left(\langle \omega_j^\varepsilon, \cdot \rangle + \theta_j^\varepsilon\right) - f \right\| < \varepsilon$$

Preuve :

D'abord,et avant de commencer, on peut voir que l'ensemble $\mathcal{N}_\sigma$ est un sous espace vectoriel de $C(I_n)$,(c'est une remarque simple, mais très importante),***notre but dans la suite est d'établir que :*** $\overline{\mathcal{N}_\sigma} = \mathbf{C(I_n)}$.
on procédera pour atteindre cet objectif par l'absurde :c'est à dire on va supposer que $\overline{\mathcal{N}_\sigma} \neq C(I_n)$,pour tomber enfin de compte sur une contradiction .
Premièrement,on sait que la fermeture d'un sous espace vectoriel est à son rôle un sous espace vectoriel[b] ce qui fait que $\overline{\mathcal{N}_\sigma}$,est,***un sous espace vectoriel*** de $C(I_n)$,***fermé*** pour la norme de la convergence uniforme (car par définition c'est l'adhérence de $\mathcal{N}_\sigma$),et en plus ***propre***, puisqu'on est parti de la supposition que : $\overline{\mathcal{N}_\sigma} \neq C(I_n)$.
on a : $\overline{\mathcal{N}_\sigma} \neq C(I_n) \implies \overline{\mathcal{N}_\sigma} \subsetneq C(I_n) \implies \left(\exists x_0 \in C(I_n)\right) : x_0 \notin \overline{\mathcal{N}_\sigma}$ (bien sûr $x_0 \neq 0$ car : $0 \in \mathcal{N}_\sigma \subset \overline{\mathcal{N}_\sigma}$).
on pose : $F = \overline{\mathcal{N}_\sigma} \oplus \mathbb{R}x_0$ et on considère l'application :

$$\begin{array}{rcccc} L & : & F = \overline{\mathcal{N}_\sigma} \oplus \mathbb{R}x_0 & \longrightarrow & \mathbb{R} \\ & & x + \lambda.x_0 & \longmapsto & L(x) = \lambda \end{array}$$

il est clair que L est bien définie et linéaire,et on peut affirmer sans problème que :$\text{Ker}\left(L\right) = \overline{\mathcal{N}_\sigma}$.
or :$\left(\forall x \in \overline{\mathcal{N}_\sigma}\right) : x = x + 0.x_0$, ce qui fait que :$\left(\forall x \in \overline{\mathcal{N}_\sigma}\right) : L(x) = 0$, donc :$\text{Ker}\left(L\right) \supset \overline{\mathcal{N}_\sigma}$ et :
$\left(\forall y = x + \lambda x_0 \in \overline{\mathcal{N}_\sigma} + \mathbb{R}x_0\right) : y \in \text{Ker}\left(L\right) \implies L(y) = \lambda = 0 \implies y = x \in \overline{\mathcal{N}_\sigma}$ donc :$\text{Ker}\left(L\right) \subset \overline{\mathcal{N}_\sigma}$.
maintenant,à travers le théorème du Noyau fermé A.1.1.1 et car $\text{Ker}\left(L\right) = \overline{\mathcal{N}_\sigma}$, on déduit que L est continue sur $F = \overline{\mathcal{N}_\sigma} \oplus \mathbb{R}x_0$,et selon le corollaire A.1.2.1,de théorème de Hahn-Banach on peut en conséquence prolonger L sur l'espace $E = C(I_n)$ tout entier, par une autre forme linéaire continue $\widetilde{L}$ et qui vérifie naturellement les deux propriétés suivantes :$\widetilde{\mathbf{L}} \neq \mathbf{0}$ et $\widetilde{\mathbf{L}}(\mathcal{N}_\sigma) = \widetilde{\mathbf{L}}(\overline{\mathcal{N}_\sigma}) = \mathbf{0}$.
si nous prolongeons une deuxième fois, par la même procédure, notre forme $\widetilde{L}$ à l'espace $L^2(I_n)$ qui est de Hilbert[c] pour le produit scalaire définie par : $\forall \left(f, g\right) \in \left(L^2\left(I_n\right)\right)^2 \quad \langle f, g \rangle := \int_{I_n} f.g \, \mathrm{d}\lambda$,
on pourra bien par suite s'appuyer sur le théorème de représentation de Riesz A.2.2.1 pour affirmer que :

a. il faut bien déterminer la norme qu'on utilise, car dans $C(I_n)$ qui est de dimension infini,la propriété de l'équivalence de toutes les normes n'est pas vérifiée (voir par exemple$\|\cdot\|_\infty$ et$\|\cdot\|_1$).

b. ceci provient tout simplement de la linéarité de la limite et du fait que dans les espace vectoriels normés, et même métriques ,l'adhérence d'une sous partie ,est l'ensemble des limites des suites convergentes à valeurs dans cette partie (Attention c'est pas valable dans les espaces topologiques quelconques).

c. ce deuxième prolongement est nécessaire car pour le produit scalaire proposé $C(I_n)$ n'est pas du tout de Hilbert.

$$\left(\exists! g \in L^2\left(I_n\right)\right) \text{ tel que : } \left(\forall h \in C(I_n)\right) : \widetilde{L}(h) = \langle g, h\rangle = \int_{I_n} h(x).g(x)\, \mathrm{d}\lambda(x)\,.$$

si on note : $\int_A g(x)\, \mathrm{d}\lambda = \mu(A) \Leftrightarrow g(x)\, \mathrm{d}\lambda(x) = \mathrm{d}\mu(x)$ cette quantité, et selon le premier point de l'exemple III.1.1.2 est une mesure extérieure signée finie [a], et le fait qu'elle est de Borel régulière peut être justifiée comme suit :

$\left(\forall B \in \mathfrak{B}_{I_n}\right)$:B est μ mesurable car :

$\left(\forall A \in \mathcal{P}(I_n)\right)$: $\mu\left(B^c \cap A\right) + \mu\left(B \cap A\right) = \int_{B^c \cap A} g(x)\, \mathrm{d}\lambda + \int_{B \cap A} g(x)\, \mathrm{d}\lambda = \int_A g(x)\, \mathrm{d}\lambda = \mu(A)$.

et : $\left(\forall A \subset I_n\right)\left(\exists B = \overline{A} \in \mathfrak{B}_{I_n}\right)$ tel que :$A \subset B$ et $\mu\left(B\right) = \int_{\overline{A}} g(x)\, \mathrm{d}\lambda = \int_A g(x)\, \mathrm{d}\lambda = \mu\left(A\right)$.

ainsi on obtient :

$$\mu \in M(I_n) \text{ et : } \boxed{\left(\forall h \in C(I_n)\right) : \widetilde{L}(h) = \int_{I_n} h(x)\, \mathrm{d}\mu(x)} \ \star\,.$$

d'après les hypothèses présentées dans l'énoncé,la fonction σ est continue ,alors :

$\left(\forall (\omega, \theta) \in \mathbb{R}^n \times \mathbb{R}\right)$: $\varphi : x \mapsto \sigma\left(\omega^\top x + \theta\right)$ est continue $\overset{\text{selon}\star}{\Longrightarrow} \left(\forall (\omega, \theta) \in \mathbb{R}^n \times \mathbb{R}\right)$: $\int_{I_n} \sigma\left(\omega^\top x + \theta\right)\, \mathrm{d}\mu = \widetilde{L}(\varphi)$

et d'une autre part :$\left(\forall (\omega, \theta) \in \mathbb{R}^n \times \mathbb{R}\right)$: $\varphi : x \mapsto \sigma\left(\omega^\top x + \theta\right) \in \mathcal{N}_\sigma$, donc et car : $\widetilde{L}(\mathcal{N}_\sigma) = \widetilde{L}(\overline{\mathcal{N}_\sigma}) = 0$,on déduit que : $\left(\forall (\omega, \theta) \in \mathbb{R}^n \times \mathbb{R}\right)$: $\int_{I_n} \sigma\left(\omega^\top x + \theta\right)\, \mathrm{d}\mu = \widetilde{L}(\varphi) = 0$,ce qui implique que $\mu = 0$ car σ est discriminatoire,alors : $\left(\forall h \in C(I_n)\right) : \widetilde{L}(h) = \int_{I_n} h(x)\, \mathrm{d}\mu(x) = 0$, cependant : $\widetilde{L} \neq 0$,***Absurde!*** .

donc $\overline{\mathcal{N}_\sigma} = C(I_n)$, d'où ce qu'il faut démontrer. ■

Ce théorème nous a démontré que les sommes de la forme (III.1) à condition que σ soit ***continue*** et ***discriminatoire***,sont denses dans $C(I_n)$,et son idée de la preuve était assez générale et peut être appliquée dans d'autres cas, comme ça sera indiqué ultérieurement.

Maintenant, nous spécialisons ce résultat en montrant que ***toute fonction sigmoïdale continue,est en particulier discriminatoire***, il convient de noter que, dans les applications de réseau neuronal,les fonctions d'activation sigmoïdale sont généralement considérées croissantes, mais ***aucune monotonie n'est requise dans nos résultats***.

Lemme 1.2.1 :

Toute fonction ***bornée mesurable sigmoïdale***,est discriminatoire dans $M(I_n)$, et en particulier toute fonction continue sigmoïdale l'est aussi, [b](voir pour cela :[13] et [12]).

Preuve :

Pour démontrer ceci,on note que pour n'importe quelle : $x, \omega \in \mathbb{R}^n$, $\theta, \varphi \in \mathbb{R}$ et $\lambda \in \mathbb{R}^+$ on a :

$$\sigma\left(\lambda(\omega^\top x + \theta) + \varphi\right) : \begin{cases} \xrightarrow[\lambda \to +\infty]{} 1 & \text{si} : \omega^\top x + \theta > 0 \\ \xrightarrow[\lambda \to +\infty]{} 0 & \text{si} : \omega^\top x + \theta < 0 \\ = \sigma(\varphi) & \text{si} : \omega^\top x + \theta = 0 \end{cases}$$

donc,la famille des fonctions :$(\sigma_\lambda)_{\lambda \in \mathbb{R}^+}$ converge simplement (ou point par point)vers la fonction :γ définie par :

$$\gamma(x) : \begin{cases} = 1 & \text{si} : \omega^\top x + \theta > 0 \\ = 0 & \text{si} : \omega^\top x + \theta < 0 \\ = \sigma(\varphi) & \text{si} : \omega^\top x + \theta = 0 \end{cases}$$

a. Cette mesure est finie car : $L^2(I_n) \subset L^1(I_n)$.

b. on reviendra à la démonstration de ce fait dans la preuve de deuxième théorème de Cybenko

on note $\Pi_{\omega,\theta}$ l'hyperplan ***affine*** donné par : $\{x \in \mathbb{R}^n \mid \omega^\top x + \theta = 0\}$ et $H_{\omega,\theta}$,le demi espace affine ***ouvert***[a] défini par : $\{x \in \mathbb{R}^n \mid y^\top x + \theta > 0\}$.
nous avons,et par hypothèse,la bornitude de la fonction σ, donc, les σ_λ sont équibornées,et en plus la convergence simple $\sigma_\lambda \xrightarrow[n\to+\infty]{} \gamma$ est vérifiée par construction,donc et en appliquant pour n'importe quelle mesure $\mu \in M(I_n)$ ***le théorème de la convergence dominée***[b],on obtient simultanément ***la μ-intégrabilité de γ et le droit de permuter l'intégrale et la limite***.

$$\left.\begin{array}{l} \bullet \ (\exists M > 0)\,(\forall \lambda > 0) : \sigma_\lambda < M. \\ \bullet \ (\forall x \in \mathbb{R}) : \sigma_\lambda(x) \xrightarrow[\lambda\to+\infty]{} \gamma(x). \\ \bullet \ \displaystyle\int_{I_n} M \,\mathrm{d}\mu = M \int_{I_n} 1 \,\mathrm{d}\mu < +\infty. \end{array}\right\} \Rightarrow \boxed{\gamma \text{ est intégrable et : } \lim_{\lambda\to+\infty} \int_{I_n} \sigma_\lambda \,\mathrm{d}\mu = \int_{I_n} \lim_{\lambda\to+\infty} \sigma_\lambda \,\mathrm{d}\mu = \int_{I_n} \gamma \,\mathrm{d}\mu.}$$

maintenant ,passons directement à montrer que σ est discriminatoire,soit $\mu \in M(I_n)$ on a :

$$\begin{aligned} (\forall \omega \in \mathbb{R}^n)\,(\forall \theta \in \mathbb{R}) : \int_{I_n} \sigma\left(\omega^\top x + \theta\right) \mathrm{d}\mu = 0 &\Rightarrow (\forall \omega \in \mathbb{R}^n)\left(\forall (\varphi,\lambda,\theta) \in \mathbb{R}^3\right) : \int_{I_n} \sigma\left(\lambda(\omega^\top x + \theta) + \varphi\right) \mathrm{d}\mu = 0 \\ &\Rightarrow (\forall \omega \in \mathbb{R}^n)\left(\forall (\varphi,\lambda,\theta) \in \mathbb{R}^3\right) : \int_{I_n} \sigma_\lambda \,\mathrm{d}\mu = 0 \\ &\Rightarrow (\forall \omega \in \mathbb{R}^n)\left(\forall (\varphi,\theta) \in \mathbb{R}^2\right) \lim_{\lambda\to+\infty} \int_{I_n} \sigma_\lambda \,\mathrm{d}\mu = \int_{I_n} \lim_{\lambda\to+\infty} \sigma_\lambda \,\mathrm{d}\mu \\ &= \int_{I_n} \gamma \,\mathrm{d}\mu = \sigma(\varphi)\cdot\mu(\Pi_{\omega,\theta}) + \mu(H_{\omega,\theta}) = 0\,. \end{aligned}$$

donc nous obtenons : $\boxed{(\forall \omega \in \mathbb{R}^n)\left(\forall (\varphi,\theta) \in \mathbb{R}^2\right) : \sigma(\varphi)\cdot\mu(\Pi_{\omega,\theta}) + \mu(H_{\omega,\theta}) = 0\,.}$ $\star$

si μ est positive, alors on pourra dire que la mesure de tout les demi plans $\left(H_{\omega,\theta}\right)_{(\omega,\theta)\in\mathbb{R}^n\times\mathbb{R}}$ de $\mathbb{R}^n$ est nulle, ce qui va nous permettre par la suite de déduire que[c] :

$$(\forall (x_1,\cdots,x_n) \in \mathbb{R}^n) : \mu\left(\prod_{i=1}^{n}]-\infty, x_i]\right) = 0.$$

et selon le lemme des classes monotones,ça va nous donner : $\mu = 0$. donc dans $M^+(I_n)$ la fonction σ est aisément discriminatoire.
Mais notre mesure n'est pas supposée positive, et ceci rendra notre tâche un peu plus difficile.
afin de régler ce petit problème, fixons $\omega \in \mathbb{R}^n$, et considérons l'application :

$$\boxed{\begin{array}{ccccc} F & : & L^\infty(\mathbb{R}) & \longrightarrow & \mathbb{R} \\ & & h & \longmapsto & \displaystyle\int_{I_n} h(\omega^\top x)\,\mathrm{d}\mu \end{array}}$$

il est facile de voir que F est bien définie et bornée car :

$$(\forall h \in L^\infty(I_n)) : \left|F(h)\right| = \left|\int_{I_n} h(\omega^\top x)\,\mathrm{d}\mu\right| \leqslant \int_{I_n} M \,\mathrm{d}\mu \leqslant M\mu(I_n) < +\infty\,.$$

et si on prend :$h = \mathbb{1}_{[\theta,+\infty[}$,alors :

$$F(h) = \int_{I_n} \mathbb{1}_{[\theta,+\infty[}(\omega^\top x)\,\mathrm{d}\mu = \mu\left(\{\omega^\top x \geqslant \theta\}\right) = \mu\left(\{\omega^\top x > \theta\}\right) + \mu\left(\{\omega^\top x = \theta\}\right) = \underbrace{\mu(H_{\omega,-\theta}) + \mu(\Pi_{\omega,-\theta}) = 0}_{\text{selon } \star}\,.$$

de même, $F(h) = 0$ si h est la fonction indicatrice de l'intervalle ouvert : $]-\infty,\theta[$,donc et par linéarité, $F(h) = 0$ pour n'importe quelle fonction indicatrice d'un intervalle donné de $\mathbb{R}$,ce qui fait que

a. il est ouvert car c'est l'image réciproque de $\mathbb{R}^{*-}$par une fonction continue$(x \mapsto \langle \omega, x\rangle + \theta)$.

b. ce théorème est valable pour les mesure signées ***finies*** et vous pouvez le démontrer en imitant la démonstration du troisième point dans la preuve de proposition III.1.1.1.

c. car on peut toujours, plonger les parties de la forme :$\prod_{i=1}^{n}]-\infty, x_i]$ dans un demi plan de $\mathbb{R}^n$,pour ceux qui ne sont pas convaincus par l'argument géométrique vous n'avez qu'a utiliser le théorème de séparation de Hahn-Banach.

pour toute fonction simple $\varphi = \sum_{i=1}^{n} \mathbb{1}_{\mathbb{I}_i}$ qui est sous forme d'une somme finie des fonctions indicatrices des intervalles $\mathbb{I}_i \subset \mathbb{R}$, on a : $F(\varphi) = 0$, et comme les fonctions simples sont denses dans $L^\infty(\mathbb{R})$, on déduit que $F = 0$.
en particulier,les fonctions mesurables bornées $c : x \mapsto \cos\big(\langle y, x\rangle\big)$ et $s : x \mapsto \sin\big(\langle y, x\rangle\big)$ nous donnent :$F(c + is) = \int_{I_n} \cos(y^\intercal x) + i \sin(y^\intercal x)\, d\mu(x) = \int_{I_n} \exp(i\, y^\intercal x)\, d\mu(x) = 0 \Rightarrow \underbrace{\widehat{\mu} = 0 \Rightarrow \mu = 0}_{\text{voir :III.1.1.1}}$.
donc,et en guise de conclusion,la fonction σ est discriminatoire, d'où ce qu'il faut prouver . ■

1.3 Application aux réseaux de neurones artificiels :

Dans cette section, nous allons appliquer les résultats précédents dans le cadre le plus intéressant, et qui est celui de la théorie des réseau neuronaux.
Une combinaison directe du théorème III.1.2.1 et du lemme III.1.2.1 montre que les réseaux avec une seule couche interne, dont ***les nœuds disposent d'une fonction d'activation sigmoïdale continue quelconque, peuvent approcher avec une précision arbitraire n'importe qu'elle fonction continue***, à condition qu'***aucune contrainte n'est imposée sur le nombre de nœuds ou la taille des poids***.

Les conséquences de ce résultat pour le rapprochement des fonctions de décisions seront bien traitées par la suite.

Théorème 1.3.1 :

Soit σ ***une fonction sigmoïdale continue***,donc les sommes ***finies*** de la forme :

$$G(x) = \sum_{j=1}^{N} \alpha_j\, \sigma(\omega_j^\intercal x + \theta_j)\,.$$

sont denses dans $C(I_n)$,et comme ça était formulé dans [13] : pour toute $f \in C(I_n)$ et $\varepsilon > 0$,il existe une somme,$G(x)$,de la forme précédente,pour laquelle :

$$(\forall x \in I_n) : \ \big|G(x) - f(x)\big| < \varepsilon.$$

Preuve :

Soit σ une fonction sigmoïdale continue,alors et sans problème on peut dire qu'elle est bornée,***or s'il n'était pas le cas*** l'ensemble : $\mathcal{I}m(\sigma) := \{\sigma(x)/\, x \in \mathbb{R}\}$ sera non borné, ce qui fait qu'il va être non majoré ou non minoré, ou les deux à la fois.
sans perte de généralité,on va traiter le cas où cet ensemble est non majorée,et pour les autres ,il suffit de procéder d'une façon analogue.
si $\mathcal{I}m(\sigma)$ est non majorée, alors il existe une suite : $(y_n)_{n\in\mathbb{N}} \in \mathcal{I}m(\sigma)^{\mathbb{N}}$ tel que : $y_n \xrightarrow[n\to+\infty]{} +\infty$
ce qui fait qu'il existe une suite :$(x_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ tel que : $\sigma(x_n) = y_n \xrightarrow[n\to+\infty]{} +\infty$.
si : $(\mathbf{x_n})_{\mathbf{n}\in\mathbb{N}}$ ***est non bornée***,alors il existe une sous suite $(x_{\phi(n)})_{n\in\mathbb{N}}$ tel que : $x_{\phi(n)} \xrightarrow[n\to+\infty]{} +\infty$ ou $x_{\phi(n)} \xrightarrow[n\to+\infty]{} -\infty$, et ceci nous donne que : $y_{\phi(n)} = \sigma(x_{\phi(n)}) \xrightarrow[n\to+\infty]{} \pm 1$,***absurde!*** , car :

$$\lim_{n\to+\infty} y_n = +\infty \Longrightarrow \lim_{n\to+\infty} y_{\phi(n)} = +\infty.$$

d'où :$(x_n)_{n\in\mathbb{N}}$ est bornée,de ce fait, et en vertu du fameux théorème de ***Bolzano-Weierstrass*** on peut extraire de $(x_n)_{n\in\mathbb{N}}$ une sous suite $(x_{\phi(n)})_{n\in\mathbb{N}}$ qui converge vers un élément $x^* \in \mathbb{R}$.
$\lim_{n\to+\infty} x_{\phi(n)} = x^* \Rightarrow \sigma\left(\lim_{n\to+\infty} x_{\phi(n)}\right) = \sigma(x^*) \Rightarrow$ [a] $\lim_{n\to+\infty} \sigma\left(x_{\phi(n)}\right) = \sigma(x^*) \Rightarrow \lim_{n\to+\infty} y_{\phi(n)} = +\infty = \sigma(x^*)$.
absurde!, donc la fonction σ est bornée.

a. grâce à la continuité de σ.

maintenant on obtient que σ est continue,sigmoïdale et bornée ce qui donne qu'elle est en particulier mesurable (car la continuité implique la mesurabilité),sigmoïdale et bornée,donc, et selon le lemme III.1.2.1,σ sera bien discriminatoire,ce qui donne enfin de compte et à travers le théorème III.1.2.1 la véracité de l'énoncé suggéré. ■

Nous allons présenter et démontrer maintenant le rôle et les implications de ces résultats dans le contexte des ***régions de décision***,et qui est vraiment un contexte très important,surtout avec son apparence dans plusieurs domaines d'applications .
soit $\mu := \lambda_{/\mathfrak{B}_{I_n}}$ la mesure de Lebesgue dans I_n ,et Soit :$P_1, P_2, \cdots, P_k$ une partition de I_n en k sous-ensembles mesurables disjoints. on définit la fonction de décision, f par :$f := \sum_{i=1}^{n} i.\mathbb{1}_{P_i}$,ce qui veut dire que : $\left(\forall i \in [\![1,k]\!]\right)\left(\forall x \in I_n\right) : f(x) = j \iff x \in P_j$.
La question qui se pose maintenant est de savoir si une telle fonction de décision peut être mise en œuvre par un réseau neuronal d'une seule couche cachée.
la réponse à cette question a été fournie dans l'article [13], et elle se présente à travers le résultat suivant :

Théorème 1.3.2 :

Soit σ ***une fonction sigmoïdale continue***, et f une fonction de décision associée à une certaine partition mesurable finie,de I_n.pour tout $\varepsilon > 0$,il existe une somme finie de la forme :

$$G(x) = \sum_{j=1}^{N} \alpha_j \, \sigma(\omega_j^{\mathsf{T}} x + \theta_j) \, .$$

et un ***compact*** $D_\varepsilon \subset I_n$ tel que :$\lambda(D_\varepsilon) \geqslant 1-\varepsilon$ et : $\left|G(x) - f(x)\right| < \varepsilon$ pour tout $x \in D_\varepsilon$,c'est à dire :

$$(\forall \varepsilon > 0)\left(\exists N_\varepsilon \in \mathbb{N}\right)\left(\exists (\alpha^\varepsilon, \theta^\varepsilon) \in \left(\mathbb{R}^{N_\varepsilon}\right)^2\right)\left(\exists (\omega_1^\varepsilon, \cdots, \omega_{N_\varepsilon}^\varepsilon) \in (\mathbb{R}^n)^{N_\varepsilon}\right)\left(\exists D_\varepsilon \in \mathfrak{B}_{I_n} \textbf{ compact tq} : \mu(D_\varepsilon^c) < \varepsilon\right)$$
$$\text{tel que :} \qquad x \in D_\varepsilon \implies \left| \sum_{j=1}^{N} \alpha_j^\varepsilon \, \sigma\left(\langle \omega_j^\varepsilon, x\rangle + \theta_j\right) - f \right| < \varepsilon \, .$$

Preuve :

On a : $f := \sum_{i=1}^{k} i.\mathbb{1}_{P_i}$ avec : $(P_i)_{1 \leqslant i \leqslant k}$ est une partition de I_n en k sous-ensembles mesurables disjoints,donc : $\int_{I_n} f \, \mathrm{d}\mu = \sum_{i=1}^{k} i.\mu(P_i) \leq \sum_{i=1}^{k} i.\mu(I_n) \leq \frac{k \times (k+1)}{2} < +\infty$,ce qui fait que : $f \in L^1\left(I_n, \mathfrak{B}_{I_n}, \mu\right)$.
donc,on peut bien appliquer le théorème de Lusin dans sa version faible I.1.3.1, pour déduire que pour tout $\varepsilon > 0$ il existe un sous ensemble ***compact*** $D \subset I_n$ avec $\mu(D^c) \leqslant \varepsilon$,et une fonction h continue sur cet ensemble tel que : $f_{|D} = h$,c'est à dire : $\left(\forall x \in D\right) : f(x) = h(x)$.
maintenant h est continue, donc,et selon le théorème III.1.3.1,on peut trouver une somme de la forme G précédente et qui satisfait : $\left(\forall x \in D\right) : \left|G(x) - f(x)\right| = \left|G(x) - h(x)\right| < \varepsilon$.

Remarque : à cause de la continuité, nous sommes toujours exposés au risque de faire quelques décisions incorrectes pour certains points , mais ce résultat nous indique que la mesure totale de ces points mal classés peut être si petite que voulu. à la lumière de cela, le théorème III.1.3.1 ***semble être le résultat le plus puissant possible de ce type***.
Nous pouvons et sans perte de généralité,restreindre un peu notre étude en considérant le problème de décision pour un seul ensemble non vide fermé $D \subset I_n$,alors dans ce cas : $f = \mathbb{1}_D$,notre but pour la suite est de trouver une fonction sous forme de la sommation (III.1) et qui approche la fonction de décision f.

soit :$\Delta(x,D) = \min\left\{|x-y|, y \in D\right\}$,selon le lemme I.1.2.1, $x \mapsto \Delta(x,D)$ est continue,ce qui fait que :

$$(\forall \varepsilon > 0): \quad f_\varepsilon(x) = \max\left\{0, \frac{\varepsilon - \Delta(x,D)}{\varepsilon}\right\}$$

est une fonction ***continue*** [a],de sorte que $f_\varepsilon(x) = 0$ pour les points x plus éloignés que ε de D alors que $f_\varepsilon(x) = 1$ pour les points $x \in D$.
par théorème III.1.3.1,nous pouvons trouver et pour n'importe quel $\eta > 0$
une certaine fonction : $G(x) = \sum_{j=1}^{N} \alpha_j \, \sigma(\omega_j^\top x + \theta_j)$,tel que : $|G(x) - f_\varepsilon(x)| < \eta$.
alors et en choisissant ce η si petit ,de préférence qu'il soit plus petit que ε, on pourra bien ajuster la fonction G à f_ε et par la suite utiliser ce G comme ***une fonction de décision approximative :*** $G(x) \leq \frac{1}{2}$ permet de deviner que $x \in D^c$ tandis que : $G(x) > \frac{1}{2}$ devine que $x \in D$.
cette procédure de décision est correcte pour tout $x \in D$ et pour tout x à distance au moins ε loin de D, mais Si x est à distance ε de D, sa classification dépend du choix particulier de G. ■

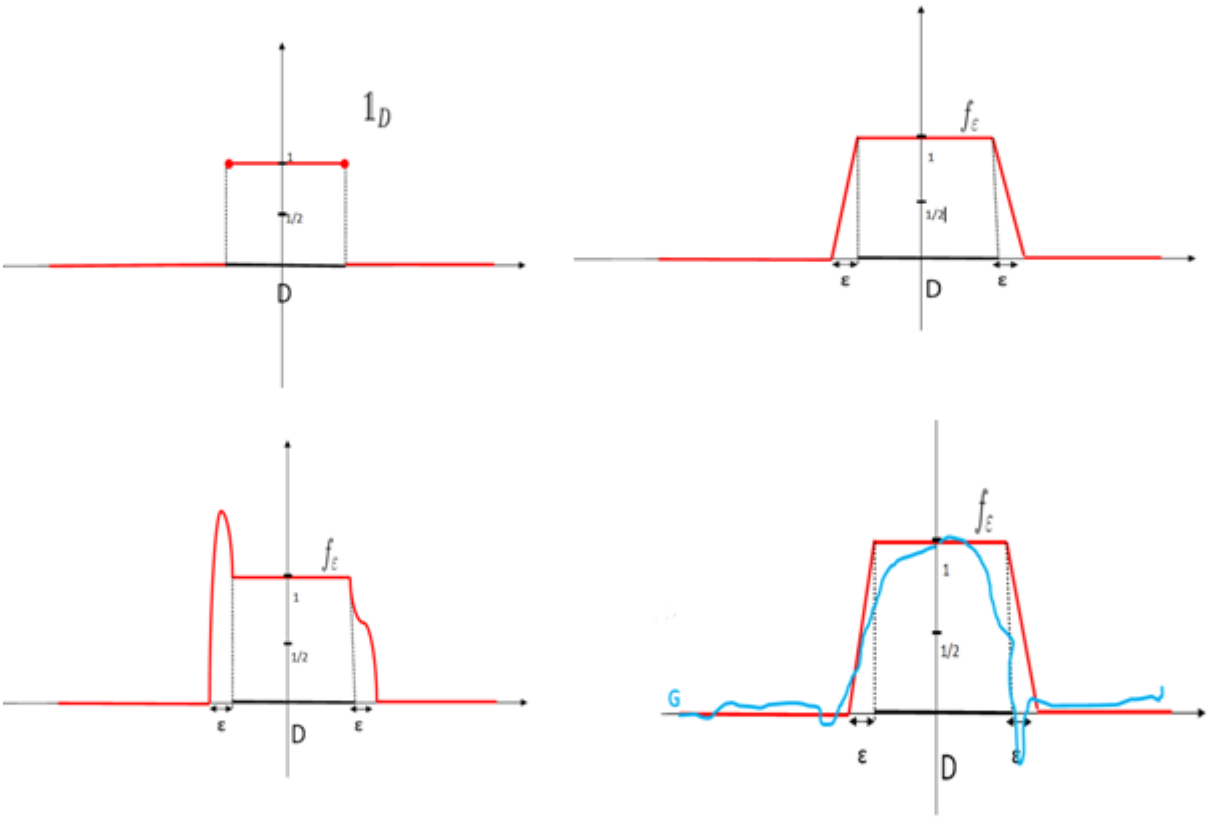


FIGURE III.2: L'idée générale de la preuve.

Remarque 1.3.1 :

En somme ce dernier théorème,montre qu'il existe un réseau qui classe correctement les points suffisamment éloignés de la région de décision, ainsi ceux qui sont à l'intérieur de cette région ,en revanche il rend la mesure des points incorrectement classés si petite que voulu ***mais ne garantit pas leur emplacement***.

a. le max entre zéro et une certaine fonction continue g,reste toujours une fonction continue,pour justifier ça il suffit de remarquer que : $(\forall x \in \mathbb{R}) : \max\big(0, g(x)\big) = \frac{|g(x)|+g(x)}{2}$.

1.4 Résultats pour d'autres fonctions d'activation :

Dans cette section, nous discutons ***d'autres classes de fonctions d'activation*** et qui ont des propriétés similaires à celles des sigmoïdales continues,mais ces exemples ***ont un intérêt un peu moins pratique***,ce qui va justifier le fait de se contenter par des preuves plus résumées,et dont la plupart vont se faire par une simple analogie avec les cas traités précédemment.

On va commencer par l'étude des ***fonctions sigmoïdales discontinues*** et qui ont vraiment une importance considérable, malgré qu'elles sont moins utilisées que les continues, ***en raison du manque d'algorithmes adéquats***.
ces fonctions présentent un intérêt théorique appréciable surtout à travers leur relation avec les perceptrons classiques.

Supposons que σ est une fonction sigmoïdale ***bornée et mesurable*** (la continuité n'est pas requise dans ce cas),nous avons un analogue du théorème III.1.3.1 et qui était évoqué pour la prmière fois en 1989 par George Cybenko dans son article [13], et il s'énonce ainsi :

Théorème 1.4.1 :(approximation des fonctions intégrables :)

Soit σ ***une fonction sigmoïdale mesurable et bornée***,donc les sommes ***finies*** de la forme :

$$G(x) = \sum_{j=1}^{N} \alpha_j \, \sigma(\omega_j^{\intercal} x + \theta_j) \, .$$

sont denses dans $L^1(I_n)$ [a],En d'autres termes :pour tout $f \in L^1(I_n)$ et $\varepsilon > 0$,il existe une somme,$G(x)$ de la forme précédente,pour laquelle :

$$\|G - f\|_1 = \int_{I_n} |G(x) - f(x)| \, \mathrm{d}x < \varepsilon.$$

Preuve

La preuve va suivre exactement les mêmes étapes que celles des théorèmes III.1.2.1 et III.1.3.1 mais avec des modifications évidentes, telles que remplacer les fonctions continues par des fonctions intégrables, et utiliser le fait que $L^{\infty}(I_n)$ est le dual de $L^1(I_n)$,la notion de discrimination d'une certaine fonction σ sera donc modifiée en conséquence de la façon suivante :

$$\left(\forall h \in L^{\infty}(I_n)\right) : \quad \left[\left(\forall \omega \in \mathbb{R}^n\right)\left(\forall \theta \in \mathbb{R}\right) : \int_{I_n} \sigma\left(\omega^{\intercal} x + \theta\right) h(x) \, \mathrm{d}x = 0\right] \Longrightarrow h(x) = 0 \text{ presque partout.}$$

sans surprise et exactement comme on l'a déjà vu au lemme III.1.2.1, les fonctions sigmoïdales mesurables bornées sont en particulier discriminatoires dans ce sens, car les mesures de la forme $h(x)\,\mathrm{d}x$ appartiennent en particulier à $M(I_n)$. ■

Lemme 1.4.1 :(l'inégalité de Markov :)

soit $(\Omega, \mathscr{F}, \mu)$ un espace mesuré, et soit $f \in \mathfrak{M}\left((\Omega, \mathscr{F}); (\mathbb{R}, \mathfrak{B}_{\mathbb{R}})\right)$ une fonction $\mathscr{F} - \mathfrak{B}_{\mathbb{R}}$ mesurables alors on a :

$$(\forall \varepsilon > 0) \; : \mu\left(\left\{|f| \geqslant \varepsilon\right\}\right) \leqslant \frac{1}{\varepsilon} \int_{\Omega} |f| \, \mathrm{d}\mu$$

Preuve :

Soit $\varepsilon > 0$ on prend $A = \left\{|f| \geqslant \varepsilon\right\}$,et $\varphi = \varepsilon \mathbb{1}_A$,il est clair que :$\varphi \leqslant |f|$ alors : $\int_{\Omega} \varphi \, \mathrm{d}\mu \leqslant \int_{\Omega} |f| \, \mathrm{d}\mu$
ce qui fait que :$\varepsilon \mu\left(\left\{|f| \geqslant \varepsilon\right\}\right) \leqslant \int_{\Omega} |f| \, \mathrm{d}\mu$ donc : $(\forall \varepsilon > 0) \; : \mu\left(\left\{|f| \geqslant \varepsilon\right\}\right) \leqslant \frac{1}{\varepsilon} \int_{\Omega} |f| \, \mathrm{d}\mu.$

a. ***Attention ! :*** cette fois ci on parle de la densité au sens de la norme L^1 ***et non plus de la norme de convergence uniforme***.

d'où ce qu'il faut prouver. ■

Proposition 1.4.1 :

Soit $(\Omega,\mathscr{F},\mu)$ un espace mesuré, et soit f et $(f_n)_{n\in\mathbb{N}^*}\in\mathfrak{M}\left((\Omega,\mathscr{F});(\mathbb{R},\mathfrak{B}_{\mathbb{R}})\right)$ des fonctions $\mathscr{F}-\mathfrak{B}_{\mathbb{R}}$ mesurables,alors on a :

$$\left((\exists r\geqslant 1)\ :\ f_n\xrightarrow[n\to+\infty]{L^r}f\right)\Longrightarrow f_n\xrightarrow[n\to+\infty]{\mu}f.$$

c'est à dire que : la convergence dans L^r,implique la convergence en mesure [a] (voir :[2]).

