

09/03/2025

Le Deep Learning, une introduction

Mattéo Eléouet

matteo.eleouet@gmail.com

YouTube : Mattéo Robino

ABSTRACT

Version 0.4.11

Table des matières

1	Introduction	7
1.1	À propos de l'auteur	7
2	Fondamentaux du Deep Learning	8
2.1	Introduction au Deep Learning	8
2.2	Structure d'un réseau de neurones artificiels (ANN)	10
2.3	Apprentissage supervisé vs non supervisé vs semi-supervisé vs apprentissage par renforcement	11
2.4	Régression et Classification	13
2.5	Les types de données : images, texte, séries temporelles	14
2.6	Comment un modèle de machine learning apprend	17
2.7	Résumé	18
2.8	Questions	20
3	Réseaux de neurones artificiels (Artificial Neural Network)	24
3.1	Le Perceptron	24
3.2	Introduction aux fonctions d'activation	26
3.3	Le Biais dans un Perceptron	32
3.4	Le Perceptron multicouches (MLP)	34
3.5	Le théorème d'approximation universelle	37
3.6	Résumé	38
3.7	Question	39
4	Processus d'apprentissage avec la rétropropagation (Backpropagation)	44
4.1	Composantes clés du processus d'apprentissage	44
4.2	La fonction de perte (loss function)	46
4.3	Descente de gradient	54
4.4	La rétropropagation (Backpropagation)	68
4.5	Résumé	79
4.6	Questions	82
4.7	Réponse	85
5	Projet: Projet avec des MLP et mesurer ses performances	87
5.1	Projet 1: Créer une calculatrice	87
5.2	Projet 2: Reconnaissance d'écriture manuscrits	105
5.3	Résumé	126
5.4	Questions	129
5.5	Réponses	132
6	La Généralisation des modèles de Deep Learning	134
6.1	Présentation des concepts clés: généralisation, surajustement et sous-ajustement ..	134
6.2	Surajustement et Sous-ajustement: Compréhension et Diagnostic	136
6.3	Principes et Méthodologies de la Cross-Validation	139
6.4	Comprendre le compromis biais-variance en deep learning	141
6.5	Résumé	146
6.6	Questions	146
6.7	Réponses	148
	Bibliographie	149

Table des matières

1	Introduction	7
1.1	À propos de l'auteur	7
2	Fondamentaux du Deep Learning	8
2.1	Introduction au Deep Learning	8
2.1.1	Qu'est-ce que le Deep Learning ?	8
2.1.2	Le lien entre le Deep Learning et le Machine Learning, en quoi le deep learning est une évolution du machine learning	9
2.2	Structure d'un réseau de neurones artificiels (ANN)	10
2.2.1	La couche d'entrée (intput layer)	11
2.2.2	Les couches cachées (hidden layer)	11
2.2.3	La couche de sortie (output layer)	11
2.3	Apprentissage supervisé vs non supervisé vs semi-supervisé vs apprentissage par renforcement	11
2.3.1	Apprentissage supervisé (Supervised Learning)	11
2.3.2	Popularité et efficacité de l'apprentissage supervisé	12
2.3.3	Apprentissage non supervisé (Unsupervised Learning)	12
2.3.4	Apprentissage semi-supervisé (Semi-supervised Learning)	13
2.3.5	Apprentissage par renforcement (Reinforcement Learning)	13
2.4	Régression et Classification	13
2.4.1	Régression	13
2.4.2	Classification	14
2.5	Les types de données : images, texte, séries temporelles	14
2.5.1	Les images	14
2.5.2	Le Texte	15
2.5.3	Les Séries Temporelles	16
2.5.4	Les Données Structurées	16
2.6	Comment un modèle de machine learning apprend	17
2.6.1	Apprentissage supervisé (supervised learning)	17
2.6.2	Apprentissage non supervisé (unsupervised learning)	17
2.6.3	Apprentissage semi-supervisé (Semi-supervised learning)	18
2.6.4	L'Apprentissage par Renforcement: Apprentissage Par Essais et Récompenses	18
2.7	Résumé	18
2.8	Questions	20
2.8.1	Correction	23
3	Réseaux de neurones artificiels (Artificial Neural Network)	24
3.1	Le Perceptron	24
3.1.1	Un perceptron avec les poids	24
3.2	Introduction aux fonctions d'activation	26
3.2.1	Introduction de Non-linéarités	26
3.2.2	Fonction d'activation sigmoïde	26
3.2.3	Décision sur la Contribution à la Sortie	28
3.2.4	Fonction tangente hyperbolique (tanh)	28

3.2.5 Fonction d'activation de Rectification (ReLU)	29
3.2.6 Fonction d'activation Softmax	30
3.2.7 Normalisation de la Sortie	32
3.3 Le Biais dans un Perceptron	32
3.4 Le Perceptron multicouches (MLP)	34
3.4.1 Propagation avant (feed-forward)	34
3.5 Le théorème d'approximation universelle	37
3.6 Résumé	38
3.7 Question	39
3.7.1 Correction	42
4 Processus d'apprentissage avec la rétropropagation (Backpropagation)	44
4.1 Composantes clés du processus d'apprentissage	44
4.2 La fonction de perte (loss function)	46
4.2.1 Introduction à la fonction de perte (loss function)	46
4.2.2 Les différents types de fonction perte	47
4.3 Descente de gradient	54
4.3.1 Descente de Gradient en 1D	55
4.3.2 Impact de la Taille du Pas (Learning Rate)	55
4.3.3 Descente de gradient en 2D	57
4.3.4 Exemple mathématique de la descente de gradient	58
4.3.5 Qu'est-ce qu'un minimum local ?	63
4.3.6 Les points de selle (saddle points)	65
4.3.7 La descente de gradient stochastique, l'apprentissage par lot	67
4.4 La rétropropagation (Backpropagation)	68
4.4.1 La backpropagation dans le processus d'apprentissage	69
4.4.2 Les Principes Fondamentaux de la Backpropagation	70
4.4.3 L'apprentissage d'un réseau avec la Backpropagation : pas à pas	71
4.4.4 Les problèmes du Vanishing Gradient et de l'Exploding Gradient	76
4.4.5 Avenir de la Backpropagation dans le Deep Learning	78
4.5 Résumé	79
4.6 Questions	82
4.7 Réponse	85
5 Projet: Projet avec des MLP et mesurer ses performances	87
5.1 Projet 1: Créer une calculatrice	87
5.1.1 Cahier des charges	87
5.1.2 Importer vos bibliothèques python et configurer votre matériel.	88
5.1.3 Le jeu de données	89
5.1.4 Architecture du réseau	89
5.1.5 Fonctions d'entraînement et de préparation	95
5.1.6 Entraînement et visualisation des résultats	97
5.1.7 Test et Visualisation: Révéler l'Âme du Modèle	101
5.2 Projet 2: Reconnaissance d'écriture manuscrits	105
5.2.1 Cahier des charges	105
5.2.2 Importation des bibliothèques Python	105
5.2.3 Configuration du Matériel de Calcul	106

5.2.4 Importation et Préparation des Données	106
5.2.5 Exploration des données	113
5.2.6 Définition de l'architecture du réseau	116
5.2.7 Entraînement du modèle	117
5.2.8 Prédiction sur l'ensemble de test	123
5.2.9 Visualisation des prédictions	125
5.3 Résumé	126
5.4 Questions	129
5.5 Réponses	132
6 La Généralisation des modèles de Deep Learning	134
6.1 Présentation des concepts clés: généralisation, surajustement et sous-ajustement .	134
6.1.1 Le surajustement (overfitting)	134
6.1.2 Le sous-ajustement (underfitting)	135
6.2 Surajustement et Sous-ajustement: Compréhension et Diagnostic	136
6.2.1 Visualiser l'Overfitting et l'Underfitting	137
6.2.2 Le biais de sélection	138
6.3 Principes et Méthodologies de la Cross-Validation	139
6.4 Comprendre le compromis biais-variance en deep learning	141
6.5 Résumé	146
6.6 Questions	146
6.7 Réponses	148
Bibliographie	149

1 Introduction

1.1 À propos de l'auteur

 Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

2 Fondamentaux du Deep Learning

2.1 Introduction au Deep Learning

2.1.1 Qu'est-ce que le Deep Learning ?

Le deep learning, est une sous-branche du machine learning lui-même étant une sous branche de l'IA, le deep learning se base sur l'imitation du fonctionnement du cerveau humain¹ pour apprendre, prendre des décisions et résoudre des problèmes. Il s'agit d'une approche avancée du machine learning, qui va au-delà des techniques traditionnelles pour atteindre une meilleure précision et une plus grande complexité dans les tâches.