Preuve :

On a : $(\forall\varepsilon>0)\ :\mu\left(\left|f_n-f\right|>\varepsilon\right)=\overbrace{\mu\left(\left|f_n-f\right|^r>\varepsilon^r\right)\leqslant\frac{1}{\varepsilon^r}\int_\Omega\left|f_n-f\right|^r\,\mathrm{d}\mu}^{\text{Selon l'inégalité de Markov}}$

donc si : $\int_\Omega\left|f_n-f\right|^r\,\mathrm{d}\mu\underset{n\to\infty}{\longrightarrow}0$ on aura :$(\forall\varepsilon>0)\ :\mu\left(\left|f_n-f\right|>\varepsilon\right)\underset{n\to\infty}{\longrightarrow}0$. ce qui achève la preuve. ■

en vertu de cette proposition,la convergence dans $\mathbf{L^1}$ ***implique la convergence en mesure*** ,et de ce fait ,nous avons un analogue du théorème III.1.3.2 et qui était énoncé dans l'article [13] comme suit :

Théorème 1.4.2 :

Soit σ ***une fonction sigmoïdale mesurable et bornée***,et f une fonction intégrable sur I_n ,alors, pour tout $\varepsilon>0$,il existe une somme finie de la forme :

$$G(x)=\sum_{j=1}^{N}\alpha_j\,\sigma(\omega_j^{\intercal}x+\theta_j)\,.$$

et un ensemble ***(pas forcément compact)*** $D\subset I_n$ tel que :$\mu(D)\geqslant 1-\varepsilon$ et : $\left|G(x)-f(x)\right|<\varepsilon\ \ (\forall x\in D)$.

Remarque 1.4.1 :

Si σ est une fonction sigmoïdale ***mesurable*** [b] et bornée, alors elle est en particulier dans $L^1(I_n)$ car :

$$(\forall x\in I_n)\ :\ \left|G(x)\right|=\left|\sum_{j=1}^{N}\alpha_j\,\sigma(\omega_j^{\intercal}x+\theta_j)\right|\leqslant\sum_{j=1}^{N}\left|\alpha_j\right|\left|\sigma(\omega_j^{\intercal}x+\theta_j)\right|\leqslant\sum_{j=1}^{N}\left|\alpha_j\right|M$$

$$\Rightarrow\int_{I_n}\left|G(x)\right|\,\mathrm{d}\lambda<\int_{I_n}\sum_{j=1}^{N}\left|\alpha_j\right|M\,\mathrm{d}\lambda=\sum_{j=1}^{N}\left|\alpha_j\right|M.\underbrace{\lambda(I_n)}_{=1}<+\infty\,.$$

Preuve :(de théorème III.1.4.2)

Selon le théorème III.1.4.1, pour toute fonction ***sigmoïdale mesurable et bornée*** σ, l'ensemble :

$$\mathcal{N}_\sigma=\left\{x\mapsto G(x)=\sum_{j=1}^{N}\alpha_j\,\sigma(\omega_j^{\intercal}x+\theta_j)\ \Big/\ N\in\mathbb{N}\,,(\alpha_1,\cdots,\alpha_N)\in\mathbb{R}^N\,,(\theta_1,\cdots,\theta_N)\in\mathbb{R}^N\,,(\omega_1,\cdots,\omega_N)\in(\mathbb{R}^n)^N\right\}$$

a. Attention!,***la réciproque est fausse*** en général.

b. le fait que σ soit mesurable est indispensable pour déduire la mesurabilité sur I_n de la fonction : $x\mapsto\sigma(\omega_j^{\intercal}x+\theta_j)$,ce qui va nous permettre par la suite de parler de l'intégrabilité de cette fonction sur I_n.

est dense dans $L^1(I_n)$ ***pour la norme*** : $\|\cdot\|_1 : f \longmapsto \int_{I_n} f(t)\,\mathrm{d}t$; donc et par la caractérisation de densité dans les espaces métriques [a] ,on a :

$$\left(\forall f \in L^1(I_n)\right)\left(\exists \left(G_n\right)_{n\in\mathbb{N}} \in \mathcal{M}^{\mathbb{N}}\right) \text{ tq : } \quad \left\|G_n - f\right\|_1 \xrightarrow[n\to+\infty]{} 0 \left(\Longleftrightarrow G_n \xrightarrow[n\to+\infty]{L^1} f \right).$$

en se basant sur cette équivalence et sur la proposition III.1.4.1 ,on pourra bien mettre en évidence que : $G_n \xrightarrow[n\to+\infty]{\lambda} f$,ce qui veut dire, et par définition, que : $(\forall \varepsilon > 0) : \lim\limits_{n\to+\infty} \lambda\left(\left|G_n - f\right| > \varepsilon\right) = 0$.
donc : $(\forall \varepsilon > 0)\left(\forall \varepsilon' > 0\right)\left(\exists \eta_{\varepsilon,\varepsilon'} > 0\right)$ tq : $(\forall n \in \mathbb{N}) : n \geqslant \eta_{\varepsilon,\varepsilon'} \Rightarrow \lambda\left(\left|G_n - f\right| > \varepsilon\right) \leqslant \varepsilon'$.
on prend en particulier : $\varepsilon = \varepsilon'$,alors : $(\forall \varepsilon > 0)\left(\exists \eta_\varepsilon > 0\right)$ tq : $(\forall n \in \mathbb{N}) : n \geqslant \eta_\varepsilon \Rightarrow \lambda\left(\left|G_n - f\right| > \varepsilon\right) \leqslant \varepsilon$.
et ceci implique que : $\boxed{(\forall \varepsilon > 0)\left(\exists N_\varepsilon = \underbrace{[\eta_\varepsilon]}_{\text{partie entière.}} + 1 \in \mathbb{N}^*\right) \text{ tq : } \lambda\left(\left|G_{N_\varepsilon} - f\right| > \varepsilon\right) \leqslant \varepsilon .}$ ⋆
on pose : $D = \left\{x \in I_n / \left|G_{N_\varepsilon}(x) - f(x)\right| \leqslant \varepsilon\right\}$, selon ⋆, $\lambda\left(\left|G_{N_\varepsilon} - f\right| > \varepsilon\right) \leqslant \varepsilon \Rightarrow \lambda(D^c) = \lambda(I_n \setminus D) \leqslant \varepsilon$.
donc : $\lambda(D) \geqslant 1 - \varepsilon$,et : $\left(\forall x \in D\right) : \left|G_{N_\varepsilon} - f\right| \leqslant \varepsilon$ d'où ce qu'il faut démontrer. ■

Remarques 1.4.2 :

1. Pour ce dernier théorème,on peut choisir en particulier la fonction f comme une fonction de décision ,et de ce fait on obtiendra un résultat qui ressemble un peu au théorème :III.1.3.2 ***mais faites attention,*** la compacité de l'ensemble D ***n'est plus garanti*** pour ce dernier cas.
2. si on ne met pas la compacité de D comme une conséquence dans le théorème III.1.3.2, alors il va être un résultat trivial de III.1.4.2, et l'appel du théorème de Lusin ne sera plus nécessaire.

2 Théorèmes d'approximation au cas d'une fonction d'activation écrasante :

2.1 Définitions et notations :

Notations :

Pour des raisons de cohérence et de simplification, on adoptera dans la suite de ce chapitre,les notations suivantes :

— pour tout $r \in \mathbb{N}^*$, on note :$\mathfrak{A}^r = \left\{A \in \left(\mathbb{R}^r\right)^{\mathbb{R}} / \left(\exists (\omega,\theta) \in \mathbb{R}^r \times \mathbb{R}\right) : A : x \mapsto \langle \omega, x\rangle + \theta\right\}$.

— pour toute fonction borélienne [b], $\sigma : \mathbb{R} \to \mathbb{R}$, on note :$\Sigma^r(\sigma) = \mathcal{N}_\sigma$.

Remarque 2.1.1 :

1. Dans ce contexte,**x** ***correspond à l'entrée d'un réseau de neurones*** ,ω ***correspond aux poids de ce réseau*** reliant la couche d'entrée à un nœud appartenant à celle intermédiaire, et **b** ***représente tout simplement le biais*** de ce nœud.

a. tenez bien compte que cette caractérisation n'est pas valable en général pour les espaces topologiques non métrisables.

b. Une fonction borélienne n'est autre qu'une fonction mesurable dont l'espace de départ est un espace topologique muni de la tribu borélienne.

2. il est facile de voir que : $\Sigma^r(\sigma)$ n'est autre que l'ensemble des fonctions qui peuvent être réalisées , par un réseau de neurones ***ordinaire***[a],dont les nœuds possèdent tous la même fonction d'activation σ.

Définition 2.1.1 :(fonction d'écrasement :)

Une fonction $\Psi : \mathbb{R} \to [0,1]$ est dite ***écrasante*** ou ***d'écrasement*** lorsqu'elle est **sigmoïdale** et ***croissante*** (voir :[34]).

Exemples 2.1.1 :

1. la fonction de Heaveside définie par : $\Psi(x) = 1_{\mathbb{R}^+}(x)$ est clairement une fonction d'écrasement.
2. la fonction "rampe" définie par : $\Psi(x) = \mathbb{1}_{[1,+\infty[}(x) + x \cdot \mathbb{1}_{[0,1]}(x)$. est encore une fonction d'écrasement.
3. le cosinus écraseur de Gallant-White défini par : $\Psi(x) = \frac{1}{2}\left(1+\cos\left(x+\frac{3\pi}{2}\right)\right)\mathbb{1}_{\left[\frac{-\pi}{2},\frac{\pi}{2}\right]} + \mathbb{1}_{\left]\frac{\pi}{2},+\infty\right[}$.

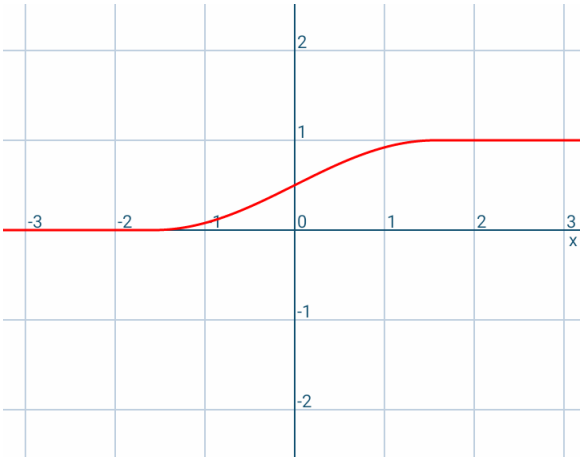


FIGURE III.3: La courbe du cosinus écraseur.

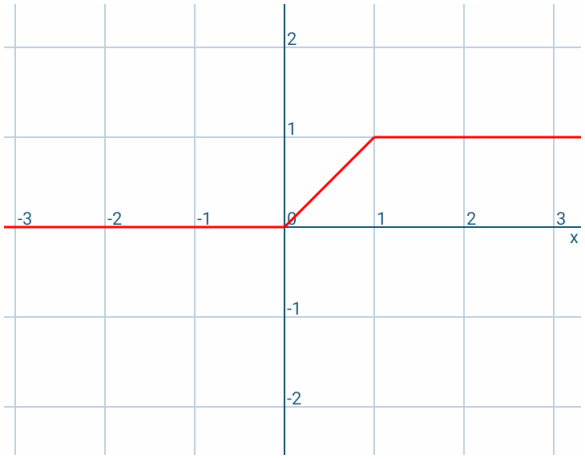


FIGURE III.4: La courbe de la fonction Rampe.

Lemme 2.1.1 :

Soit $f : I \to \mathbb{R}$ une fonction ***croissante*** sur un intervalle : $I \subset \mathbb{R}$,alors ,elle admet une limite à gauche et à droite en tout point $a \in I$,et de plus on a :

$$\boxed{\lim_{x \to a^-} f(x) \leqslant f(a) \leqslant \lim_{x \to a^+} f(x).}$$

Preuve :

La preuve est vraiment très classique,or,et selon [7] il suffit de considérer pour tout $a \in I$ les ensembles :

$$\mathbb{F}_a^+ = \left\{x \in I / f(x) \geqslant f(a)\right\} \text{ et : } \mathbb{F}_a^- = \left\{x \in I / f(x) \leqslant f(a)\right\}.$$

sans problème,ces deux ensembles sont ***non vides*** (car I est ouvert) et par la croissance de notre fonction,on obtient respectivement que le premier est minoré tandis que le deuxième est majoré

a. on désigne par le terme ordinaire ici que tous les nœuds constituant notre réseau sont doté d'une fonction d'entrée totale qui est la somme pondérée,et d'une fonction de sortie confondue avec l'identité.

donc les quantités : $\inf \mathbb{F}_a^+$,$\sup \mathbb{F}_a^-$ auront bien un sens, et en plus :

$$\left(\forall x \in \mathbb{F}_a^-\right) : f(x) \leqslant f(a) \Rightarrow \sup \mathbb{F}_a^- \leqslant f(a) \text{ et } \left(\forall x \in \mathbb{F}_a^+\right) : f(x) \geqslant f(a) \Rightarrow \inf \mathbb{F}_a^+ \geqslant f(a).$$

donc et à travers la monotonie de la fonction f et la définition de la borne supérieure ainsi que l'inférieure :on peut déduire que : $\lim\limits_{x \to a^-} f(x) = \sup \mathbb{F}_a^- \leqslant f(a)$ et $\lim\limits_{x \to a^-} f(x) = \inf \mathbb{F}_a^+ \geqslant f(a)$
d'où : $\lim\limits_{x \to a^-} f(x) \leqslant f(a) \leqslant \lim\limits_{x \to a^+} f(x)$. ■

Théorème 2.1.1 :(voir : [7])

Si une fonction f est croissante sur un intervalle ouvert I,alors l'ensemble des points où elle n'est pas continue est au plus dénombrable [a].

Preuve :

Soient a et b deux réels de I tels que $a < b$, supposons que f est non continue en ces deux points. alors : $\lim\limits_{x \to a^-} f(x) < \lim\limits_{x \to a^+} f(x) \leqslant \lim\limits_{x \to b^-} f(x) < \lim\limits_{x \to b^+} f(x)$ ce qui fait que les intervalles ouverts :

$$\left]\lim_{x \to a^-} f(x), \lim_{x \to a^+} f(x)\right[\text{ et } \left]\lim_{x \to b^-} f(x), \lim_{x \to b^+} f(x)\right[$$

sont disjoints et non vides,par la densité de $\mathbb{Q}$ dans $\mathbb{R}$,chacun d'eux contient au moins un rationnel. on peut donc et ***par l'axiome de choix*** construire une application injective, associant à tout point de discontinuité de f, un rationnel. comme l'ensemble des rationnels est dénombrable, le résultat s'ensuit. ■

Remarque 2.1.2 :

On peut étendre les les deux résultats précédents pour les fonctions décroissantes à travers un simple remplacement de : f par $-f$.

Corollaire 2.1.1 :

1. Toute fonction $f : I \to \mathbb{R}$ croissante,sur un intervalle ouvert I de $\mathbb{R}$ est $\left(\mathfrak{B}_{\mathbb{R}} \sqcap I, \mathfrak{B}_{\mathbb{R}}\right)$ mesurable.
2. en particulier ***les fonctions d'écrasement sont toujours boréliennes***.

Preuve :

1. Afin de démontrer ce corollaire , on note :$\mathfrak{D}$ l'ensemble des points de discontinuité de f et $\mathfrak{C} = \mathfrak{D}^c$ l'ensemble de ses points de continuité,selon le théorème III.2.1.1 :$\mathfrak{D}$ est au plus dénombrable,ce qui fait qu'il est mesurable [b] donc $\mathfrak{C} = \mathfrak{D}^c$ sera à son tour mesurable comme étant le complémentaire d'un ensemble mesurable, et par suite la fonction indicatrice :$\mathbb{1}_{\mathfrak{C}}$ sera immédiatement une fonction mesurable ce qui implique que : $f \cdot \mathbb{1}_{\mathfrak{C}}$ le sera aussi,car c'est le produit de deux fonctions mesurables (la fonction f est continue sur $\mathfrak{C}$).
   et d'une autre part :l'application : $\varphi : x \mapsto f \cdot \mathbb{1}_{\mathfrak{D}}$ est mesurable car :
   $$\left(\forall A \in \mathfrak{B}_{\mathbb{R}}\right) : \varphi^{-1}(A) \subset \mathfrak{D} \Longrightarrow \operatorname{card}\left(\varphi^{-1}(A)\right) \leqslant \aleph_0 \Longrightarrow \varphi^{-1}(A) \in \mathfrak{B}_{\mathbb{R}} \sqcap I.$$
   donc et sans aucun problème la fonction : $f = f \cdot \mathbb{1}_{\mathfrak{C}} + f \cdot \mathbb{1}_{\mathfrak{D}}$ sera mesurable.
2. par définition les fonctions d'écrasement sont croissantes donc et selon le premier point ça va impliquer qu'elles sont mesurables. ■

a. c'est à dire de cardinalité $\leq \aleph_0$.

b. On laisse au lecteur le soin de vérifier que tout sous ensemble au plus dénombrable de $\mathbb{R}$ est mesurable en particulier .

Définition 2.1.2 :(les réseaux de neurones $\Sigma\Pi$:)

Pour toute fonction $\sigma : \mathbb{R} \to \mathbb{R}$,et $r \in \mathbb{N}^*$ on définit la classe des fonctions $\Sigma\Pi^r(\sigma)$ par :

$$\Sigma\Pi^r(\sigma) = \left\{ x \mapsto G(x) = \sum_{j=1}^{q} \beta_j \cdot \prod_{k=1}^{l_j} \sigma(A_{jk}(x)) \Big/ q \in \mathbb{N}^* \, , l_j \in \mathbb{N} \, , (\beta_1, \cdots, \beta_q) \in \mathbb{R}^q \, , A_{jk}(x) \in \mathfrak{A}^r \right\}.$$

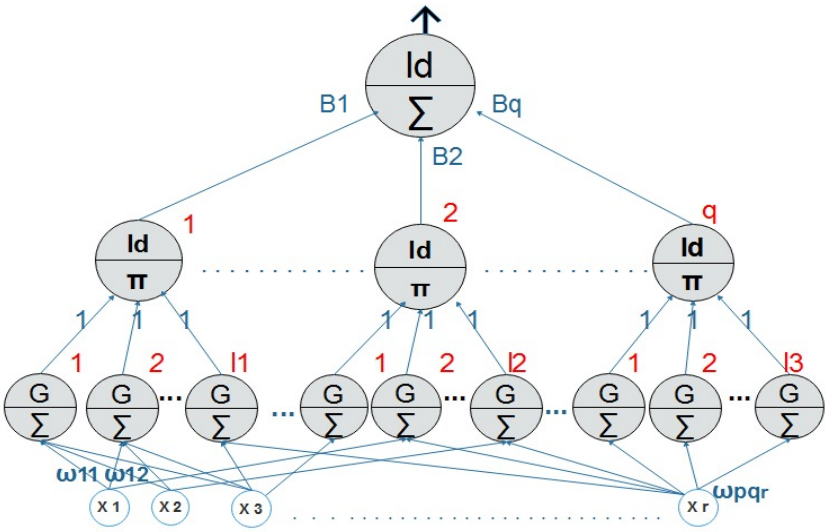


FIGURE III.5: L'architecture d'un réseaux de neurones $\Sigma\Pi^r(G)$.

Remarque 2.1.3 :

Á l'instar de [34],on prouvera dans la suite, nos résultats généraux pour les réseaux $\Sigma\Pi$,et ensuite on va les étendre aux réseaux Σ,qui ne sont qu'un simple cas particulier des $\Sigma\Pi$, pour lesquels : $l_j = 1$ pour tout $j \in \{1, \cdots, q\}$.

Définition 2.1.3 :(la densité uniforme compacte :)

1. Un sous-ensemble S de $C(\mathbb{R}^r)$ est dit ***uniformément compactement dense*** dans $C(\mathbb{R}^r)$, si ***pour tout compact*** $K \subset \mathbb{R}^r$, $S_K = \left\{ f_{|K} / f \in S \right\}$ est ρ_K- dense [a] dans $C(K)$
   avec : $\left(\forall (f,g) \in C(K)\right) : \rho_K(f,g) = \sup_{x \in K} \left| f(x) - g(x) \right|$.
2. Une suite de fonctions $\left(f_n\right)_{n \in \mathbb{N}} \in C(\mathbb{R}^r)^{\mathbb{N}}$ ***converge*** vers une fonction $f \in C(\mathbb{R}^r)$ ***uniformément compactement*** sur $C(\mathbb{R}^r)$ si pour tout compact $K \subset \mathbb{R}^r$: $\rho_K(f_n, f) \xrightarrow[n \to +\infty]{} 0$.

Remarque 2.1.4 :

On peut voir facilement que la densité uniforme d'un ensemble $S \subset C\left(\mathbb{R}^r\right)$,implique directement sa densité uniforme compacte et ceci provient du fait que lorsque :$\sup_{x \in \mathbb{R}^r} \left| f(x) - g(x) \right| < \varepsilon$ est vérifiée alors : $\sup_{x \in K} \left| f(x) - g(x) \right| < \sup_{x \in \mathbb{R}^r} \left| f(x) - g(x) \right| < \varepsilon$ le sera aussi sans doute.
cette petite remarque va faire l'objet d'un petit passage dans la preuve du théorème III.2.2.3.

a. il est très important de spécifier la norme, car dans les espaces de dimension infinie ***l'équivalence de toute les normes n'est pas vérifiée.***

Définition 2.1.4 :(la μ-équivalence :)

Soit μ une mesure ***quelconque*** sur $\left(\mathbb{R}^r, \mathfrak{B}_{\mathbb{R}^r}\right)$,si f et g appartiennent à $\overbrace{\mathfrak{M}\left(\left(\mathbb{R}^r, \mathfrak{B}_{\mathbb{R}^r}\right); \left(\mathbb{R}, \mathfrak{B}_{\mathbb{R}}\right)\right)}^{\text{l'espace de fonctions } \mathfrak{B}_{\mathbb{R}^r}-\mathfrak{B}_{\mathbb{R}} \text{ mesurables}}$,alors on dit qu'elles sont μ-équivalentes ,ssi : $\mu\left\{x \in \mathbb{R}^r / f(x) \neq g(x)\right\} = 0$, et on note dans ce cas : $f \mathcal{R}_\mu g$.

Remarque 2.1.5 :

1. Si μ est finie avec $\mu\left(\mathbb{R}^r\right) = m$, (pensez aux mesures de probabilité par exemple) ,alors la définition précédente peut être exprimée ainsi :
$$f \mathcal{R}_\mu g \iff \mu\left\{x \in \mathbb{R}^r / f(x) = g(x)\right\} = m .$$
2. comme son nom l'indique la relation " $\mathcal{R}_\mu$" ***est bien une relation d'équivalence*** sur : $\mathfrak{M}\left(\left(\mathbb{R}^r, \mathfrak{B}_{\mathbb{R}^r}\right); \left(\mathbb{R}, \mathfrak{B}_{\mathbb{R}}\right)\right)$,ce qui fait qu'on peut bien donner un sens à l'espace quotient :
$$M\left(\left(\mathbb{R}^r, \mathfrak{B}_{\mathbb{R}^r}\right); \left(\mathbb{R}, \mathfrak{B}_{\mathbb{R}}\right)\right) = \mathfrak{M}\left(\left(\mathbb{R}^r, \mathfrak{B}_{\mathbb{R}^r}\right); \left(\mathbb{R}, \mathfrak{B}_{\mathbb{R}}\right)\right) \Big/ \mathcal{R}_\mu$$
et ce dernier fera bien l'objet de plusieurs résultats qui seront présentés par la suite.

Notation :

On opte pour la suite les notations suivantes :
$$M^r = M\left(\left(\mathbb{R}^r, \mathfrak{B}_{\mathbb{R}^r}\right); \left(\mathbb{R}, \mathfrak{B}_{\mathbb{R}}\right)\right) \text{ et : } \mathfrak{M}^r = \mathfrak{M}\left(\left(\mathbb{R}^r, \mathfrak{B}_{\mathbb{R}^r}\right); \left(\mathbb{R}, \mathfrak{B}_{\mathbb{R}}\right)\right) .$$

Définition 2.1.5 :(distance associée à une mesure :)

De toute mesure ***finie*** μ sur $\left(\mathbb{R}^r, \mathfrak{B}_{\mathbb{R}^r}\right)$,provient une métrique ρ_μ sur M^r définie par :

$$\begin{array}{rccl} \rho_\mu & : & M^r \times M^r & \longrightarrow & \mathbb{R}^+ \\ & & (f,g) & \longmapsto & \inf\left\{\varepsilon > 0 \Big/ \mu\left\{x \in \mathbb{R}^r / |g(x) - f(x)| > \varepsilon\right\} < \varepsilon\right\}. \end{array}$$

Remarque 2.1.6 :

1. On peut justifier facilement le fait que ρ_μ est une distance sur M^r, mais faites attention!,elle ne l'est pas du tout sur : $\mathfrak{M}^r$,car la propriété de séparation ne devient plus valable [a].
2. l'hypothèse de la finitude de μ est très importante, car elle nous permet de justifier l'existence de l'image :
or si on note : $\left(\forall (f,g) \in M^r\right) : \mathcal{A}_{f,g} = \left\{\varepsilon > 0 \Big/ \mu\left\{x \in \mathbb{R}^r / |g(x) - f(x)| > \varepsilon\right\} < \varepsilon\right\} \neq \emptyset$
on aura : $\mu\left\{x \in \mathbb{R}^r / |g(x) - f(x)| > \varepsilon\right\} \leqslant \mu\left(\mathbb{R}^r\right) \in \mathbb{R}^+ \Longrightarrow \mu\left(\mathbb{R}^r\right) + 1 \in \mathcal{A}_{f,g} \Longrightarrow \mathcal{A}_{f,g} \neq \emptyset$
$\Longrightarrow \rho_\mu(f,g) = \inf \mathcal{A}_{f,g}$ existe.

2.2 Résultats fondamentaux :

Théorème 2.2.1 :

Soit σ une fonction ***continue non constante quelconque***,alors et comme ça était mentionné dans [34] , l'ensemble $\Sigma\Pi^r(\sigma)$ est uniformément compactement dense dans $C\left(\mathbb{R}^r\right)$.

a. Dans ce cas : $\rho_\mu(f,g) = 0 \iff f = g$ $\mu-$presque partout.

Preuve :

Nous appliquons pour démontrer ceci le théorème de Stone-Weierstrass I.2.2.1.
soit $K \subset \mathbb{R}^r$ un compact[a], pour toute fonction σ,$\Sigma\Pi^r(\sigma)$ est clairement une sous algèbre dans $C^r(K)$, et si on prend $(x,y) \in K$ avec : $x \neq y$, alors on trouvera forcement un $A \in \mathfrak{A}^r$ tel que : $\sigma\big(A(x)\big) \neq \sigma\big(A(y)\big)$, pour voir ceci,choisissez $a,b \in \mathbb{R}$ avec $a \neq b$ et : $\sigma(a) \neq \sigma(b)$. (de tels éléments existent car ***la fonction*** σ ***est supposée non constante***) et prenez un $A(\cdot) \in \mathfrak{A}^r$ satisfaisant : $A(x) = a, A(y) = b$. alors : $\sigma\big(A(x)\big) \neq \sigma\big(A(y)\big)$, ce qui garantit que : $\Sigma\Pi^r(\sigma)$ sépare les points de K.
d'une autre part, il y en a des fonctions $\sigma(A(\cdot))$ qui sont constantes et non égales à zéro, pour voir cela choisissez $b \in \mathbb{R}$ tel que : $\sigma(b) \neq 0$ et définir $\sigma\big(A(x)\big) = \langle 0.x\rangle + b$,donc :
$\big(\forall x \in K\big) : \sigma\big(A(x)\big) = \sigma(b) \neq 0$.
alors toutes les hypothèses du théorème de Stone-Weierstrass I.2.2.1 sont vérifiées,d'où ce qu'il faut prouver. ■

Remarques 2.2.1 :

1. En d'autres termes,ce théorème énonce que les réseaux de neurones $\Sigma\Pi$ sont capables d'approximer avec une précision arbitraire toute fonction continue à valeurs réelles sur un ensemble compact.
2. l'exigence de la compacité apparait chaque fois où les valeurs possibles des entrées x sont bornées .
3. une caractéristique intéressante de ce résultat est que la fonction d'activation σ peut être n'importe qu'elle fonction continue non constante.
4. nos résultats ultérieurs découleront tous de ce théorème.

Lemme 2.2.1 :(caractérisation de la ρ_μ-convergence :)

Soit μ ***une mesure de probabilité*** sur $\big(\mathbb{R}^r, \mathfrak{B}_{\mathbb{R}^r}\big)$,et soit ρ_μ sa distance associée,(voir la définition :III.2.1.5,alors pour toute suite :$(f_n)_{n\in\mathbb{N}\in(M^r)^{\mathbb{N}}}$ et pour toute fonction :$f \in M^r$, les assertions suivantes sont équivalentes (voir : [34]) :

a) $\rho_\mu\big(f_n, f\big) \xrightarrow[n\to+\infty]{} 0$.

b) $(\forall \varepsilon > 0)$: $\mu\Big\{x \in \mathbb{R}^r / \big|f_n(x) - f(x)\big| > \varepsilon\Big\} \xrightarrow[n\to+\infty]{} 0$.

c) $\displaystyle\int_{\mathbb{R}^r} \min\Big\{\big|f_n(x) - f(x)\big|, 1\Big\}\, \mathrm{d}\mu \xrightarrow[n\to+\infty]{} 0$.

Preuve :

a) $\Longleftrightarrow$ **b) :** Évidente,il suffit d'écrire la définition de la limite pour les deux assertions et ça ira de soi.

b) $\Longrightarrow$ **c) :** si :$(\forall \varepsilon > 0)$: $\mu\Big\{x \in \mathbb{R}^r / \big|f_n(x) - f(x)\big| > \varepsilon\Big\} \xrightarrow[n\to+\infty]{} 0$ alors :à partir d'un certain rang on aura :$\mu\Big\{x \in \mathbb{R}^r / \big|f_n(x) - f(x)\big| > \frac{\varepsilon}{2}\Big\} < \frac{\varepsilon}{2}$ donc :

$$\int_{\mathbb{R}^r} \min\Big\{\big|f_n - f\big|, 1\Big\}\, \mathrm{d}\mu < \int_{\mathbb{R}^r} \big|f_n - f\big|\, \mathrm{d}\mu < \overbrace{\int_{\mathbb{R}^r} \big|f_n - f\big| \times \mathbb{1}_{\left\{|f_n - f| \leqslant \frac{\varepsilon}{2}\right\}}\, \mathrm{d}\mu}^{<\frac{\varepsilon}{2}} + \overbrace{\int_{\mathbb{R}^r} \big|f_n - f\big| \times \mathbb{1}_{\left\{|f_n - f| > \frac{\varepsilon}{2}\right\}}\, \mathrm{d}\mu}^{<\frac{\varepsilon}{2}} < \varepsilon\,.$$

c) $\Longrightarrow$ **b) :** on obtient cette implication directement de la proposition : III.1.4.1· ■

a. selon le théorème de Borel-Lebesgue, les compacts dans $\mathbb{R}^r$ sont exactement les fermés bornés.

Remarque 2.2.2 :

On peut remplacer dans le point **"b)"** du lemme précédent la différence en valeur absolue par n'importe quelle distance générant la topologie Euclidienne dans $\mathbb{R}$,ainsi, la quantité intégrée au point **"c)"** peut être substitué par une métrique bornée sur :$\mathbb{R}$.

Lemme 2.2.2 :(La convergence uniforme compacte $\Rightarrow$ la $\rho_\mu-$convergence :)

Soit μ une mesure de probabilité sur $\left(\mathbb{R}^r, \mathfrak{B}_{\mathbb{R}^r}\right)$ et soit $\left(f_n\right)_{n\in\mathbb{N}} \in (M^r)^{\mathbb{N}}$ une suite de fonctions mesurables qui converge uniformément compactement vers une fonction $f \in M^r$,alors : $\rho_\mu\left(f_n, f\right) \xrightarrow[n\to+\infty]{} 0$.

Preuve :

Choisissons un $\varepsilon > 0$, d'après le lemme III.2.2.1,et pour qu'on puisse déduire que : $\rho_\mu\left(f_n, f\right) \underset{n\to\infty}{\longrightarrow} 0$ il suffit de trouver un $N_\varepsilon \in \mathbb{N}$ tel que : pour tout $n \geqslant N_\varepsilon$ on a : $\int_{\mathbb{R}^r} \min\left\{\left|f_n(x) - f(x)\right|, 1\right\} \mathrm{d}\mu < \varepsilon$.

soit $n \in \mathbb{N}$, on note :$B_n = B(0,n] = \left\{ x \in \mathbb{R}^r \Big/ \sqrt{\sum_{i=1}^{r} x_i^2} \leqslant n \right\} \subset \mathbb{R}^r$ la boule ***fermée*** de centre 0 et de rayon n pour la norme Euclidienne dans $\mathbb{R}^r$.

il est facile de voir que : $\mathbb{R}^r = \bigcup_{i\in\mathbb{N}} B_n$, donc :$\mu\left(\mathbb{R}^r\right) = \mu\left(\bigcup_{i\in\mathbb{N}} B_n\right)$,et d'une autre part,la suite $\left(B_n\right)_{n\in\mathbb{N}}$ est clairement croissante car : $(\forall n \in \mathbb{N})$:$B_n \subset B_{n+1}$,aussi, et sans aucune difficulté,les $\left(B_n\right)_{n\in\mathbb{N}}$ sont mesurables car : B_n fermé $\Longrightarrow B_n^c \in \Theta_{\mathbb{R}^r} \Longrightarrow B_n^c \in \tau\left(\Theta_{\mathbb{R}^r}\right) := \mathfrak{B}_{\mathbb{R}^r} \Longrightarrow B_n \in \mathfrak{B}_{\mathbb{R}^r}$, donc de tout cela ***et selon la continuité croissante de la mesure***, on obtient : $\mu\left(\mathbb{R}^r\right) = \mu\left(\bigcup_{i\in\mathbb{N}} B_n\right) = \lim_{n\to+\infty} \mu\left(B_n\right)$.

ce qui se traduit comme suit :

$$(\forall \varepsilon > 0)\left(\exists N_\varepsilon \in \mathbb{R}^+\right) \text{ tq } : (\forall n \in \mathbb{N}) : \ n \geqslant N_\varepsilon \Longrightarrow \left|\mu\left(B_n\right) - 1\right| = 1 - \mu\left(B_n\right) \leqslant \varepsilon .$$

et ceci implique que : $(\forall \varepsilon > 0)\left(\exists \underbrace{M_\varepsilon}_{=[N_{\varepsilon/2}]+1} \in \mathbb{N}\right) \text{ tq } : \left|\mu\left(B_{M_\varepsilon}\right) - 1\right| = 1 - \mu\left(B_{M_\varepsilon}\right) = \mu\left(\mathbb{R}^r \setminus B_{M_\varepsilon}\right) \leqslant \frac{\varepsilon}{2}$. ★

or, $\left(f_n\right)_{n\in\mathbb{N}}$ converge uniformément compactement vers $f \in C(\mathbb{R}^r)$,alors : pour tout compact $K \subset \mathbb{R}^r$ on a : $\sup_{x\in K}\left|f_n(x) - f(x)\right| \underset{n\to\infty}{\longrightarrow} 0$ donc, et en particulier [a] : $\sup_{x\in B_{M_\varepsilon}} \left|f_n(x) - f(x)\right| \underset{n\to\infty}{\longrightarrow} 0$

ce qui fait que : $\left(\forall \varepsilon' > 0\right)\left(\exists \eta_{\varepsilon'} \in \mathbb{R}^+\right) \text{ tq } : (\forall n \in \mathbb{N}) : \ n \geqslant \eta_{\varepsilon'} \Longrightarrow \sup_{x\in B_{M_\varepsilon}} \left|f_n(x) - f(x)\right| \leqslant \varepsilon'$.

on prend en particulier $\varepsilon' = \frac{\varepsilon}{2}$,alors :

$$\left(\exists \eta_{\varepsilon'} \in \mathbb{R}^+\right) \text{ tq } : (\forall n \in \mathbb{N}) : \ n \geqslant \eta_{\varepsilon'} \Longrightarrow \sup_{x\in B_{M_\varepsilon}} \left|f_n(x) - f(x)\right| \leqslant \varepsilon' = \frac{\varepsilon}{2} .$$

d'où :

$$\begin{aligned}
(\forall n \in \mathbb{N}) : \int_{\mathbb{R}^r} \min\left\{\left|f_n(x) - f(x)\right|, 1\right\} \mathrm{d}\mu &= \int_{\mathbb{R}^r \setminus B_{M_\varepsilon}} \min\left\{\left|f_n(x) - f(x)\right|, 1\right\} \mathrm{d}\mu + \int_{B_{M_\varepsilon}} \min\left\{\left|f_n(x) - f(x)\right|, 1\right\} \mathrm{d}\mu \\
&\leqslant \int_{\mathbb{R}^r \setminus B_{M_\varepsilon}} 1 \, \mathrm{d}\mu + \int_{B_{M_\varepsilon}} \left|f_n(x) - f(x)\right| \mathrm{d}\mu \\
&\leqslant \mu\left(\mathbb{R}^r \setminus B_{M_\varepsilon}\right) + \int_{B_{M_\varepsilon}} \sup_{x\in B_{M_\varepsilon}} \left|f_n(x) - f(x)\right| \mathrm{d}\mu \\
&\leqslant \underbrace{\mu\left(\mathbb{R}^r \setminus B_{M_\varepsilon}\right)}_{\leqslant \frac{\varepsilon}{2} \text{ selon } :\star} + \int_{\mathbb{R}^r} \sup_{x\in B_{M_\varepsilon}} \left|f_n(x) - f(x)\right| \mathrm{d}\mu \leqslant \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon
\end{aligned}$$

a. la boule $B_{M_\varepsilon} = B(0, M_\varepsilon]$ est un fermé borné de $\mathbb{R}^r$ donc, elle est compact selon le théorème de ***Borel-Lebesgue***.

donc : $\int_{\mathbb{R}^r} \min\left\{\left|f_n(x)-f(x)\right|,1\right\}\,\mathrm{d}\mu \leqslant \varepsilon$, d'où ce qu'il faut prouver. ■

Lemme 2.2.3 :(la ρ_μ densité de $C\left(\mathbb{R}^r\right)$dans M^r,[34])

Si μ est une mesure ***finie*** sur : $\left(\mathbb{R}^r,\mathfrak{B}_{\mathbb{R}^r}\right)$,alors l'ensemble $C\left(\mathbb{R}^r\right)$ sera $\rho_\mu-$dense dans M^r.

Preuve :

Pour montrer que l'ensemble $C\left(\mathbb{R}^r\right)$ est $\rho_\mu-$ dense dans ***l'espace métrique*** $\left(M^r,\rho_\mu\right)$,***il faut et il suffit*** de trouver pour n'importe quel fonction :$f\in M^r$, une certaine suite : $\Phi_{n\in\mathbb{N}}\in C\left(\mathbb{R}^r\right)$ tel que : $\Phi_n \xrightarrow[n\to+\infty]{\rho_\mu} f$,on rappelle que selon le lemme III.2.2.1,cette convergence revient à dire que :

$$\int_{\mathbb{R}^r}\min\left\{\left|\Phi_n(x)-f(x)\right|,1\right\}\mathrm{d}\mu \xrightarrow[n\to+\infty]{} 0$$

soit : $f\in M^r$,on pose : $(\forall n\in\mathbb{N}):\varphi_n:=f\times\mathbb{1}_{\{|f|\leqslant n\}}\in L^1\left(\mathbb{R}^r,\mathfrak{B}_{\mathbb{R}^r},\mu\right)$, il est bien claire que :

$$\lim_{n\to+\infty}\int_{\mathbb{R}^r}\min\left\{\left|\varphi_n-f\right|,1\right\}\mathrm{d}\mu=0\ ^{a}$$

donc,pour tout : $\varepsilon>0$ on peut trouver un N_ε suffisamment grand tel que :

$$\left(\forall n\geqslant N_\varepsilon\right):\int_{\mathbb{R}^r}\min\left\{\left|\varphi_n-f\right|,1\right\}\mathrm{d}\mu<\frac{\varepsilon}{2}.\quad\bigstar$$

D'une autre part, et selon la proposition I.3.1.1 μ sera régulière (voir la définition I.3.1.1) ce qui fait que $\left(\mathbb{R}^r,\mathfrak{B}_{\mathbb{R}^r},\mu\right)$ sera en particulier ***un espace de Radon,*** alors et par le théorème de Lusin généralisé I.3.2.1 qui nous garantit la possibilité de restreindre toutes les fonctions mesurables sur un espace de Radon à des autres qui sont continues sur une partie ***fermée*** ,aussi grande que voulue,on peut trouver forcement une fonction continue Φ_n tel que :$\varphi_n=\Phi_n$ sur un fermé F qui vérifie :$\mu\left(F\right)<1-\frac{\varepsilon}{2}$. et dans ce cas on obtient pour tout $n\geqslant N_\varepsilon$:

$$\begin{aligned}\int_{\mathbb{R}^r}\min\left\{\left|\varphi_n-\Phi_n\right|,1\right\}\mathrm{d}\mu&=\int_F\min\left\{\left|\varphi_n-\Phi_n\right|,1\right\}\mathrm{d}\mu+\int_{F^c}\min\left\{\left|\varphi_n-\Phi_n\right|,1\right\}\mathrm{d}\mu\\&<\int_{F^c}1\,\mathrm{d}\mu<\mu\left(F^c\right)=<1-\mu\left(F\right)=1-\left(1-\frac{\varepsilon}{2}\right)=\frac{\varepsilon}{2}\quad\bigstar\bigstar\end{aligned}$$

alors et par une simple application de l'inégalité triangulaire on obtient pour tout $n\geqslant N_\varepsilon$:

$$\int_{\mathbb{R}^r}\min\left\{\left|f-\Phi_n\right|,1\right\}\mathrm{d}\mu\leqslant\int_{\mathbb{R}^r}\min\left\{\left|f-\varphi_n\right|,1\right\}\mathrm{d}\mu+\int_{\mathbb{R}^r}\min\left\{\left|\varphi_n-\Phi_n\right|,1\right\}\mathrm{d}\mu<\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon$$

d'où ce qu'il faut établir. ■

Théorème 2.2.2 :

Pour toute fonction ***continue non constante*** σ,tout $r\in\mathbb{N}^*$,et toute ***mesure de probabilité*** μ sur $\left(X,\mathfrak{B}_X\right)$, l'ensemble $\Sigma\Pi^r(\sigma)$ est $\rho_\mu-$dense dans M^r.

Autrement dit,et en adoptant la même formulation que [34] ,les réseaux de neurones $\Sigma\Pi$ à une seule couche cachée et à propagation en avant,peuvent bien approcher n'importe quelle fonction mesurable,indépendamment de la fonction continue non constante σ utilisée,de la dimension de l'espace d'entrée r, et d'environnement spatial d'entrée [b] μ, en l'occurrence les réseaux $\Sigma\Pi$ sont des approximateurs universels.

a. Cette convergence peut être justifiée géométriquement par le fait que lorsque n devient plus grand les courbes de φ_n et f deviennent plus confondues.

b. La mesure adoptée pour l'espace d'entrée s'appelle l'environnement spatial d'entrée,cette terminologie était introduite pour la première fois en 1988 par Halbert White.

Preuve :

Prenons une fonction continue non constante σ , selon le théorème III.2.2.1 l'ensemble $\Sigma\Pi^r(\sigma)$ est uniformément compactement dense dans $C\left(\mathbb{R}^r\right)$,en outre, notre mesure μ est de probabilité donc on peut déduire en vertu du lemme III.2.2.2 que $\Sigma\Pi^r(\sigma)$ est $\rho_\mu-$ dense dans M^r et en conséquence $\Sigma\Pi^r(\sigma)$ sera $\rho_\mu-$ dense dans M^r,ce qui achève bien notre preuve. ■

Remarque 2.2.3 :

On doit signaler quand même, que la dernière partie de la preuve précédente a été basée sur un résultat simple mais très important de la topologie classique et qui s'énonce ainsi :
Soit $A\subset B\subset C$ ***trois parties emboitées d'un espace métrique*** (X,d) ***si*** A ***est dense dans*** B ***et*** B ***est dense dans*** C***. alors*** A ***sera dense dans*** C.

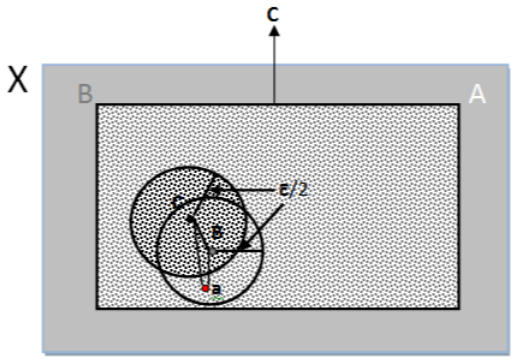


FIGURE III.6: Indication de la démonstration .

La preuve de ce résultat est très facile et elle découle essentiellement de l'inégalité triangulaire,on laissera au lecteur le plaisir de la démontrer avec détails en lui donnant comme indication la figure :III.6 .

Lemme 2.2.4 :(l'approximation des fonctions d'écrasement par des Σ réseaux de neurones :)

Soit $F:\mathbb{R}\to[0,1]$ une fonction d'écrasement continue, et $\Psi:\mathbb{R}\to[0,1]$ une fonction d'écrasement quelconque alors on a :

$$(\forall\varepsilon>0)\left(\exists H_\varepsilon\in\Sigma^1(\Psi)\right)\text{ tel que : }\quad\sup_{\lambda\in\mathbb{R}}\left|F(\lambda)-H_\varepsilon(\lambda)\right|<\varepsilon\,.$$

Preuve :

Soit $\varepsilon>0$, sans perte de généralité on peut considérer que : $\varepsilon<1$,notre but est de trouver un entier $Q\in\mathbb{N}$,une collection finie des réels, $(\beta_1,\cdots,\beta_Q)\in\mathbb{R}^Q$,et des fonctions affines : $(A_1,\cdots,A_Q)\in\left(\mathfrak{A}^1\right)^Q$ tel que :

$$\sup_{\lambda\in\mathbb{R}}\left|F(\lambda)-\sum_{j=1}^{Q}\beta_j\Psi\left(A_j(\lambda)\right)\right|<\varepsilon\,.$$

prenons pour cela un entier $Q\in\mathbb{N}$ de sorte que :$\dfrac{1}{1+Q}<\dfrac{\varepsilon}{2}$,et pour tout $j\in\{1,\cdots,Q\}$ posons $\beta_j=\dfrac{1}{Q+1}$,on choisit un $M>0$ tel que :$\Psi(-M)<\dfrac{\varepsilon}{2Q}$, et $\Psi(M)>1-\dfrac{\varepsilon}{2Q}$,(un tel M existe toujours car : $\lim\limits_{\lambda\to+\infty}\Psi(\lambda)=1$ et $\lim\limits_{\lambda\to-\infty}\Psi(\lambda)=0$), et on considère ensuite le Q+1-uplet :$(r_1,\cdots,r_{Q+1})$

défini par :

$$\begin{cases} r_j = \sup\left\{\lambda\in\mathbb{R}/F(\lambda)=\frac{j}{Q+1}\right\} & \text{si : } j\in\{1,\cdots,Q\}. \\ r_j = \sup\left\{\lambda\in\mathbb{R}/F(\lambda)=1-\frac{1}{2(Q+1)}\right\} & \text{si : } j=Q+1. \end{cases}$$

► ***Note explicative*** : il est indispensable de mentionner que les deux bornes supérieures précédentes ont bien un sens et ceci se justifie comme suit :
F est continue avec : $\lim\limits_{\lambda\to-\infty} F(\lambda)=0$ et $\lim\limits_{\lambda\to+\infty} F(\lambda)=1$,donc et ***selon le théorème des valeurs intermédiaires*** :

$$\left(\forall\delta\in]0,1[\right)\left(\exists\lambda_\delta\in\mathbb{R}\right)\text{ tq : } F(\lambda_\delta)=\delta\,.$$

en posant en particulier $\delta=\frac{j}{Q+1}$ pour tout $j\in[\![1,Q]\!]$, on pourra affirmer facilement que :

$$\left\{\lambda\in\mathbb{R}/F(\lambda)=\frac{j}{Q+1}\right\}\neq\varnothing\,.$$

d'une autre part : $\boxed{\left(\forall\, j\in[\![1,Q]\!]\right):\left\{\lambda\in\mathbb{R}/F(\lambda)=\frac{j}{Q+1}\right\}\text{ est borné.}}$ car s'il ne l'est pas on aura :

$$(\forall n\in\mathbb{N})\left(\exists\alpha_n\geqslant n\right)\text{ tel que : } F(\alpha_n)=\frac{j}{Q+1}\,.$$

et ceci va impliquer que : $\lim\limits_{n\to+\infty}\alpha_n=+\infty$ et $\lim\limits_{n\to+\infty}F\left(\alpha_n\right)=\frac{j}{Q+1}<1$, ce qui est absurde car $\lim\limits_{\lambda\to+\infty}F(\lambda)=1$.
pour l'ensemble $\left\{\lambda\in\mathbb{R}/F(\lambda)=1-\frac{1}{2(Q+1)}\right\}$ la bornitude et la non videsse [a] peuvent être établies d'une façon similaire. □
pour tout $\theta<\theta'$ notons :$A_{\theta,\theta'}\in\mathfrak{A}^1$ l'unique forme affine vérifiant : $A_{\theta,\theta'}(\theta)=M$ et $A_{\theta,\theta'}(\theta')=-M$.
l'approximation désirée est donc : $H_\varepsilon(\lambda)=\sum\limits_{j=1}^{Q}\beta_j\Psi\left(A_{\theta,\theta'}(\lambda)\right)$ puisqu'on peut voir facilement que sur les intervalles :$]-\infty,r_1],]r_1,r_2],\cdots,]r_Q,+\infty[$ on a :$\left|F(\lambda)-H_\varepsilon\right|<\varepsilon$ [b] ■

Théorème 2.2.3 :

Pour toute fonction ***d'écrasement*** Ψ,tout $r\in\mathbb{N}^*$,et toute ***mesure de probabilité*** μ sur $\left(X,\mathfrak{B}_X\right)$, l'ensemble $\Sigma\Pi^r(\Psi)$ est uniformément compactement dense dans $C\left(\mathbb{R}^r\right)$ et au même temps il est $\rho_\mu-$dense dans M^r.

c'est à dire que les théorèmes :III.2.2.1 et III.2.2.2 ***restent valables pour les fonctions d'écrasement,*** [c]et c'est exactement ce qui était établi dans l'article : [34].

Preuve :

Nous allons démontrer en premier lieu que :l'ensemble $\Sigma\Pi^r(\Psi)$ est uniformément compactement dense dans $C\left(\mathbb{R}^r\right)$,et le fait de la ρ_μ densité dans M^r sera déduite immédiatement via les lemmes :III.2.2.2,III.2.2.3 et la remarque III.2.2.3 ·
rappelons que selon le théorème III.2.2.1 , on a :l'ensemble $\Sigma\Pi^r(\sigma)$ est uniformément compactement dense dans $C\left(\mathbb{R}^r\right)$ pour toute fonction ***continue non constante*** σ, alors si on établit que :$\Sigma\Pi^r(\Psi)$ est uniformément compactement dense dans $\Sigma\Pi^r(\sigma)$ pour une certaine fonction continue non constante σ, le résultat désiré sera obtenu immédiatement grâce à la remarque III.2.2.3·
prenons $\varepsilon>0$ et $\ell\in\mathbb{N}^*$, car la multiplication est continue [d] et $[0,1]^\ell$ est compact, on pourra dé-

a. Ce terme a été introduit en 2015 pour remplacer le mot vuidesse qui est devenu moins utilisé et presque oublié dans le vocabulaire et le lexique français voir pour cela le dictionnaire DMF dans le site officiel du centre national de ressources textuelles et lexicales : www.cnrtl.fr.

b. Ça découle des inégalité précédentes, de la croissance des fonctions F et Ψ et la décroissance de la forme affine $A_{\theta,\theta'}$.

c. On rappelle que les fonctions d'écrasement ne sont pas forcement continues .

d. C'est à dire que : $(x_1,\cdots,x_\ell)\mapsto\prod\limits_{k=1}^{\ell}x_i$ est une fonction continue.

duire selon le fameux théorème de Heine[a], qu'elle est uniformément continue sur l'intervalle $[0,1]^{\ell}$,ce qui fait qu'il existe un $\delta_{\varepsilon}>0$ tel que :

$$\left(\forall\left((a_1,\cdots,a_\ell),(b_1,\cdots,b_\ell)\right)\in\left([0,1]^{\ell}\right)^2\right):\ \left[\left(\forall k\in\{1,\cdots,\ell\}\right):|a_k-b_k|<\delta_{\varepsilon}\right]\Rightarrow\left|\prod_{k=1}^{\ell}a_k-\prod_{k=1}^{\ell}b_k\right|<\varepsilon.$$

par le lemme :III.1.3.4, on peut trouver une fonction :$H_{\delta_{\varepsilon}}(\cdot)=\sum_{t=1}^{T}\beta_t\Psi\left(A_t(\cdot)\right)$ tel que : $\sup_{\lambda\in\mathbb{R}^r}\left|F(\lambda)-H_{\delta_{\varepsilon}}(\lambda)\right|<\delta_{\varepsilon}$
ce qui fait que :

$$\sup_{x\in\mathbb{R}}\left|\prod_{k=1}^{\ell}\sigma\left(A_k(x)\right)-\prod_{k=1}^{\ell}H_{\delta_{\varepsilon}}\left(A_k(x)\right)\right|=\sup_{x\in\mathbb{R}}\left|\prod_{k=1}^{\ell}\sigma\left(A_k(x)\right)-\prod_{k=1}^{\ell}\sum_{t=1}^{T}\beta_t\Psi\left(A_t\left(A_k(x)\right)\right)\right|<\varepsilon.$$

et car : $A_t\left(A_k(x)\right)\in\mathfrak{A}^r$,on peut déduire que : $\prod_{k=1}^{\ell}H_{\delta_{\varepsilon}}\left(A_k(\cdot)\right)\in\Sigma\Pi^r(\Psi)$.[b]

donc la fonction : $\prod_{k=1}^{\ell}\sigma\left(A_k(x)\right)$ peut être uniformément approchée par des éléments de $\Sigma\Pi^r(\Psi)$,et selon la remarque III.2.1.4 ceci implique qu'elle peut être uniformément compactement approchée par des éléments de $\Sigma\Pi^r(\Psi)$ donc, et sans aucun problème,$\Sigma\Pi^r(\Psi)$ sera uniformément compactement dense dans $\Sigma\Pi^r(\sigma)$ d'où ce qu'il faut établir. ■

Lemme 2.2.5 :

Pour toute fonction d'écrasement Ψ,tout $\varepsilon>0$, et tout $M>0$, il existe une fonction $\varphi_{M,\varepsilon}\in\Sigma^1(\Psi)$ tel que : $\sup_{\lambda\in[-M,M]}\left|\varphi_{M,\varepsilon}(\lambda)-\cos(\lambda)\right|<\varepsilon$.

Preuve :

Soit Γ la fonction du cosinus écraseur de Gallant-White[c],on peut remarquer que :

$$2\Gamma\left(x-\frac{3\pi}{2}\right)-1=\cos(x)\times\mathbb{1}_{[\pi,2\pi]}-\mathbb{1}_{]-\infty,\pi[}+\mathbb{1}_{]2\pi,+\infty[}\,.$$

Donc en ajoutant,soustrayant et multipliant par un scalaire un nombre fini de versions translatées

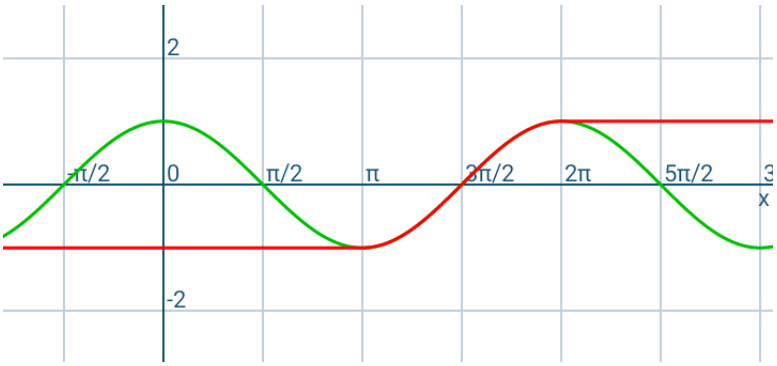


FIGURE III.7: La courbe de la fonction : $x\mapsto 2\Gamma\left(x-\frac{3\pi}{2}\right)-1$.

de Γ on ***peut obtenir une fonction***$\varphi_{M,\varepsilon}\in\Sigma^1(\Psi)$ qui se restreint à la fonction cosinus sur n'importe quel intervalle $[k\pi,k'\pi]$, et par la suite sur n'importe quel intervalle $[-M,M]$[d].
Le résultat découle maintenant du lemme III.2.2.3 et de ***l'inégalité triangulaire***. ■

a. Un théorème démontré par Eduard Heine en 1872 et qui énonce que toute application continue d'un ***espace métrique compact*** à valeurs dans un espace métrique quelconque est uniformément continue.
b. Il suffit de développer l'expression de note fonction pour rendre ce fait plus clair.
c. voir le troisième point de l'exemple :III.2.1.1·
d. il suffit de plonger cet intervalle dans un autre de la forme :$[k\pi,k'\pi]$.

Lemme 2.2.6 :

Soit $g:\mathbb{R}^r \to \mathbb{R}$ une certaine fonction numérique définie par : $g: x \mapsto \sum_{j=1}^{N} \beta_j \cos\left(A_j(x)\right)$,avec : $N \in \mathbb{N}^*, (\beta_1, \cdots, \beta_N) \in \mathbb{R}^N$ et : $A_j \in \mathfrak{A}^r$,alors, pour toute fonction d'écrasement Ψ,tout compact $K \subset \mathbb{R}^r$, et tout $\varepsilon > 0$, il existe un $f \in \Sigma^r(\Psi)$ tel que : $\sup_{x\in K} |g(x) - f(x)| < \varepsilon$.

Preuve :

Prenons un $M > 0$ tel que :pour tout : $j \in \{1, \cdots, N\}$ $A_j(K) \subset [-M, M]$, (car : $N < +\infty$, K est compact, et les A_j sont continues un tel M existe toujours,et ceci revient au fait qu'une fonction continue sur un compact est toujours bornée et atteint ses bornes).
on pose : $\varepsilon' = \frac{\varepsilon}{\sum_{j=1}^{N} |\beta_j|}$,par le lemme III.2.2.5,il existe une fonction $\varphi_{M,\varepsilon'} \in \Sigma^1(\Psi)$ tel que :

$$\sup_{\lambda\in[-M,M]} |\varphi_{M,\varepsilon}(\lambda) - \cos(\lambda)| < \varepsilon' .$$

ce qui fait que : $\left(\forall j \in \{1, \cdots, N\}\right)\left(\forall x \in K\right) : \left|\varphi_{M,\varepsilon}(A_j(x)) - \cos(A_j(x))\right| < \frac{\varepsilon}{N'}$.

$$\begin{aligned}
\text{donc : } \left(\forall x \in K\right) : \left|\sum_{j=1}^{N} \beta_j \varphi_{M,\varepsilon}(A_j(x)) - g(x)\right| &= \left|\sum_{j=1}^{N} \beta_j \varphi_{M,\varepsilon}(A_j(x)) - \sum_{j=1}^{N} \beta_j \cos\left(A_j(x)\right)\right| \\
&= \left|\sum_{j=1}^{N} \beta_j \left(\varphi_{M,\varepsilon}(A_j(x)) - \cos\left(A_j(x)\right)\right)\right| \\
&\leqslant \sum_{j=1}^{N} \left|\beta_j\right| \times \left|\varphi_{M,\varepsilon}(A_j(x)) - \cos\left(A_j(x)\right)\right| < \sum_{j=1}^{N} \left|\beta_j\right| \times \varepsilon' = \varepsilon .
\end{aligned}$$

ce qui achève notre preuve. ■

Théorème 2.2.4 :

Pour toute fonction ***d'écrasement*** Ψ,tout $r \in \mathbb{N}^*$,et toute ***mesure de probabilité*** μ sur $(X, \mathfrak{B}_X)$, l'ensemble $\Sigma^r(\Psi)$ est uniformément compactement dense dans $C(\mathbb{R}^r)$ et au même temps il est $\rho_\mu-$dense dans M^r.

c'est à dire que les théorèmes :III.2.2.1 et III.2.2.2 ***restent valables pour les*** Σ***réseaux de neurones.***(voir :[34] et [33])

Preuve :

On note Υ l'ensemble des polynômes trigonométriques défini par :

$$\Upsilon = \Sigma\Pi^r(\cos(\cdot)) = \left\{ x \mapsto G(x) = \sum_{j=1}^{q} \beta_j \cdot \prod_{k=1}^{l_j} \cos(A_{jk}(x)) \Big/ q \in \mathbb{N}^* \text{ , } l_j \in \mathbb{N} \text{ , } (\beta_1, \cdots, \beta_q) \in \mathbb{R}^q \text{ , } A_{jk}(x) \in \mathfrak{A}^r \right\} .$$

Par le théorème III.2.2.1cet ensemble est uniformément compactement dense dans : $C(\mathbb{R}^r)$ [a] donc et en appliquant plusieurs fois l'identité trigonométrique [b] :

$$\left(\forall (a,b) \in \mathbb{R}^2\right) : \quad \cos(a) \times \cos(b) = \frac{1}{2}\left(\cos(a+b) + \cos(a-b)\right) .$$

nous pouvons réécrire les éléments de Υ sous la forme : $\sum_{i=1}^{T} \alpha_i \cos\left(A_i(\cdot)\right)$ où : $\alpha_\epsilon \mathbb{R}$ et $A_i \in \mathfrak{A}^r$.

maintenant ,la densité uniforme compacte de Υ dans $C(\mathbb{R}^r)$ provient directement du lemme :III.2.2.6, et par la suite la ρ_μ densité de Υ dans M^r découlera immédiatement des lemmes :III.2.2.1,III.2.2.2 et la remarque III.2.2.2 · ■

a. Car la fonction $x \mapsto \cos(x)$ est continue et non constante sur $\mathbb{R}$.
b. Cette formule symbolise un peu la nostalgie de l'époque de lycée , n'est-ce pas cher lecteur ?! .

3 Théorèmes d'approximation de Funahashi :

3.1 Mise en situation :

On complétera notre liste de théorèmes d'approximation par cette section dans laquelle on va présenter, et d'une façon un peu résumée, les résultats principaux qui ont été proposés par le mathématicien japonais Ken-Ichi Funahashi $(1952,\cdots)$, la plupart des énoncés de ces résultats ne vont pas surprendre notre lecteur, car elles ne sont que des cas particuliers et des applications directes des résultats cités précédemment, mais l'ajout de cette section, vient comme une sorte d'appréciation, et de gratitude envers ce mathématicien qui se considère parmi les pionniers et les fondateurs de la théorie d'approximation, et à qui revient l'originalité des idées de la plupart des théorèmes traités auparavant.

3.2 Résultats principaux :

Théorème 3.2.1 :(Premier théorème de Funahashi , [21] :)

Soit $\phi(x)$ une fonction non constante, bornée, continue, et croissante.Soit $K \subset \mathbb{R}^n$ un compact et $f\left(x_l,\cdots,x_n\right)$ une fonction réelle continue sur K ,alors pour tout $\varepsilon > 0$, il existe un entier $N_{\varepsilon,K}$ des vecteurs réels : $\left(c_1,\cdots,c_n\right) \in \mathbb{R}^n$ et $\left(\theta_1,\cdots,\theta_n\right) \in \mathbb{R}^n$,et une matrice $\left(\omega_{i,j}\right)_{\substack{1\leqslant i\leqslant n\\ 1\leqslant j\leqslant N}}$ tel que :

$$\sup_{x\in K}\left|\sum_{j=1}^{N} c_j\,\phi\left(\sum_{i=1}^{n}\omega_{i,j}x_j-\theta_i\right)-f\right|<\varepsilon\,.\ ^{a}$$

Preuve :

La fonction ϕ est bornée et croissante donc on peut déduire que :

$$(\forall\, x\in\mathbb{R}):\ \lim_{n\to-\infty}\phi(x):=m\leqslant\phi(x)\leqslant M:=\lim_{n\to+\infty}\phi(x)$$

alors si on reprend les théorèmes précédents pour ce type de fonctions à la place des fonctions d'écrasement on obtiendra un résultat exactement analogue au théorème :III.2.2.4 et de cette façon le résultat sera aisément prouvé. ■

Remarque 3.2.1 :

On peut remarquer que l'hypothèse de la continuité de ϕ n'a pas été évoquée dans la dernière démonstration ,et ceci revient au fait que ***le résultat de Hornik-White-Stinchcombe pour les*** Σ ***réseaux de neurones est très fort, et il généralise celui énoncé dans ce dernier théorème*** ,mais faites attention!,avant les travaux de ces premiers, cette Hypothèse a joué un grand rôle dans la preuve proposée par Funahashi et qui était basée essentiellement sur le théorème de Paley-Wiener et la formule d'Irie-Miyake.

Théorème 3.2.2 :(Deuxième théorème de Funahashi, [21] :)

Pour n'importe quelle fonction $g \in M^n$,et pour tout $\varepsilon > 0$ il existe un compact $K \subset \mathbb{R}^n$ et une fonction $f \in \Sigma\left(\Psi\right)$ tel que : $\mu\left(K\right) = 1-\varepsilon$ et : $\sup_{x\in K}\left|f(x)-g(x)\right| < \varepsilon$,et ceci indépendamment du choix de la fonction d'écrasement Ψ ,de l'entier n, et de la mesure de probabilité :μ.

a. bien sûr les c_i, les θ_j et les $\omega_{i,j}$ dépendent du choix de ε et K .

Preuve :

Fixons un $\varepsilon > 0$,selon le théorème de Lusin I.1.3.1, il existe un certain compact $K^1 \subset \mathbb{R}$ tel que : $\mu(\mathbb{K}_1) > 1 - \frac{\varepsilon}{2}$ et : $g_{/\mathbb{K}_1}$ est continue sur $\mathbb{K}_1$, et selon le théorème de Tietze I.4.2.1 on peut prolonger $g_{/\mathbb{K}_1}$ par une certaine fonction continue $g' \in C(\mathbb{R}^r)$, ce qui fait qu'on aura :

$$\sup_{x\in\mathbb{K}_1} g'(x) = \sup_{x\in\mathbb{K}_1} g(x).$$

par le théorème : III.2.2.4 $\Sigma^r(\Psi)$ est uniformément compactement dense dans $C(\mathbb{R}^r)$,donc si on choisi un compact :$\mathbb{K}_2 \subset \mathbb{R}$ tel que :$\mu(\mathbb{K}) = 1 - \frac{\varepsilon}{2}$,on trouvera forcement une fonction $f \in \Sigma(\Psi)$ qui vérifie :
$\sup_{x\in\mathbb{K}_2} |g(x) - f(x)| < \varepsilon.$
donc : $\sup_{x\in\mathbb{K}\cap\mathbb{K}_1} |g(x) - f(x)| \leqslant \sup_{x\in\mathbb{K}_2} |g(x) - f(x)| < \varepsilon$,et d'une autre part :
$\mu(\mathbb{K}_1) > 1-\frac{\varepsilon}{2}$ et $\mu(\mathbb{K}_2) > 1-\frac{\varepsilon}{2} \Rightarrow \mu(\mathbb{K}_1{}^c) < \frac{\varepsilon}{2}$ et $\mu(\mathbb{K}_2^c) < \frac{\varepsilon}{2} \Rightarrow \mu(\mathbb{K}_1{}^c \cup \mathbb{K}_2^c) \leqslant \mu(\mathbb{K}_2^c) + \mu(\mathbb{K}_1{}^c) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2}$
alors :$\mu(\mathbb{K}_1{}^c \cup \mathbb{K}_2^c) = \mu((\mathbb{K}_1 \cap \mathbb{K}_2)^c) < \varepsilon \Longrightarrow \mu(\mathbb{K}_1 \cap \mathbb{K}_2) > 1 - \varepsilon$.
d'où le résultat. ■

Remarque 3.2.2 :

Ce dernier théorème nous a présenté un résultat très important et qui reflète la puissance des réseaux de neurones d'approcher n'importe quelle application mesurable en restant dans un cadre d'erreur tolérable,mais il a exigé d'adopter pour les nœuds constituants notre réseau une fonction d'activation d'écrasement ,une hypothèse que Cybenko a fait du bon travail pour l'affaiblir et de considérer juste des fonctions sigmoïdales mais et à son tour il était obligé d'ajouter la condition de continuité et d'affaiblir le type de la fonction à approcher et la rendre juste une fonction de décision et c'est ce qu'on voit au théorème :III.1.3.2 .

4 Propriété de parcimonie chez les réseaux de neurones :

4.1 Mise en contexte :

Dans la pratique, le nombre de paramètres nécessaires pour réaliser une approximation est un critère très important, ce qui fait que dans la recherche d'un approximateur, le concepteur du modèle doit toujours prendre ce critère en considération, et ceci dans l'objectif de ***construire ce modèle d'une sorte que le nombre de paramètres ajustables soit le plus faible possible***, on dit dans ce cas qu'on cherche à faire l'approximation la plus parcimonieuse .
Après la justification de leur capacité approximative, dans cette partie nous allons montrer, que certains types de réseaux neuronaux peuvent vraiment se présenter comme des ***approximateurs convenables et parcimonieux***, ainsi on insistera une autre fois sur l'importance de la fonction d'activation adoptée, et qui sera dans ce cas aussi, un élément clé et décisif pour la réalisation de la parcimonie.

4.2 Résultats principaux :

Le mathématicien américain **Andrew.R.Barron** a prouvé en 1993 dans son article [4] que, l'approximation par des modèles qui dépendent, non linéairement, des paramètres ajustables , est beaucoup plus parcimonieuse que celle qui se fait par des modèles dépendants linéairement de ces paramètres et plus précisément on a l'énoncé suivant :

Théorème 4.2.1 :(Barron,[4] :)

Pour une précision donnée, le nombre de paramètres ajustables, croît exponentiellement par rapport au nombre de variables dans le cas des ***approximateurs linéaires par rapport à leurs paramètres,*** alors qu'il croît juste linéairement par rapport à ce nombre pour des approximateurs non linéaires. ce qui reflète la grande puissance de ces derniers.

Remarque 4.2.1 :

1. Á la lumière du théorème précédent on peut remarquer que la parcimonie devient donc plus importante lorsque le nombre d'entrées du modèle est grand, autrement dit, si on a juste une ou deux entrées, on peut utiliser indifféremment un modèle linéaire comme les polynômes par exemple ou également un autre non linéaire (par rapport à ses paramètres) comme les réseaux de neurones artificiels.
2. exactement comme la propriété d'approximation universelle, la parcimonie n'est pas vérifiées pour tous les types des réseaux neuronaux, or, l'importance du choix de la fonction d'activation s'impose une autre fois dans ce cadre comme un critère critique qui peut améliorer la capacité du réseau si elle est bien choisie, ou l'entraver et le rendre par conséquent non parcimonieux, dans le cas contraire.
   voir à titre d'exemple le réseau Madaline [a] à plusieurs entrées ci dessous, et qui est linéaire par rapport à ces paramètres.

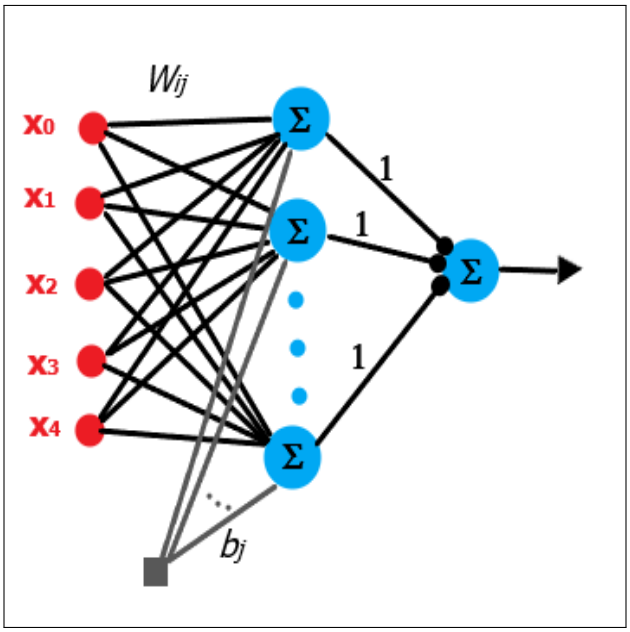


FIGURE III.8: Exemple d'un réseau non parcimonieux.

Théorème 4.2.2 :(Hornik,[35])

Pour les réseaux de neurones ***à fonction d'activation sigmoïdale***, l'erreur commise dans l'approximation varie comme l'inverse du nombre de neurones cachés, et elle est donc ***indépendante du nombre de variables de la fonction à approcher***.
Par conséquent, pour une précision fixée, et donc pour un nombre de neurones cachés donné, le nombre de paramètres du réseau est proportionnel au nombre de variables de la fonction à approcher.

a. Un Madaline (Many adaline) est tout simplement un réseau à trois couches constitué de plusieurs Adaline

Remarque 4.2.2 :

Ce résultat s'applique aux réseaux de neurones à fonction d'activation sigmoïdale puisque la sortie de ces neurones n'est pas linéaire par rapport aux poids synaptiques.
Cette propriété montre l'intérêt des réseaux de neurones sigmoïdes par rapport à d'autres approximateurs comme les polynômes dont la sortie est une fonction linéaire des paramètres ajustables, et pour mieux éclaircir les choses on peut dire que :
pour un même nombre d'entrées, le nombre de paramètres ajustables à déterminer est plus faible pour un réseau de neurones que pour un polynôme.

Cette propriété est indispensable dans le cas du filtrage des textes car le nombre d'entrées est typiquement très élevé.

Conclusion :

Les résultats de ce chapitre ont établi que les réseaux de neurones artificiels feed forward à une seule couche cachée**,** sont capables d'approcher et **,**au degré souhaité de précision**,** plusieurs types de fonctions**,**y compris les fonctions continues et mesurables**,** et on a vu que la capacité de ces réseaux **,** et d'un point de vue théorique **,** dépend essentiellement du sens désiré d'approximation et du type de la fonction d'activation adoptée**,** ce qui va nous permettre enfin de compte de déduire que ces réseaux sont des ***approximateurs universels,*** et d'affirmer que ***tout manque de succès dans le cadre applicatif doit survenir forcement d'un apprentissage inadéquat, un nombre insuffisant d'unités cachées ou d'une absence complète de la relation déterministe entre l'entrée et la sortie désirée (cas aléatoire)***.