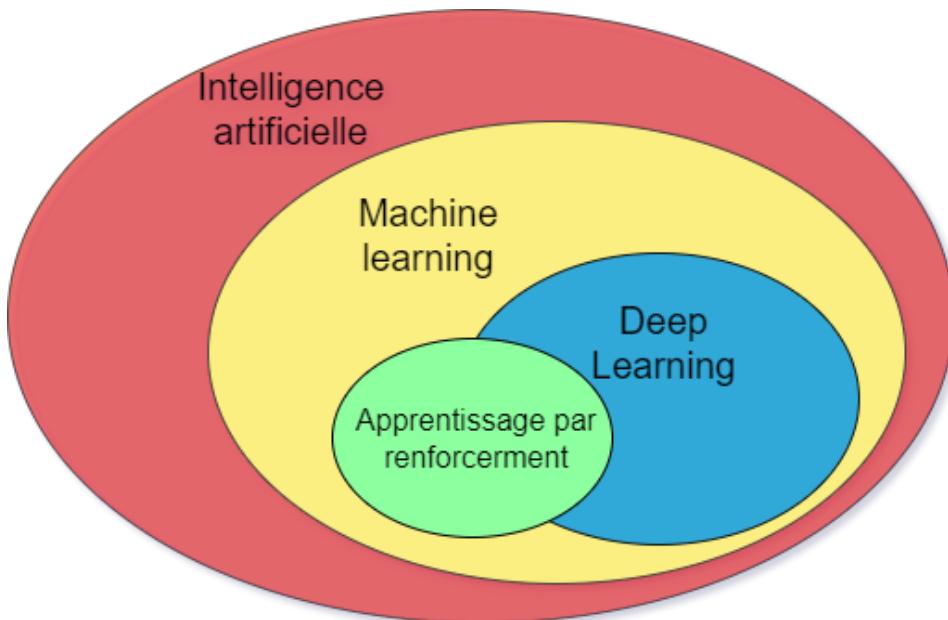


Fig. 1. – Relation entre l’Intelligence artificielle, le Machine Learning, l’apprentissage par renforcement et le Deep Learning. l’IA, le Machine Learning, l’Apprentissage par Renforcement, et le Deep Learning comme des ensembles imbriqués.

Ce diagramme illustre comment l’Intelligence Artificielle (IA), le Machine Learning (ML), l’Apprentissage par Renforcement (RL) et le Deep Learning (DL) sont liés entre eux.

L’IA est le domaine le plus vaste qui englobe toutes les techniques permettant à des machines de simuler des comportements « intelligents ». Le Machine Learning est une sous-catégorie de l’IA qui se concentre sur les méthodes permettant aux algorithmes de s’améliorer à partir de données, un « apprentissage ».

L’Apprentissage par Renforcement est une autre sous-catégorie de l’IA qui implique l’apprentissage par essai-erreur, où un agent apprend à réaliser une tâche en maximisant la récompense obtenue à travers ses interactions avec son environnement. Il convient de noter que l’Apprentissage par Renforcement peut être vu à la fois comme une branche du Machine Learning et comme une technique qui peut être combinée avec le Deep Learning, donnant naissance à ce que l’on appelle l’Apprentissage par Renforcement Profond (Deep reinforcement learning).

¹Ouais attention, c'est grossier comme statement

Le deep learning est un sous-ensemble du machine learning qui s'appuie sur les réseaux de neurones artificiels, qui sont des modèles informatiques inspirés des connexions neuronales du cerveau humain. Ces modèles sont composés de différentes couches de neurones artificiels, chacune étant capable de traiter une partie spécifique de l'information. L'information est introduite dans la première couche (la couche d'entrée), puis est traitée et transmise de couche en couche (couche cachée) jusqu'à la dernière (la couche de sortie). Chaque couche « apprend » de l'information de la couche précédente, la modifie et la transmet à la suivante. C'est ce processus d'apprentissage qui permet à un réseau de neurones de faire des prédictions précises ou de prendre des décisions en fonction des données d'entrée.

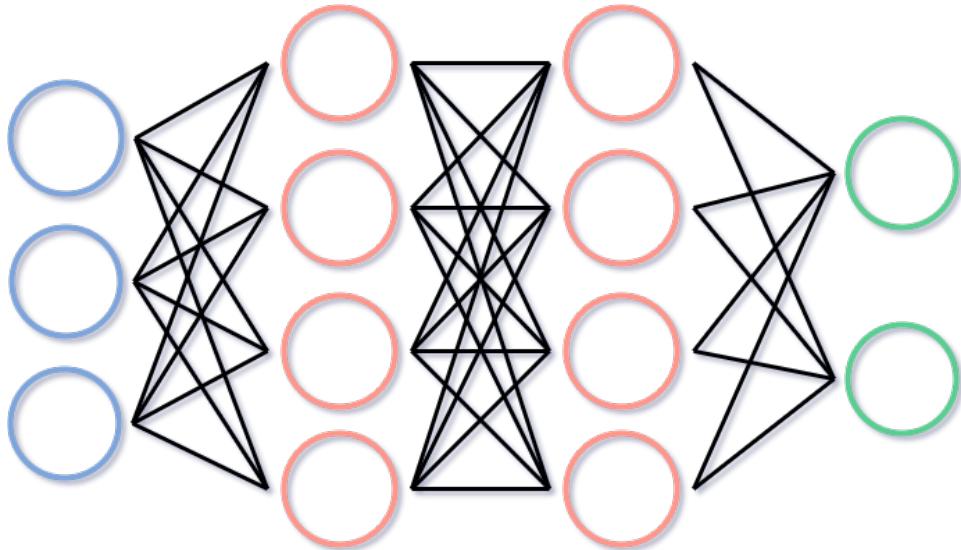


Fig. 2. – diagramme représentant un réseau de neurones artificiels

Ce diagramme illustre la structure typique d'un réseau de neurones artificiels. La couche d'entrée, représentée en bleu, reçoit des informations dérivées de données d'entrée, comme les valeurs de pixels d'une image. Ces informations sont ensuite traitées par les couches cachées (en rose), avec chaque neurone (cercle) recevant des informations de la couche précédente, effectuant un calcul simple et transmettant le résultat à la couche suivante via des connexions neuronales (lignes noires). Ce processus se poursuit jusqu'à la couche de sortie. Du point de vue du diagramme, l'information circule de gauche à droite à travers le réseau.

Dans ce livre, le terme « machine learning » sera utilisé pour englober l'ensemble des techniques de machine learning, y compris le deep learning. Cependant, si je souhaite spécifiquement me référer au machine learning traditionnel en excluant le deep learning, j'emploierai le terme de « machine learning classique ».

2.1.2 Le lien entre le Deep Learning et le Machine Learning, en quoi le deep learning est une évolution du machine learning

Le deep learning est une sous-branche du machine learning classique, ils utilisent tous les deux le même paradigme, celui de l'apprentissage à partir de données.

Le machine learning n'est pas explicitement programmé par un humain, c'est-à-dire, qu'il n'y a pas d'humain pour mettre des conditions à toutes étapes pour créer un algorithme comme le font les systèmes experts. Le machine learning classique apprend à partir de données, par

exemple on pourrait entraîner à prédire le prix d'une maison en fonction de ses caractéristiques (comme la superficie, le nombre de chambres, l'emplacement, etc.), ou à classer les courriels comme « spam » ou « non-spam ».

Le deep learning lui aussi apprend à partir de données, il ne sera alors pas programmé, explicité par un humain, mais utilise un type d'algorithme propre à lui, les réseaux de neurones avec plusieurs couches (avec une seule couche, cela n'est pas considéré comme « deep »). À l'origine créée par G. Hinton, Y. LeCun et Y. Bengio et ses collègues, ces réseaux de neurones, tentent de simuler le comportement du cerveau humain - d'où l'emploi du terme « neurone artificiel ».

Un avantage du deep learning est qu'il peut identifier lui-même les caractéristiques importantes dans les données, une tâche généralement appelée « feature engineering » en machine learning classique, auquel il faut créer les variables qui seront importantes, que l'ont défini à l'aide de méthodes statistiques afin de retirer celles qui le seront moins dans le but d'aider le modèle à apprendre. Grâce à ses couches cachées, le deep learning est capable de modéliser des structures de données complexes sans intervention humaine préalable. Il apprend lui-même quelles variables sont importantes ou non en attribuant une pondération à chacune.

La possibilité d'ajouter plusieurs couches cachées crée de la « profondeur », permettant de réaliser des algorithmes d'une complexité croissante. Chaque couche apprend une représentation de plus en plus complexe à partir des données d'entrée, et les représentations apprises par chaque couche sont ensuite utilisées par la couche suivante pour modéliser un résultat. Cette capacité à gérer des données non structurées, comme des images ou du texte, est l'avantage clé qui distingue le deep learning du machine learning classique. En général, pour des données structurées, le machine learning classique peut généralement être une solution plus appropriée en utilisant une bibliothèque de machine learning classique comme XGBoost.

Des algorithmes ayant une forte complexité nécessitent de grandes quantités de données pour l'entraînement des modèles. Les entraînements sont gourmands en puissance de calculs et en temps . Un exemple d'un cas extrême : l'équivalent « ChatGPT » conçu par Méta, LLaMA [1], a été entraîné sur 2048 cartes graphiques (Nvidia Tesla A100 ayant une mémoire vive de 80 Go et d'une valeur de 20'000€ sur le commerce en 2023), durant cinq mois d'affilée.

Le machine learning classique et le deep learning s'appuient tous les deux sur l'apprentissage à partir de données afin d'identifier des motifs, des schémas, etc, le deep learning n'est pas forcément la meilleure solution à un problème de machine learning. En effet, le choix du modèle dépend de la complexité algorithmique nécessaire, de la quantité de données et de plusieurs autres facteurs tels que la criticité du domaine ciblé, le système sur lequel il sera déployé, ...

2.2 Structure d'un réseau de neurones artificiels (ANN)

Un réseau de neurones artificiels (ANN, pour Artificial Neural Network) est un modèle computationnel qui s'inspire de la manière dont les réseaux neuronaux biologiques du cerveau humain traitent l'information. Composé de neurones interconnectés, un ANN apprend à partir des données d'entrée pour effectuer une variété de tâches, allant de la simple classification à la génération de contenu complexe.