Après toutes ces affirmations qu'on a présenté jusqu'à l'instant dans l'espoir de justifier théoriquement la grande aptitude des réseaux de neurones artificiels, on mettra dans le chapitre suivant cette machine à l'épreuve et ceci dans le cadre précis de la résolution des équations différentiels fractionnaires.

Chapitre IV

Application : Résolution numérique des équations différentielles fractionnaires par les Réseaux de Neurones Artificiels :

Résumé :

Dans ce chapitre, nous allons commencer par un aperçu sur la dérivation fractionnaire, et à travers lequel on découvrira comment cette notion et qui était définie habituellement pour des ordres entiers, s'est étendue pour englober des ordres quelconques, ensuite on verra que cette généralisation sera à l'origine d'apparition d'un nouveau type d'équations différentielles, appelées les équations différentiels fractionnaires.
L'invalidité des schémas numériques classiques pour ce types d'équations va nous obliger de se mettre dans le cadre d'apprentissage automatique, dans lequel on montrera que ***les réseaux de neurones artificiels*** adoptant une fonction sigmoïdale en tant que fonction d'activation, et une architecture feed-forward à trois couches, sont ***capables de générer des schémas itératifs pour résoudre les équations différentielles*** linéaires d'ordre ***fractionnaire***.

L'algorithme classique d'apprentissage du ***rétro-propagation*** et qui était basé sur la méthode de descente de gradient comme on a déjà vu au deuxième chapitre, sera, et après l'introduction de quelques modifications, capable d'***approcher, et au degré souhaité de précision, la solution*** de notre problème et même sur n'importe quel intervalle d'étude donné.

Enfin de compte, et pour être plus précis, certains problèmes de test seront également présentés lors d'une étude illustrative des résultats numériques fournis par la méthode qu'on va proposer.

Mots clés : Réseaux de neurones, dérivation fractionnaire, équations différentielles d'ordre fractionnaire, algorithme de rétro-propagation.

Introduction :

Les équations différentielles sont des outils permettant d'étudier des phénomènes naturels dans un large éventail d'applications, et au fil des ans, des recherches scientifiques fondamentales ont conclu que celles qui se posent dans le cadre des problèmes du monde réel ont rarement des solutions analytiques exactes, par conséquent, nous étions obligés de les analyser numériquement via des techniques et des simulations sur l'ordinateur, ce qui va nous offrir un moyen puissant pour les résoudre plus facilement et rapidement.

En raison de leur complexité particulière, la classe des équations différentielles d'ordre fractionnaire s'est marquée par le comportement de ses solutions qui sont très difficile à trouver, à décrire et à comprendre, et pour cela on a proposé d'utiliser des modèles d'apprentissage pour estimer cette solution sur un domaine d'étude donné.

Dans ce sens le modèle le plus classique qui peut être proposé, sera les réseaux de neurones artificiels bien évidemment, la particularité de ces réseaux et comme on déjà vu au chapitre III réside dans le fait que chaque structure régulière peut être considérée comme un approximateur universel parcimonieux, ce qui fait, que ce modèle est approprié au problème mentionné précédemment et qu'il peut estimer avec une précision approximative très élevée la fonction inconnue sur le domaine dont on dispose.

L'objectif de ce chapitre est d'utiliser un réseau neuronal feed-forward à trois couches pour résoudre le problème de l'équation différentielle ordinaire fractionnaire (E.D.F.O), et afin d'arriver à ce but nous allons procéder comme suit :
dans un premier temps, on va discrétiser l'intervalle d'étude, bien sûr, il existe plusieurs façons pour le faire mais ici, nous utiliserons juste le schéma de discrétisation standard, ensuite l'équation différentielle d'origine sera reformulée dans le cas discret sous la forme d'un système d'équations dont les inconnus sont les images des points nodaux, et enfin ce système va être convertie en un problème d'optimisation qui sera résolu par l'auto-apprentissage des réseaux de neurones artificiels et qu'on le fera dans ce cadre par l'algorithme de rétropropagation du gradient qu'on a vu au deuxième chapitre.

1 Initiation à l'analyse fractionnaire :

L'***analyse fractionnaire*** est une discipline des mathématiques qui s'intéresse à l'étude des différentes possibilités permettant de définir des puissances numériques réelles [a] pour l'opérateur de différenciation, et en d'autres termes, c'est une branche qui cherche à attribuer un sens à la dérivation selon un ordre qui n'est pas forcement entier.

La première apparition du concept de la dérivée fractionnaire était en 1694 dans une lettre écrite par le mathématicien allemand **Wilhelm Leibniz** (1646 – 1716) à son collègue **Guillaume de l'Hôpital** (français, 1661, 1704), qui n'a pas exclu la possibilité de définir une telle notion.
le fondement effective de cette théorie s'est fait deux siècles après et il revient à **Joseph Liouville** (1809 – 1882) et à **Oliver Heaviside** (1850 – 1925) qui a introduit l'utilisation pratique de ces opérateurs différentiels fractionnaires dans l'analyse des lignes de transport d'électricité .

1.1 Quelques fonctions spéciales :

1.1.1 La fonction Gamma :

Définition 1.1.1 : (la fonction Gamma,[23] :)

On appelle ***"fonction Gamma"*** la fonction définie par :

$$\begin{array}{rccl} \Gamma & : & \mathbb{R}^{+*} & \longrightarrow & \mathbb{R}^{+} \\ & & x & \longmapsto & \displaystyle\int_0^{+\infty} t^{x-1} \times e^{-t}\, \mathrm{d}t \end{array}$$

a. Attention !, ne vous laisser pas tromper par le terme "fractionnaire", l'ordre de dérivation qu'on adoptera dans cette théorie sera réel.

Proposition 1.1.1 :

La fonction Gamma Γ présentée précédemment est bien définie, continue et même de classe C^{∞} sur $\mathbb{R}^{*+}$, en outre, elle est convexe et vérifie les caractérisations suivantes :

1. $\Gamma(1)=1$, et $\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$.
2. $(\forall x>0)\ :\ \Gamma(x+1)=x\times\Gamma(x)$.
3. la fonction : $x\mapsto \ln\big(\Gamma(x)\big)$ est convexe.

Preuve :

On note :

$$f:\ \begin{array}{ccc}\mathbb{R}^{+}\times\mathbb{R}^{+*} & \longrightarrow & \mathbb{R}^{+}\\ (x,t) & \longmapsto & t^{x-1}\times e^{-t}\end{array}\quad\text{et :}\quad \big(\forall x\in\mathbb{R}^{+*}\big):\ f_x:\ \begin{array}{ccc}\mathbb{R}^{+*} & \longrightarrow & \mathbb{R}^{+}\\ t & \longmapsto & t^{x-1}\times e^{-t}\end{array}$$

quel que soit $x\in\mathbb{R}$, f_x est continue, positive sur $]0,+\infty[$ donc $\Gamma(x)$ est définie lorsque f_x est intégrable sur $]0,+\infty[$, et bien sûr il est bien le cas car :

- $f_x(t)\underset{0}{\sim}t^{1-x}\Longrightarrow f_x$ est intégrable sur $]0,1[$.
- $f_x(t)\underset{+\infty}{=}o\left(\frac{1}{t^2}\right)\Longrightarrow f_x$ est intégrable sur $]0,+\infty[$.

en plus on a : $\Big(\forall\,[a,b]\subset\mathbb{R}\Big)\big(\forall x\in[a,b]\big)\,(\forall\,t>0)\ :\ f(x,t)\leqslant\varphi(t)$ avec : $\varphi(t)=\begin{cases}t^{a-1} & \text{si : } t\in\,]0,1]\\ e^{-t}\times t^{b-1} & \text{si : }]1,+\infty[\end{cases}$

donc et car φ est positive, continue par morceaux et intégrable sur : $]0,+\infty[$, on déduit et selon le théorème de continuité des intégrales paramétriques que : $x\mapsto\Gamma(x)$ est continue.
pour le fait que $\Gamma\in C^{\infty}\left(\mathbb{R}^{+*}\right)$, on va établir que : $\left(\forall k\in\mathbb{N}^{*}\right):\Gamma\in C^{k}\left(\mathbb{R}^{+*}\right)$.
d'abord : $\left(\forall k\in\mathbb{N}^{*}\right):f\in C^{k}\left(\left(\mathbb{R}^{+*}\right)^{2}\right)$ et :

$$\left(\forall(x,t)\in\left(\mathbb{R}^{+*}\right)^{2}\right):\frac{\partial^{k}f}{\partial x^{k}}(x,t)=\big(\ln(t)\big)^{k}\times e^{-t}\times t^{x-1}$$

Considérons de nouveau un intervalle compact $[a,b]\subset\mathbb{R}^{+*}$ et soit $\Psi_h:]0,+\infty[\to\mathbb{R}$ telle que :

$$\Big(\forall\,t\in\,]0,+\infty[\Big):\Psi(t)=\big|\ln(t)\big|^{k}\times\varphi(t)$$

on a : $\Psi_k\underset{0}{=}o\left(t^{\frac{a}{2}-1}\right)$ et $\Psi_k\underset{+\infty}{=}o\left(t^{-2}\right)$ donc Ψ_k est positive, continue par morceaux, intégrable sur $]0,+\infty[$ et vérifie :

$$\left(\forall(x,t)\in[a,b]\times\mathbb{R}^{+*}\right):\left|\frac{\partial^{k}f}{\partial x^{k}}(x,t)\right|\leqslant\Psi_k(t)$$

d'où et selon le théorème de dérivabilité des intégrales paramétriques on déduit que :

$$\left(\forall k\in\mathbb{N}^{*}\right)\left(\forall x\in\mathbb{R}^{+*}\right):\ \Gamma^{(k)}(x)=\frac{\mathrm{d}^{k}}{\mathrm{d}x^{k}}\int_{0}^{+\infty}f(x,t)\,\mathrm{d}t=\int_{0}^{+\infty}\frac{\partial^{k}f}{\partial x^{k}}(x,t)\,\mathrm{d}t=\int_{0}^{+\infty}\big(\ln(t)\big)^{k}\times e^{-t}\times t^{x-1}\,\mathrm{d}t$$

de cette dernière formule on peut constater que la dérivée seconde de Γ sera strictement positive, ce qui implique immédiatement que cette fonction est strictement convexe.

► pour les caractérisation :

- $\Gamma(1) = \int_0^{+\infty} e^{-t}\,\mathrm{d}t = \left[-e^{-t}\right]_0^{+\infty} = 1$ et $\Gamma\left(\frac{1}{2}\right) = \int_0^{+\infty} \frac{e^{-t}}{\sqrt{t}}\,\mathrm{d}t$ en posant : $u = \sqrt{t}$ on aura : $\Gamma\left(\frac{1}{2}\right) = 2 \times \int_0^{+\infty} e^{-t^2}\,\mathrm{d}t = \sqrt{\pi}$ (intégrale de Gauss).
- $(\forall x > 0)\ :\ \Gamma(x+1) = \int_0^{+\infty} t^x \times e^{-t}\,\mathrm{d}t = \underbrace{\left[-e^{-t} \times t^x\right]_0^{+\infty}}_{=0} - \int_0^{+\infty} -e^{-t} \times x t^{x-1}\,\mathrm{d}t = x \times \Gamma(x)$
- $x \longmapsto \ln\left(\Gamma(x)\right)$ est convexe comme étant le composé d'une fonction convexe et d'un autre croissante $\ln(\cdot)$. ■

Remarques 1.1.1 :

1. Le théorème de **Liouville** en algèbre différentielle montre en particulier que de nombreuses primitives de fonctions usuelles, telle que la fonction Γ ***ne peuvent jamais s'exprimer comme combinaison de fonctions élémentaires***.
2. le théorème de **Bohr-Mollerup**, (voir : [24]) caractérise la fonction gamma, comme la seule fonction définie sur $x > 0$ et qui vérifie simultanément les trois caractérisations mentionnées précédemment.
3. la fonction Γ est appelée parfois la fonction factorielle, car elle vérifie :

$$(\forall n \in \mathbb{N})\ :\ \Gamma(n+1) = n \times \Gamma(n) = n \times n-1 \times \Gamma(n-1) = \cdots = n! \times \Gamma(1) = n!$$

et ceci nous permet de la considérer comme une généralisation de cette notion, ce qui va nous permettre de l'étendre au cas des nombres réels,et on peut écrire dans ce cas :

$$(\forall x > 0)\ : x! = \Gamma(x+1)\,.$$

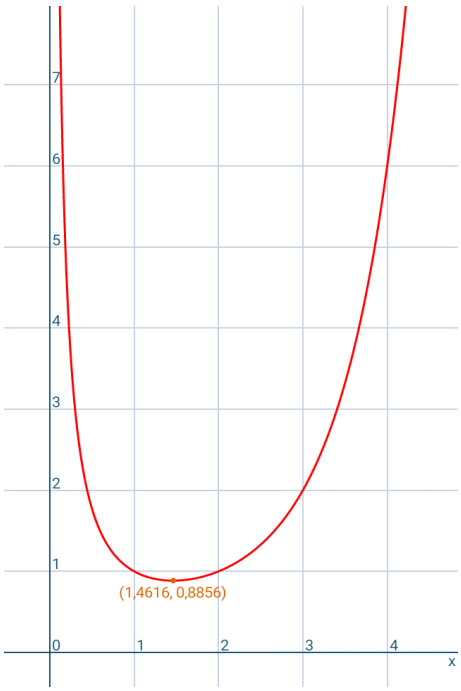


FIGURE IV.1: La courbe de la fonction Gamma.

1.1.2 La fonction Beta :

Définition 1.1.2 : (la fonction Bêta,[56] :)

On appelle ***"fonction Bêta"***, (ou fonction d'Euler) la fonction définie par :

$$\begin{array}{rccl} B & : & \mathbb{R}^{+*}\times\mathbb{R}^{+*} & \longrightarrow & \mathbb{R}^{+} \\ & & (x,y) & \longmapsto & \displaystyle\int_0^1 t^{x-1}\times(1-t)^{y-1}\,\mathrm{d}t \end{array}$$

Proposition 1.1.2 :

La fonction Bêta présentée précédemment est bien définie, symétrique, et elle vérifie l'égalité fondamentale suivante :

$$\left(\forall (x,y)\in\left(\mathbb{R}^{+*\,2}\right)\,:\; B(x,y)=\frac{\Gamma(x)\times\Gamma(y)}{\Gamma(x+y)}\right.$$

Preuve :

- ***la bonne définition de la fonction Bêta*** **:** pour démontrer ce fait, on va vérifier en premier lieu que l'application φ donnée ainsi :

$$\begin{array}{rccl} \varphi & : & \mathbb{R}^{+*}\times\mathbb{R}^{+*} & \longrightarrow & \mathbb{R}^{+} \\ & & (x,y) & \longmapsto & \displaystyle\int_0^{+\infty}\frac{t^{x-1}}{(1+t)^{x+y}}\,\mathrm{d}t \end{array}$$

  a bien un sens, puis, on établira qu'elle coïncide avec notre fonction B.
  on a :

$$\frac{t^{x-1}}{(1+t)^{x+y}}\underset{0^+}{\sim}t^{x-1}\;\text{ et :}\;\frac{t^{x-1}}{(1+t)^{x+y}}\underset{+\infty}{\sim}\frac{1}{t^{1+y}}$$

  donc et selon les critères de Riemann, et d'équivalence, l'intégrale généralisée $\displaystyle\int_0^{+\infty}\frac{t^{x-1}}{(1+t)^{x+y}}\,\mathrm{d}t$ sera convergente pour tout $x,y\in\mathbb{R}^{+*}$, ce qui fait que φ est bien définie.
  -en effectuant le changement de variable : $\rho=\dfrac{t}{t+1}$ on aura : $\quad t=\dfrac{\rho}{1-\rho}\;$ et $\;\mathrm{d}t=\dfrac{\mathrm{d}\rho}{(1-\rho)^2}$.
  donc pour tout $x>0$ et $y>0$ on a :

$$\begin{aligned}\varphi(x,y)=\int_0^{+\infty}\frac{t^{x-1}}{(1+t)^{x+y}}\mathrm{d}t&=\int_0^1\left(\frac{\rho}{1-\rho}\right)^{x-1}\left(1+\frac{\rho}{1-\rho}\right)^{-x-y}\frac{1}{(1-\rho)^2}\,\mathrm{d}\rho\\&=\int_0^1\left(\frac{\rho}{1-\rho}\right)^{x-1}(1-\rho)^{x+y}\frac{1}{(1-\rho)^2}\,\mathrm{d}\rho\\&=\int_0^1\rho^{x-1}\times(1-\rho)^{y-1}\,\mathrm{d}\rho=B(x,y)\end{aligned}$$

  et ceci nous donne en conséquence que la fonction bêta est bien définie sur $\left(\mathbb{R}^{+*}\right)^2$, car φ l'est.

- ***la symétrie*** **:** Il suffit d'effectuer dans l'expression de notre fonction le changement de variable $\rho=1-t$ pour obtenir l'égalité voulue :

$$\left(\forall\,(x,y)\in\left(\mathbb{R}^{+*}\right)^2\right)\,:\;B(x,y)=B(y,x)\,.$$

► ***la relation avec la fonction Gamma*** **:** soit a, b deux nombres strictement positifs, on a :

$$\Gamma(a)\times\Gamma(b)=\int_0^{+\infty}\int_0^{+\infty}e^{-x-y}\times x^{a-1}\times y^{b-1}\,\mathrm{d}x\,\mathrm{d}y\,.$$

on fait le changement de variables : $\rho=x+y$ et $x=\rho\cdot w$, donc $\rho\in\mathbb{R}^+$, $w\in[0,1]$ et :

$$\mathrm{d}\rho=\mathrm{d}x+\mathrm{d}y\,,\ \mathrm{d}x=w\,\mathrm{d}\rho+\rho\,\mathrm{d}w\,,\ \mathrm{d}y=(1-w)\,\mathrm{d}\rho-\rho\,\mathrm{d}w$$

donc $\mathrm{d}x\mathrm{d}y=\rho\,\mathrm{d}w\,\mathrm{d}\rho$, ce qui fait que :

$$\Gamma(a)\times\Gamma(b)=\int_0^1 w^{a-1}(1-w)^{b-1}\,\mathrm{d}w\times\int_0^{+\infty}e^{-\rho}\cdot\rho^{a+b-1}\,\mathrm{d}\rho=B(a,b)\times\Gamma(a+b)\,.$$

d'où ce qu'il faut établir. ■

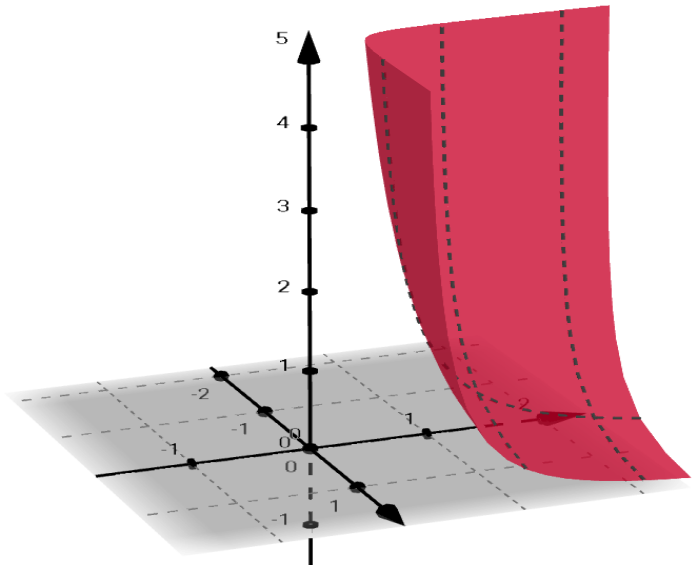


FIGURE IV.2: La courbe de la fonction bêta.

1.2 Intégrale fractionnaire de Riemann- Liouville :

Proposition 1.2.1 : (Formule de Cauchy pour l'intégration itérée :)

Soit f une fonction continue sur $\mathbb{R}$, alors on a pour tout $a\in\mathbb{R}$ fixé, pour tout $n\in\mathbb{N}^*$**,** et pour tout $(x,y_1,\cdots,y_{n-1})\in([a,+\infty[)^n$:

$$\boxed{\int_a^x\int_a^{y_1}\cdots\int_a^{y_{n-1}}f(y_n)\,\mathrm{d}y_n\cdots\mathrm{d}y_2\,\mathrm{d}y_1=\frac{1}{(n-1)!}\int_a^x(x-t)^{n-1}f(t)\,\mathrm{d}t}$$

Preuve :

On va procéder par récurrence sur n :

- pour $n=1$: l'égalité est triviale ce qui fait que la raison de récurrence est vérifiée.
- soit $n\in\mathbb{N}^*$, on suppose que notre proposition est vérifiée pour n, et on montre qu'elle l'est

aussi pour $n+1$:

$$\begin{aligned}\int_a^x\int_a^{y_1}\cdots\int_a^{y_n} f(y_n)\,\mathrm{d}y_{n+1}\cdots\mathrm{d}y_2\,\mathrm{d}y_1 &= \int_a^x \frac{1}{(n-1)!}\times\int_a^{y_1}\left(y_1-t\right)^{n-1} f(t)\,\mathrm{d}t\\ &= \int_a^x\left[\int_a^{y_1}\frac{1}{n!}\times\frac{\mathrm{d}}{\mathrm{d}y_1}\left(y_1-t\right)^n f(t)\right]\mathrm{d}t\\ &= \int_a^x\frac{1}{n!}\times\frac{\mathrm{d}}{\mathrm{d}y_1}\left[\int_a^{y_1}\left(y_1-t\right)^n f(t)\right]\mathrm{d}t\\ &= \frac{1}{n!}\times\int_a^x\left(y_1-t\right)^n f(t)\,\mathrm{d}t\end{aligned}$$

d'où :$\left(\forall n\in\mathbb{N}^*\right):\int_a^x\int_a^{y_1}\cdots\int_a^{y_{n-1}} f(y_n)\,\mathrm{d}y_n\cdots\mathrm{d}y_2\,\mathrm{d}y_1=\frac{1}{(n-1)!}\int_a^x(x-t)^{n-1}f(t)\,\mathrm{d}t$ ■

▷ La dérivée usuelle qu'on a eu l'habitude de l'utiliser dans le cadre de l'analyse réelle classique, peut être décrite plus correctement comme un opérateur défini de l'ensemble $\mathfrak{D}(\mathbb{R})$ des fonctions dérivable vers l'ensemble des fonctions numériques $\mathcal{F}\left(\mathbb{R},\mathbb{R}\right)$, de la façon suivante :

$$\begin{array}{rccc}\Delta: & \mathfrak{D}(\mathbb{R}) & \longrightarrow & \mathcal{F}\left(\mathbb{R},\mathbb{R}\right)\\ & f & \longmapsto & f'\end{array}$$

maintenant et à ce stade une question assez naturelle qui peut se poser, peut-on trouver un opérateur linéaire H , ou une ***semi-dérivé***, tel que :

$$H^2:=H\circ H=\Delta\,?$$

pour garder en principe la cohérence de cette interrogation, et surtout pour des ordres supérieurs à deux, on sera obligé de rendre l'espace de départ et d'arrivée de l'opérateur Δ identiques, pour cela on va le prendre et sans perte de généralité $C^\infty(\mathbb{R})$, l'espace des fonctions continument dérivable.

▷ Considérant une fonction $x\mapsto f(x)$ qui est ***définie et continue*** sur $\mathbb{R}^+$, nous pouvons et sans aucun problème exprimer sa primitive (qui s'annule en 0), par l'intégrale impropre définie de a à x par :

$$\left(\forall x>0\right):\left[\mathrm{P}\left(f\right)\right](x)=\int_0^x f(t)\,\mathrm{d}t\,.$$

selon la proposition 2.2.1, une application itérée de cet opérateur pour la fonction f nous donne :

$$\left(\forall n\in\mathbb{N}^*\right)\left(\forall x>0\right):\ \left[\mathrm{P}^n\left(f\right)\right](x):=\left[\left(P\circ\cdots\circ P\right)\left(f\right)\right](x)=\frac{1}{(n-1)!}\times\int_0^x\underbrace{(x-t)^{n-1}}_{\geq 0}\times f(t)\,\mathrm{d}t\,.$$

l'utilisation de la fonction gamma $x\mapsto\Gamma(x)$ pour franchir la nature entière de la factorielle sera un choix naturel et évident pour donner un sens aux applications fractionnaires de l'opérateur intégral, et ainsi on peut écrire :

$$\boxed{\left(\forall f\in C^0(\mathbb{R}^+)\right)\left(\forall\alpha>0\right)\left(\forall x>0\right):\ \left[\mathrm{P}^\alpha\left(f\right)\right](x):=\frac{1}{\Gamma(\alpha)}\times\int_0^x(x-t)^{\alpha-1}\times f(t)\,\mathrm{d}t\,.}\quad(\blacklozenge)$$

Définition 1.2.1 : (intégrale fractionnaire de Riemann- Liouville :)

On appelle ($\blacklozenge$) décrite ci-dessus la formule fondamentale du calcul fractionnaire, ou encore l'intégrale fractionnaire de Riemann-Liouville.

Proposition 1.2.2 :(la propriété du semi-groupe pour l'opérateur « P » :)

Soit $f \in C^0(\mathbb{R}^+)$, l'opérateur d'intégration P satisfait la propriété suivante pour tout α et β ***strictement positifs :***

$$\left(\forall x \in \left(\mathbb{R}^{+*}\right)\right) : \left[P^{\alpha}\left[P^{\beta}\left(f\right)\right]\right](x) = \left[P^{\beta}\left[P^{\alpha}\left(f\right)\right]\right](x) = \left[P^{\alpha+\beta}\left(f\right)\right](x) = \frac{1}{\Gamma(\alpha+\beta)}\int_0^x (x-t)^{\alpha+\beta-1} f(t)\,\mathrm{d}t.$$

Preuve :

Soit α et β deux nombres positifs strictement, on a et pour tout $x > 0$:

$$\begin{aligned}\left[P^{\alpha}\left[P^{\beta}\left(f\right)\right]\right](x) &= \frac{1}{\Gamma(\alpha)}\int_0^x (x-t)^{\alpha-1}\left[P^{\beta}\left(f\right)\right](t)\,\mathrm{d}t \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)}\int_0^x\int_0^t (x-t)^{\alpha-1}(t-s)^{\beta-1} f(s)\,\mathrm{d}s\,\mathrm{d}t \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)}\int_0^x f(s)\left(\int_s^x (x-t)^{\alpha-1}(t-s)^{\beta-1}\,\mathrm{d}t\right)\mathrm{d}s\end{aligned}$$

et en effectuant le changement de variables : $\rho = \dfrac{t-s}{x-s}$,on obtient :

$$\left[P^{\alpha}\left[P^{\beta}\left(f\right)\right]\right](x) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)}\int_0^x (x-s)^{\alpha+\beta-1} f(s)\left(\int_0^1 (1-r)^{\alpha-1} r^{\beta-1}\,\mathrm{d}r\right)\mathrm{d}s \quad (\blacktriangle)$$

on remarque que l'intégrande dans la dernière expression n'est autre que la fameuse fonction bêta et qui vérifie la propriété suivante :

$$\int_0^1 (1-\rho)^{\alpha-1}\rho^{\beta-1}\,\mathrm{d}\rho = B(\alpha,\beta) = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}\,.$$

en remplaçant cela dans l'équation (▲) on obtient :

$$\left[P^{\alpha}\left[P^{\beta}\left(f\right)\right]\right](x) = \frac{1}{\Gamma(\alpha+\beta)}\int_0^x (x-s)^{\alpha+\beta-1} f(s)\,\mathrm{d}s = \left[P^{\alpha+\beta}\left(f\right)\right](x)\,.$$

d'où ce qu'il faut prouver. ■

1.3 Dérivée fractionnaire de Caputo :

Le mathématicien italien **Michele Caputo** $(1927,\cdots)$ s'est inspiré des travaux de Riemann et Liouville pour améliorer et généraliser les notions d'intégration et de dérivation dans le cadre fractionnaire, et en 1967 il a proposé dans son article [8] une idée intelligente basée sur la propriété du semi groupe mentionnée à la proposition IV.2.2.2 pour introduire une généralisation fractionnaire de la dérivation usuelle et qui porte jusqu'à l'instant son nom : ***"la dérivation de Caputo"***.

Notations :

On note pour la suite : $\widehat{\mathbb{R}} = \mathbb{R} \setminus \mathbb{Z}$ et $\widehat{\mathbb{R}}^+ = \mathbb{R}^+ \setminus \mathbb{N}$.

Définitions 1.3.1 : (intégrale fractionnaire de Caputo :)

- Soit $f \in C^0(\mathbb{R})$ ***une fonction continue*** sur $\mathbb{R}$, et $a \in \mathbb{R}$ un nombre réel fixé , l'intégrale fractionnaire de Caputo ***pour un ordre positif*** est une généralisation de celui de Riemann-Liouville qui s'exprime sous les deux formes suivantes :

$$\begin{cases} \text{-supérieure :} \quad (\forall \alpha > 0)\,(\forall x > a) : \left[{}_aP_x^{\alpha}(f)\right](x) = \dfrac{1}{\Gamma(\alpha)} \displaystyle\int_a^x (x-t)^{\alpha-1} f(t)\,\mathrm{d}t. \\ \qquad \text{et :} \\ \text{-inférieure :} \quad (\forall \alpha > 0)\,(\forall x < a) : \left[{}_xP_a^{\alpha}(f)\right](x) = \dfrac{1}{\Gamma(\alpha)} \displaystyle\int_x^a (t-x)^{\alpha-1} f(t)\,\mathrm{d}t. \end{cases}$$

- afin d'étendre la formulation précédente aux ordres réels négatifs [a], Caputo s'est inspiré de la propriété du semi groupe chez l'intégrale de Riemann-Liouville pour nous fournir la définition suivante :
  - d'une part : $\left(\forall n \in \mathbb{N}^*\right) : \left[{}_xP_a^{-n}(f)\right](\cdot) = \dfrac{\mathrm{d}^n f}{\mathrm{d}x^n}(\cdot) = f^{(n)}(\cdot)$
  - et d'une autre :

$$\begin{aligned} \left(\forall \alpha \in \widehat{\mathbb{R}}^+\right)\left(\forall f \in C^{\lceil\alpha\rceil}(\mathbb{R})\right)(\forall x > a) : \left[{}_aP_x^{-\alpha}(f)\right](x) :&= \left[{}_aP_x^{\lceil\alpha\rceil-\alpha}\left[{}_aP_x^{-\lceil\alpha\rceil}(f)\right]\right](x) \\ &= \frac{1}{\Gamma\left(\lceil\alpha\rceil-\alpha\right)} \int_a^x (x-t)^{\lceil\alpha\rceil-\alpha-1} f^{(\lceil\alpha\rceil)}(t)\,\mathrm{d}t \\ &= \frac{1}{\Gamma\left(\lceil\alpha\rceil-\alpha\right)} \int_a^x \frac{f^{(\lceil\alpha\rceil)}(t)}{(x-t)^{\alpha-\lceil\alpha\rceil+1}}\,\mathrm{d}t. \end{aligned}$$

  et d'une façon analogue à la forme supérieure présentée en dessus, on obtient pour la forme inférieure :

$$\left(\forall \alpha \in \widehat{\mathbb{R}}^+\right)\left(\forall f \in C^{\lceil\alpha\rceil}(\mathbb{R})\right)(\forall x > a) : \left[{}_xP_a^{-\alpha}(f)\right](x) := \frac{1}{\Gamma\left(\lceil\alpha\rceil-\alpha\right)} \int_x^a \frac{f^{(\lceil\alpha\rceil)}(t)}{(t-x)^{\alpha-\lceil\alpha\rceil+1}}\,\mathrm{d}t\,.$$

Remarques 1.3.1 :

1. Dans la définition précédente, on a adopté la notation anglo-saxonne pour la partie entière et selon laquelle on écrit :
$$\underbrace{\lfloor\alpha\rfloor}_{\text{parte entière de } \alpha.} \leqslant \alpha < \lceil\alpha\rceil = \lfloor\alpha\rfloor + 1\,.$$
2. il est bien facile de justifier que : $\left(\forall n \in \mathbb{N}^*\right) : \left[{}_xP_a^{-n}(f)\right](\cdot) = \dfrac{\mathrm{d}^n f}{\mathrm{d}x^n}(\cdot) = f^{(n)}(\cdot)$ et ceci par le simple fait que la dérivation n'est autre que l'opérateur réciproque de l'intégration.
3. ***faites attention!*** , dans le cas où $\alpha \in \widehat{\mathbb{R}} = \left(\mathbb{R}^+ \setminus \mathbb{N}\right)$, la définition d'intégration fractionnaire proposée par Caputo ne s'adapte plus avec la propriété du commutativité (voir[3] page :22), c'est à dire que :
$$\left(\forall \alpha \in \widehat{\mathbb{R}}^+\right)(\forall x > a) : \left[{}_aP_x^{-\alpha}(f)\right](x) := \left[{}_aP_x^{\lceil\alpha\rceil-\alpha}\left[{}_aP_x^{-\lceil\alpha\rceil}(f)\right]\right](x) \neq \underbrace{\left[{}_aP_x^{-\lceil\alpha\rceil}\left[{}_aP_x^{\lceil\alpha\rceil-\alpha}(f)\right]\right](x)}_{\text{La dérivée fractionnaire de Riemann-Liouville.}}$$

a. On rappelle que le problème dans ce cas se pose au niveau de la fonction Gamma qui est définie juste pour des valeurs strictement positifs.

Définition 1.3.2 : (dérivation fractionnaire de Caputo :)

Soit $a \in \mathbb{R}$ et $\alpha \in \mathbb{R}$ des réels fixés, et soit $f \in C^{\lceil \alpha \rceil}(\mathbb{R})$ une fonction de classe $C^{\lceil \alpha \rceil}$ sur $\mathbb{R}$, on définit la dérivée de Caputo (supérieure) de f pour un l'ordre réel α, comme suit :

$$(\forall x > a) : \quad \left[{}_aD_x^{\alpha}(f)\right](x) = \begin{cases} f^{(\alpha)}(x) & \text{si : } \alpha \in \mathbb{Z}. \\ \left[{}_aP_x^{-\alpha}(f)\right](x) & \text{si : } \alpha \in \widehat{\mathbb{R}}. \end{cases}$$

ce qui fait que :

$$(\forall x > a) : \quad \left[{}_aD_t^{\alpha}(f)\right](x) = \begin{cases} f^{(\alpha)}(x) & \text{si : } \alpha \in \mathbb{N}. \\ \dfrac{1}{\Gamma\left(\lceil \alpha \rceil - \alpha\right)} \displaystyle\int_a^x \frac{f^{(\lceil \alpha \rceil)}(t)}{(x-t)^{\alpha - \lceil \alpha \rceil + 1}}\, \mathrm{d}t & \text{si : } \alpha \in \widehat{\mathbb{R}}^{+}. \\ \dfrac{1}{\Gamma(-\alpha)} \displaystyle\int_a^x (x-t)^{-\alpha-1} f(t)\, \mathrm{d}t & \text{si : } \alpha < 0. \end{cases}$$

et pour ce qui concerne la dérivée de Caputo inférieure, elle sera et sans surprise définie d'une façon analogue, juste il faut inverser les bornes de l'intégrale et remplacer $(x-t)$ par $(t-x)$ dans l'intégrande.