2.2.1 La couche d'entrée (intput layer)

La couche d'entrée, comme son nom l'indique, est le point de départ de tout réseau de neurones. Chaque neurone dans cette couche correspond à une caractéristique spécifique de la donnée d'entrée. Si le réseau est dédié au traitement d'images, chaque neurone pourrait représenter la valeur d'un pixel. Dans un contexte de prévision météorologique, chaque neurone pourrait correspondre à des facteurs variés tels que la température, l'humidité ou la vitesse du vent.

2.2.2 Les couches cachées (hidden layer)

Après la couche d'entrée, viennent les couches cachées d'un réseau de neurones. Ces couches sont nommées « cachées » parce qu'elles ne sont pas visibles à l'extérieur du réseau. Les couches cachées sont le lieu où la majorité du traitement se déroule. Chaque noeud dans une couche cachée reçoit des entrées de tous les noeuds de la couche précédente. Ces entrées sont ensuite pondérées, sommées, et transmises à une fonction qui génère la sortie du noeud.

Bien que les couches cachées soient nommées ainsi, elles ne sont pas secrètes ni inaccessibles. Le terme « caché » est plutôt utilisé pour indiquer qu'elles ne sont pas directement visibles dans l'interface utilisateur ou dans les sorties finales du modèle, c'est le côté « black box » du deep learning.

Un réseau de neurones peut avoir n'importe quel nombre de couches cachées, et chaque couche peut avoir n'importe quel nombre de noeuds. Le choix de la structure exacte dépend de la complexité du problème à résoudre. En règle générale, plus le problème est complexe, plus il y aura de couches cachées et de noeuds.

2.2.3 La couche de sortie (output layer)

La dernière couche d'un réseau de neurones est la couche de sortie. Cette couche a pour mission de produire les résultats ou prédictions finales du réseau. Comme dans les couches cachées, chaque noeud de la couche de sortie reçoit des entrées de tous les noeuds de la couche précédente. Cependant, le nombre de noeuds dans la couche de sortie dépend généralement du type de problème que le réseau est destiné à résoudre. Par exemple, pour une tâche de classification binaire, il n'y aurait qu'un seul noeud dans la couche de sortie, tandis que pour une tâche de classification multiclasse, il y aurait un noeud pour chaque classe possible.

2.3 Apprentissage supervisé vs non supervisé vs semi-supervisé vs apprentissage par renforcement

2.3.1 Apprentissage supervisé (Supervised Learning)

L'apprentissage supervisé est un paradigme du machine learning dans lequel un modèle est entraîné à partir d'un ensemble de données étiquetées. Dans cet ensemble de données, chaque exemple (échantillon) comporte des entrées jumelées à la réponse attendue, dénommée « étiquette ». Deux types principaux de problèmes sont résolus par l'apprentissage supervisé : la classification et la régression.

La classification consiste à prédire une catégorie ou une classe à partir d'un ensemble défini (par exemple, déterminer si un email est un spam ou non), tandis que la régression consiste

à prédire une valeur continue (par exemple, prédire le prix d'une maison à partir de ses caractéristiques).

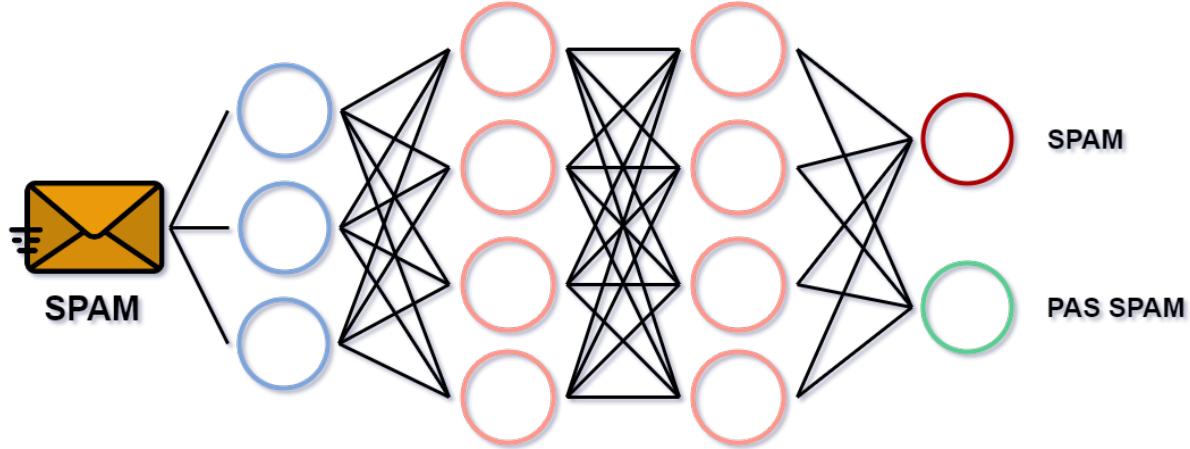


Fig. 3. – diagramme représentant un réseau de neurones artificiels faisant de la classification en supervised learning

Description du diagramme: Ce diagramme représente le processus d'apprentissage supervisé. Les données d'entrée sont associées à des étiquettes spécifiques, et ces paires d'entrées et d'étiquettes sont utilisées pour entraîner le modèle. Le modèle apprend alors à prédire l'étiquette correcte à partir des données d'entrée.

2.3.2 Popularité et efficacité de l'apprentissage supervisé

L'apprentissage supervisé est souvent considéré comme le plus populaire parmi les différentes méthodes d'apprentissage du machine learning, principalement en raison de sa performance. Cette méthode a démontré sa capacité à résoudre une grande variété de problèmes complexes dans divers domaines, allant de la reconnaissance d'images à la prédition financière.

La nature des données étiquetées utilisées en apprentissage supervisé offre un avantage considérable : la possibilité de fournir au modèle des informations claires et précises sur les résultats attendus, ce qui améliore considérablement son habileté à réaliser des prédictions précises et fiables. Par conséquent, bien que d'autres formes d'apprentissage machine continuent d'évoluer et d'apporter des contributions significatives, l'apprentissage supervisé reste le pilier du domaine en raison de sa puissance et de son efficacité démontrées.

2.3.3 Apprentissage non supervisé (Unsupervised Learning)

L'apprentissage non supervisé est un autre paradigme de machine learning où, contrairement à l'apprentissage supervisé, les modèles sont entraînés sur des données non étiquetées. Le but est de découvrir des structures ou des relations cachées dans les données. Les problèmes d'apprentissage non supervisé incluent généralement la réduction de dimensionnalité, la détection d'anomalies, et le regroupement (clustering).

Le clustering, par exemple, vise à diviser l'ensemble de données en groupes distincts de manière que les données dans chaque groupe soient similaires entre elles et différentes des données des autres groupes. Par exemple, vous pourrez faire du clustering pour regrouper un groupe d'utilisateur pour un système de recommandation.

2.3.4 Apprentissage semi-supervisé (Semi-supervised Learning)

L'apprentissage semi-supervisé est un paradigme de machine learning qui combine les éléments des méthodes supervisées et non supervisées. Dans cette approche, un modèle est entraîné sur un ensemble de données partiellement étiquetées - un grand ensemble de données d'entrée non étiquetées, avec un petit ensemble de données d'entrée étiquetées.

L'idée est que le modèle peut apprendre de la structure globale des données non étiquetées pour aider à prédire les étiquettes pour ces données, tout en utilisant les données étiquetées pour guider le processus d'apprentissage. C'est une approche efficace lorsque l'étiquetage des données est coûteux ou chronophage, la difficulté de trouvé des données étiquetées sera le nerf de la guerre dans l'industrie.

2.3.5 Apprentissage par renforcement (Reinforcement Learning)

L'apprentissage par renforcement est un paradigme de machine learning où un agent apprend à prendre des décisions en interagissant avec son environnement. Dans cette approche, un agent effectue certaines actions dans un environnement et reçoit des récompenses ou des punitions (feedback) en retour. L'objectif de l'agent est d'apprendre une politique, qui est une stratégie pour choisir des actions qui maximisent la récompense cumulative à long terme.

Contrairement à l'apprentissage supervisé ou non supervisé, l'apprentissage par renforcement est basé sur l'idée d'apprendre par l'expérience et par l'interaction. Il est souvent utilisé pour les problèmes où un agent doit prendre une série de décisions et où la récompense pour une action peut être retardée. Les exemples d'applications comprennent les jeux (comme les échecs ou le Go), la navigation de robots, et le trading algorithmique.

2.4 Régression et Classification

2.4.1 Régression

La régression, est utilisée lorsque la variable de sortie est une quantité continue. Dans ce cas, l'objectif est de prédire une valeur numérique précise. Les exemples courants de problèmes de régression comprennent la prédition des prix des maisons, la prévision des températures, ou l'estimation de la durée d'un événement.

L'un des modèles de régression les plus couramment utilisés dans le deep learning est le réseau de neurones à régression (Regression Neural Network). La dernière couche de ce réseau de neurones est un seul neurone avec une fonction d'activation linéaire, produisant une sortie continue.

L'erreur entre les prédictions et les valeurs réelles est mesurée par une fonction de coût, comme l'erreur quadratique moyenne (Mean Squared Error - MSE) ou l'erreur absolue moyenne (Mean Absolute Error - MAE).

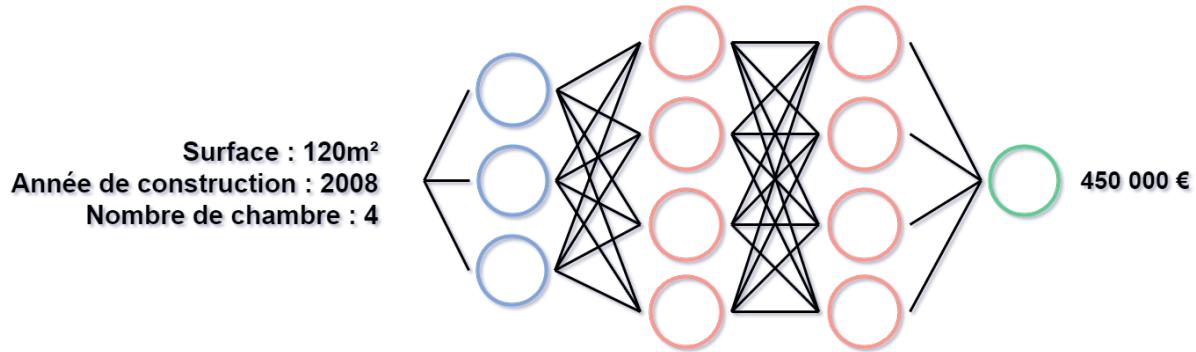


Fig. 4. – diagramme représentant un réseau de neurones artificiels faisant de la régression.

Description du diagramme: Ce diagramme illustre un réseau de neurones à régression. Il prend plusieurs entrées, les fait passer par des couches cachées (où l'apprentissage a lieu), puis produit une sortie continue par le biais d'un seul neurone dans la couche de sortie.

2.4.2 Classification

La classification en deep learning, en revanche, est utilisée lorsque la variable de sortie est une catégorie. Dans ce cas, l'objectif est de prédire la classe ou la catégorie à laquelle appartient une entrée donnée. Les exemples courants de problèmes de classification comprennent la détection de spam, la reconnaissance d'images, ou la prédiction de la survie des passagers du Titanic.

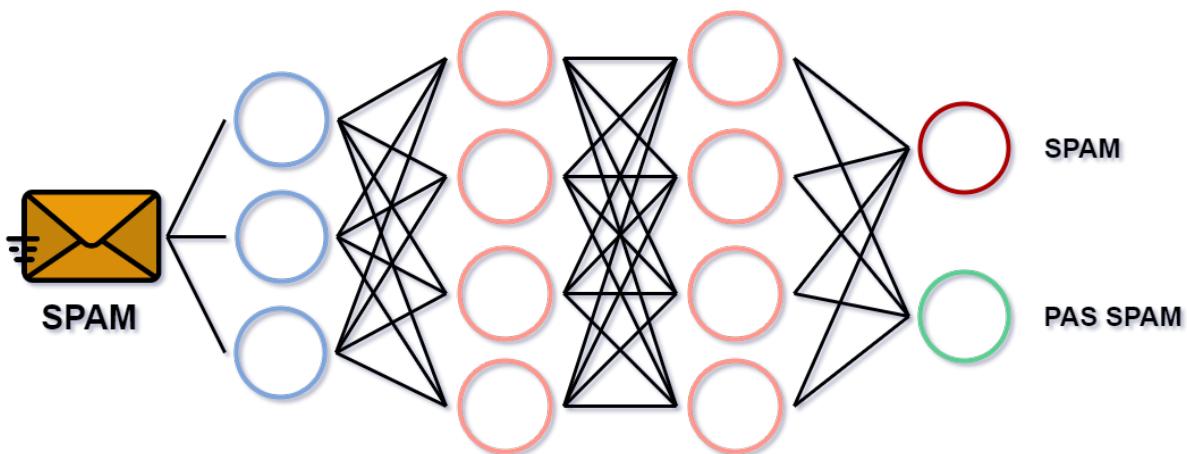


Fig. 5. – diagramme représentant un réseau de neurones artificiels faisant de la classification en supervised learning

Description du diagramme: Ce diagramme représente le processus d'apprentissage supervisé. Les données d'entrée sont associées à des étiquettes spécifiques, et ces paires d'entrées et d'étiquettes sont utilisées pour entraîner le modèle. Le modèle apprend alors à prédire l'étiquette correcte à partir des données d'entrée.

2.5 Les types de données : images, texte, séries temporelles

2.5.1 Les images

Les images sont l'un des types de données les plus couramment utilisés en deep learning. Elles sont généralement représentées sous forme de matrices, où chaque cellule de la matrice correspond à un pixel de l'image. Les images en niveaux de gris sont représentées par une matrice 2D, où chaque cellule contient une valeur entre 0 et 255 représentant l'intensité de

gris. Les images en couleur sont généralement représentées en format RGB (Red, Green, Blue), qui est une matrice 3D où chaque pixel est représenté par trois valeurs correspondant aux intensités des couleurs rouge, verte et bleue.

Dans le cas des images, les packages Python comme PIL ou OpenCV sont souvent utilisés pour le prétraitement des images. Pour PyTorch, une image est transformée en tenseur, avec des dimensions correspondant à (Channels, Height, Width), où Channels correspond aux canaux de couleur (3 pour RGB, 1 pour les images en niveaux de gris), et Height et Width correspondant à la hauteur et à la largeur de l'image en pixels.

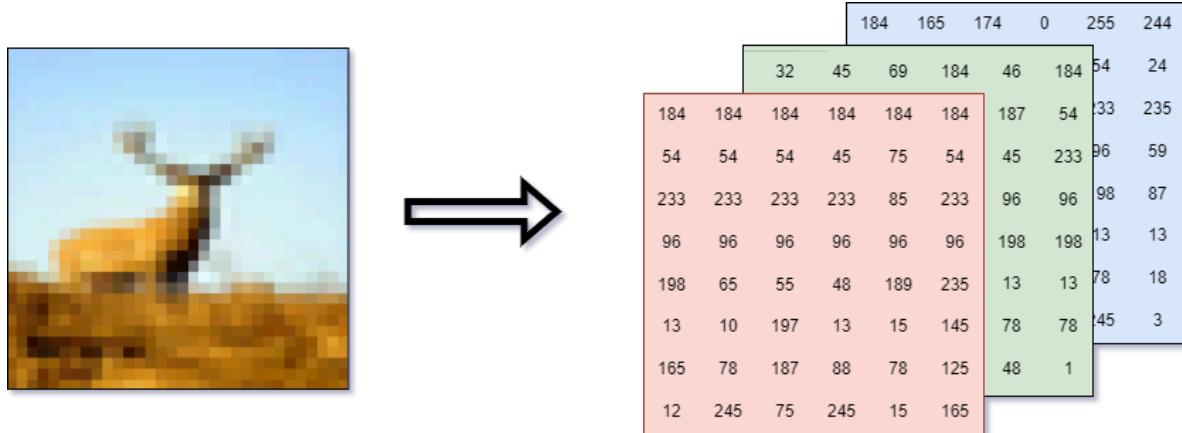


Fig. 6. – Image du jeu de données CIFAR-10, les valeurs sont des valeurs aléatoires.

Description du diagramme: Le diagramme représente une image en couleur sous forme de matrice 3D. Chaque pixel est présenté par trois valeurs, chacune correspondant à l'intensité des couleurs rouge, verte et bleue.

2.5.2 Le Texte

Le texte est un autre type de données largement utilisé en deep learning. Les données textuelles sont généralement représentées sous forme de séquences de mots ou de caractères. Pour l'encodage, des techniques comme l'encodage one-hot, l'embedding de mots (Word Embedding) ou l'encodage par des vecteurs de mots (Word2Vec, GloVe) sont souvent utilisées.



Fig. 7. – Illustration de l'encodage d'une séquence de texte

Description du diagramme: Le diagramme illustre comment une séquence de texte est encodée. Chaque mot est transformé en un vecteur à travers une opération d'embedding.

2.5.3 Les Séries Temporelles

Les séries temporelles sont des données collectées à intervalles réguliers sur une période de temps. Elles sont largement utilisées dans de nombreux domaines tels que la finance, la météorologie et la santé. En deep learning, les séries temporelles sont majoritairement traitées avec des modèles de type séquence tels que les réseaux de neurones récurrents (RNN) ou les Transformers, nous verrons ces architectures de deep learning plus loin dans le livre.

2.5.4 Les Données Structurées

Les données structurées sont un type de données qui suit un modèle ou un format prédéfini, facilitant ainsi leur analyse et leur traitement. Elles peuvent être stockées dans des bases de données relationnelles ou des fichiers csv, xlsx et sont généralement organisées en colonnes et en lignes. Chaque colonne représente une caractéristique (ou un « attribut ») et chaque ligne représente un exemple (ou une « observation »).

Un exemple classique de données structurées est un tableau de données contenant des informations sur des individus, où chaque colonne représente une caractéristique telle que l'âge, le sexe, le revenu, etc., et chaque ligne représente une personne spécifique.

Les données structurées peuvent être traitées avec des réseaux de neurones, qui modélise des relations entre les caractéristiques. Pour préparer ces données pour l'apprentissage, les caractéristiques numériques sont souvent normalisées (c'est-à-dire mises à l'échelle pour avoir une moyenne de 0 et un écart-type de 1), tandis que les caractéristiques catégorielles sont communément encodées en utilisant des techniques comme l'encodage one-hot.

ID	Age	Sexe	Profession	Revenu annuel (K€)
1	25	F	Ingénieur	50
2	32	M	Médecin	70
3	45	F	Professeur	45
4	23	M	Étudiant	13
5	55	F	Retraité	40

Fig. 8. – Illustration d'un jeu de données structurées

Description du diagramme: Le diagramme illustre une table de données structurées, avec chaque colonne représentant une caractéristique différente et chaque ligne représentant une observation distincte. Certaines des caractéristiques sont numériques, tandis que d'autres sont catégorielles.