Proposition 1.3.1 :

Soit $\alpha \in \widehat{\mathbb{R}}$ et $f \in C^{\lceil \alpha \rceil}(\mathbb{R})$ tel que $n-1 < \alpha < n = \lceil \alpha \rceil \in \mathbb{N}$, alors et dans ce cas, l'opérateur de dérivation fractionnaire de Caputo $\left[{}_aD_x^{\alpha}(\bullet)\right]$ vérifie les propriétés suivantes :

$$\lim_{\alpha \to n} \left[{}_aD_x^{\alpha}(f)\right](x) = f^{(n)}(x) \text{ et } \lim_{\alpha \to n-1} \left[{}_aD_x^{\alpha}(f)\right](x) = f^{(n-1)}(x) - f^{(n-1)}(a)$$

preuve :

On va juste utiliser une intégration par partie :

$$\begin{aligned} \left[{}_aD_x^{\alpha}(f)\right](x) &= \frac{1}{\Gamma\left(\lceil \alpha \rceil - \alpha\right)} \int_a^x \frac{f^{(n)}(t)}{(x-t)^{\alpha - \lceil \alpha \rceil + 1}}\, \mathrm{d}t \\ &= \frac{1}{\Gamma\left(\lceil \alpha \rceil - \alpha\right)} \times \left(\left[-f^{(n)}(t) \cdot \frac{(x-t)^{\lceil \alpha \rceil - \alpha}}{\lceil \alpha \rceil - \alpha} \right]_{t=a}^{t=x} - \int_a^x -f^{(n+1)}(t) \times \frac{(x-t)^{\lceil \alpha \rceil - \alpha}}{n-\alpha}\, \mathrm{d}t \right) \\ &= \frac{1}{\Gamma\left(\lceil \alpha \rceil - \alpha\right)} \times \left(-f^{(n)}(a) \cdot \frac{(x-a)^{\lceil \alpha \rceil - \alpha}}{\lceil \alpha \rceil - \alpha} + \int_a^x f^{(n+1)}(t) \times \frac{(x-t)^{\lceil \alpha \rceil - \alpha}}{n-\alpha}\, \mathrm{d}t \right) \\ &= \frac{1}{\Gamma\left(\lceil \alpha \rceil - \alpha + 1\right)} \times \left(-f^{(n)}(a) \cdot (x-a)^{\lceil \alpha \rceil - \alpha} + \int_a^x f^{(n+1)}(t) \times \frac{(x-t)^{\lceil \alpha \rceil - \alpha}}{n-\alpha}\, \mathrm{d}t \right) \end{aligned}$$

en faisant tendre respectivement $\alpha \to n$ et $\alpha \to n-1$, on obtiendra :

$$\lim_{\alpha \to n} \left[{}_aD_x^{\alpha}(f)\right](x) = f^{(n)}(a) + \left[f^{(n)}(x)\right]_{t=a}^{t=x} = f^{(n)}(x)$$

et :

$$\begin{aligned} \lim_{\alpha \to n} \left[{}_aD_x^{\alpha}(f)\right](x) &= f^{(n)}(a).(x-a) + \left[f^{(n)}(x).(x-t)\right]_{t=a}^{t=x} + \int_a^x f^{(n)}(t)\, \mathrm{d}t \\ &= \left[f^{(n-1)}(t)\right]_{t=a}^{t=x} \\ &= f^{(n-1)}(x) - f^{(n-1)}(a) \end{aligned}$$

d'où ce qu'il faut établir. ■

Exemple 1.3.1 :

Supposons que f est un monôme de la forme $f(x)=x^n$ définie sur $\mathbb{R}^+$, c'est à dire :

$$\boxed{\begin{array}{ccccc} f & : & \mathbb{R}^+ & \longrightarrow & \mathbb{R}^+ \\ & & t & \longmapsto & t^n \end{array}}$$

on a clairement : $f\in C^\infty\left(\mathbb{R}^+\right)$, et si on pose $a=0$ et on prend $\alpha\in\widehat{\mathbb{R}}^+$ on obtiendra :

$$\left(\forall x\in\mathbb{R}^+\right):\quad \left[{}_aD_t^\alpha\left(f\right)\right](x)=\frac{1}{\Gamma\left(\lceil\alpha\rceil-\alpha\right)}\int_a^x\frac{f^{(\lceil\alpha\rceil)}(t)}{(x-t)^{\alpha-\lceil\alpha\rceil+1}}\,\mathrm{d}t\,.$$

- si $\lceil\alpha\rceil>n$ alors : $f^{(\lceil\alpha\rceil)}(t)=0$, ce qui fait que :

$$\left(\forall x\in\mathbb{R}^+\right):\quad \left[{}_0D_t^\alpha\left(t^n\right)\right](x)=0\,.$$

- si $\lceil\alpha\rceil<n$ alors : $f^{(\lceil\alpha\rceil)}(t)=n\times(n-1)\times\cdots\times\left(n-\lceil\alpha\rceil+1\right)\times t^{n-\lceil\alpha\rceil}$, ce qui fait que :

$$\begin{aligned}
\left(\forall x\in\mathbb{R}^+\right):\quad \left[{}_0D_t^\alpha\left(t^n\right)\right](x)&=\frac{1}{\Gamma\left(\lceil\alpha\rceil-\alpha\right)}\times\int_0^x\frac{f^{(\lceil\alpha\rceil)}(t)}{(x-t)^{\alpha-\lceil\alpha\rceil+1}}\,\mathrm{d}t\\
&=\frac{1}{\Gamma\left(\lceil\alpha\rceil-\alpha\right)}\times\int_0^x\frac{n\times(n-1)\times\cdots\times\left(n-\lceil\alpha\rceil+1\right)\times t^{n-\lceil\alpha\rceil}}{(x-t)^{\alpha-\lceil\alpha\rceil+1}}\,\mathrm{d}t\\
&=\frac{1}{\Gamma\left(\lceil\alpha\rceil-\alpha\right)}\times\int_0^x\frac{n!\times t^{n-\lceil\alpha\rceil}}{\left(n-\lceil\alpha\rceil\right)!\times(x-t)^{\alpha-\lceil\alpha\rceil+1}}\,\mathrm{d}t\\
&=\frac{\Gamma(n+1)}{\left(n-\lceil\alpha\rceil\right)!\times\Gamma\left(\lceil\alpha\rceil-\alpha\right)}\times\int_0^x t^{n-\lceil\alpha\rceil}\times(x-t)^{\lceil\alpha\rceil-\alpha-1}\,\mathrm{d}t\\
\left(\text{on pose : }\rho=\frac{t}{x}\right)\qquad&=\frac{\Gamma(n+1)}{\left(n-\lceil\alpha\rceil\right)!\times\Gamma\left(\lceil\alpha\rceil-\alpha\right)}\times\int_0^1(x.\rho)^{n-\lceil\alpha\rceil}\times\left(x-\rho.x\right)^{\lceil\alpha\rceil-\alpha-1}x\,\mathrm{d}\rho\\
&=\frac{\Gamma(n+1)\times x^{n-\alpha}}{\left(n-\lceil\alpha\rceil\right)!\times\Gamma\left(\lceil\alpha\rceil-\alpha\right)}\times\int_0^1\rho^{n-\lceil\alpha\rceil}\times\left(1-\rho\right)^{\lceil\alpha\rceil-\alpha-1}\,\mathrm{d}\rho\\
&=\frac{\Gamma(n+1)\times x^{n-\alpha}}{\left(n-\lceil\alpha\rceil\right)!\times\Gamma\left(\lceil\alpha\rceil-\alpha\right)}\times B\left(n-\lceil\alpha\rceil+1,\lceil\alpha\rceil-\alpha\right)\\
&=\frac{\Gamma(n+1)\times x^{n-\alpha}}{\cancel{\left(n-\lceil\alpha\rceil\right)!}\times\cancel{\Gamma\left(\lceil\alpha\rceil-\alpha\right)}}\times\frac{\cancel{\Gamma\left(\left(n-\lceil\alpha\rceil+1\right)\times\Gamma\left(\lceil\alpha\rceil-\alpha\right)\right)}}{\Gamma\left(n-\cancel{\lceil\alpha\rceil}+1+\cancel{\lceil\alpha\rceil}-\alpha\right)}\\
&=\boxed{\frac{\Gamma(n+1)}{\Gamma(n+1-\alpha)}\times x^{n-\alpha}}
\end{aligned}$$

donc :
$$\boxed{\left(\forall\alpha\in\widehat{\mathbb{R}}^+\right)\left(\forall x\in\mathbb{R}^+\right):\ \lceil\alpha\rceil<n\Rightarrow\left[{}_0D_t^\alpha\left(t^n\right)\right](x)=\frac{\Gamma(n+1)}{\Gamma(n+1-\alpha)}\times x^{n-\alpha}\,.}\quad(\bigstar)$$

Dans ($\bigstar$), si on affecte à α une valeur entière, alors on tombera sur la formule classique de la dérivation répétée d'un monôme, et qui est donnée par :

$$\left(\forall\alpha\in\widehat{\mathbb{R}}^+\right)\left(\forall x\in\mathbb{R}^+\right):\ \alpha\leqslant n\Rightarrow\left[{}_0D_t^\alpha\left(t^n\right)\right](x)=\frac{n!}{(n-\alpha)!}\times x^{n-\alpha}=\frac{\mathrm{d}^\alpha}{\mathrm{d}x^\alpha}x^n\,.$$

▷ ***La semi dérivée de la fonction identité*** **:** pour expliciter son expression on va poser dans ($\star$) $k=1$ et $\alpha=0.5\in\widehat{\mathbb{R}}^+$, et ainsi on obtiendra :

$$\frac{\mathrm{d}^{\frac{1}{2}}}{\mathrm{d}x^{\frac{1}{2}}}x:=\left[{}_0D_t^\alpha(t)\right](x)=\frac{\Gamma(1+1)}{\Gamma\left(1+1-\frac{1}{2}\right)}\times x^{1-\frac{1}{2}}=\frac{\Gamma(2)}{\Gamma\left(\frac{3}{2}\right)}\times x^{\frac{1}{2}}=\frac{\Gamma(2)}{\Gamma\left(1+\frac{1}{2}\right)}\times\sqrt{x}=\frac{\Gamma(2)}{\frac{1}{2}\cdot\Gamma\left(\frac{1}{2}\right)}\times\sqrt{x}=\frac{1}{\frac{\sqrt{\pi}}{2}}\times\sqrt{x}.$$

Pour garantir qu'elle s'agit bien de la ***"demi-dérivée"*** nous allons la réappliquer une seconde fois pour obtenir :

$$\frac{\mathrm{d}^{\frac{1}{2}}}{\mathrm{d}x^{\frac{1}{2}}}\left(\frac{\mathrm{d}^{\frac{1}{2}}}{\mathrm{d}x^{\frac{1}{2}}}x\right)=\frac{\mathrm{d}^{\frac{1}{2}}}{\mathrm{d}x^{\frac{1}{2}}}\frac{2x^{\frac{1}{2}}}{\sqrt{\pi}}=\frac{2}{\sqrt{\pi}}\times\frac{\Gamma(1+\frac{1}{2})}{\Gamma\left(\frac{1}{2}-\frac{1}{2}+1\right)}x^{\frac{1}{2}-\frac{1}{2}}=\frac{2}{\sqrt{\pi}}\times\frac{\Gamma\left(\frac{3}{2}\right)}{\Gamma(1)}x^{0}=\frac{2\frac{\sqrt{\pi}}{2}\times x^{0}}{\sqrt{\pi}}=1$$

$$\Rightarrow\frac{\mathrm{d}^{\frac{1}{2}}}{\mathrm{d}x^{\frac{1}{2}}}\left(\frac{\mathrm{d}^{\frac{1}{2}}}{\mathrm{d}x^{\frac{1}{2}}}x\right)=1=\frac{\mathrm{d}}{dx}x.$$

et qui est bien le résultat attendu.

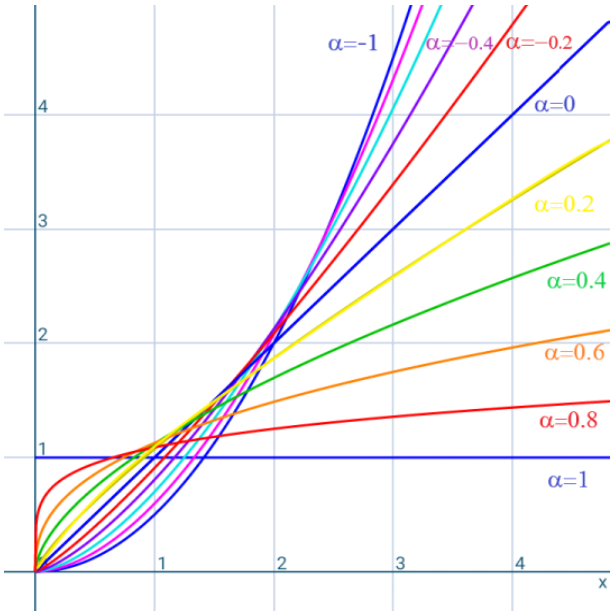


FIGURE IV.3: La dérivée fractionnaire de $x\mapsto x$ pour certains ordres $\alpha\in[-1,1]$.

Remarques 1.3.2 :

1. La $k^{\text{ème}}$ dérivée d'une fonction f ***est une propriété locale uniquement lorsque*** **k** ***est un nombre entier***, mais pour des ordres qui ne le sont pas ce n'est plus le cas, autrement dit,la dérivée fractionnaire d'une fonction f en un point x_0 pour des ordres non entières ne dépend pas de son comportement au voisinage de ce point, contrairement aux dérivées de puissance entière qui ont été fondées sur la notion du limite, et ceci rend bien évidemment ***l'approximation de ce type de dérivée par des formulations numériques une tâche compliquée et très difficile***[a].
2. la dérivée d'ordre fractionnaire d'une fonction nécessite sa connaissance sur tout l'intervalle d'étude, alors que dans le cas entier, seule la connaissance locale de f autour d'un point est nécessaire,cette propriété s'interprète par le fait que les systèmes d'ordres fractionnaires sont des systèmes à mémoire longue, tandis que les systèmes entiers sont vus comme des systèmes à mémoire courte.
3. dans ce qui précède, on a présenté la définition la plus facile ou la plus naturelle pour la dérivation fractionnaire et qui est tout simplement celle de Caputo, mais dans la littérature beaucoup d'autre choix sont proposés (voir [17]), par exemple : la dérivée fractionnaire de Riesz, de Liouville, de Miller,...etc.

a. Malgré sa difficulté une telle tâche n'est pas complètement impossible, et vous pouvez trouver dans la littérature quelques shémas qui le font, voir par exemple les pages : 27−28 de [17] (Schéma de Grünwald-Letnikov et schéma G^{α}).

4. la proposition IV.1.3.1 se traduit par le fait que l'opérateur de dérivation de Caputo est toujours continue à gauche des ordres entiers, en revanche, sa continuité à droite d'un $n \in \mathbb{N}$ est conditionnée par la réalisation de la propriété $f^{(n-1)}(a) = 0$, ce qui rend cet opérateur un peu moins avantageux que celui de Riemann-Liouville,(voir Remarque IV.1.3.1), et qui est continue dans les deux côtés de n'importe quel élement de $\mathbb{N}$ sans aucune exigence supplémentaire. [a].

5. par contre à l'opérateur de différentiation fractionnaire de Riemann-Liouville et qui généralise automatiquement la dérivation usuelle pour les ordres entiers, l'opérateur de Caputo nécessite toujours une définition par parties pour le renforcer à s'adapter avec cette dérivation, et ceci provient essentiellement du fait que le premier est basé sur une différentiation après intégration, ce qui lui permet de franchir le problèmes des constantes supplémentaires, tandis que le deuxième est basé sur une intégration après différentiation et qui ne lui permet pas de surmonter ce problème.

2 Description et discrétisation du problème :

2.1 Description du problème :

Notre but principal dans la suite est d'utiliser le réseau neuronal multicouches feed-forward qu'on a étudié au deuxième chapitre, (voir :II.3.2.1 page : 45), pour trouver une approximation convenable, convaincante, ou au moins acceptable de la solution de l'équation différentielle fractionnaire suivante :

$$\boxed{\left(\mathcal{E}_\alpha\right)\begin{cases}\left[{}_aD_t^\alpha\left(V(t)\right)\right] + h(t)\times V(t) = s(t). & \text{si : } a<t<b.\\ V(t) = V(a) = \beta \in \mathbb{R}. & \text{si : } t=a.\end{cases}} \tag{IV.1}$$

dans cette dernière équation, $V(\cdot) \in C^1\left([a,b]\right)$ joue le rôle de ***la fonction inconnue*** qu'on cherche à trouver, tandis que $h(\cdot)$ et $s(\cdot)$ sont des fonctions continues à valeurs réelles, fixées et connues préalablement, c'est à dire qu'elles ***sont considérées comme des paramètres*** pour notre équation, en plus, et afin de simplifier le processus de calcul, on choisira $\alpha \in \left]0,1\right]$, et on considère l'intervalle d'étude comme $[a,b] = [0,T]$, et même s'il ne l'était pas on peut le rendre ainsi, car tout intervalle de $\mathbb{R}$ peut être transformé à cette forme via une correspondance affine simple.

Maintenant, ***l'idée qu'on propose consiste à remplacer*** $V(y)$ ***sur le domaine*** $\Omega = [0,T]$ ***par une architecture neuronale bien conçue***, et puis la modifier au fur et à mesure pour aboutir une bonne approximation de notre solution, et on signale que cette façon de procéder peut être facilement utilisée pour n'importe quelles relations fractionnaires, même si elles sont d'une nature un peu plus compliquée que (IV.1).

2.2 Discrétisation du problème :

La première étape dans l'approche itérative suggérée pour résoudre ce problème, est de le rendre d'abord plus adaptable à des reformulations discrètes, et ceci en combinant sa condition initiale avec la solution éventuelle qu'on lui propose [b], ce qui fait que dans le cadre de l'équa-

a. On laisse au lecteur le soin de prouver ce fait d'une façon détaillée, et pour quelques indications, on peut l'orienter vers [3] page 21.

tion $(\mathcal{E}_\alpha)$, cette fonction qu'on présente comme un bon candidat pour être la solution effective de notre problème, doit concilier son inspiration de l'architecture neuronale feed-forward avec la vérification de la condition initiale, et naturellement, ceci va nous conduire à la considération de la fonction : $\widetilde{V}(t)$ définie par :

$$\left(\forall\, t \in [0,T]\right) : \quad \widetilde{V}(t) = \beta + t \times N(t) = \beta + t \times \sum_{i=1}^{I} \omega_i^{(2)} \times \sigma\left(\omega_i^{(1)} \times t + b_i\right) \qquad (*)$$

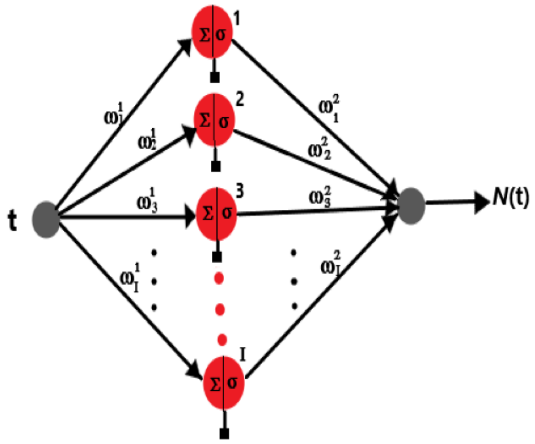


FIGURE IV.4: Le réseau feed-forward réalisant $N(\cdot)$.

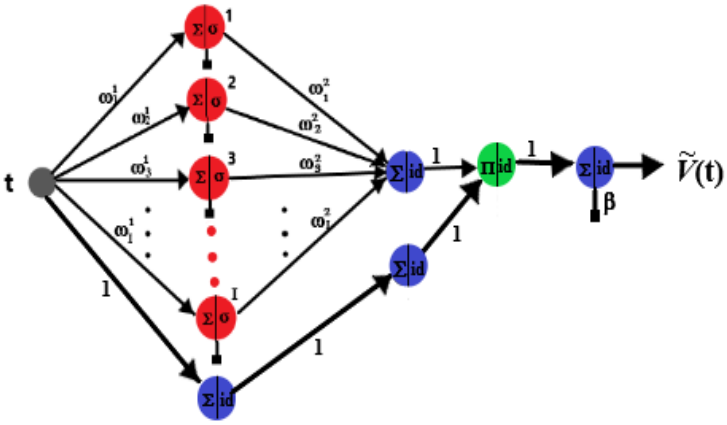


FIGURE IV.5: Le ΣΠ réseau réalisant $\widetilde{V}(\cdot)$.

Le choix de cette formulation peut être justifiée par le fait que $\widetilde{V}(\cdot)$ contient l'architecture d'un réseau de neurones feed-forward, et qui se reflète par $N(\cdot)$, et au même temps, elle satisfait la condition de départ : $\widetilde{V}(0) = \beta$.

En principe la fonction $\widetilde{V}(\cdot)$ présente une solution possible, et non forcement effective, qui fait intervenir les trois vecteurs de paramètres ajustables $\omega^1 \in \mathbb{R}^I$, $\omega^2 \in \mathbb{R}^I$ et $b \in \mathbb{R}^I$.

une substitution directe de $V(t)$ par $(*)$ dans l'équation (IV.1) transformera notre problème de sa forme d'origine à une autre plus simple ,plus pratique, non soumise à des contraintes ou des conditions de départ, et qui se présente ainsi :

$$\left(\forall\, t \in [0,T]\right) : \left[{}_0D_t^{\alpha}\left(t \times \sum_{i=1}^{I} \omega_i^{(2)} \times \sigma\left(\omega_i^{(1)} \times t + b_i\right)\right)\right] + h(t) \times \left(\beta + t \times \sum_{i=1}^{I} \omega_i^{(2)} \times \sigma\left(\omega_i^{(1)} \times t + b_i\right)\right) = s(t) \qquad (\blacktriangledown)$$

Généralement, un réseau neuronal ne peut pas prendre facilement la place de l'inconnue dans une équation différentielle fractionnaire, mais avec le type particulier de la fonction d'activation choisie ,et qui était sigmoïdale dans ce cas, ce fait devient justifiable par la propriété d'approximation universelle qu'on a traité au chapitre III, et selon laquelle l'espace des réseau neuronaux sigmoïdes est dense dans celui des fonctions continues, et dans ce sens, on profite de l'occasion pour insister sur l'importance pratique tant que théorique du type de la fonction d'activation choisie.

b. Le recours à cette idée revient essentiellement à l'absence de la notion de la dérivée fractionnaire à gauche ou à droite d'un point.

À ce stade, le problème principal qui se pose encore réside dans la grande difficulté de déterminer explicitement l'expression de la dérivée fractionnaire $D_t^{\alpha}\left(\widetilde{V}(t)\right)$, et spécialement lorsque cette fonction $\widetilde{V}(\cdot)$ est non polynomiale, ceci va nous pousser à chercher un moyen qui peut nous aider à franchir le calcul directe de cette quantité, et surtout à cause de la complexité pratique qu'il présente.

L'approche la plus classique qu'on peut proposer pour résoudre le problème mentionné ci-dessus, sera de substituer la fonction d'activation σ qu'on la suppose de classe C^{∞} dorénavant par ***son développement en séries entières***, et qui peut être obtenu de la formule de **Maclaurin** par exemple.

► Avant de continuer, on suppose d'abord que le rayon de convergence de la série entière [a]

$$\boxed{\Phi = \sum_{n\in\mathbb{N}} \frac{\sigma^{(n)}(0)}{n!} \times x^n}$$

est égale à $+\infty$, c'est à dire que :

$$R_{\Phi} := \sup\left\{|x| \;/\; x\in\mathbb{R} \text{ et : } \sum_{n\geqslant 0} \frac{\sigma^{(n)}(0)}{n!} \times x^n \text{ converge}\right\} = \sup\left\{|x| \;/\; x\in\mathbb{R} \text{ et : } \frac{\sigma^{(n)}(0)}{n!} \times x^n \text{ bornée}\right\} = +\infty\,.$$

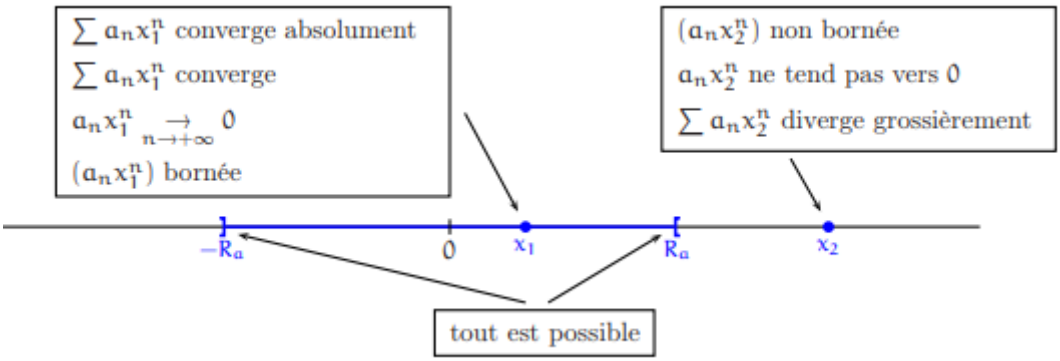


FIGURE IV.6: Le comportement d'une série entière de rayon de convergence R.

pratiquement, et pour garantir la réalisation de cette condition, il suffit de s'assurer que l'une des conditions suivantes est vérifiée :

- $(\forall x\in\mathbb{R})$ $\left(\frac{\sigma^{(n)}(0)}{n!} \times x^n\right)_{n\in\mathbb{N}}$ est bornée. (*critère d'Abel.*)
- $\limsup_{n\to+\infty} \sqrt[n]{\left|\frac{\sigma^{(n)}(0)}{n!}\right|} = 0\,.$ (*critère de Cauchy-Hadamard,*[26].)
- $\lim_{n\to+\infty} \left|\frac{\sigma^{(n+1)}(0)}{(n+1)!}\right| \times \left|\frac{n!}{\sigma^{(n)}(0)}\right| = \frac{1}{n+1} \times \left|\frac{\sigma^{(n+1)}(0)}{\sigma^{(n)}(0)}\right| = 0\,.$ (*critère d'Alembert.*)

a. C'est une série de fonctions réelles (ou complexes) φ_n de définies par : $(\forall z\in\mathbb{R})$: $\varphi_n(z) = a_n \times z^n$.
pour un z donné $\sum_{n\geqslant 0} a_n \times z^n$, ***n'est pas nécessairement convergente,*** et s'il était le cas on pose : $f(z) = \sum_{n=0}^{+\infty} a_n \times z^n$

le fait que $R_\Phi = +\infty$ va nous permettre d'attribuer un sens, à la fonction somme ϕ sur tout l'ensemble $\mathbb{R}$, et qui est effectivement le disque de convergence de Φ dans ce cas, et on écrit donc :

$$\boxed{\begin{array}{rccl} \phi & : &]-\infty, \overbrace{+\infty}^{=R_\Phi}[& \longrightarrow & \mathbb{R} \\ & & x & \longmapsto & \displaystyle\sum_{k=0}^{+\infty} \frac{\sigma^{(k)}(0)}{k!} \times x^k \end{array}} \quad (\diamondsuit)$$

Définition 2.2.1 :(fonction développable en séries entières :)

Soit $x_0 \in \mathbb{R}$ un réel, et $\rho \in \mathbb{R}^+ \cup \{+\infty\}$ un rayon strictement positif, on dit qu'une fonction g est développable en série entière en x_0 sur $]x_0-\rho, x_0+\rho[$, si et seulement si :

$$\boxed{\rho \leqslant R_S \quad \text{et} \quad \Big(\forall x \in]x_0-\rho, x_0+\rho[\Big) : \; g(x) = \sum_{n=1}^{+\infty} a_n \times (x-x_0)^n = g(x_0) + \sum_{n=1}^{+\infty} a_n \times (x-x_0)^n .}$$

et dans ce cas là cette identité s'appelle le développement en série entière de g en x_0 sur $]x_0-\rho, x_0+\rho[$.

Remarques 2.2.1 :

1. L'hypothèse $\rho \leqslant R_S$ où S est la série entière définie par :
$$S = \sum_{n \geqslant 0} a_n \times (x-x_0)^n$$
est nécessaire pour attribuer un sens à sa somme pour tout $x \in]x_0-\rho, x_0+\rho[$.
2. Une fonction g sera développable en série entière en x_0 sur $]x_0-\rho, x_0+\rho[$, si et seulement si sa translatée $\tau_{x_0}(g) : x \longmapsto g(x_0+x)$ l'est en 0 sur : $]-\rho, \rho[$, ce qui fait qu'on pourra toujours se ramener à des développements en série entière en 0.

Lemme 2.2.1 :

Soit $\sum_{n \geqslant 0} a_n \cdot x^n$ une série entière, de rayon de convergence R, et soit ρ un réel tel que $0 < \rho < R$, alors :

$$(\forall \varepsilon > 0)\,(\exists n_0 \in \mathbb{N}) \text{ tel que : } (\forall n \geqslant n_0) \quad \sup_{|x| \leqslant \rho} \left| \sum_{k=0}^{n} a_n \cdot x^n - \sum_{k=0}^{+\infty} a_n \cdot x^n \right| < \varepsilon .$$

c'est à dire qu'une série entière converge uniformément sur n'importe quel intervalle borné inclus dans son disque de convergence.

Preuve :

Fixons un ρ' dans $]\rho, R[$,pour tout $n \in \mathbb{N}$ on a :

$$|a_n \cdot x^n| \leqslant \underbrace{|a_n| (\rho')^n}_{\text{bornée}} \times \frac{\rho^n}{(\rho')} \leqslant M \left(\frac{\rho^n}{\rho'}\right)^n .$$

alors, pour tout réel $x \in]-\rho, \rho[$:

$$\left| \sum_{k=0}^{n} a_n \cdot z^n - \sum_{k=0}^{+\infty} a_n \cdot z^n \right| \leqslant \sum_{k=0}^{+\infty} M \left(\frac{\rho^n}{\rho'}\right)^n = \left(\frac{Mr}{r'-r}\right) \left(\frac{\rho}{\rho'}\right)^{n+1}$$

cette dernière majoration est indépendante de x, donc la convergence est bien uniforme, ce qui achève notre preuve ■

Théorème 2.2.1 :

Soit f une fonction développable en série entière au voisinage de 0, telle que :

$$\left(\forall x \in \left]-\rho,\rho\right[\right) : \quad f(x) = \sum_{n=0}^{+\infty} a_n \cdot x^n \qquad \left(\text{avec : } a_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}\right)$$

alors :

- la fonction f est indéfiniment dérivable sur $\left]-\rho,\rho\right[$, et sa dérivée sur cet intervalle n'est autre que la somme de la série des dérivées terme à terme, et on écrit :

$$\left(\forall x \in \left]-\rho,\rho\right[\right) : \ f'(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left(\sum_{n=0}^{+\infty} a_n \cdot x^n\right) = \sum_{n=0}^{+\infty} \frac{\mathrm{d}}{\mathrm{d}x}\left(a_n \cdot x^n\right) = \sum_{n=1}^{\infty} n \cdot a_n \cdot x^{n-1}$$

- la primitive de f sur $\left]-\rho,\rho\right[$ et qui s'annule en 0 n'est autre que la somme de la série intégrée terme à terme, et on écrit :

$$\left(\forall x \in \left]-\rho,\rho\right[\right) : \ \int_0^x f(t)\,\mathrm{d}t = \int_0^x \sum_{n=0}^{+\infty} a_n \cdot t^n \,\mathrm{d}t = \sum_{n=0}^{+\infty} \int_0^x a_n \cdot t^n \,\mathrm{d}t = \sum_{n=0}^{\infty} \frac{a_n}{n+1} \cdot x^{n+1}$$

Preuve :

- Pour montrer le premier résultat, on commence d'abord par remarquer que la série $\sum_{n\geqslant 0} a_n \cdot z^n$ et la série dérivée $\sum_{n\geqslant 0} n \cdot a_n \cdot z^{n-1}$ ont des rayons de convergence identiques, en effet :

$$\lim_{n\to\infty} \frac{n+1}{n} = 1 \ \text{ et } \ \lim_{n\to\infty} \frac{a_{n+1}}{a_n} = R \implies \lim_{n\to\infty} \frac{(n+1)\cdot a_{n+1}}{n \cdot a_n} = \text{"}1 \times R\text{"} = R$$

  le lemme 3.2.1 entraîne que la convergence est uniforme sur tout intervalle $\left[-\rho,\rho\right]$ inclus dans $\left]-R,R\right[$, et bien sûr si une suite de fonctions dérivables converge uniformément sur un intervalle, ainsi que la suite de ses dérivées, alors la limite de la suite sera dérivable à l'intérieur de cet intervalle, et sa dérivée ne va être que la limite des dérivées.
  ceci implique que la fonction f est dérivable sur tout intervalle $\left]-\rho,\rho\right[\subset \left]-R,R\right[$ donc sur $\left]-R,R\right[$ tout entier, et par récurrence, f sera donc indéfiniment dérivable sur $\left]-R,R\right[$.
- pour le second point du théorème il suffit d'appliquer le résultat de dérivabilité à la série primitive.

d'où ce qu'il faut prouver. ■

Définition 2.2.2 : (série de Maclaurin :)

Si f est une fonction indéfiniment dérivable sur un intervalle $\left]-R,R\right[$, alors on appelle ***série de Maclaurin*** associée à f en 0 la série entière définie par :

$$S_f = \sum_{n\geqslant 0} \frac{f^{(n)}(0)}{n!} \times x^n \,.$$

Corollaire 2.2.1 :

Si f est ***une fonction développable en série entière*** sur un intervalle $\left]-\rho,\rho\right[$, alors :

1. elle sera indéfiniment dérivable au voisinage de 0, et son développement en 0 sur cet intervalle est :

$$f(x) = \sum_{n=0}^{+\infty} \frac{f^{(n)}(0)}{n!} x^n = f(0) + f'(0)\,x + \cdots + \frac{f^{(n)}(0)}{n!} x^n + \cdots$$

2. elle admettra un développement limité en 0 à tout ordre, et dans ce cas là, et pour un certain ordre fixé : $n \in \mathbb{N}$ il sera la somme partielle d'ordre n de la série de Maclaurin de f :

$$f(x) = f(0) + f'(0) \cdot x + \cdots + \frac{f^{(n)}(0)}{n!} \cdot x^n + o(x^n)\,.$$

Preuve :

1. Posons :

$$\left(\forall x \in \left]-\rho, \rho\right[\right) : \quad f(x) = \sum_{n=0}^{+\infty} a_n \cdot x^n$$

par une application répétée du premier point de théorème IV.3.2.1, on déduit que pour tout $k \in \mathbb{N}$ la $k^{\text{ème}}$ dérivée de f est donnée par :

$$f^{(k)}(z) = \sum_{n=k}^{+\infty} n \times (n-1) \cdots \times (n-k+1) \times a_n \times z^{n-k}\,.$$

et dans ce cas, le premier terme est : $k! \times a_k = f^{(k)}(0)$, ce qui fait que $f \in C^{\infty}(\mathbb{R})$ et $a_k = \dfrac{f^{(k)}(0)}{k!}$.

2. Il suffit d'écrire le développement en séries entières correctement et ça ira de soi par la suite :

$$\left(\forall x \in \left]-\rho, \rho\right[\right) : f(x) = \sum_{k=0}^{+\infty} \frac{f^{(k)}(0)}{n!} \times x^n = f(0) + f'(0) \cdot x + \cdots + \frac{f^{(n)}(0)}{n!} \cdot x^n + \underbrace{\left(x^n \times \overbrace{\sum_{k=n+1}^{+\infty} \frac{f^{(k)}(0)}{n!} \cdot x^{k-n}}^{<+\infty}\right)}_{=o(x^n)}.$$

d'où le résultat désiré. ■

Remarque 2.2.2 :

la réciproque du deuxième point dans le dernier théorème est fausse en général, or, ***il se peut qu'une fonction f admette un développement limité à tous ordres au voisinage de*** 0***, sans qu'elle soit somme de sa série de Maclaurin***, un exemple classique de ce fait est la fonction définie par :

$$(\forall x \in \mathbb{R}) : \quad \varphi(x) = \begin{cases} \operatorname{sign}(x) \cdot \exp\left(-\frac{1}{x^2}\right) & \text{si } x \neq 0 \\ 0 & \text{sinon.} \end{cases}$$

pour tout : $n \geqslant 0$, on a :

$$\left|x^{-n} \times \varphi(x)\right| \leqslant |x|^{-n} \times \exp\left(-\frac{1}{x^2}\right) \xrightarrow[x \to 0]{} 0\,.$$

on en déduit donc que φ ainsi que toutes ses dérivées sont nulles en 0, et que le développement limité de f à tout ordre est nul aussi, pourtant f n'est pas du tout identiquement nulle [a].

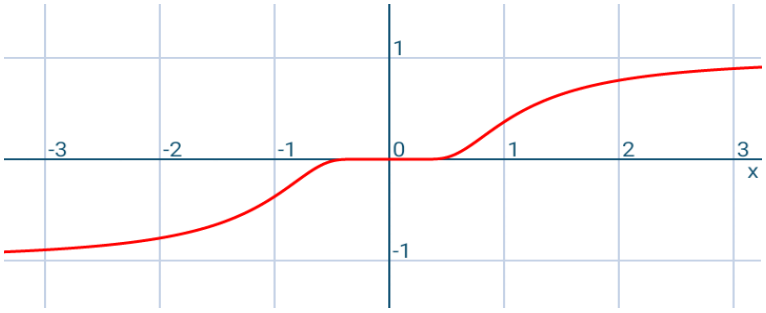


FIGURE IV.7: La courbe de la fonction :$x \mapsto \varphi(x)$.

a. Dans le cas où vous n'êtes pas bien convaincus, une deuxième lecture de cette remarque sera bien souhaitable.

Maintenant, et pour garder la véracité de la réciproque nous allons essayer de donner une condition suffisante pour laquelle une fonction indéfiniment dérivable sera forcement développable en série entière.
pour cela on fera appel à la formule classique de Taylor avec reste intégral.