Les bibliothèques Python comme Pandas sont utilisées pour le prétraitement des données structurées. Pour PyTorch, une table de données structurées est transformée en tenseur, avec des dimensions correspondant à (Nombre d'exemples, Nombre de caractéristiques).

2.6 Comment un modèle de machine learning apprend

2.6.1 Apprentissage supervisé (supervised learning)

Imaginez que vous essayez d'apprendre à jouer au golf pour la première fois. Au début, vos coups sont aléatoires et imprécis. Vous pouvez comparer cela à un modèle de machine learning avant son entraînement : il fait des prédictions, mais elles sont souvent très éloignées de la vérité.

L'essentiel de l'apprentissage, que ce soit pour vous en tant que golfeur ou pour un modèle de machine learning, réside dans la correction des erreurs.

Quand vous jouez au golf, vous obtenez immédiatement un feedback. Si la balle a dévié à gauche, vous savez que vous devez ajuster votre swing pour la prochaine fois. Dans un modèle de machine learning, ce feedback prend la forme de ce que nous appelons une « fonction de perte » (pour une compréhension intuitive, vous pouvez simplement l'appeler « score d'erreur »). C'est une mesure qui permet au modèle de savoir à quel point il s'est trompé. Plus ce score est élevé, plus le modèle s'est trompé dans sa prédiction. Nous parlerons plus en détails de la fonction de perte au chapitre du processus d'apprentissage.

2.6.2 Apprentissage non supervisé (unsupervised learning)

Bien que l'apprentissage non supervisé puisse sembler mystérieux au premier abord, il peut être illustré de manière simple. Imaginez que vous donnez à un enfant un ensemble de blocs de différentes formes et couleurs sans lui fournir d'instructions précises. Avec le temps, vous remarquerez que l'enfant commence à classer les blocs par couleur ou par forme, à les empiler selon leur taille, ou encore à les aligner par ordre de taille croissante ou décroissante. L'enfant, en observant et en découvrant des motifs, apprend de ses propres observations. C'est exactement le processus que suit un modèle de machine learning dans le cadre de l'apprentissage non supervisé.

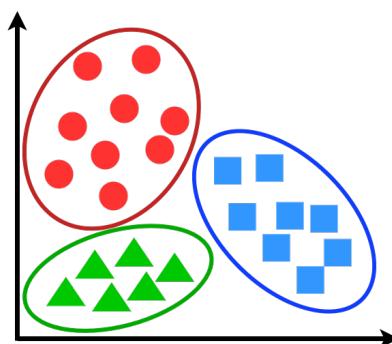


Fig. 9. – Diagramme de clustering à 3 classe

Description du diagramme: Le diagramme montre la classification d'un cluster d'un apprentissage non supervisé via les formes.

Lorsqu'un modèle de machine learning « apprend », il ne fait pas vraiment ce que nous, en tant qu'êtres humains, faisons lorsque nous apprenons quelque chose. Au lieu de cela, il optimise une fonction - une formule mathématique, si vous voulez - qui décrit la « performance » du modèle sur les données d'entraînement. Cette fonction est communément appelée « loss function », ou fonction de coût.

La fonction de coût est un peu comme un GPS pour le modèle. Si vous pensez à votre position actuelle comme étant l'état actuel du modèle, alors la fonction de coût est la distance entre votre position actuelle et votre destination - l'objectif d'apprentissage du modèle.

2.6.3 Apprentissage semi-supervisé (Semi-supervised learning)

L'apprentissage semi-supervisé est un concept intéressant qui se situe quelque part entre l'apprentissage supervisé et l'apprentissage non supervisé. Pour comprendre ce concept, imaginez un enfant jouant avec des blocs de différentes formes et couleurs. Cette fois, cependant, un adulte est présent et donne occasionnellement à l'enfant des conseils sur comment jouer avec les blocs. Par exemple, l'adulte pourrait montrer à l'enfant comment empiler les blocs de la même taille ou les aligner par couleur. Toutefois, la majeure partie du temps, l'enfant est libre d'explorer et d'apprendre par lui-même. C'est essentiellement ce que fait un modèle de machine learning dans un contexte d'apprentissage semi-supervisé.

Dans le contexte de l'apprentissage semi-supervisé, une partie des données d'entraînement est étiquetée, tandis que l'autre partie ne l'est pas. Le modèle apprend à partir des deux : il utilise les données étiquetées pour comprendre certaines structures des données, et il utilise les données non étiquetées pour explorer et découvrir d'autres structures par lui-même.

2.6.4 L'Apprentissage par Renforcement: Apprentissage Par Essais et Récompenses

Imaginez un enfant qui apprend à jouer à un jeu vidéo pour la première fois. Au début, l'enfant ne connaît pas les règles du jeu et fait beaucoup d'erreurs. Mais avec le temps, il commence à comprendre comment le jeu fonctionne. Il apprend que certaines actions entraînent des points ou des récompenses, tandis que d'autres peuvent entraîner des pénalités ou des pertes de vie. Ce processus d'apprentissage par essais et erreurs, guidé par des récompenses et des pénalités, est l'essence de l'apprentissage par renforcement.

Dans le cadre de l'apprentissage par renforcement, un agent d'apprentissage (le modèle) interagit avec un environnement (comme le jeu vidéo), fait des actions, reçoit des feedbacks sous forme de récompenses ou de pénalités, et apprend à travers ce processus. L'objectif de l'agent est de maximiser la somme totale des récompenses qu'il reçoit au fil du temps. Pour ce faire, l'agent doit apprendre quelle est la meilleure action à faire dans chaque situation ou état de l'environnement.

Contrairement à l'apprentissage supervisé et semi-supervisé, l'apprentissage par renforcement n'a pas de paires d'entrées/sorties étiquetées à partir desquelles apprendre directement. Au lieu de cela, l'agent apprend de l'expérience - il fait des actions, voit les résultats, et ajuste son comportement en conséquence.

Un concept clé dans l'apprentissage par renforcement est la politique. Une politique est essentiellement une stratégie que l'agent utilise pour décider quelle action faire dans chaque état. L'agent commence souvent par une politique aléatoire, puis l'ajuste au fil du temps pour favoriser les actions qui ont tendance à donner de meilleures récompenses.

2.7 Résumé

Dans ce chapitre, nous avons vu qu'est-ce que c'était le deep learning, qu'il était une sous-division de l'IA et du machine learning. Les modèles de deep learning ont une profondeur et une complexité surpassant largement celles de leurs homologues du machine learning. De

plus, le deep learning permet une extraction de caractéristiques de manière automatique et hiérarchique, contrairement au machine learning traditionnel qui nécessite une extraction manuelle des caractéristiques.

Le chapitre explore également les quatre types d'apprentissage en machine learning : supervisé, non supervisé, semi-supervisé et par renforcement. L'apprentissage supervisé et non supervisé fonctionne avec des données étiquetées et non étiquetées respectivement. L'apprentissage semi-supervisé combine les deux, tandis que l'apprentissage par renforcement fonctionne par interaction avec l'environnement.

La régression et la classification sont expliquées comme deux types de tâches courantes en apprentissage supervisé, la régression prévoyant des valeurs continues et la classification prévoyant des catégories.

Enfin, le chapitre met l'accent sur les différents types de données traitées en deep learning : les images, le texte et les séries temporelles. Les images sont généralement représentées sous forme de matrices, où chaque cellule correspond à un pixel. Le texte, quant à lui, est généralement représenté sous forme de séquences de mots ou de caractères

L'apprentissage supervisé, non supervisé, semi-supervisé et par renforcement sont les quatre grandes catégories d'apprentissage dans le domaine du machine learning. Chacun d'eux se réfère à un ensemble unique de stratégies et de techniques d'apprentissage, illustrées par des analogies telles que jouer au golf, jouer avec des blocs de construction ou jouer à un jeu vidéo.

Dans l'apprentissage supervisé, un modèle apprend à partir de données étiquetées, tout comme un golfeur apprend en recevant un feedback après chaque coup. L'apprentissage non supervisé, en revanche, ressemble plus à un enfant qui joue avec des blocs sans instructions explicites, découvrant les structures et les motifs par lui-même. L'apprentissage semi-supervisé est un équilibre entre les deux, comme un enfant jouant avec des blocs, mais recevant occasionnellement des conseils d'un adulte. Ces trois types d'apprentissage s'appuient sur l'idée d'une fonction de coût, qui guide l'apprentissage du modèle en lui fournissant un score d'erreur pour chaque prédiction.

Enfin, l'apprentissage par renforcement est un processus d'apprentissage dynamique guidé par des récompenses et des pénalités. C'est comme un enfant apprenant à jouer à un jeu vidéo, ajustant constamment son comportement pour maximiser le score total. Dans cette approche, l'agent d'apprentissage (le modèle) interagit activement avec son environnement, apprend de l'expérience et ajuste progressivement sa politique - une stratégie pour décider quelle action entreprendre dans chaque état.

2.8 Questions

1. **Quelle est la description la plus précise du Deep Learning ?**

- a. C'est une approche avancée du Machine Learning qui vise à imiter le fonctionnement du cerveau humain pour apprendre et résoudre des problèmes.
- b. C'est un enchaînement de condition if et des else vrai ou faux pour résoudre des problématiques.
- c. C'est une méthode de statistique qui utilise des algorithmes pour modéliser les données.
- d. C'est un sous-domaine de l'IA qui se concentre uniquement sur l'analyse du langage naturel de type ChatGPT.

2. **Le diagramme de la figure 1 des relations entre l'IA, le ML et le DL illustre que :**

- a. Le Machine Learning et le Deep Learning sont complètement séparés.
- b. Le Machine Learning est un sous-ensemble du Deep Learning.
- c. L'IA, le Machine Learning et le Deep Learning sont des sous-ensembles les uns des autres.
- d. Le Deep Learning et l'IA ne sont pas liés.

3. **Quelle est la différence principale entre le Machine Learning et le Deep Learning ?**

- a. Le Deep Learning nécessite des données manuellement étiquetées tandis que le Machine Learning ne le fait pas.
- b. Le Deep Learning utilise des réseaux de neurones artificiels plus complexes et peut effectuer une extraction de caractéristiques de manière automatique et hiérarchique.
- c. Le Machine Learning est une approche plus moderne que le Deep Learning.
- d. Le Machine Learning peut traiter des données de grande dimension et non structurées, tandis que le Deep Learning ne le peut pas.

4. **Qu'est-ce que l'apprentissage supervisé ?**

- a. C'est une méthode où un modèle apprend à prendre des décisions en interagissant avec son environnement.
- b. C'est une méthode qui combine les éléments des méthodes supervisées et non supervisées.
- c. C'est une méthode où un modèle est entraîné sur des données non étiquetées pour découvrir des structures ou des relations cachées dans les données.
- d. C'est une méthode où un modèle est entraîné à partir d'un ensemble de données étiquetées.

5. **En quoi consiste la classification dans le contexte de l'apprentissage supervisé ?**

- a. La classification implique de prédire l'appartenance d'un nouvel échantillon à l'une des catégories prédefinies, basée sur un apprentissage à partir de données étiquetées.
- b. Prédire une valeur continue.
- c. La classification est le processus de regroupement de données similaires en catégories ou clusters sans étiquettes prédefinies.
- d. Maximiser la récompense cumulative à long terme.

6. **Qu'est-ce qui décrit le mieux l'apprentissage non supervisé ?**

- a. Les modèles sont entraînés sur des données étiquetées pour résoudre des problèmes spécifiques.
- b. Les modèles sont entraînés pour prendre des décisions en interagissant avec leur environnement.
- c. Les modèles sont entraînés sur des données non étiquetées pour découvrir des structures cachées.
- d. Les modèles sont entraînés à partir d'un ensemble de données partiellement étiquetées.

7. Quel type de tâche en deep learning utiliserait-on pour prédire le prix d'une maison?

- a. Classification
- b. Régression
- c. Les deux
- d. Aucun des deux

8. Comment les images en couleur sont-elles représentées pour le deep learning

- a. Par une matrice 3D où chaque pixel est représenté par trois valeurs correspondant aux intensités des couleurs rouge, verte et bleue
- b. Par une matrice 2D où chaque cellule contient une valeur entre 0 et 255
- c. Par une seule valeur correspondant à l'intensité de la couleur la plus dominante
- d. Elles ne sont pas représentées, elles sont utilisées telles quelles

9. Qu'est-ce que l'apprentissage supervisé ?

- a. Un modèle apprend sans aucune guidance ou label
- b. Un modèle apprend en recevant des feedbacks sur ses erreurs et en ajustant ses prédictions
- c. Un modèle apprend par essai et erreur dans un environnement avec récompenses et pénalités
- d. Un modèle apprend en recevant des feedbacks occasionnels tout en explorant et découvrant par lui-même

10. Qu'est-ce que l'apprentissage non supervisé ?

- a. Un modèle apprend sans aucune guidance ou label, découvrant les patterns dans les données par lui-même
- b. Un modèle apprend par essai et erreur dans un environnement avec récompenses et pénalités
- c. Un modèle apprend en recevant des feedbacks sur ses erreurs et en ajustant ses prédictions
- d. Un modèle apprend en recevant des feedbacks occasionnels tout en explorant et découvrant par lui-même

11. Qu'est-ce que l'apprentissage par renforcement ?

- a. Un modèle apprend en recevant des feedbacks sur ses erreurs et en ajustant ses prédictions
- b. Un modèle apprend par essai et erreur dans un environnement avec récompenses et pénalités

- c. Un modèle apprend sans aucune guidance ou label, découvrant les patterns dans les données par lui-même
- d. Un modèle apprend en recevant des feedbacks occasionnels tout en explorant et découvrant par lui-même

2.8.1 Correction

1. Quelle est la description la plus précise du Deep Learning ?

Réponse: a C'est une approche avancée du Machine Learning qui vise à imiter le fonctionnement du cerveau humain pour apprendre et résoudre des problèmes

2. Le diagramme avec l'image « Relation_IA_ML_DL.png » illustre que :

Réponse: c L'IA, le Machine Learning et le Deep Learning sont des sous-ensembles les uns des autres.

3. Quelle est la différence principale entre le Machine Learning et le Deep Learning ?

Réponse: b Le Deep Learning utilise des réseaux de neurones artificiels plus complexes et peut effectuer une extraction de caractéristiques de manière automatique et hiérarchique.

4. Qu'est-ce que l'apprentissage supervisé ?

Réponse: d C'est une méthode où un modèle est entraîné à partir d'un ensemble de données étiquetées.

5. En quoi consiste la classification dans le contexte de l'apprentissage supervisé ?

Réponse: a La classification implique de prédire l'appartenance d'un nouvel échantillon à l'une des catégories prédefinies, basée sur un apprentissage à partir de données étiquetées.

6. Qu'est-ce qui décrit le mieux l'apprentissage non supervisé ?

Réponse: c Les modèles sont entraînés sur des données non étiquetées pour découvrir des structures cachées.

7. Quel type de tâche en deep learning utiliserait-on pour prédire le prix d'une maison?

Réponse: b Régression

8. Comment les images en couleur sont-elles généralement représentées pour le deep learning

Réponse: a Par une matrice 3D où chaque pixel est représenté par trois valeurs correspondant aux intensités des couleurs rouge, verte et bleue

9. Qu'est-ce que l'apprentissage supervisé ?

Réponse: b Un modèle apprend en recevant des feedbacks sur ses erreurs et en ajustant ses prédictions

10. Qu'est-ce que l'apprentissage non supervisé ?

Réponse: a Un modèle apprend sans aucune guidance ou label, découvrant les patterns dans les données par lui-même

11. Qu'est-ce que l'apprentissage par renforcement ?

Réponse: b Un modèle apprend par essai et erreur dans un environnement avec récompenses et pénalités

3 Réseaux de neurones artificiels (Artificial Neural Network)

Dans ce chapitre nous rentrons plus en détails dans les réseaux de neurones. Nous parlerons du perceptron qui est un neurone artificiel, des différentes fonctions d'activations utilisées qui permettent de casser la linéarité des perceptrons et d'avoir des relations complexes dans les données, nous parlerons de la structure d'un réseau de neurones, du fonctionnement des perceptrons entre eux et du théorème d'approximation universelle, qui démontre la capacité des réseaux de neurones à modéliser une vaste gamme de fonctions continues.

3.1 Le Perceptron

Un perceptron est un neurone artificiel ou une unité de traitement dans un réseau neuronal. Les perceptrons sont utilisés comme un bloc de base dans les réseaux de neurones profonds formant la base des modèles de deep learning actuels.

3.1.1 Un perceptron avec les poids

La structure d'un perceptron est relativement simple. Il comprend plusieurs entrées, chacune étant pondérée par un poids spécifique. Ces poids déterminent l'importance relative de chaque entrée pour la sortie. Un biais est également ajouté pour ajuster la sortie indépendamment des entrées. La sortie d'un perceptron est obtenue en sommant le produit de chaque entrée pondérée par son poids. Nous aborderons le rôle du biais en détail dans la suite, commençons sans pour faciliter les calculs.

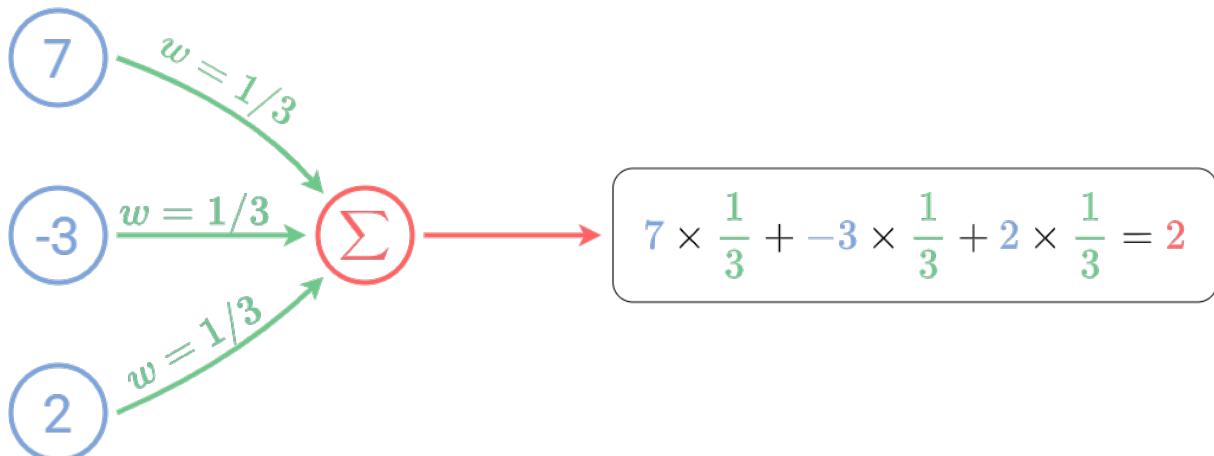


Fig. 10. – Perceptron sans biais qui calcule la moyenne des entrées.