Théorème 2.2.2 : (formule de Taylor avec reste intégrale :)

Soit f une fonction de classe C^{n+1} sur un intervalle $]-R,R[$ alors on a :

$$\left(\forall x\in\,]-R,R[\right):\quad f(x)=f(0)+f'(0)\cdot x+\cdots+\frac{f^{(n)}(0)}{n!}\cdot x^n+\int_0^x\frac{(x-t)^n}{n!}f^{(n+1)}(t)\,\mathrm{d}t\,.$$

Preuve :

On procédera par récurrence sur n, pour $n=0$, la formule est vraie car :

$$\left(\forall x\in\,]-R,R[\right):\quad f(x)=f(0)+\big(f(x)-f(0)\big)=f(0)+\int_0^x f'(t)\,\mathrm{d}t\,.$$

soit maintenant $n\geqslant 0$, on suppose que :

$$\left(\forall x\in\,]-R,R[\right):\quad f(x)=f(0)+f'(0)\cdot x+\cdots+\frac{f^{(n)}(0)}{n!}\cdot x^n+\int_0^x\frac{(x-t)^n}{n!}f^{(n+1)}(t)\,\mathrm{d}t$$

alors et pour tout $x\in\,]-R,R[$:

$$\begin{aligned}\int_0^x\frac{(x-t)^{n+1}}{(n+1)!}\times f^{(n+2)}(t)\,\mathrm{d}t&=\left[\frac{(x-t)^{n+1}}{(n+1)!}\times f^{(n+1)}(t)\right]_0^x-\int_0^x\frac{-(x-t)^n}{n!}\times f^{(n+1)}(t)\,\mathrm{d}t\\&=-\frac{f^{(n+1)}(0)}{(n+1)!}\times x^{n+1}+\int_0^x\frac{(x-t)^n}{n!}\times f^{(n+1)}(t)\,\mathrm{d}t\\\big(\text{et selon l'hypothèse de réccurence}\big)\quad&=-\frac{f^{(n+1)}(0)}{(n+1)!}\times x^{n+1}+f(x)-\left(f(0)+f'(0)\cdot x+\cdots+\frac{f^{(n)}(0)}{n!}\cdot x^n\right)\end{aligned}$$

donc :

$$\left(\forall x\in\,]-R,R[\right):\quad f(x)=f(0)+f'(0)\cdot x+\cdots+\frac{f^{(n+1)}(0)}{(n+1)!}\cdot x^n+\int_0^x\frac{(x-t)^{n+1}}{(n+1)!}f^{(n+2)}(t)\,\mathrm{d}t$$

d'où ce qu'il faut démontrer. ■

Corollaire 2.2.2 :

Si f est indéfiniment dérivable sur $]-R,R[$ et s'il existe $M>0$ et $a\in\,]0,+\infty[$ positifs tels que pour tout n et pour tout $x\in\,]-R,R[$,

$$|f^{(n)}(x)|\leqslant M\times a^n\,.$$

alors f est développable en série entière en 0 sur $]-R,R[$, et son développement sera bien sûr la somme de sa série de Maclaurin.

Preuve :

L'idée de la preuve consiste à utiliser la formule de Taylor avec reste intégral et de majorer ce reste comme suit :

$$(\forall x>0):\left|\int_0^x \frac{(x-t)^n}{n!}f^{(n+1)}(t)\,\mathrm{d}t\right|\leqslant\int_0^x (x-t)^n\frac{M\cdot a^n}{n!}\,\mathrm{d}t=\frac{x^{n+1}M}{(n+1)!}\times a^n\leqslant\underbrace{\frac{(R.a)^{n+1}}{(n+1)!}}_{\xrightarrow[n\to+\infty]{}0}\times\frac{M}{a}\xrightarrow[n\to+\infty]{}0\,.$$

et pour le cas de $x\leqslant 0$ un raisonnement analogue peut garantir que cet intégrale reste encore convergeant vers 0.
et ceci entraînera le résultat désiré [a]. ■

► Maintenant, et en restant encore dans le cadre de la reformulation de notre problème d'origine $(\mathcal{E}_a)$, on supposera pour la suite que :

$$\boxed{(\exists M>0)\left(\exists a\in\,]0,+\infty[\right)\left(\forall x\in\,]-\infty,+\infty[\right):\quad\left|\sigma^{(n)}(x)\right|\leqslant M\times a^n\,.}\qquad(\diamondsuit\diamondsuit)$$

on a et selon $(\diamondsuit)$ la fonction : $x\mapsto\sum_{k=0}^{+\infty}\frac{\sigma^{(k)}(0)}{k!}\times x^k$ est bien définie sur $\mathbb{R}$ toute entière.

et selon $(\diamondsuit\diamondsuit)$ et le corollaire : IV.2.2.2, σ sera développable en série entière en 0 sur $\mathbb{R}=\,]-\infty,\overbrace{+\infty}^{=R_\Phi}[$, et son développement ne va être que la somme de sa série de Maclaurin, donc on peut écrire dans ce cas :

$$\boxed{(\forall x\in\mathbb{R}):\quad\sigma(x)=\sum_{k=0}^{+\infty}\frac{\sigma^{(k)}(0)}{k!}\times x^k.}$$

alors, et en remplaçant σ par ce développement dans $(\blacktriangledown)$ on aura, et pour tout $t\in[0,T]$:

$$\boxed{\left[{}_0D_t^{\alpha}\left(t\times\sum_{i=1}^{I}\omega_i^{(2)}\times\sum_{k=0}^{+\infty}\frac{\sigma^{(k)}(0)}{k!}\times\left(\omega_i^{(1)}\times t+b_i\right)^k\right)\right]+h(t)\times\left(\beta+t\times\sum_{i=1}^{I}\omega_i^{(2)}\times\sum_{k=0}^{+\infty}\frac{\sigma^{(k)}(0)}{k!}\times\left(\omega_i^{(1)}\times t+b_i\right)^k\right)=s(t)}$$

or, et selon la formule de binôme de newton on a [b] :

$$\left(\omega_i^{(1)}\times t+b_i\right)^k=\sum_{j=0}^{k}C_k^j\cdot\left(\omega_i^{(1)}\right)^j\cdot t^j\cdot\left(b_i\right)^{n-j}.$$

d'où on obtient d'une part que :

$$\begin{aligned}t\times\sum_{i=1}^{I}\omega_i^{(2)}\times\sum_{k=0}^{+\infty}\frac{\sigma^{(k)}(0)}{k!}\times\left(\omega_i^{(1)}\times t+b_i\right)^k&=t\times\sum_{i=1}^{I}\omega_i^{(2)}\times\sum_{k=0}^{+\infty}\overbrace{\frac{\sigma^{(k)}(0)}{k!}}^{:=\lambda_k}\times\sum_{j=0}^{k}C_k^j\cdot\left(\omega_i^{(1)}\right)^j\cdot t^j\cdot\left(b_i\right)^{n-j}\\&=\sum_{i=1}^{I}\sum_{k=0}^{+\infty}\sum_{j=0}^{k}\lambda_k\cdot\omega_i^{(2)}\cdot C_k^j\cdot\left(\omega_i^{(1)}\right)^j\cdot t^{j+1}\cdot\left(b_i\right)^{n-j}\end{aligned}$$

a. On rappelle que : $(\forall a\in\mathbb{R}^{+*}):\ \frac{a^n}{n!}\xrightarrow[n\to+\infty]{}0$ (il suffit d'utiliser le critère d'Alembert pour les suites.).

b. Ici C_k^j désigne le nombre de combinaisons de k éléments parmi n et : $C_k^j=\frac{k!}{j!\times(k-j)!}$.

et d'une autre :

$$\left[{}_0D_t^{\alpha}\left(t\cdot\sum_{i=1}^{I}\omega_i^{(2)}\cdot\sum_{k=0}^{+\infty}\lambda_k\cdot\left(\omega_i^{(1)}\cdot t+b_i\right)^k\right)\right]=\left[{}_0D_t^{\alpha}\left(\sum_{i=1}^{I}\sum_{k=0}^{+\infty}\sum_{j=0}^{k}\lambda_k\cdot\omega_i^{(2)}\cdot C_k^j\cdot\left(\omega_i^{(1)}\right)^j\cdot t^{j+1}\cdot\left(b_i\right)^{n-j}\right)\right]$$

$$=\sum_{i=1}^{I}\sum_{k=0}^{+\infty}\sum_{j=0}^{k}\lambda_k\cdot\omega_i^{(2)}\cdot C_k^j\cdot\left(\omega_i^{(1)}\right)^j\cdot\left(b_i\right)^{n-j}\left[{}_0D_t^{\alpha}\left(t^{j+1}\right)\right]$$

$$\left(\text{de l'exemple : IV.1.3.1 et car : }\overbrace{\lceil\alpha\rceil=1}^{\alpha\in[0,1[}\leqslant n\right)=\sum_{i=1}^{I}\sum_{k=0}^{+\infty}\sum_{j=0}^{k}\lambda_k\cdot\omega_i^{(2)}\cdot C_k^j\cdot\left(\omega_i^{(1)}\right)^j\cdot\left(b_i\right)^{n-j}\cdot\left(\frac{\Gamma(j+2)}{\Gamma(j+2-\alpha)}\cdot t^{j+1-\alpha}\right)$$

donc d'après (▼), et en vertu de tout ce qui était présenté auparavant, on peut déduire que quel que soit $t\in[0,T]$ on a :

$$\boxed{\sum_{i=1}^{I}\sum_{k=0}^{+\infty}\sum_{j=0}^{k}\lambda_k.\omega_i^{(2)}C_k^j\left(\omega_i^{(1)}\right)^j\left(b_i\right)^{n-j}\left(\frac{\Gamma(j+2)}{\Gamma(j+2-\alpha)}t^{j+1-\alpha}\right)+h(t).\left(\beta+\sum_{i=1}^{I}\sum_{k=0}^{+\infty}\sum_{j=0}^{k}\lambda_k.\omega_i^{(2)}C_k^j\left(\omega_i^{(1)}\right)^j t^{j+1}\left(b_i\right)^{n-j}\right)=s(t)}$$

et par une simple factorisation on peut réécrire cette équation ainsi :

$$\boxed{\sum_{i=1}^{I}\sum_{k=0}^{+\infty}\sum_{j=0}^{k}\lambda_k.\omega_i^{(2)}C_k^j\left(\omega_i^{(1)}\right)^j\left(b_i\right)^{n-j}\left(\frac{\Gamma(j+2)}{\Gamma(j+2-\alpha)}t^{j+1-\alpha}+h(t).t^{j+1}\right)+\beta.h(t)=s(t)}$$

où $\lambda_k\in\mathbb{R}$ désigne ici : $\dfrac{\sigma^{(k)}(0)}{k!}$.

en notant : $\zeta_{i,j}^k=\lambda_k.\omega_i^{(2)}C_k^j\left(\omega_i^{(1)}\right)^j\left(b_i\right)^{n-j}$ on obtient :

$$\boxed{\sum_{i=1}^{I}\sum_{k=0}^{+\infty}\sum_{j=0}^{k}\zeta_{i,j}^k\times\left(\frac{\Gamma(j+2)}{\Gamma(j+2-\alpha)}t^{j+1-\alpha}+h(t).t^{j+1}\right)+\beta.h(t)=s(t)}\qquad(\blacktriangledown\blacktriangledown)$$

▷ Afin d'achever notre procédure de discrétisation, on va essayer bien évidemment de construire un maillage convenable pour notre intervalle d'étude $\Omega=[0,T]$, et pour cela, on prendra un entier positif $P\in\mathbb{N}^*$, et on adopte une partition régulière[a] et qui va se déterminer tout simplement via les points nodaux suivants :

$$\left(\forall m\in[\![0,P]\!]\right):\quad t_m=m\times\frac{T}{P}.$$

et dans ce cas, notre maillage sera clairement donné par la famille : $(\Omega_m)_{0\leqslant m\leqslant P-1}$ avec :

$$\left(\forall m\in\{0,\cdots,P-1\}\right):\ \Omega_m=[t_m,t_{m+1}].$$

on rappelle que les éléments Ω_m, et qui portent habituellement le nom des mailles, sont moins nombreux d'une seule unité que les point nodaux, voir pour cela la figure IV.8.

a. C'est à dire que toutes les mailles constituant cette partition sont de même amplitude.

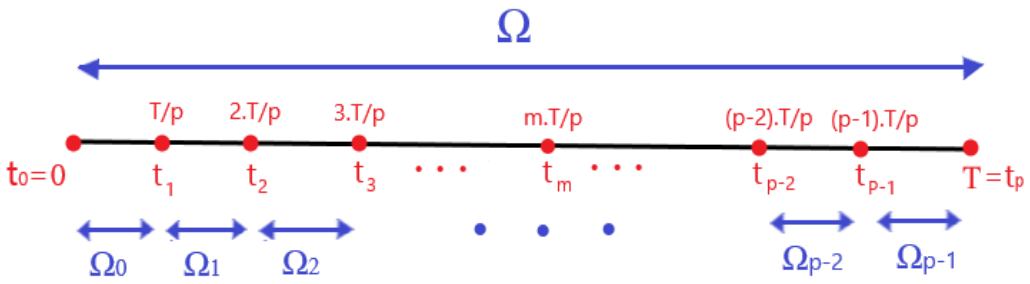


FIGURE IV.8: Maillage proposé du domaine d'étude $\Omega = [0, T]$.

à l'instar des procédures numériques usuelles, la substitution de $+\infty$ par un ordre de limitation [a] acceptablement grand, et l'affectation des valeurs nodales t_m à la variable t dans (▼▼) va ramener notre problème d'origine $(\mathcal{E}_\alpha)$ à sa forme finale, et qui se présente dans notre cas par le système des $P+1$ équations suivant :

$$\left(\forall m \in \{0,\cdots,P\}\right) : \quad \sum_{i=1}^{I}\sum_{k=0}^{N}\sum_{j=0}^{k} \zeta_{i,j}^{k} \times \left(\frac{\Gamma(j+2)}{\Gamma(j+2-\alpha)}\left(t_m\right)^{j+1-\alpha} + h\left(t_m\right).\left(t_m\right)^{j+1}\right) + \beta.h\left(t_m\right) = s\left(t_m\right) \qquad (\mathfrak{S})$$

avec : $\zeta_{i,j}^{k} = \lambda_k.\omega_i^{(2)} C_k^j \left(\omega_i^{(1)}\right)^j \left(b_i\right)^{n-j}$ et $\lambda_k = \dfrac{\sigma^{(k)}(0)}{k!}$.

en principe, notre but dans la suite sera de trouver les pondérations : $\widehat{\omega}_i^{(1)} \in \mathbb{R}^I$ et $\widehat{\omega}_i^{(2)} \in \mathbb{R}^I$ qui résolvent ce système [b].

3 Principe de résolution par un processus d'apprentissage :

3.1 Passage à un problème d'optimisation :

Dans ce qui précède, nous avons vu comment on a pris notre problème d'origine $(\mathcal{E}_\alpha)$, et on l'a ramené, après une longue procédure de discrétisation, à une forme plus simplifiée relativement, et qui s'exprime par le système d'équations $(\mathfrak{S})$, mais malgré cela, notre but de trouver une bonne approximation à la solution présumée de notre problème reste encore non achevé, et pour le faire il faut trouver comme on a déjà dit les vecteurs résolvant ce système, or et même au cas de leur existence, cette dernière tâche reste très difficile, si elle n'est pas impossible, à réaliser analytiquement, ce qui fait qu'on sera obligés de procéder numériquement à travers la générations d'une suite itérative de pondérations qui convergent vers les solutions : $\widehat{\omega}_i^{(1)} \in \mathbb{R}^I$ et $\widehat{\omega}_i^{(2)} \in \mathbb{R}^I$ qu'on les suppose existantes dans ce contexte.
maintenant, la question qui se pose encore, est comment déterminer la façon qu'on doit adopter pour effectuer cette génération, et surtout que le système $(\mathfrak{S})$ qu'on traite n'est pas linéaire, ou d'un autre type classique dont lequel on dispose déjà d'une gamme d'algorithmes de résolution prête à l'avance [c], et consacrée pour ce fait ?.
la réponse qu'on propose sera de réécrire notre système d'équations sous la forme d'un problème d'optimisation convenable, puis le résoudre par un algorithme classique comme celui de descente de gradient par exemple.

Dans la suite on va justifier d'abord l'existence d'une solution exacte pour notre problème d'origine $(\mathcal{E}_\alpha)$, et après on rentrera dans les détails du passage de $(\mathfrak{S})$ à un problème d'optimisation $(\mathbb{P})$ qu'on le déterminera ultérieurement, et on verra que l'application d'un algorithme de

a. On désigne ici l'ordre de limitation des séries de Maclaurin.
b. On signale que le problème d'existence de ces pondérations, est bien présent dans ce cas.
c. Comme l'algorithme de Jacobi ou de Gauss-Seidel pour les systèmes linéaires usuels (voir :[7]).

minimisation itérative dans le cadre de ce problème ***va nous conduire vers une bonne approximation de la solution exacte de*** $(\mathcal{E}_\alpha)$ ***et ceci même si ce problème*** $(\mathbb{P})$ ***n'admet pas de solutions.***

3.1.1 Existence d'une solution exacte :

Théorème 3.1.1 : (d'existence d'une solution pour un problème différentiel fractionnaire,[28] :)

Considérons le problème différentiel défini par :

$$(\mathcal{D}_\alpha)\begin{cases}\left[{}_0D_t^\alpha\left(V(t)+\Psi\left(t,V(t)\right)\right)\right]=\left\langle A\left(t,V(t)\right),V(t)\right\rangle+f\left(t,V(t)\right) & \text{si : } t>0.\\ V(0)=V_0\in\mathbb{R}. & \text{si : } t=0.\end{cases}$$

avec :

- $(X,\|\cdot\|)$ est un espace de Banach .
- $\left[{}_0D_t^\alpha(\cdot)\right]$ désigne la dérivée fractionnaire supérieure d'ordre $\alpha\in\left]0,1\right]$ au sens de Caputo.
- $f:J\times X\longrightarrow X$, $\Psi:J\times X\longrightarrow X$ sont des fonctions données où $J=\left[0,+\infty\right[$.
- $A(\cdot,\cdot):J\times X\longrightarrow\mathcal{B}(X)$ est une application qui associe à chaque couple $(t,x)\in J\times X$ un opérateur linéaire borné [a] $A(t,x)\in\mathcal{B}(X)$ définie de X dans X.

si les hypothèses suivantes sont vérifiées :

$(\mathbf{H_1})$ **:** la fonction $(t,x)\mapsto A(t,x)\in\mathcal{B}(X)$ est continue et il existe deux fonctions continues, bornées et non négatives $\varphi(\cdot)$ et $\psi(\cdot)$ définies sur $J=\mathbb{R}^+$ telles que :

$$\begin{cases}(\forall x\geqslant 0):\quad 0\leqslant\psi(x)\leqslant x+1.\\ \text{et :}\\ (\forall t\geqslant 0)\left(\forall x\in X\right):\quad \left\|A(t,x)\right\|_{\mathcal{B}(X)}\leqslant\dfrac{\varphi(t)}{1+t^{\alpha+1}}\times\psi\left(\dfrac{\|x\|}{1+t^{\alpha+1}}\right).\end{cases}$$

$(\mathbf{H_2})$ **:** la fonction $f:J\times X\longrightarrow X$ est continue, et il existe deux fonctions continues, bornées et non négatives $a(\cdot)$, $b(\cdot)$ définies sur J telles que :

$$(\forall t\geqslant 0)\left(\forall x\in X\right):\quad \left\|f(t,x)\right\|\leqslant\frac{a(t)}{1+t^{\alpha+1}}\|x\|+b(t).$$

$(\mathbf{H_3})$ **:** la fonction $\Psi:J\times X\longrightarrow X$ est continue et il existe une constante $\varsigma\geqslant 0$ telle que :

$$\begin{cases}(\forall t\geqslant 0)\left(\forall(u,v)\in X^2\right):\quad \left\|\Psi(t,u)-\Psi(t,v)\right\|\leqslant\varsigma\times\|u-v\|.\\ \text{et :}\\ \delta=\sup\limits_{t\geqslant 0}\left\|\Psi(t,0)\right\|<+\infty.\end{cases}$$

$(\mathbf{H_4})$ **:** il existe $\rho>0$ tel que :

$$\begin{cases}C_0(\rho)+C_1(\rho)<\rho\quad\text{où :}\\ C_0(\rho)=\|V_0\|+\left\|\Psi\left(0,V(0)\right)\right\|+\delta+\varsigma\times\rho.\\ C_1(\rho)=\dfrac{1}{\Gamma(\alpha+1)}\left(\|\varphi\|_\infty\rho^2+\|a\|_\infty\rho+\|b\|_\infty\right).\end{cases}$$

alors le problème $(\mathcal{D}_\alpha)$ ***admet au moins une solution.***

a. C'est à dire que c'est une application linéaire entre deux espaces vectoriels normés X et Y (dans ce cas $X=Y$) telle que l'image de la boule unité de X est une partie bornée de Y, on peut montrer facilement qu'ils s'identifient aux applications linéaires continues de X dans Y.

Preuve :

La démonstration de ce résultat se base essentiellement sur deux idées principales, la première est de mettre le système $\left(\mathcal{D}_\alpha\right)$ sous la forme d'un problème du point fixe, et la deuxième consiste à appliquer le fameux ***théorème du point fixe de Schauder***[a] pour garantir l'existence d'une solution à ce problème.

► ***se ramener à un problème du point fixe :*** supposons que $V(\cdot)$ est une solution du problème $\left(\mathcal{D}_\alpha\right)$ alors on a :

$$(\forall\, t>0):\quad \left[{}_0D_t^\alpha\left(V(t)+\Psi\left(t,V(t)\right)\right)\right]=\left\langle A\left(t,V(t)\right),V(t)\right\rangle+f\left(t,V(t)\right).$$

en appliquant l'opérateur d'intégration fractionnaire $\left[{}_0P_x^\alpha(\cdot)\right]$ aux deux membres de l'équation ci-dessus nous obtenons :

$$\begin{aligned}(\forall\, x>0):\quad V(x)+\Psi\left(x,V(x)\right)+\overbrace{\lambda}^{\text{constante}}&=\left[{}_0P_x^\alpha\left(\left\langle A\left(t,V(t)\right),V(t)\right\rangle+f\left(t,V(t)\right)\right)\right]\\&=\frac{1}{\Gamma(\alpha)}\int_0^x(x-t)^{\alpha-1}\times\left(\left\langle A\left(t,V(t)\right),V(t)\right\rangle+f\left(t,V(t)\right)\right)\mathrm{d}t.\end{aligned}$$

maintenant, et en utilisant la condition initiale : $V(0)=V_0$ on trouve :

$$\lambda=-V(0)-\Psi\left(0,V(0)\right)=-V_0-\Psi\left(0,V(0)\right)$$

ce qui fait que :

$$\begin{aligned}(\forall\, x>0):\quad V(x)+\Psi\left(x,V(x)\right)+\lambda&=V(x)+\Psi\left(x,V(x)\right)-V_0-\Psi\left(0,V(0)\right)\\&=\frac{1}{\Gamma(\alpha)}\int_0^x(x-t)^{\alpha-1}\times\left(\left\langle A\left(t,V(t)\right),V(t)\right\rangle+f\left(t,V(t)\right)\right)\mathrm{d}t.\end{aligned}$$

donc nous obtenons enfin de compte, et pour tout $x>0$ l'égalité suivante :

$$V(x)=-\Psi\left(x,V(x)\right)+V_0+\Psi\left(0,V(0)\right)+\frac{1}{\Gamma(\alpha)}\int_0^x(x-t)^{\alpha-1}.\left(\left\langle A\left(t,V(t)\right),V(t)\right\rangle+f\left(t,V(t)\right)\right)\mathrm{d}t$$

et qui s'appelle ***l'équation intégrale de Volterra,*** (Vito Volterra, italien, 1860 1940).

Réciproquement, si $V(\cdot)$ est une solution de l'équation de Volterra, alors on doit juste procéder inversement en appliquant l'opérateur différentiel $\left[{}_0D_t^\alpha(\cdot)\right]$ à ses deux membres, pour déduire que notre fonction $V(\cdot)$ sera forcement une solution de l'équation différentielle $\left(\mathcal{D}_\alpha\right)$ donc on peut dire qu'il y a une équivalence entre ce problème et l'équation de Volterra.

► ***utiliser le théorème du point fixe de Schauder :*** dans le but de démontrer l'existence d'une solution de l'équation intégrale de Volterra on introduit le sous-ensemble :

$$B_\rho:=\left\{V\in C^1\left(\mathbb{R}^+\right)/\,\|V\|_\infty\leqslant\rho\right\}.$$

avec $\rho>0$ est la constante positive introduite dans l'hypothèse : $(\mathbf{H_4})$.
on considère sur B_ρ l'opérateur $\mathcal{F}$ défini par :

$$\begin{aligned}\mathcal{F}\ :\ B_\rho&\longrightarrow C^1\left(\mathbb{R}^+\right)\\ g&\longmapsto \mathcal{F}(g)=V_0+\Psi\left(0,g(0)\right)-\Psi\left(\bullet,g(\bullet)\right)+\frac{1}{\Gamma(\alpha)}\int_0^\bullet(\bullet-t)^{\alpha-1}\times\Phi\left(t,g(t)\right)\,\mathrm{d}t\end{aligned}$$

a. Ce théorème a été présenté pour la première fois en 1930 par le mathématicien polonais **Juliusz Pawel Schauder**(1899-1943) dans son article[55], et il énonce que si C est un ***convexe fermé*** non vide d'un $\mathbb{R}-$ espace vectoriel topologique séparé $\left(E,+,\cdot,\tau\right)$, alors toute application L continue de C dans C telle que $L\left(C\right)$ est relativement compact, c'est à dire $\overline{L\left(C\right)}$ est compact, admettra forcement un point fixe.
l'importance de ce théorème (dans cette version) réside dans le fait qu'***il ne nécessite pas que l'espace*** E ***soit de Banach,*** (voir [9] page : 79).

avec :

$$\Phi\left(t,g(t)\right)=\left\langle A\left(t,g(t)\right),g(t)\right\rangle+f\left(t,g(t)\right)$$

sur cet ensemble, l'équation intégrale de Volterra peut être écrite sous la forme de problème du point fixe suivant :

$$\boxed{\text{trouver un } V\in B_\rho \text{ tel que : } \quad \mathcal{F}(V)=V.}$$

et qu'on va montrer le fait qu'il admet au moins une solution en s'appuyant sur le théorème du point fixe de Schauder dont on a parlé précédemment.
maintenant et pour appliquer ce théorème, et déduire l'existence d'une solution, il faut vérifier la réalisation des trois hypothèses suivantes :

i. B_ρ est un convexe fermé non vide.

ii. B_ρ est stable par $\mathcal{F}$, ou autrement dit : $\mathcal{F}\left(B_\rho\right)\subset B_\rho$.

iii. $\mathcal{F}$ est continue et $\mathcal{F}\left(B_\rho\right)$ est relativement compact.

— **La première condition :** il est clair que l'ensemble B_ρ est une boule fermée de centre 0 et de rayon ρ pour la norme de la convergence uniforme, donc elle sera clairement une partie non vide (car elle contiendra au moins la fonction nulle), convexe et fermée de $C^1\left(\mathbb{R}^+\right)$ muni de cette norme.

— **La deuxième condition :** d'abord notons pour tout $g\in B_\rho$ $\mathcal{F}(g)$ par $\left(\mathcal{F}g\right)$, alors dans ce cas on a :

$$\begin{aligned}
\left\|\frac{\left(\mathcal{F}g\right)(x)}{1+x^{\alpha+1}}\right\| &= \left\|\frac{V_0+\Psi\left(0,g(0)\right)-\Psi\left(x,g(x)\right)}{1+x^{\alpha+1}}+\frac{1}{\Gamma(\alpha)}\int_0^x\frac{(x-t)^{\alpha-1}}{1+x^{\alpha+1}}\times\Phi\left(t,g(t)\right)\,\mathrm{d}t\right\| \\
&\leqslant \frac{1}{1+x^{\alpha+1}}\left(\left\|V_0\right\|+\left\|\Psi\left(0,g(0)\right)\right\|+\left\|\Psi\left(x,g(x)\right)-\Psi\left(x,0\right)\right\|+\left\|\Psi\left(x,0\right)\right\|\right) \\
&\quad+\frac{1}{\Gamma(\alpha)}\int_0^x\frac{(x-t)^{\alpha-1}}{1+x^{\alpha+1}}\times\left\|\Phi\left(t,g(t)\right)\right\|\,\mathrm{d}t \\
\text{selon }\left(\mathbf{H_3}\right) \quad &\leqslant \frac{1}{1+x^{\alpha+1}}\left(\left\|V_0\right\|+\left\|\Psi\left(0,g(0)\right)\right\|+\varsigma\times\left\|g(x)-0\right\|+\delta\right) \\
&\quad+\frac{1}{\Gamma(\alpha)}\int_0^x\frac{(x-t)^{\alpha-1}}{1+x^{\alpha+1}}\times\left\|\left\langle A\left(t,g(t)\right),g(t)\right\rangle+f\left(t,g(t)\right)\right\|\,\mathrm{d}t \\
&\leqslant \frac{1}{1+x^{\alpha+1}}\left(\left\|V_0\right\|+\left\|\Psi\left(0,g(0)\right)\right\|+\varsigma\times\rho+\delta\right) \\
&\quad+\frac{1}{\Gamma(\alpha)}\int_0^x\frac{(x-t)^{\alpha-1}}{1+x^{\alpha+1}}\times\left(\left\|A\left(t,g(t)\right)\right\|_{\mathcal{B}(X)}.\left\|g(t)\right\|+\left\|f\left(t,g(t)\right)\right\|\right)\,\mathrm{d}t
\end{aligned}$$

Nous obtenons grâce aux hypothèses $\left(\mathbf{H_1}\right)$, $\left(\mathbf{H_2}\right)$ et $\left(\mathbf{H_3}\right)$ l'estimation suivante :

$$\left\|\frac{\left(\mathcal{F}g\right)(x)}{1+x^{\alpha+1}}\right\|\leqslant\frac{C_0\left(\rho\right)}{1+x^{\alpha+1}}+C_1\left(\rho\right).\frac{t^\alpha}{1+t^{\alpha+1}}.$$

d'où :

$$\lim_{t\to+\infty}\frac{\left(\mathcal{F}g\right)(x)}{1+x^{\alpha+1}}=0.$$

et selon l'hypothèse $\left(\mathbf{H_4}\right)$:

$$\left\|\left(\mathcal{F}g\right)\right\|_\infty\leqslant C_0\left(\rho\right)+C_1\left(\rho\right)\leqslant\rho.$$

ce qui prouve que : $\quad g\in B_\rho\Rightarrow\left(\mathcal{F}g\right)\in B_\rho$.

— **La troisième condition :** cette dernière condition est un peu difficile à démontrer, et surtout qu'elle se base à son tour sur un lemme un peu compliqué et qui donne des caractérisation de la compacité relative dans l'espace des fonction continues , pour plus de détails sur la démonstration de cette hypothèse veuillez voir [28] pages :27 – 32.

et ceci achève bien notre preuve . ■

Corollaire 3.1.1 :

Dans le cadre de l'équation différentielle fractionnaire suivante :

$$\boxed{\left(\mathcal{E}_\alpha\right)\begin{cases}\left[{}_0D_t^\alpha\left(V(t)\right)\right]+h(t)\times V(t)=s(t). & \text{si : } 0<t<T.\\ V(t)=V(0)=\beta\in\mathbb{R}. & \text{si : } t=0.\end{cases}}$$

et qui fait effectivement l'objet de notre étude dans ce chapitre, l'existence d'une solution est toujours garantie si la condition suivante est satisfaite :

$$\left(\left\|\varphi\right\|_\infty\cdot\left(\left\|s\right\|_\infty+\Gamma(\alpha+1).\left|\beta\right|\right)\right)<\frac{\Gamma(\alpha+1)}{4}$$

où φ ici désigne la fonctions définie sur $[0,T]$ par : $\varphi(t)=\left(1+t^{\alpha+1}\right)\times\left|h(t)\right|$

Preuve :

L'idée pour démontrer ce corollaire est de vérifier en premier lieu la réalisation des quatre hypothèses $(\mathbf{H_1})-(\mathbf{H_4})$ mentionnées au théorème IV.3.1.1, et puis appliquer ce dernier pour aboutir immédiatement le résultat désiré.
d'abord commençons par remarquer que le système fractionnaire $(\mathcal{E}_\alpha)$ n'est autre qu'un cas particulier de $(\mathcal{D}_\alpha)$ et pour lequel on a :

- $(X,\|\cdot\|)=(\mathbb{R},|\cdot|)$ et qui est bien un espace de Banach .
- $V_0=\beta\in\mathbb{R}$.
- $f:[0,T]\times\mathbb{R}\to\mathbb{R}$, $\Psi:[0,T]\times\mathbb{R}\to\mathbb{R}$ sont les fonctions définies par : $f(t,x)=s(t)$ et $\Psi(t,x)=0$.
- $A(\cdot,\cdot):[0,T]\times\mathbb{R}\longrightarrow\mathcal{B}(\mathbb{R})$ est l'application qui associe à chaque couple $(t,x)\in[0,T]\times\mathbb{R}$ l'application linéaire $A(t,x):=\left\langle A(t,x),\bullet\right\rangle=\left(-h(t)\times\bullet\right)\in\mathcal{B}(\mathbb{R})\cong\mathbb{R}$.

alors et dans ce cas on a :

$(\mathbf{H_1})$ **:** la fonction $(t,x)\mapsto A(t,x)=\left(-h(t)\times\bullet\right)\in\mathcal{B}(\mathbb{R})$ est continue car, et par hypothèse $h(\cdot)$ est continue, en plus et en considérant les fonctions définies sur $[0,T]$ par : $\varphi(t)=\left(1+t^{\alpha+1}\right)\times\left|h(t)\right|$ et $\psi(t)=\mathbb{1}$ on aura :

$$\begin{cases}\left(\forall x\in[0,T]\right): \quad 0\leqslant\psi(x)=1\leqslant x+1.\\ \text{et :}\\ \left(\forall t\in[0,T]\right)(\forall x\in\mathbb{R}): \quad \overbrace{\left\|A(t,x)\right\|_{\mathcal{B}(\mathbb{R})}}^{=|h(t)|}\leqslant\frac{\varphi(t)}{1+t^{\alpha+1}}\times\psi\left(\frac{\|x\|}{1+t^{\alpha+1}}\right)=\frac{\overbrace{\left|h(t)\right|.\left(1+t^{\alpha+1}\right)}^{=|h(t)|}}{1+t^{\alpha+1}}\times1.\end{cases}$$

$(\mathbf{H_2})$ **:** puisque : $s(\cdot)$ est continue la fonction $f:[0,T]\times\mathbb{R}\to\mathbb{R}$ le sera aussi, et en considérant les fonctions définies sur $[0,T]$ par : $a(\cdot)=0$ et $b(\cdot)=|s(\cdot)|$ on aura :

$$\left(\forall t\in[0,T]\right)(\forall x\in\mathbb{R}): \quad \underbrace{\left\|f(t,x)\right\|}_{=|s(t)|}\leqslant\frac{a(t)}{1+t^{\alpha+1}}\|x\|+b(t)=|s(t)|.$$

$(\mathbf{H_3})$: puisqu'elle est identiquement nulle, la fonction $\Psi : [0,T] \times \mathbb{R} \to \mathbb{R}$ est continue, et en posant $\varsigma = 0$ on aura facilement :

$$\begin{cases} \left(\forall\, t \in [0,T]\right)\left(\forall\,(u,v) \in \mathbb{R}^2\right) : \quad \left\|\Psi(t,u) - \Psi(t,v)\right\| \leqslant \varsigma \times \|u - v\|. \\ \text{et :} \\ \delta = \sup\limits_{0 \leqslant t \leqslant T} \left\|\Psi(t,0)\right\| = 0 < +\infty. \end{cases}$$

$(\mathbf{H_4})$: Dans ce cas particulier nous avons :

$$\begin{cases} C_0(\rho) = \overbrace{\|V_0\|}^{=|\beta|} + \overbrace{\left\|\Psi(0,V(0))\right\|}^{=0} + \overbrace{\delta}^{=0} + \overbrace{\varsigma \times \rho}^{=0} = |\beta|. \\ C_1(\rho) = \dfrac{1}{\Gamma(\alpha+1)}\left(\|\varphi\|_\infty \rho^2 + \underbrace{\|a\|_\infty}_{=0}\rho + \|b\|_\infty\right). \end{cases}$$

or, et d'une autre part on a :

$$\left(\|\varphi\|_\infty \cdot \left(\|s\|_\infty + \Gamma(\alpha+1).|\beta|\right)\right) < \frac{\Gamma(\alpha+1)}{4} \Rightarrow 1 - 4 \times \left(\frac{1}{\Gamma(\alpha+1)}\|\varphi\|_\infty \cdot \left(\frac{1}{\Gamma(\alpha+1)} \times \|s\|_\infty + |\beta|\right)\right) < 0$$

$$\Rightarrow \left(\exists x_0 > 0\right) : \frac{1}{\Gamma(\alpha+1)}\|\varphi\|_\infty x_0^2 - x_0 + \frac{1}{\Gamma(\alpha+1)} \times \|s\|_\infty + |\beta| = 0$$

ce qui fait qu'il existera forcement un $\rho > 0$ au voisinage de x_0 tel que :

$$C_0(\rho) + C_1(\rho) = |\beta| + \frac{1}{\Gamma(\alpha+1)}\left(\|\varphi\|_\infty \rho^2 + \|s\|_\infty\right) < \rho$$

donc et selon le théorème IV.3.1.1, l'équation $(\mathcal{E}_\alpha)$ admettra forcement une solution. ■

Remarques 3.1.1 :

1. Le corollaire précédent donne une condition suffisante pour l'existence d'une solution au problème $(\mathcal{E}_\alpha)$, mais faites attention elle n'est pas du tout nécessaire, ce qui fait qu'on peut bien trouver des cas où la solution existe sans que cette condition soit vérifiée.
2. malgré sa nécessité théorique, la question d'unicité de cette solution, n'est pas très inportante dans notre cadre d'étude, et surtout qu'on cherche juste à fournir une approximation numérique d'une fonction qui résout notre problème, mais si le lecteur est interessé par cette question, on peut lui suggérer d'aller voir le théorème 2.3.1 page 36 de la source [28] et qui se base sur six hypothèses pour garantir à la fois l'existence et l'unicité d'une solution au problème différentiel quasi-linéaire $(\mathcal{D}_\alpha)$ et même dans un cas plus général où la condition initiale est non locale [a].

3.1.2 Détails du passage de système $(\mathcal{S})$ à un problème d'optimisation :

Dans la section précédente on a remis notre problème d'origine $(\mathcal{E}_\alpha)$ sous sa forme discrétisée, et qui se présente comme on a déjà signalé par le système des $P+1$ équations suivant :

$$\left(\forall\, m \in \{0,\cdots,P\}\right) : \quad \sum_{i=1}^{I}\sum_{k=0}^{N}\sum_{j=0}^{k} \zeta_{i,j}^{k} \times \left(\frac{\Gamma(j+2)}{\Gamma(j+2-\alpha)}\left(t_m\right)^{j+1-\alpha} + h\left(t_m\right).\left(t_m\right)^{j+1}\right) + \beta.h\left(t_m\right) = s\left(t_m\right)$$

a. La condition non locale est une condition donnée sur la solution en un nombre fini ou infini d'instants, celle ci est présentée sous forme de la relation $V(0) = V_0 + g(V)$ liant $V(0)$ et d'autres valeurs de la solution,par exemple : $g(V) = \sum\limits_{k=0}^{p} c_i V(\tau_i)$ où $(c_i)_{1 \leqslant i \leqslant p}$ sont des constantes données et $0 < \tau_1 < \cdots < \tau_p < T$ sont des instants connus.

avec :

$$\zeta_{i,j}^{k}=\lambda_k.\omega_i^{(2)}C_k^j\left(\omega_i^{(1)}\right)^j\left(b_i\right)^{n-j}\text{ et }\lambda_k=\frac{\sigma^{(k)}(0)}{k!}.$$

la première chose à connaitre sur ce système, c'est ses inconnus, et qui sont dans ce cas les pondération $\omega^{(1)}\in\mathbb{R}^I$,$\omega^{(2)}\in\mathbb{R}^I$ et les biais $b\in\mathbb{R}^I$, la deuxième chose, est de garder à l'esprit que ce système n'admet pas toujours des solutions et même s'il est le cas l'unicité ne sera pas forcement garantie, la troisième et la dernière chose est de remarquer qu'il n'est pas linéaire, ce qui rend sa résolution d'un point de vu numérique une tâche un peu plus difficile et surtout avec l'absence d'un arsenal algorithmique consacré pour cela, ce fait va nous mettre sous l'obligation de ramener ce système à un problème d'optimisation de la façon suivante :
d'abord et afin d'alléger nos prochaines écritures, on adoptera et pour tout : $m\in[\![0,P]\!]$ la notation suivante :

$$\left(\forall\left(\omega^{(1)},\omega^{(2)},b\right)\in\left(\mathbb{R}^I\right)^3\right):\quad\Lambda_m\left(\omega^{(1)},\omega^{(2)},b\right)=\sum_{i=1}^{I}\sum_{k=0}^{N}\sum_{j=0}^{k}\zeta_{i,j}^{k}\times\left(\frac{\Gamma(j+2)}{\Gamma(j+2-\alpha)}\left(t_m\right)^{j+1-\alpha}+h\left(t_m\right).\left(t_m\right)^{j+1}\right)+\beta.h\left(t_m\right)$$

nous avons :

$$\begin{aligned}\left(\omega^{(1)},\omega^{(2)},b\right)\in\left(\mathbb{R}^I\right)^3\text{ est une solution de }(\mathfrak{S})&\iff\left(\forall m\in\{0,\cdots,P\}\right):\quad\Lambda_m\left(w^{(1)},w^{(2)},b\right)=s\left(t_m\right)\\&\iff\left(\forall m\in\{0,\cdots,P\}\right):\quad\Lambda_m\left(\omega^{(1)},\omega^{(2)},b\right)-s\left(t_m\right)=0\\&\iff\left(\forall m\in\{0,\cdots,P\}\right):\quad\left(\Lambda_m\left(\omega^{(1)},\omega^{(2)},b\right)-s\left(t_m\right)\right)^2=0\\&\iff\sum_{m=0}^{P}\left(\Lambda_m\left(\omega^{(1)},\omega^{(2)},b\right)-s\left(t_m\right)\right)^2=0\\&\Longrightarrow\left(\omega^{(1)},\omega^{(2)},b\right)\in\left(\mathbb{R}^I\right)^3\text{ est une solution du problème}\\&\qquad\qquad\text{d'optimisation : }(\mathbb{P})\end{aligned}$$

avec $(\mathbb{P})$ est donné par :

$$(\mathbb{P})\begin{cases}\min E\left(\omega^{(1)},\omega^{(2)},b\right)=\dfrac{1}{2}\times\displaystyle\sum_{m=0}^{P}\underbrace{\left(\Lambda_m\left(\omega^{(1)},\omega^{(2)},b\right)-s\left(t_m\right)\right)^2}_{\text{l'erreur locale : }E_m(\omega)}\\\left(\omega^{(1)},\omega^{(2)},b\right)\in\left(\mathbb{R}^I\right)^3\end{cases}$$

à ce stade, la résolution de ce problème, fera notre objectif pour la suite, et ceci car la détermination de son ensemble des points optimaux fournit une très bonne localisation des pondérations résolvants $(\mathfrak{S})$ dans $(\mathbb{R}^I)^3$, or et comme on peut le remarquer facilement, l'inclusion, $\mathbb{S}_{\mathfrak{S}}\subset\mathbb{S}_{\mathbb{P}}$ est toujours vérifiée [a].

Remarque 3.1.2 :

On peut avoir la réciproque de la dernière étape du notre passage de $(\mathfrak{S})$ à $(\mathbb{P})$, mais en ajoutant l'hypothèse d'existence d'une solution à $(\mathfrak{S})$, c'est à dire que :

$$\begin{cases}\left(w^{(1)},w^{(2)},b\right)\in\left(\mathbb{R}^I\right)^3\text{ est une solution du problème }(\mathbb{P})\\(\mathfrak{S})\text{ admet une solution.}\end{cases}\Rightarrow\left(w^{(1)},w^{(2)},b\right)\text{ est une solution de }(\mathfrak{S})$$

a. ***Attention !***, l'autre inclusion est fausse en général, et on remarque que la dernière étape dans le passage de $(\mathfrak{S})$ à $(\mathbb{P})$ était juste en implication et non plus en équivalence comme les précédentes, et si vous avez encore des doutes, pensez par exemple à l'équation $(E):\ x^2+1=0$.

Maintenant, si notre problème d'origine $(\mathcal{E}_\alpha)$ vérifie certaines conditions qui le rendent admettant une solution exacte $V(\cdot)$, par exemple, ceux qu'on a traité au corollaire IV.3.1.1, alors, et à la lumière de ce qu'on a vu au troisième chapitre, il existera forcement un réseaux neuronal qui approche cette solution, ce qui fait qu'on trouvera sans doute des pondérations $\left(w^{(1)}, w^{(2)}\right) \in \left(\mathbb{R}^I\right)^2$ et des biais $b \in \mathbb{R}^I$ qui résolvent le système $(\mathfrak{S})$ mais bien sûr en commettant une certaine erreur $\varepsilon(t)$ dont la norme uniforme $\|\varepsilon\|_\infty$ est négligeable, ceci rendra logiquement ces vecteurs très proches de la valeur minimale de $E(\omega)$, et tout cela va nous permettre d'approcher ces pondération, en faisant l'appel à des algorithmes d'apprentissage convenables qui vont, et en vertu de ce qui précède, nous conduire vers une très bonne approximation de la solution exacte de $(\mathcal{E}_\alpha)$, et cela même si ce problème $(\mathbb{P})$ n'admet pas de solutions.

3.2 Algorithme d'apprentissage proposé :

Le processus d'apprentissage dans les réseaux de neurones artificiels consiste à modifier les paramètres du réseau avec une règle de correction d'erreur appropriée, et ceci dans le but de trouver une collection de poids et de biais convenable pour la résolution du problème qu'on traite.

Jusqu'à présent, plusieurs algorithmes d'apprentissage ont été construits pour réaliser ce processus, et ils se basent tous sur la règle du delta classique, et qui provient en principe de la méthode de descente de gradient.

Dans notre cas, et afin d'avoir plus de performance pratique, on adoptera la vision habituelle d'apprentissage en ligne et qui s'appuie en général sur l'optimisation de l'erreur après la présentation de chaque exemple de notre ensemble de données, ou plus clairement elle se base sur une minimisation progressive de l'erreur local $E_m(\omega)$ avec $m \in [\![0, P]\!]$, pour approcher les pondérations optimales de notre système [a].

Afin d'atteindre cet objectif l'algorithme de rétro-propagation de gradient avec un élan donné est choisi dans ce cas pour s'occuper de la modification quantitative et qualitative des paramètres $\omega_i^{(1)}$, $\omega_i^{(2)}$ et b_i qui apparaissent clairement dans la formulation du problème $(\mathbb{P})$, ces paramètres seront sélectionnés dans un premier temps d'une façon aléatoire ensuite, le signal d'entrée traversera notre réseau pour générer une sortie nette à valeur réelle, et qui sera comparée à celle souhaitée pour calculer en fin de compte l'erreur en sortie.

Maintenant, la fonction du coût $E_m(\cdot)$, et qui est heureusement différentiable dans ce cas, doit être minimisée dans la direction opposée à sa dérivée, or, et comme on l'a déjà dit, l'algorithme d'apprentissage supervisé choisi en l'occurrence est celui de rétro-propagation avec élan et qui modifie le paramètre $\omega_i^{(2)}$ comme suit :

$$\left(\forall i \in \{1, \cdots, I\}\right) : \quad \begin{cases} \omega^{(2)}(\tau+1) = \omega_i^{(2)}(\tau) + \Delta\,\omega_i^{(2)}(\tau) \\ \Delta\,\omega_i^{(2)}(\tau) = -\eta \times \dfrac{\partial E_m}{\partial \omega_i^{(2)}} + \gamma \times \Delta\,\omega_i^{(2)}(\tau-1) \end{cases} \qquad (\bigstar)$$

dans ce qui précède, l'indice τ désigne l'ordre de l'itération courante, tandis que l'indice i est l'étiquette de la connexion du poids, de plus, $\omega_i^{(2)}(\tau+1)$ et $\omega_i^{(2)}(\tau)$ décrivent respectivement les paramètres de poids ajusté et actuel.
pour les deux paramètres η et γ, ils sont respectivement le taux d'apprentissage et le terme d'élan, et qui contrôlent en réalité la vitesse de convergence de cet algorithme.Afin d'accomplir la procé-

a. De l'autre part il y a ce qu'on appelle un apprentissage en batch ou aussi off ligne et qui consiste à minimiser l'erreur après la présentation de tout les exemples de la base, ou plus clairement il consiste à minimiser l'erreur global $E(\omega) = \sum_{m=0}^{p} E_m(\omega)$.

dure d'apprentissage, la dérivée partielle ci-dessus s'exprime ainsi :

$$\begin{aligned}
\frac{\partial E_m}{\partial \omega_i^{(2)}}\left(\omega^{(1)},\omega^{(2)},b\right) &= \frac{\partial E_m}{\partial \Lambda_m}\left(\omega^{(1)},\omega^{(2)},b\right) \times \frac{\partial \Lambda_m}{\partial \omega_i^{(2)}}\left(\omega^{(1)},\omega^{(2)},b\right) \\
&= \left(\Lambda_m\left(\omega^{(1)},\omega^{(2)},b\right) - s\left(t_m\right)\right) \\
&\quad \times \left[\sum_{k=0}^{N}\sum_{j=0}^{k} \lambda_k . C_k^j \left(\omega_i^{(1)}\right)^j \left(b_i\right)^{n-j} \times \left(\frac{\Gamma(j+2)}{\Gamma(j+2-\alpha)}\left(t_m\right)^{j+1-\alpha} + h\left(t_m\right).\left(t_m\right)^{j+1}\right)\right] \\
&= \left(\sum_{i=1}^{I}\sum_{k=0}^{N}\sum_{j=0}^{k} \zeta_{i,j}^k \times \left(\frac{\Gamma(j+2)}{\Gamma(j+2-\alpha)}\left(t_m\right)^{j+1-\alpha} + h\left(t_m\right).\left(t_m\right)^{j+1}\right) + \beta.h\left(t_m\right) - s\left(t_m\right)\right) \\
&\quad \times \left[\sum_{k=0}^{N}\sum_{j=0}^{k} \lambda_k . C_k^j \left(\omega_i^{(1)}\right)^j \left(b_i\right)^{n-j} \times \left(\frac{\Gamma(j+2)}{\Gamma(j+2-\alpha)}\left(t_m\right)^{j+1-\alpha} + h\left(t_m\right).\left(t_m\right)^{j+1}\right)\right]
\end{aligned}$$

le réglage des paramètres $\omega_i^{(1)}$ et b_i se fait par une procédure similaire à (★), et le vecteur gradient dans ce cas se présent comme suit :

pour les $\omega_i^{(1)}$:

$$\begin{aligned}
\frac{\partial E_m}{\partial \omega_i^{(1)}}\left(\omega^{(1)},\omega^{(2)},b\right) &= \frac{\partial E_m}{\partial \Lambda_m}\left(\omega^{(1)},\omega^{(2)},b\right) \times \frac{\partial \Lambda_m}{\partial \omega_i^{(1)}}\left(\omega^{(1)},\omega^{(2)},b\right) \\
&= \left(\Lambda_m\left(\omega^{(1)},\omega^{(2)},b\right) - s\left(t_m\right)\right) \\
&\quad \times \left[\sum_{k=0}^{N}\sum_{j=0}^{k} \lambda_k . C_k^j . \omega_i^{(2)} . j . \left(\omega_i^{(1)}\right)^{j-1} \left(b_i\right)^{n-j} \times \left(\frac{\Gamma(j+2)}{\Gamma(j+2-\alpha)}\left(t_m\right)^{j+1-\alpha} + h\left(t_m\right).\left(t_m\right)^{j+1}\right)\right] \\
&= \left(\sum_{i=1}^{I}\sum_{k=0}^{N}\sum_{j=0}^{k} \zeta_{i,j}^k \times \left(\frac{\Gamma(j+2)}{\Gamma(j+2-\alpha)}\left(t_m\right)^{j+1-\alpha} + h\left(t_m\right).\left(t_m\right)^{j+1}\right) + \beta.h\left(t_m\right) - s\left(t_m\right)\right) \\
&\quad \times \left[\sum_{k=0}^{N}\sum_{j=0}^{k} \lambda_k . C_k^j . \omega_i^{(2)} . j . \left(\omega_i^{(1)}\right)^{j-1} \left(b_i\right)^{n-j} \times \left(\frac{\Gamma(j+2)}{\Gamma(j+2-\alpha)}\left(t_m\right)^{j+1-\alpha} + h\left(t_m\right).\left(t_m\right)^{j+1}\right)\right]
\end{aligned}$$

et pour les b_i :

$$\begin{aligned}
\frac{\partial E_m}{\partial b_i}\left(\omega^{(1)},\omega^{(2)},b\right) &= \frac{\partial E_m}{\partial \Lambda_m}\left(\omega^{(1)},\omega^{(2)},b\right) \times \frac{\partial \Lambda_m}{\partial b_i}\left(\omega^{(1)},\omega^{(2)},b\right) \\
&= \left(\Lambda_m\left(\omega^{(1)},\omega^{(2)},b\right) - s\left(t_m\right)\right) \\
&\quad \times \left[\sum_{k=0}^{N}\sum_{j=0}^{k} \lambda_k . C_k^j . \omega_i^{(2)} . \left(\omega_i^{(1)}\right)^j \left(n-j\right) . \left(b_i\right)^{n-j-1} \times \left(\frac{\Gamma(j+2)}{\Gamma(j+2-\alpha)}\left(t_m\right)^{j+1-\alpha} + h\left(t_m\right).\left(t_m\right)^{j+1}\right)\right] \\
&= \left(\sum_{i=1}^{I}\sum_{k=0}^{N}\sum_{j=0}^{k} \zeta_{i,j}^k \times \left(\frac{\Gamma(j+2)}{\Gamma(j+2-\alpha)}\left(t_m\right)^{j+1-\alpha} + h\left(t_m\right).\left(t_m\right)^{j+1}\right) + \beta.h\left(t_m\right) - s\left(t_m\right)\right) \\
&\quad \times \left[\sum_{k=0}^{N}\sum_{j=0}^{k} \lambda_k . C_k^j . \omega_i^{(2)} . \left(\omega_i^{(1)}\right)^j \left(n-j\right) . \left(b_i\right)^{n-j-1} \times \left(\frac{\Gamma(j+2)}{\Gamma(j+2-\alpha)}\left(t_m\right)^{j+1-\alpha} + h\left(t_m\right).\left(t_m\right)^{j+1}\right)\right]
\end{aligned}$$

3.3 Exemple illustratif :

Dans cette section, l'efficacité de la méthode itérative proposée précédemment sera examinée sur un problème de test concret, de plus, les résultats numériques obtenus seront présentés graphiquement à travers des courbes qui rendront l'interprétation des résultats une opération plus facile.

Au cours des prochaines simulations, et pour plus de facilité, nous utilisons les paramètres suivants et qui peuvent être modifiés ultérieurement pour des raisons comparatives, ou pour montrer l'effet d'un certain paramètre sur la qualité d'apprentissage :

1. Nombre de neurones cachés : $I = 6$.
2. La fonction d'activation (logistique dans ce cas [a]) : $\sigma(\bullet) = \text{Logsig}(\bullet) = \dfrac{\exp(\bullet)}{1+\exp(\bullet)}$.

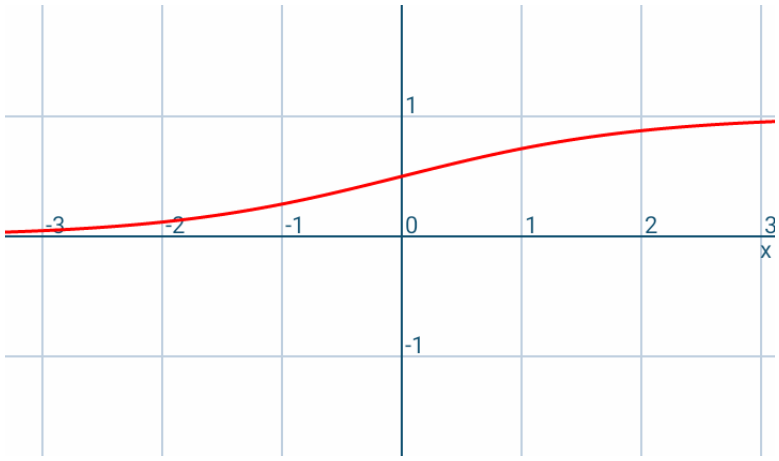


FIGURE IV.9: La courbe de la fonction "Logsig".

3. Le taux d'apprentissage (constant dans ce cas) : $\eta = 0.01$.
4. L'élan (constant dans ce cas) :$\gamma = 0.05$.
5. L'ordre de limitation pour les séries de Maclaurin : $N = 5$.
6. Nombre des points nodaux : $P = 11$.

►***Description du problème :*** soit $\alpha \in]0,1]$, considérons l'équation différentielle fractionnaire suivante :

$$\left(\mathcal{P}_\alpha\right)\begin{cases}\left[{}_0D_t^\alpha\left(V(t)\right)\right] + V(t) = t^2 + \dfrac{\Gamma(3)}{\Gamma(3-\alpha)} \times t^{2-\alpha} + 1. & \text{si : } 0 < t < 1.\\ V(t) = V(0) = 1. & \text{si : } t = 0.\end{cases}$$

d'abord nous pouvons remarquer que ce système n'est autre qu'un cas particulier, de $\left(\mathcal{E}_\alpha\right)$ et pour lequel on a :

a. Les fonctions logistiques sont en général de la forme : $f(t) = \dfrac{K}{1+a.e^{-kt}}$ où K, a et k sont des réels positifs quelconques.
ces fonctions sont des solutions en temps continu du modèle de **Verhulst** (belge,1804-1849) et qui se présente par :

$$y' = k.y\left(1-\frac{y}{K}\right) \text{avec : } k > 0 \text{ et : } K > 0$$

ce modèle et qui s'exprime en temps discret par : $u_{n+1} - u_n = a.u_n\left(1-\dfrac{u_n}{K}\right)$, imagine que le taux de natalité et le taux de mortalité d'une population donné sont des fonctions affines de la taille de la population, respectivement décroissante et croissante, autrement dit, plus la taille de la population augmente, plus son taux de natalité diminue et son taux de mortalité augmente, ce modèle est très utile et il a été proposé dans le cadre de la dynamique des populations pour remplacer celui de **Malthus** (britannique,1766-1834), et qui proposait un taux d'accroissement constant conduisant à une croissance exponentielle de la population.

— $\Omega = [0,1]$.

— $h(t) = \mathbb{1}$.

— $s(t) = t^2 + \dfrac{\Gamma(3)}{\Gamma(3-\alpha)} \times t^{2-\alpha} + 1$.

— $\beta = 1$.

et aussi il est bien facile de prédire sa solution exacte, et qui est définie dans ce cas par :

$$\boxed{\left(\forall\, t \in [0,1]\right) : \quad V(t) = t^2 + 1}$$

notre but maintenant c'est de vérifier si cette approche de résolution numérique par une procédure d'apprentissage peut nous fournir une bonne approximation de cette solution exacte.

►***Discrétisation du problème :*** Dans le cadre de ce problème, la procédure de discrétisation adoptée est celle qui était décrite dans la section IV.2.2 et dans laquelle la technique du maillage utilisée pour le domaine d'étude : $\Omega = [0,1]$ était basée sur la subdivision régulière décrite dans la figure IV.7.

►***Étapes de résolution par un processus d'apprentissage :*** Dans un premier temps, le processus d'apprentissage incrémentiel, et qui est basé dans ce cas sur un algorithme de rétro-propagation avec élan, commence par une affectation de petites constantes réelles aléatoires aux paramètres $\omega_i^{(1)}$, $\omega_i^{(2)}$ et b_i du réseau, ensuite, la règle de correction $(\star)$ est appliquée pour ajuster successivement les poids de connexion et les termes de biais jusqu'à l'obtention d'une solution appropriée.

Á ce stade et pour justifier la précision de cette technique, on exposera dans un premier temps les valeurs de l'erreur commise pour l'ordre $\alpha = 0.25$, et qui seront présentés dans le tableau IV.1, après, et vers la fin de cette section on fera du même pour d'autres ordres fractionnaires et qui seront également présentés pour un nombre de neurones cachés $N = 6$ et un nombre maximal d'itérations $\tau_{\max} = 500$ dans un autre tableau IV.3, puis, et pour une bonne visualisation de ces résultats, la fonction d'erreur correspondante à ces tableaux sera tracée, et pour des différents nombres d'itérations à la figure IV.10, en plus, les solutions exactes et approximatives seront à leurs tour illustrées sur la figure IV.11.
Les erreurs absolues entre les solutions exactes et approximatives sur les points nodaux vont être lues à l'aide de la fonction de l'erreur locale absolue :

$$\boxed{E_{\text{abs}}(m) = \left| V\left(t_m\right) - \widetilde{V}\left(t_m\right) \right|}$$

et qu'on la présentera pour $\tau_{\max} = 100$ dans la figure IV.12.
Enfin, l'efficacité de la structure neuronale proposée sera étudiée, et pour les différents éléments de contrôle (L'ordre de limitation N,et le nombre des neurones I.) à travers la fonction de l'erreur moyenne absolue E_{mid} définie par :

$$\boxed{E_{\text{mid}} = \sum_{m=0}^{P} \left| V\left(t_m\right) - \widetilde{V}\left(t_m\right) \right| = \sum_{m=0}^{P} E_{\text{abs}}(m)}$$

et qui va être présentée par le diagramme :IV.13 (pour $\tau_{\max} = 100$).

	α=0.25		
t_m	Solution exacte	Solution approchée	Erreur locale
0.1	1.0100	1.040000000090579	9.0579×10^{-11}
0.2	1.0400	1.010000000081472	8.1472×10^{-11}
0.3	1.0900	1.090000000012699	1.2698×10^{-11}
0.4	1.1600	1.160000000091338	9.1337×10^{-11}
0.5	1.2500	1.250000000063236	6.3235×10^{-11}
0.6	1.3600	1.360000000009754	9.7540×10^{-11}
0.7	1.4900	1.490000000027850	2.7849×10^{-11}
0.8	1.6400	1.640000000054688	5.4688×10^{-11}
0.9	1.8100	1.810000000095751	9.5750×10^{-11}

TABLE IV.1: Les solutions exactes et approchées dans le cas où : $\alpha = 0.25$, $I = 6$ et $\tau_{\max} = 500$.

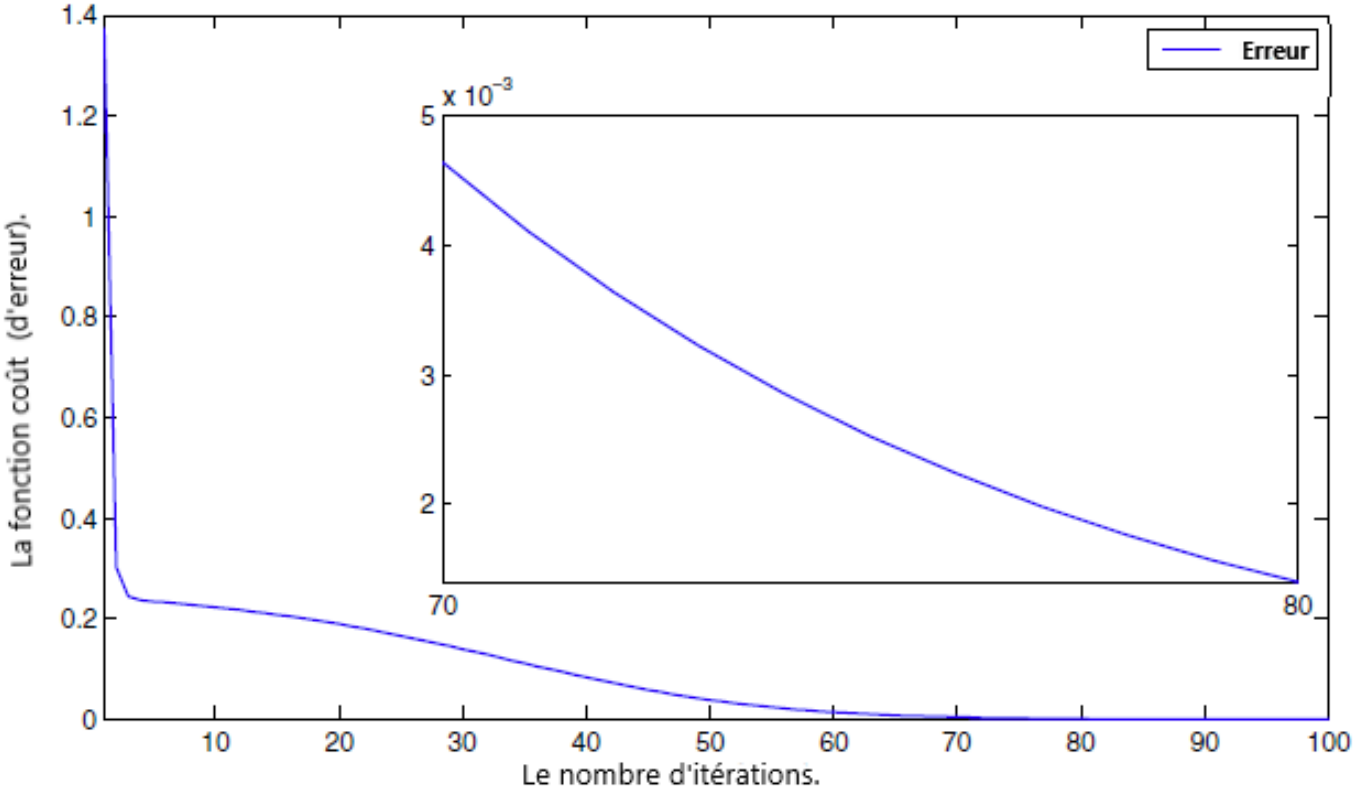


FIGURE IV.10: L'erreur moyenne en fonction du nombre d'itérations τ pour : $\tau \in \{0, \cdots, 100\}$.

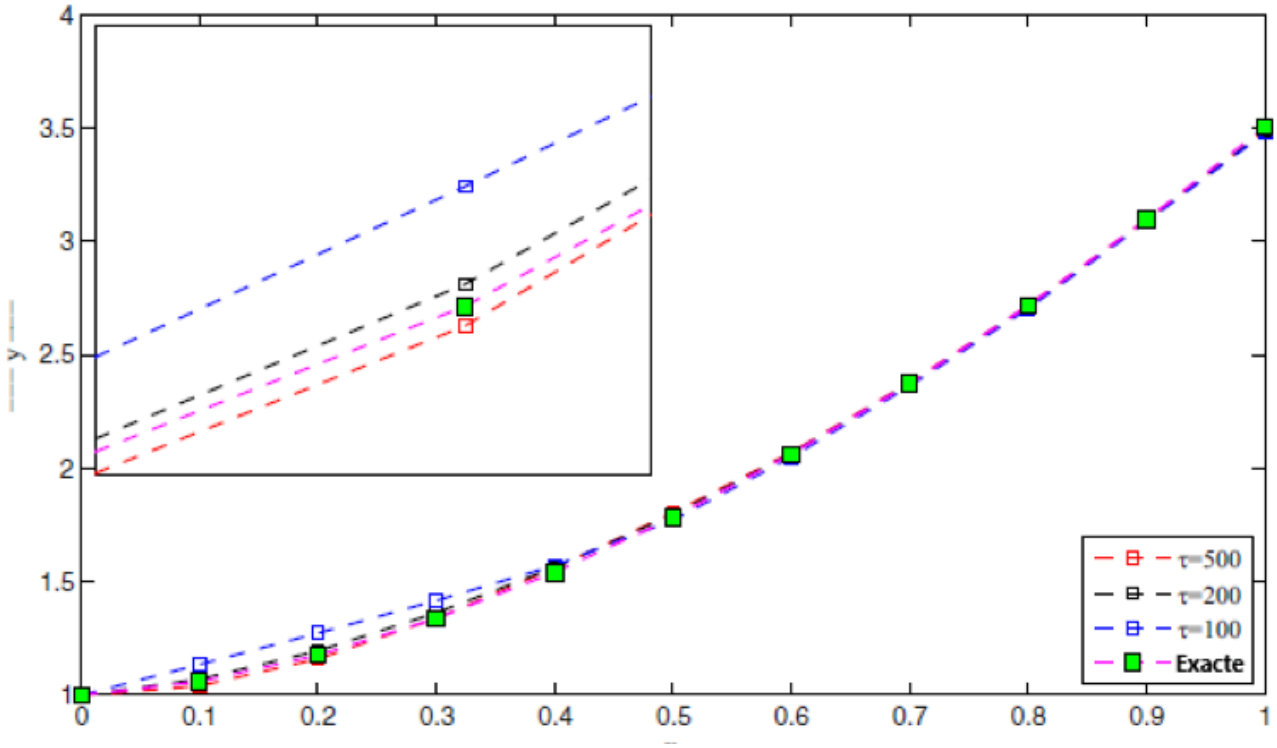


FIGURE IV.11: L'erreur moyenne en fonction du nombre d'itérations τ pour : $\tau = 100, 200$ et 500.

	Le nombre des neurones cachés.			
t_m	$I = 6$	$I = 8$	$I = 10$	$I = 12$
0.1	6.9483×10^{-11}	4.8976×10^{-14}	9.1900×10^{-15}	6.9908×10^{-17}
0.2	9.1710×10^{-12}	8.4559×10^{-13}	4.9836×10^{-16}	8.9090×10^{-17}
0.3	9.5022×10^{-12}	8.4631×10^{-13}	1.5974×10^{-16}	9.5929×10^{-16}
0.4	3.3445×10^{-11}	7.0936×10^{-14}	3.4039×10^{-16}	5.4722×10^{-17}
0.5	4.3874×10^{-11}	5.8527×10^{-16}	5.8527×10^{-16}	1.3862×10^{-17}
0.6	3.8156×10^{-11}	2.7603×10^{-14}	2.2381×10^{-16}	1.4929×10^{-17}
0.7	7.5952×10^{-11}	6.7970×10^{-14}	7.5127×10^{-16}	2.5751×10^{-17}
0.8	7.5920×10^{-11}	6.5510×10^{-14}	2.5510×10^{-16}	8.4072×10^{-17}
0.9	1.8687×10^{-11}	1.6261×10^{-14}	5.0596×10^{-16}	2.5428×10^{-17}

TABLE IV.2: Les valeurs de l'erreur locale absolue pour des différents nombres de neurones cachés.

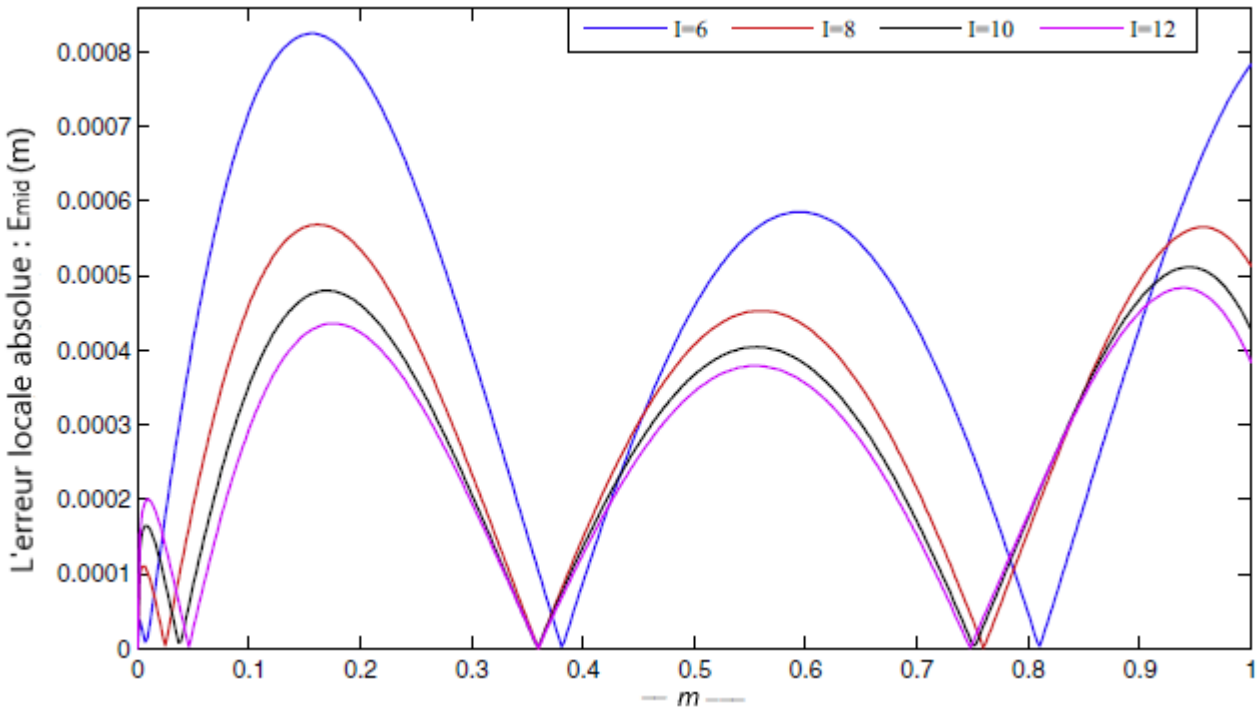


FIGURE IV.12: La courbe de l'erreur locale absolue pour des différents nombres de neurones cachés.