Ici, les valeurs d'entrées sont : 7, -3 et 2, elles ont chacune un poids de $\frac{1}{3}$, ce qui transforme ce perceptron en une machine... à calculer la moyenne ! $7 \times \frac{1}{3} + -3 \times \frac{1}{3} + 2 \times \frac{1}{3} = 2$, ici la sortie de ce perceptron est 2.

```
x = [7, -3, 2] # Valeurs d'entrée x
w = [1/3, 1/3, 1/3] # Poids w
output = sum(x[i] * w[i] for i in range(len(x)))
print(output) # L'output du perceptron est égal à 2
```

Voici un deuxième exemple où les valeurs sont différentes.

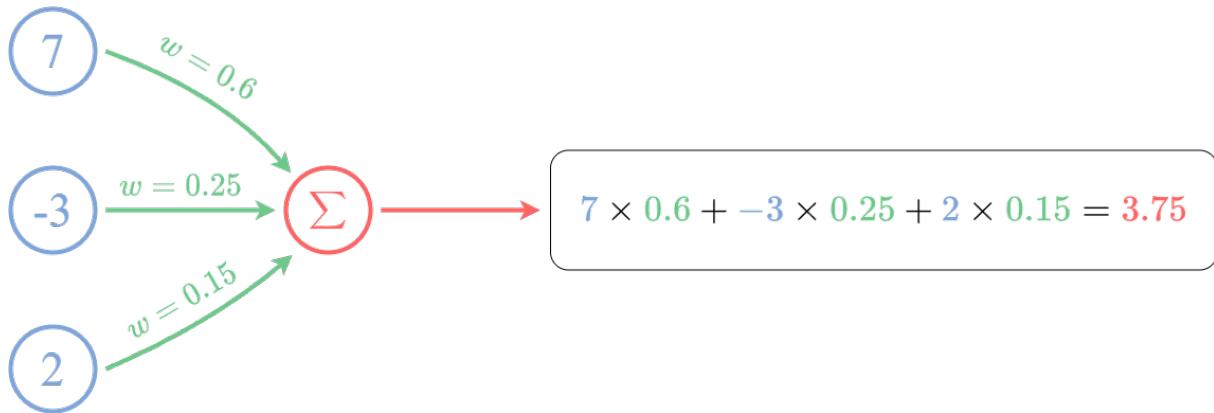


Fig. 11. – Perceptron sans biais qui calcule la sortie avec des poids différent.

Ici, les valeurs d'entrées sont les mêmes, mais elles ont des poids différents. Le calcul est $7 \times 0.6 + -3 \times 0.25 + 2 \times 0.15 = 3.75$, ici la sortie de ce perceptron est 3.75.

les entrées sont $x = \begin{pmatrix} 7 \\ -3 \\ 2 \end{pmatrix}$, les poids sont $w = \begin{pmatrix} 0.6 \\ 0.25 \\ 0.15 \end{pmatrix}$, on remplace les inconnus par les valeurs de x et $w \rightarrow \hat{y} = \begin{pmatrix} 7 \\ -3 \\ 2 \end{pmatrix}^T \times \begin{pmatrix} 0.6 \\ 0.25 \\ 0.15 \end{pmatrix} = 3.75$. Nous venons d'effectuer un produit scalaire.

```

x = np.array([7, -3, 2]) # Vecteur d'entrée x
w = np.array([0.6, 0.25, 0.15]) # Vecteur de poids w

output = np.dot(x.T, w) # sur numpy cela s'appelle dot car un produit scalaire
en anglais se nomme "dot product"
print(output) # L'output du perceptron est égal à 3.75

```

La formule mathématique de ces calculs serait $\hat{y} = x^T w$ où \hat{y} la sortie du perceptron et x^T est la transposée de x

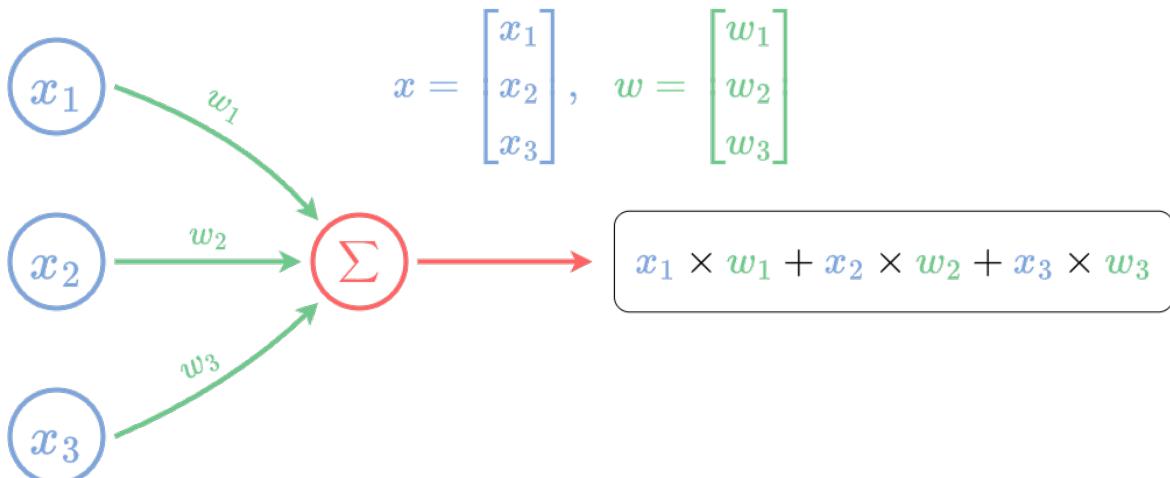


Fig. 12. – Calcule mathématique d'un perceptron sans biais.

Ce qu'il faut comprendre c'est plus la valeur d'un poids est élevée plus son influence sur le perceptron sera grande, c'est de cette manière-là que dans le deep learning s'effectue le *feature engineering* qui consiste à choisir les variables importantes. Par exemple, si vous deviez définir le prix d'un bien immobilier à partir de plusieurs variables, vous pourriez penser aux nombres de pièces, la couleur des murs, mais la variable du nombre du m^2 serait probablement très importante pour définir le prix d'un bien immobilier, elle aurait alors plus d'importance que celle de la couleur des murs.

3.2 Introduction aux fonctions d'activation

Une fonction d'activation est un élément essentiel des réseaux de neurones artificiels, notamment des perceptrons. Elle est utilisée pour transformer la sortie pondérée de la somme des entrées d'un neurone. Cette transformation peut aider à introduire des non-linéarités dans le modèle, permettre à chaque neurone de prendre une décision sur sa contribution à la sortie globale et aider à normaliser la sortie d'un neurone.

3.2.1 Introduction de Non-linéarités

Jusque-là nos perceptrons étaient linéaires, pour faire simple notre perceptron ce sont des additions et des multiplications, ce qui est mathématiquement linéaire, graphiquement, la linéarité, c'est une droite.

Les fonctions d'activation sont non-linéaires. C'est une caractéristique clé qui permet aux réseaux de neurones d'apprendre à partir de données complexes et non linéaires. Sans les non-linéarités introduites par les fonctions d'activation, un réseau de neurones serait essentiellement un modèle linéaire, limité dans sa capacité à traiter des relations plus complexes entre les entrées et les sorties.

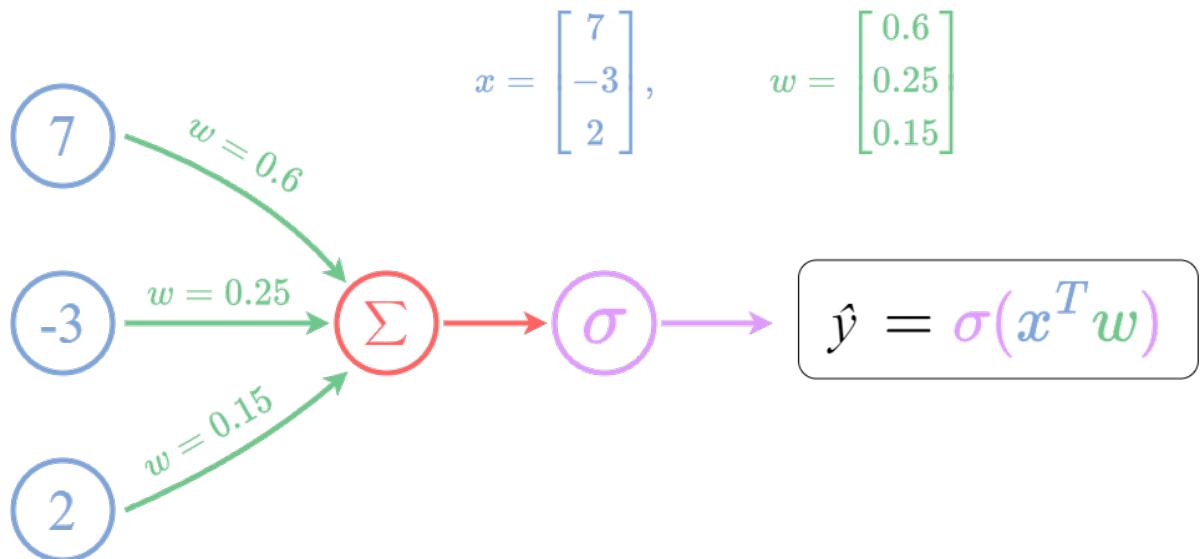


Fig. 13. – Perceptron sans biais avec une fonction d'activation en sortie représentée par σ .