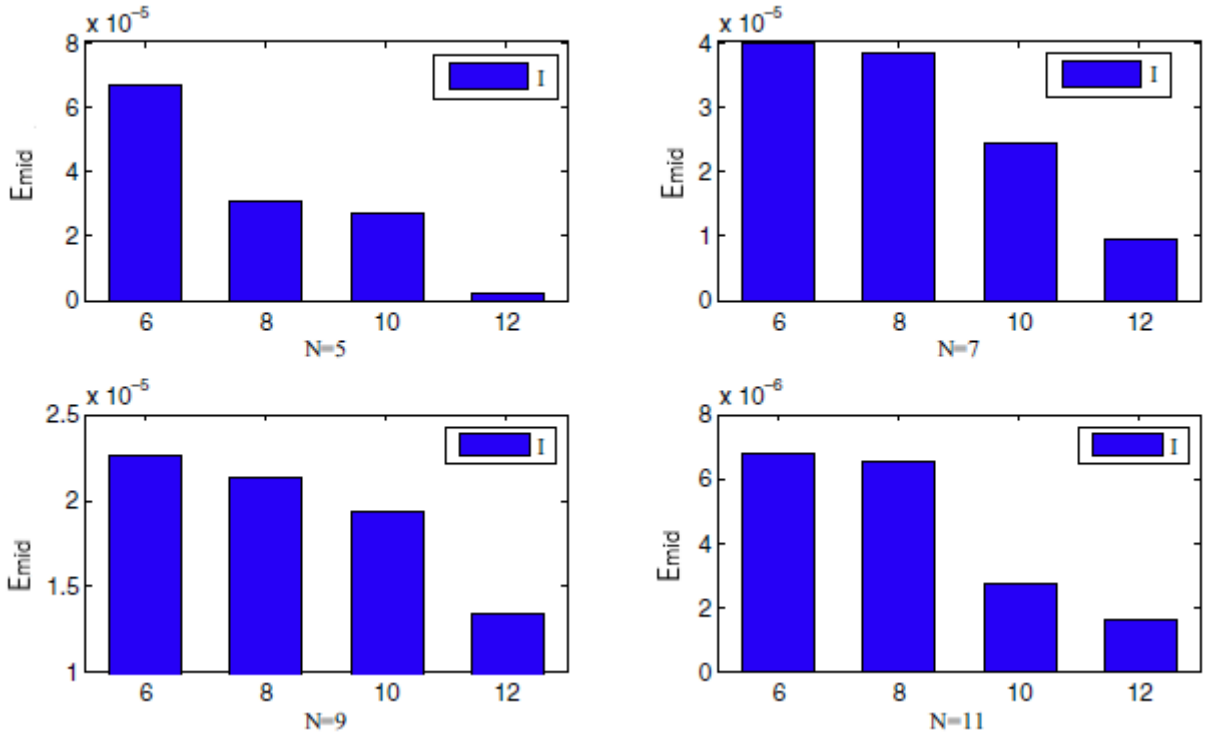


FIGURE IV.13: L'effet des paramètres N et I sur la qualité d'apprentissage.

	Les valeurs de α.		
t_m	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
0.1	9.6488×10^{-11}	4.5283×10^{-11}	2.5602×10^{-11}
0.2	9.5716×10^{-11}	3.5472×10^{-11}	1.4518×10^{-11}
0.3	8.0028×10^{-11}	3.0180×10^{-11}	9.2834×10^{-12}
0.4	8.8530×10^{-11}	9.2143×10^{-12}	2.7308×10^{-11}
0.5	5.4188×10^{-11}	5.7390×10^{-11}	8.9179×10^{-12}
0.6	4.2176×10^{-11}	5.8903×10^{-11}	1.7321×10^{-11}
0.7	7.3576×10^{-11}	6.4140×10^{-11}	3.8641×10^{-11}
0.8	7.1573×10^{-11}	6.3517×10^{-11}	2.4577×10^{-11}
0.9	8.7059×10^{-11}	7.3820×10^{-11}	3.8429×10^{-11}
	L'erreur locale absolue.		

TABLE IV.3: Les valeurs de l'erreur locale absolue pour des différents choix de α.

4 Quelques domaines d'applications des systèmes fractionnaires :

Les systèmes fractionnaires apparaissent de plus en plus dans les différents domaines de la recherche, toutefois, l'intérêt progressif que l'on porte à ces systèmes et surtout grâce à leurs applications qui apparaissent en sciences fondamentales et appliquées à la fois.
On peut noter que pour la majeure partie des domaines présentés ci dessous, les opérateurs fractionnaires sont utilisés pour prendre en compte des effets de mémoire .
Mentionnons à cet égard les ouvrages [51, 52] et qui regroupent diverses applications du calcul fractionnaire.

4.1 En Physique :

Une des applications les plus remarquables du calcul fractionnaire en physique était dans le contexte de la mécanique classique.
Le physicien Fred Riewe a montré en 1996 dans son article [49] que le Lagrangien contenant des dérivées temporelles d'ordres fractionnaires ***conduit à une modélisation pertinente du mouvement avec des forces non-conservatives*** comme les frottements àtitre d'exemple.
Ce résultat est remarquable du fait que les forces de frottement ou même les forces non-conservatives [a] en général sont essentielles dans le traitement variationnel macroscopique habituel, et par conséquent, dans les méthodes les plus avancées de la mécanique classique.
Riewe a généralisé le calcul des variations habituel au Lagrangien qui dépend des dérivées fractionnaires afin de s'adapter avec les forces non-conservatives habituelles.

4.2 En automatique :

En automatique, ce n'est qu'au début des années 1990 que le régulateur CRONE (commande robuste d'ordre non entier) a été proposé par Oustaloup [45].
En profitant des propriétés avantageuses des systèmes d'ordres fractionnaires, ce régulateur permet d'assurer la robustesse de la commande dans une bande de fréquence donnée, la réussite de

a. Par contre aux forces concervatives, le travail produit par ce type de forces dépend du chemin suivi par son point d'action.

cette approche fut énorme, et plusieurs variantes de cette commande ont vu le jour ($1^{ère}$, $2^{ème}$ et $3^{ème}$ générations).

4.3 En Acoustique :

Pour certains instruments de musique à vent, les pertes visco-thermiques peuvent être modélisées efficacement à l'aide de dérivées fractionnaires temporelles [30].

Conclusion :

Dans ce chapitre on a utilisé les réseaux de neurones artificiels, et qui se présentent comme des approximateurs universels, pour approcher la solution d'une équation différentielle fractionnaire à valeur initiale.
Pour arriver à cet objectif, une structure neuronale à trois couches avec une fonction d'activation sigmoïde a été utilisée lors de la modélisation de ce problème.
Le placement de cette structure neuronale dans l'équation différentielle concernée, le recours à la discrétisation, l'emploi du développement de Maclaurin pour franchir la complexité pratique du calcul de la dérivée fractionnaire pour la fonction d'activation, et l'utilisation de l'erreur quadratique moyenne, nous ont conduit à un problème d'optimisation relativement classique et qui sera traité numériquement par l'algorithme d'apprentissage de rétro-propagation, et qui est basé sur la règle de correction delta.
Enfin, un problème de test a été fait pour mieux illustrer la méthodologie proposée, les résultats numériques et les simulations informatiques obtenus mettent en évidence le fait que cette technique itérative a introduit vraiment une puissante amélioration sur la résolution des problèmes fractionnaires à valeurs initiales.

Conclusion générale et perspectives

Au cours de ce travail, nous avons traité le problème de la résolution numérique des équations différentielles fractionnaires ordinaires par des réseaux de neurones formels, et plus précisemment par des perceptrons multicouches.

En premier lieu on a commencé par un prérequis essentiel d'analyse fonctionnelle, puis on est passé à une étude générale de l'apprentissage automatique et du modèle des réseaux de neurones artificiels, ensuite on a utilisé ce prérequis pour justifier l'aptitude de ces derniers, et qui se considèrent pour un choix particulier de la fonction d'activation comme des approximateurs universels parcimonieux, enfin on a investi cette propriété fondamentale dans le cadre de l'étude numérique des équations différentielles fractionnaires et qui seront résolus dans ce contexte à l'aide d'un processus d'apprentissage, fondé du côté algorithmique sur la méthode de rétro-propagation de gradient.

Comme perspectives de notre travail, nous souhaitons d'abord avoir plus de choix et de possibilités pour la fonction d'activation [a], et qui peuvent garder la vérification de la propriété d'approximation parcimonieuse de notre réseau neuronal, ainsi on espère bien de généraliser la forme du problème différentiel qu'on traite ,et ceci en considérant par exemple un ordre de dérivation $\alpha \geqslant 1$, ou un domaine d'étude Ω non bornée, ou même des modèles non linéaires, ainsi la recherche d'une gamme algorithmique plus puissante sera vraiment désirable, enfin on signale que l'extension de cette méthode pour équations aux dérivée fractionnaire partielles reste encore une question très importante, et qui se pose jusqu'à l'instant dans le cadre de la recherche.

a. Récemment, quelques travaux de recherches ont fait vraiment du bon effort dans ce sens, et on cite à titre exemple l'article [25] de Vigor Ismailov et Namig Guliyev qui ont proposé en 2018 une construction algorithmique d'une fonction d'activation qui peut garantir la vérification de la propriété d'approximation, et ceci en faisant appel à la notion des fonctions $\lambda-$croissantes.

Annexe A

Prérequis essentiel pour la lecture du troisième chapitre :

Dans cette annexe, on met à la disposition du lecteur un rappel résumé de quelques théorèmes de topologie et d'analyse fonctionnelle, ainsi que leurs preuves et leurs corollaires, et ceci dans le but de stimuler et de renforcer son bagage et son pré-acquis mathématique.

La lecture de cette partie n'est pas du tout obligatoire ou indispensable, mais elle sera en revanche souhaitable, et surtout pour une bonne compréhension de quelques preuves dans lesquelles l'importance de ces théorèmes apparait clairement,(voir par exemple la démonstration du résultat : III.1.2.1).

1 Théorèmes du Noyau fermé et de Hahn-Banach :

1.1 Théorème du noyau fermé :

Théorème 1.1.1 :

Une forme linéaire sur un $\mathbb{R}$-espace vectoriel E est continue **si et seulement si** son noyau est fermé.

Preuve :

Il est clair que si $f \in E^*$ est continue $H = f^{-1}(\{0\})$ est fermé comme image réciproque d'un singleton.
Pour la réciproque considérons $a \notin H$ tel que : $f(a) = 1$, comme H est fermé il en est de même de :$\{a\} \times H$ et par suite de $a + H$,qui est l'image réciproque du fermé $\{a\} \times H$ par l'application continue $(x, y) \mapsto x + y$, et comme $0 \notin a + H$, il existe une boule $B(0, r[$ qui ne rencontre pas $a + H$, par suite, $x \in B(0, r[\Rightarrow f(x) \neq 1$.
montrons maintenant que : $x \in B(0, r[\Rightarrow \left|f(x)\right| \leqslant 1$, on suppose le contraire et on note : $f(x) = \alpha$ avec $x \in B(0, r[$ et $|\alpha| > 1$; alors $\left\|\frac{x}{\alpha}\right\| = \left(\frac{1}{|\alpha|}\right)\|x\| < r$ et $f\left(\frac{x}{\alpha}\right) = 1$ ce qui contredit la définition de $B(0, r[$. f étant bornée sur la boule $B(0, r[$ est bornée sur la boule unité et donc continue. ■

Remarque 1.1.1 :

Ce théorème est classique et très célèbre, et il nous servira dans la suite pour légitimer quelques passages au niveau de la démonstration du théorème universel d'approximation II.1.2.1.

1.2 Théorème de Hahn-Banach :

Lemme 1.2.1 :(de Zorn,[6] :)

Tout ensemble ordonné, ***inductif***[a] et non vide admet un élément maximal.

Théorème 1.2.1 :(Hahn-Banach Analytique,[7] :)

Soit $(E,+,.)$ un $\mathbb{R}$-espace vectoriel,et $p:E\to\mathbb{R}^+$ une semi norme sur cet espace,c'est à dire une application vérifiant : • $p(\lambda\cdot x)=\lambda\times p(x)\qquad(\forall x\in E),(\forall\lambda\geq 0)$.

• $p(x+y)\leq p(x)+p(y)\qquad\left(\forall(x,y)\in E^2\right)$.

-**Si** G est un sous-espace vectoriel de E, et $g:G\to\mathbb{R}$ est une forme linéaire telle que :

$(\forall x\in G):g(x)\leq p(x)$

alors il existe une forme linéaire $f:E\to\mathbb{R}$ qui prolonge g et qui vérifie :

$(\forall x\in E):f(x)\leq p(x)$.

Preuve :

On considère :

$\mathfrak{P}:=\left\{(S,h)\diagup \text{ avec } S \text{ sev de } E, h\in E^*, G\subset S, h \text{ prolonge } g \textbf{ et } \forall x\in D(h), h(x)\leq p(x)\right\}$[b].

On munit $\mathfrak{P}$ de la relation d'ordre :

$$(S_1,h_1)\leq(S_2,h_2)\iff S_1\subset S_2 \text{ et } h_2 \text{ prolonge } h_1.$$

► $\mathfrak{P}\neq\emptyset$ car $g\in\mathfrak{P}$.

►$\mathfrak{P}$ **est inductif** : soit $Q\subset P$ un sous-ensemble totalement ordonné, on note $Q=(S_i,h_i)_{i\in I}$, on définit $D:=\bigcup_{i\in I}S_i$ et $h(x):=h_i(x)$ si $x\in S_i$.

il est facile de voir que :(D,h) est un majorant de Q. Donc d'après le lemme de Zorn, $\mathfrak{P}$ possède un élément maximal (H,f). Prouvons que $H=E$:

Supposons que $H\neq E$ et soit $x_0\notin H$, on pose $F=H+\mathbb{R}x_0$ et :

$$\begin{array}{rccc} h & : & F=H\oplus\mathbb{R}x_0 & \longrightarrow & \mathbb{R} \\ & & x+\lambda.x_0 & \longmapsto & h(x)=f(x)+\lambda.\alpha \end{array}$$

pour un certain α, le but étant que $h\in P$. On doit donc avoir :

$$f(x)+t\alpha\leq p(x+tx_0)\qquad(\forall x\in H),(\forall t\in\mathbb{R}).$$

c'est à dire :

$$\begin{cases} f(x)+\alpha\leq p(x+x_0) & (\forall x\in H) \\ f(x)-\alpha\leq p(x-x_0) \end{cases}$$

par (i). Il suffit donc de choisir α tel que :

$$\sup_{y\in H}(f(y)-p(y-x_0))\leq\alpha\leq\inf_{x\in H}(p(x+x_0)-f(x))$$

Ceci est possible car :

$$\begin{aligned} f(x)+f(y)&\leq p(x+y)\\ &\leq p(x+x_0)+p(y-x_0)\end{aligned}$$

Donc $f(y)-p(y-x_0)\leq p(x+x_0)-f(x)\qquad\forall(x,y)\in H$.

Ainsi, $h\in P$ et $f\leq h, f\neq h$, on obtient une contradiction. ■

a. un ensemble ordonné, $(\xi,\preccurlyeq)$ est dit ***inductif*** si toutes ses ***chaînes*** (sous ensembles totalement ordonnées) admettent un élement maximal.

b. E^* désigne le ***dual algébrique*** de E , c'est à dire l'espace des formes linéaires sur E.

Corollaire 1.2.1 :

Soit $(E,+,\cdot,\|.\|)$ un espace vectoriel normé, G un sous-espace vectoriel de E. Soit $g : G \to \mathbb{R}$ linéaire et ***continue***. Alors il existe $f \in E'$ ***prolongeant*** g telle que $\|f\| = \|g\|$.

Preuve :

On pose $p(x) := g\|x\|$, les hypothèses du théorème sont bien vérifiées et on a $|f(x)| \leq \|g\|.\|x\|$, d'où $\|f\| \leq \|g\|$, donc $\|f\| = \|g\|$ car f prolonge g. ■

Corollaire 1.2.2 :

Soit $(E,+,\cdot,\|.\|)$ un espace vectoriel normé, Pour tout $x_0 \in E$, il existe $f_0 \in E'$ [a] tel que $\|f_0\| = \|x_0\|$ et $\langle f_0, x_0 \rangle = \|x_0\|^2$.

Preuve :

Appliquer le corollaire avec $G := \mathbb{R}x_0$ et $g(tx_0) := t\|x_0\|^2$. ■

Remarques 1.2.1 :

1. Pour l'application qu'on a choisi dans la preuve,on a :

$$\|g\| := \sup_{x \in E\setminus\{0\}} \frac{\|g(x)\|}{\|x\|} = \sup_{t \in \mathbb{R}^*} \frac{\left\|t\|x_0\|^2\right\|}{\|t\|x_0\|\|} = \|x_0\|$$

2. ce corollaire va jouer un rôle indispensable,dans la démonstration du théorème III.1.2.1 .

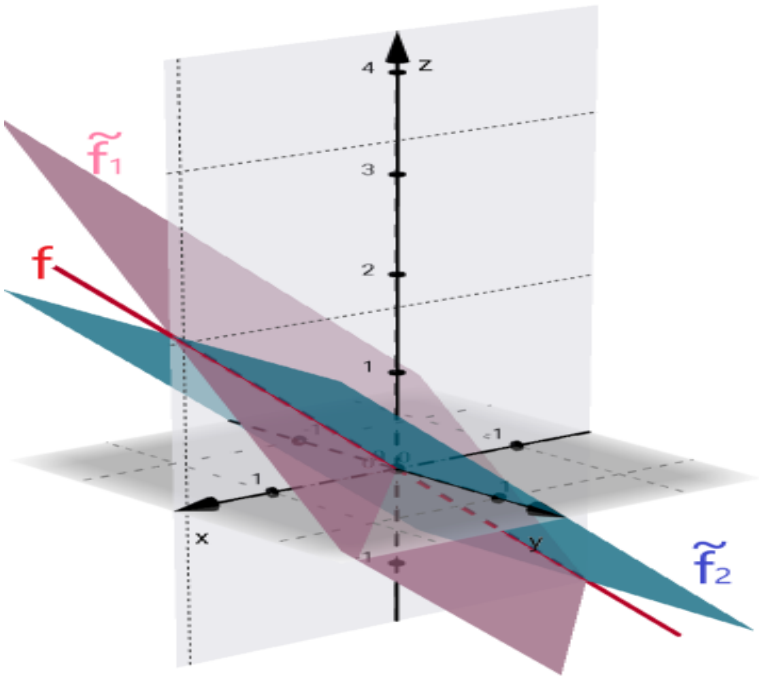


FIGURE A.1: L'idée du théorème de Hahn-Banach pour $E = \mathbb{R}^2$, $F = \mathbb{R} \times \{0\}$ et $f(x) = 2x$.

a. E' désigne le ***dual topologique*** de E , c'est à dire l'espace des formes linéaires ***continues*** sur E, juste comme remarque :il est facile de voir que $E' \subseteq E^*$.

2 Espaces Hilbertiens et théorème de représentation de Riesz :

2.1 Espaces Hilbertiens :

Définition 2.1.1 :(Espace de Hilbert) :

Soit $(E,+,\cdot,\langle\ ,\ \rangle)$ ***un espace préhilbertien*** réel,c'est à dire,un $\mathbb{R}$-espace vectoriel munit d'un ***produit scalaire***,[a]on dit que cet espace est de Hilbert ou Hilbertien s'il est complet pour la norme $\|.\|$ définie par : $(\forall x\in E):\ \|x\|=\sqrt{\langle x,x\rangle}$ [b] .

Exemples 2.1.1 :

1. Sans surprise,l'espace $E=\mathbb{R}^n$ munit du produit scalaire usuel définit par :

$$\boxed{\begin{array}{rcccc}\langle\cdot,\cdot\rangle & : & E\times E & \longrightarrow & \mathbb{R}\\ & & (x,y) & \longmapsto & \sum_{i=0}^{n}x_i.y_i\end{array}}$$

   est un espace de Hilbert,et en général :***tout espace euclidien*** [c] ***est de Hilbert***.

2. soit $(\Omega,\mathscr{F},\mu)$ un espace mesuré, on note :$\mathscr{L}^2(\Omega,\mathscr{F},\mu)$ l'espace des fonctions ***réelles*** de carré μ-intégrable ,c'est à dire :

$$\mathscr{L}^2(\Omega,\mathscr{F},\mu):=\left\{\underbrace{f\in\mathfrak{M}\left((\Omega,\mathscr{F});(\mathbb{R},\mathfrak{B}_{\mathbb{R}})\right)}_{\text{l'espace de fonctions }\mathscr{F}-\mathfrak{B}_{\mathbb{R}}\text{ mesurables}}\ \Big/\ \int_\Omega|f|^2\ \mathrm{d}\mu<+\infty\right\}$$

   -on définit ensuite l'espace $L^2(\Omega,\mathscr{F},\mu):=\mathscr{L}^2(\Omega,\mathscr{F},\mu)\Big/\Re$,où $\Re$ est la relation d'équivalence sur$\mathscr{L}^2(\Omega,\mathscr{F},\mu)$ définie par l'égalité μ-presque partout, cet espace quotient va nous permettre de corriger l'application :

$$\begin{array}{rcccc}\|\cdot\|_2 & : & \mathscr{L}^2(\Omega,\mathscr{F},\mu) & \longrightarrow & \mathbb{R}^+\\ & & f & \longmapsto & \left(\int_\Omega|f|^2\ \mathrm{d}\mu\right)^{\frac{1}{2}}\end{array}$$

   et qui est juste une ***semi-norme*** [d] sur $\mathscr{L}^2(\Omega,\mathscr{F},\mu)$,et la rendre une norme sur $L^2(\Omega,\mathscr{F},\mu)$. Maintenant,et sur l'espace $L^2(\Omega,\mathscr{F},\mu)$,on va définir un produit scalaire comme suit :

$$\boxed{\forall(f,g)\in\left(L^2(\Omega,\mathscr{F},\mu)\right)^2\quad\langle f,g\rangle:=\int_\Omega f.g\ \mathrm{d}\mu}$$

   il est facile de remarquer que : $\left(\forall f\in L^2(\Omega,\mathscr{F},\mu)\right):\ \|f\|_2=\sqrt{\langle f,f\rangle}$ donc et en se basant sur le fameux ***théorème de Riesz-Ficher*** [e] on peut déduire facilement que :
   $\left(L^2(\Omega,\mathscr{F},\mu),+,.,\langle.,.\rangle\right)$ est un espace de Hilbert.

a. ***un produit scalaire***(réel) est une forme bilinéaire réelle,symétrique,positive et définie positive.
b. cette norme s'appelle la ***norme Euclidienne*** ou la ***norme induite*** du produit scalaire de notre espace.
c. c'est à dire un $\mathbb{R}$-espace vectoriel préhilbertien ***de dimension fini.***
d. l'axiome de séparation n'est pas vérifiée.
e. ce'théorème affirme que pour tout $p\geq1$, $L^p(\Omega,\mathscr{F},\mu)$ est un espace de Banach pour la norme :
$$\|f\|_p:=\left(\int_\Omega|f|^p\ \mathrm{d}\mu\right)^{\frac{1}{p}}$$

Définition 2.1.2 :(l'orthogonalité) :

Soit $(E,+,\cdot,\langle\ ,\ \rangle)$ ***un espace préhilbertien réel*** :

1. On dit que deux vecteurs x et y de E sont ***orthogonaux*** $(\text{pour le produit scalaire}\langle\ ,\ \rangle)$, et on note $\mathbf{x}\perp\mathbf{y}$,si et seulement si : $\boxed{\langle x,y\rangle = 0}$.
2. **L'orthogonal d'une partie** $X \subset E$, noté $X^{\perp}$, est le sous ensemble de E constituée des vecteurs orthogonaux à tous les vecteurs de X,autrement dit :

$$X^{\perp} = \left\{ y \in F \mid \forall x \in X, \quad \langle x,y\rangle = 0 \right\}.$$

Proposition 2.1.1 :

Soit $(E,+,\cdot,\langle\ ,\ \rangle)$ ***un espace préhilbertien réel***,et soit $X \subset E$ une partie de E [a] alors on a :

1. $X^{\perp}$ est toujours ***un sous espace vectoriel fermé*** de E.
2. l'égalité $\mathbf{X}^{\perp} = \left(\text{vect}(\mathbf{X})\right)^{\perp}$ est toujours vérifiée.

Preuve :

Soit $(E,+,\cdot,\langle\ ,\ \rangle)$ un espace préhilbertien réel,et soit $X \subset E$ une partie de E

1. Montrons que :$X^{\perp}$ est toujours ***un sous espace vectoriel fermé*** de E :
   -d'abord,on remarque que :
   $$\begin{aligned} X^{\perp} &= \left\{ y \in F \mid \forall x \in X, \quad \langle x,y\rangle = 0 \right\} \\ &= \bigcap_{x\in X} \text{Ker}\left(\langle x,\cdot\rangle\right) \end{aligned}$$
   avec $\langle x,\cdot\rangle$ est la forme linéaire [b]définie par :
   $$\begin{array}{rccc} \langle x,\cdot\rangle & : & E & \longrightarrow & \mathbb{R} \\ & & y & \longmapsto & \langle x,y\rangle \end{array}$$
   et selon ***l'inégalité de Cauchy-Schwartz*** on a : $\boxed{\left(\forall (x,y) \in X^2\right) : \langle x,y\rangle \leqslant \|x\|\|y\|}$,donc :
   $\left(\forall x \in X\right) : y \mapsto \langle x,y\rangle$ est ***continue***,ce qui fait que :
   $\left(\forall x \in X\right)$:$\text{Ker}\left(\langle x,\cdot\rangle\right)$ est un sous espace vectoriel fermé de E.
   Rappel : le noyau d'une application linéaire,est toujours un sous espace vectoriel et si de plus cette application est continue ,alors il sera sans doute fermé car c'est l'image réciproque de$\{0\}$ qui est fermé par une application continue.
   donc et en guise de conclusion :
   $$X^{\perp} = \bigcap_{x\in X} \text{Ker}\left(\langle x,\cdot\rangle\right) \text{ est un sous espace vectoriel fermé de } E \text{ [c]}.$$
2. montrons que :$X^{\perp} = \left(\text{vect}(X)\right)^{\perp}$ est vraie :
   d'abord on a :$X \subset \left(\text{vect}(X)\right)$ donc :
   $$\left(\text{vect}(X)\right)^{\perp} \subset X^{\perp} \quad \star .$$
   et d'une autre part : si on prend $u \in X^{\perp}$:
   -on a :$\left(\forall v \in \text{vect}(X)\right) \left(\exists \left(\lambda_1,\lambda_2,\cdots,\lambda_n\right) \in \mathbb{R}^n\right) \left(\exists \left(x_1,x_2,\cdots,x_n\right) \in X^n\right) : \ v = \sum_{i=1}^{n} \lambda_i x_i.$

a. ***on n'exige pas de cette partie qu'elle soit un sous espace vectoriel de*** **E**
b. elle est clairement linéaire car le produit scalaire ,et par définition,est bilinéaire.
c. on signale que l'intersection quelconque des sous espaces vectoriels est un sous espace vectoriel et cette propriété reste valable pour les ensembles fermés aussi

donc :$\left(\forall v \in \text{vect}(X)\right) : \langle u, v\rangle = \left\langle u, \sum_{i=1}^{n} \lambda_i x_i \right\rangle = \sum_{i=1}^{n} \lambda_i \langle u, x_i\rangle = 0$ (car $u \in X^{\perp}$). donc :$u \in \text{vect}(E)^{\perp}$,ce qui fait que :

$$X^{\perp} \subset \left(\text{vect}(X)\right)^{\perp} \quad \star\star.$$

donc de $\star$ et $\star\star$ on déduit que : $X^{\perp} = \left(\text{vect}(X)\right)^{\perp}$. d'où le résultat ■

2.2 Théorème de représentation de Riesz :

Théorème 2.2.1 :(Riesz,[6] :)

Soit $\left(E, +, \cdot, \langle\ ,\ \rangle\right)$ ***un espace de Hilbert réel***,pour toute forme linéaire f ***continue*** sur E,***il existe un unique*** $y \in E$ tel que :

$$\left(\forall x \in E\right) : f(x) = \langle y, x\rangle .$$

Preuve :

$F := \ker f$ est fermé ,***car*** f ***est continue***.

— ***Existence*** : Si $F^{\perp} = \{0\}$, alors : $\underbrace{(F^{\perp})^{\perp}}_{=\overline{F}} = \{0\}^{\perp} = E$,or F est un sous espace vectoriel ***fermé*** donc $f = 0$ et on prend $y = 0$. Sinon, soit $w \in F^{\perp}$ tel que $\|w\| = 1$.
on a $f(w) \neq 0$ et pour : $x \in E$: $\left(x - \frac{f(x)}{f(w)}.w\right) \in \ker f = F$ donc :$\left\langle w, x - \frac{f(x)}{f(w)}.w \right\rangle = 0$ d'où, pour $x \in E$,

$$\underbrace{f(x) = f(x)\langle w, w\rangle}_{\text{car } w \text{ est un vecteur unitaire}} = f(w)\langle w, x\rangle = \langle y, x\rangle$$

en posant $y = f(w).w$.

— ***Unicité*** : Soit $y_1, y_2 \in E$ tel que pour tout $x \in E$, $\langle y_1, x\rangle = f(x) = \langle y_2, x\rangle$.
Alors,en prenant $x = y_1 - y_2$, on obtient :

$$\langle y_1 - y_2, y_1 - y_2\rangle = \|y_1 - y_2\|^2 = 0$$

d'où $y_1 = y_2$. ■

Remarque 2.2.1 :

1. Selon la propriété de représentation matricielle des applications linéaires définies sur des espaces de dimension fini [a] vers d'autres qui sont à leurs tours de dimension fini,on peut voir que ce théorème devient presque trivial.
2. ce résultat joue un rôle important ,surtout en calcul différentiel,car il nous permet de ***généraliser*** quelque notions,comme le ***vecteur gradient*** en ***dimension infini*** [b].
3. signalons que ce théorème n'est pas là pour faire joli ! ,mais il sera d'une très grande importance pour la justification et la preuve du fameux théorème d'approximation par les réseaux neuronaux.

a. on rappelle que si E est de dimension fini,alors le dual algébrique E^* et le dual topologique E' deviennent ***confondus***.

b. en fait on peut voir le vecteur gradien comme un représentant de la différentielle.

Index

Bibliographie

[1] BJ Albert. Novikov. on convergence proofs for perceptrons. Technical report, DTIC Document, 1963.

[2] Robert B Ash. *Real Analysis and Probability : Probability and Mathematical Statistics : a Series of Monographs and Textbooks*. Academic press, 2014.

[3] Benaissa Assia. Quelques propriétés et applications de l'opérateur fractionnaire de caputo. Master's thesis, Université Dr Tahar Moulay - Saïda, Algérie, 2016/2017.

[4] A Barron. Universal approximation bounds for superposition of a sigmoidal function. *IEEE Transaction on Information Theory*, 19 :930–944, 1993.

[5] Pierre Borne, Mohamed Benrejeb, and Joseph Haggège. *Les réseaux de neurones : présentation et applications*, volume 15. Editions OPHRYS, 2007.

[6] Nicolas Bourbaki. les structures fondamentales de l'analyse : livre iii : topologie générale. *Eléments de mathématique : première partie :*, 1949.

[7] Franck Boyer. Agrégation externe de mathématiques analyse numérique. *ellipses*, 2013.

[8] Michele Caputo. Linear models of dissipation whose q is almost frequency independent—ii. *Geophysical Journal International*, 13(5) :529–539, 1967.

[9] Antoine Chambert-Loir, Stéfane Fermigier, and Vincent Maillot. *Exercices de mathématiques pour l'agrégation : analyse 1 : Antoine Chambert-Loir, Stéfane Fermigier, Vincent Maillot*. masson, 1995.

[10] David E Chauvin, Yves et Rumelhart. *Backpropagation : théorie, architectures et applications*. Psychology Press, 2013.

[11] E.W. Cheney. *Introduction to approximation theory*. International series in pure and applied mathematics. McGraw-Hill Book Co., 1966.

[12] G. Cybenko. Correction :approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 5(4) :455–455, Dec 1992.

[13] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4) :303–314, 1989.

[14] Claude Dellacherie. Un complément au théorème de weierstrass-stone. *Séminaire de probabilités de Strasbourg*, 1 :52–53, 1967.

[15] Jean-Pierre Demailly. *Analyse numérique et équations différentielles-4ème Ed.* EDP sciences, 2016.

[16] Gérard Dreyfus. *Réseaux de neurones : méthodologie et applications*. Eyrolles, 2004.

[17] François Dubois, Ana Cristina Galucio, and Nelly Point. Introduction à la dérivation fractionnaire-théorie et applications. *Techniques de l'Ingènieur*, 2008.

[18] James Dugundji et al. An extension of tietze's theorem. *Pacific Journal of Mathematics*, 1(3) :353–367, 1951.

[19] Mohamed Ettaouil. Cours master :optimisation non linéaire et modélisation., 2017-2018.

[20] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[21] Ken-Ichi Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural networks*, 2(3) :183–192, 1989.

[22] Federico Girosi and Tomaso Poggio. Networks and the best approximation property. *Biological cybernetics*, 63(3) :169–176, 1990.

[23] Maurice Godefroy. *La fonction gamma : théorie, histoire, bibliographie*. Gauthier-Villars, 1901.

[24] Detlef Gronau and Janusz Matkowski. Geometrical convexity and generalizations of the bohr-mollerup theorem on the gamma function. *Mathematica Pannonica*, 4(2) :153–160, 1993.

[25] Namig J Guliyev and Vugar E Ismailov. Approximation capability of two hidden layer feedforward neural networks with fixed weights. *Neurocomputing*, 316 :262–269, 2018.

[26] Jacques Hadamard. *Sur le rayon de convergence des series ordonnees suivant les puissances d'une variable*. Ecole royale polytechnique, 1888.

[27] Paul R Halmos. *Measure theory*, volume 18. Springer, 2013.

[28] Medjekal Hamza. *Existence et unicité de la solution d'une équation différentielle fractionnaire impulsive de temps infini dans un espace de Banach*. PhD thesis, Université Badji Mokhtar Annaba,Algérie, 2015.

[29] D Hebb. The organization of behavior john wiley & sons. *New York*, 1949.

[30] Th Hélie and Denis Matignon. Diffusive representations for the analysis and simulation of flared acoustic pipes with visco-thermal losses. *Mathematical Models and Methods in Applied Sciences*, 16(04) :503–536, 2006.

[31] Francis Hirsch and Gilles Lacombe. *Eléments d'analyse fonctionnelle : cours et exercices*. Masson, 1997.

[32] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8) :2554–2558, 1982.

[33] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2) :251–257, 1991.

[34] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5) :359–366, 1989.

[35] Kurt Hornik, Maxwell Stinchcombe, Halbert White, and Peter Auer. Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. *Neural Computation*, 6(6) :1262–1275, 1994.

[36] Ahmad Jafarian, Safa Measoomy Nia, Alireza Khalili Golmankhaneh, and Dumitru Baleanu. On artificial neural networks approach with new cost functions. *Applied Mathematics and Computation*, 339 :546–555, 2018.

[37] Jean-Pierre Kahane and Pierre Gilles Lemarié. *Séries de Fourier et ondelettes*. Presse Universitaire, 1998.

[38] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1) :59–69, 1982.

[39] Ph Leray. *Réseaux bayésiens : Apprentissage et diagnostic de systemes complexes*. PhD thesis, Université de Rouen, 2006.

[40] K Ming Leung. Adaline for patern classification. Technical report, Polytechnic institute of new york, 2008.

[41] N Lusin. Sur les propriétés des fonctions mesurables. *CR Acad. Sci. Paris (Projet Numdam)*, 154(25) :1688–1690, 1912.

[42] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4) :115–133, 1943.

[43] Marvin Minsky and Seymour Papert. Perceptrons : An lntroduction to computational geometry. *MITPress, Cambridge, Massachusetts*, 1969.

[44] Patrick Naïm, Pierre-Henri Wuillemin, Philippe Leray, Olivier Pourret, and Anna Becker. *Réseaux bayésiens*. Editions Eyrolles, 2011.

[45] Alain Oustaloup. *La dérivation non entière*. Number 2. Hermes, 1995.

[46] K Parthasarathy. Probability measure on metric spaces. *Journal of the American Statistical Association*, 63, 09 1968.

[47] George Pólya and Gabor Szegö. *Problems and Theorems in Analysis II : Theory of Functions. Zeros. Polynomials. Determinants. Number Theory. Geometry*. Springer Science & Business Media, 1997.

[48] M.J.D. Powell. *Approximation Theory and Methods*. Cambridge University Press, 1981.

[49] Fred Riewe. Nonconservative lagrangian and hamiltonian mechanics. *Physical Review E*, 53(2) :1890, 1996.

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

[51] Hilfer Rudolf. *Applications of fractional calculus in physics*. World Scientific, 2000.

[52] JATMJ Sabatier, Ohm Parkash Agrawal, and JA Tenreiro Machado. *Advances in fractional calculus*, volume 4. Springer, 2007.

[53] Arthur Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3) :210–219, July 1959.

[54] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 44(1.2) :206–226, 2000.

[55] Juliusz Schauder. Der fixpunktsatz in funktionalraümen. *Studia Mathematica*, 2(1) :171–180, 1930.

[56] Vadim Schechtman. Introduction aux fonctions spéciales. *cours de calcul fractionnaire*, 2010.

[57] Basel Solaiman and Lepage Richard. *Les réseaux de neurones artificiels et leurs applications en imagerie et en vision par ordinateur*. Presse Universitaire du Québec, 2003.

[58] Marshall Harvey Stone. Applications of the theory of boolean rings to general topology. *Transactions of the American Mathematical Society*, 41(3) :375–481, 1937.

[59] Claude Touzet. *les réseaux de neurones artificiels, introduction au connexionnisme*. EC2, 1992.

[60] Onno van Gaans. Probability measures on metric spaces. Technical report, Delft University of Technology, 2002.

[61] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5) :988–999, 1999.

[62] Bernard Widrow and Marcian E Hoff. Adaptive switching circuits. Technical report, Stanford Univ Ca Stanford Electronics Labs, 1960.