3.2.2 Fonction d'activation sigmoïde

La fonction d'activation sigmoïde, également appelée fonction logistique, est une fonction couramment utilisée en deep learning, elle introduit une non-linéarité et produit des sorties dans une plage spécifique entre 0 et 1, graphiquement, la sigmoïde a une forme en « S »

caractéristique. Sa forme aplati les valeurs extrêmes tout en amplifiant les valeurs proches de zéro, offrant ainsi une interprétation probabiliste des activations.

La formule mathématique qui définit cette fonction s'exprime comme suit :

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

```
def sigmoid(x):
    return 1 / (1 + np.exp(-x))
print(sigmoid(3.75))

>>> 0.978
```

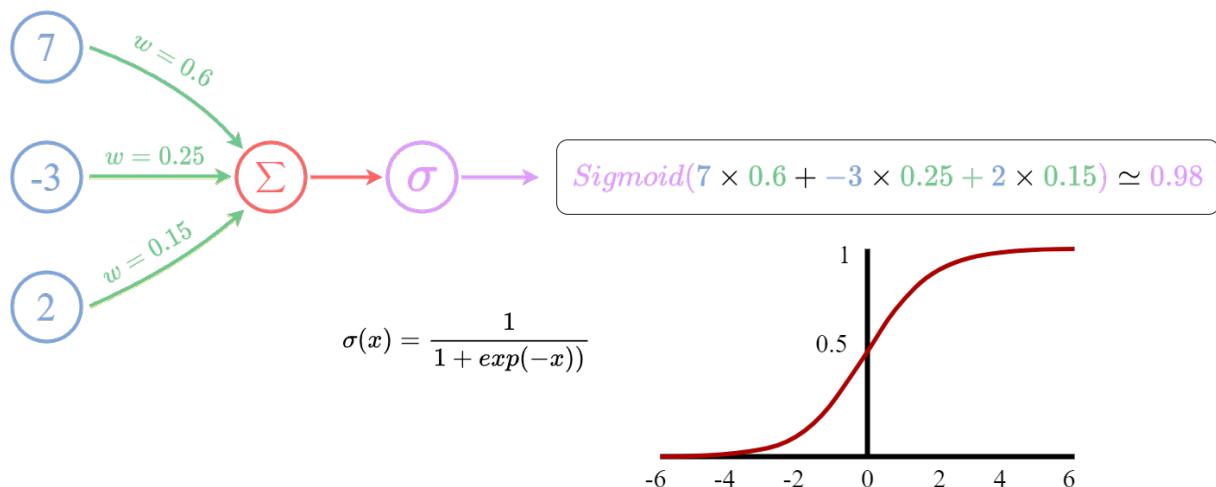


Fig. 14. – Représentation d'un perceptron sans biais utilisant la fonction d'activation sigmoïde pour transformer une sortie pondérée en une probabilité.

La figure associée illustre ce processus visuellement. Elle présente un perceptron avec trois noeuds d'entrée, chacun représentant **7**, **-3**, et **2**, connectés au noeud central par des arêtes étiquetées avec les poids **0.6**, **0.25**, et **0.15**. Le noeud central, marqué par un Σ en rouge, effectue la somme pondérée $7 \times 0.6 + -3 \times 0.25 + 2 \times 0.15 = 3.75$. Cette valeur est transmise à un noeud supplémentaire, coloré en violet et marqué par σ , symbolisant la fonction d'activation sigmoïde. Ce noeud transforme **3.75** en **0.98**. À droite, une courbe tracée en rouge montre la fonction $\sigma(x) = \frac{1}{1 + \exp(-x)}$, avec un axe des abscisses variant de -6 à 6 et un axe des ordonnées de 0 à 1, affichant clairement la forme en « S » caractéristique de la sigmoïde. Cette représentation graphique aide à visualiser comment la sigmoïde compresse les valeurs extrêmes vers 0 ou 1, tout en centrant les valeurs proches de zéro autour de 0,5.

La fonction sigmoïde était largement utilisée dans les anciennes architectures de réseaux de neurones. Cependant, elle a été supplantée par d'autres fonctions d'activation plus performantes dans de nombreux scénarios. La sigmoïde peut entraîner des problèmes de saturation lorsque les valeurs d'entrée sont très élevées ou très basses, ce qui peut conduire à une convergence lente et à une atténuation du gradient lors de l'entraînement. Nous aborderons l'apprentissage en détail dans le chapitre suivant dédié à ce sujet. Elle reste néanmoins utilisée pour des tâches comme la classification binaire, où la sortie peut représenter la probabilité qu'une instance appartienne à une classe donnée.

3.2.3 Décision sur la Contribution à la Sortie

Les fonctions d'activation permettent également à chaque neurone de prendre une décision concernant sa contribution à la sortie du réseau. Par exemple, dans le cas de la fonction d'activation sigmoïde, qui transforme chaque entrée en une valeur entre 0 et 1, un neurone peut décider d'activer ou non sa sortie en fonction de la valeur de la sortie pondérée.

Si la sortie pondérée est très positive, la fonction sigmoïde produira une valeur proche de 1, ce qui signifie que le neurone est « activé » et contribue pleinement à la sortie du réseau. Si la sortie pondérée est très négative, la fonction sigmoïde produira une valeur proche de 0, ce qui signifie que le neurone est « désactivé » et ne contribue pas vraiment à la sortie du réseau.

3.2.4 Fonction tangente hyperbolique (tanh)

La fonction d'activation Tangente Hyperbolique, abrégée en Tanh. Elle est utilisée pour transformer la somme pondérée des entrées d'un neurone en une sortie qui peut être ensuite transmise aux autres neurones du réseau.

La principale caractéristique de la tanh est sa plage sortie, entre -1 et 1 contrairement à la fonction d'activation sigmoïde, qui produit une sortie entre 0 et 1 . Cette plage, centrée autour de zéro, est bénéfique lors de l'apprentissage de notre modèle, si les sorties sont toutes positives ou négatives, ce qui produit un phénomène appelé le « biais de notification » (reporting bias), ralentissant l'apprentissage du réseau.

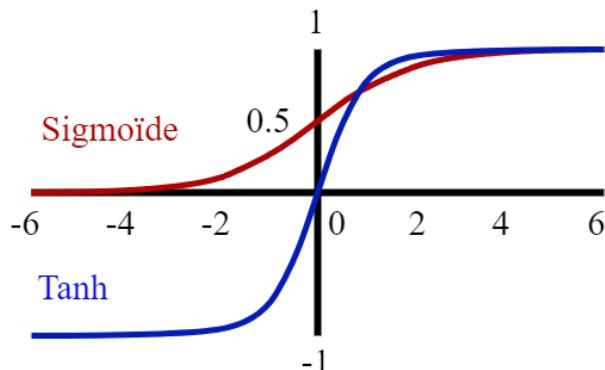


Fig. 15. – Courbe de la tanh et de la sigmoïde.

Une bonne fonction d'activation a une dérivée facile à calculer, celle de la tanh est simple à calculer pour un ordinateur ce qui sera très utile lors de l'algorithme de la rétropropagation (backpropagation) qui utilise les dérivées des fonctions d'activations pour ajuster les paramètres du modèle lors de l'apprentissage, nous verrons cela plus en détail plus tard.

Cependant, Tanh n'est pas sans défauts. En particulier face au problème du gradient qui disparaît (vanishing gradient), qui arrive souvent sur des réseaux ayant de nombreuses couches cachées, surtout quand la dérivée de la fonction d'activation est proche de 0. Ce qui arrive à la tanh lorsque les sorties sont proches de 1 ou de -1 . Vous comprendrez mieux cela au chapitre prochain lorsque nous verrons le processus d'apprentissage.

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

```
# x est la sortie du perceptron
def tanh(x):
    return (np.exp(x) - np.exp(-x)) / (np.exp(x) + np.exp(-x))
```

3.2.5 Fonction d'activation de Rectification (ReLU)

La fonction Rectified Linear Unit, appelée ReLU, est probablement la fonction d'activation la plus utilisée dans le deep learning. La ReLU est définie comme la fonction qui renvoie l'entrée si elle est positive et zéro dans le cas contraire. En d'autres termes, elle « rectifie » les valeurs négatives en les mettant à zéro.

La formule mathématique est très simple :

$$\sigma(x) = \max(0, x)$$

```
def relu(x):
    return np.maximum(0, x)
print(relu(10))
print(relu(-5))

>>> 10
>>> 0
```

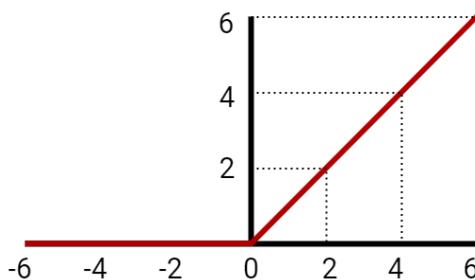


Fig. 16. – Courbe de la ReLU.

Cette simplicité a plusieurs avantages qui font de la ReLU une fonction d'activation extrêmement populaire dans le deep learning. Premièrement, ReLU est très facile à calculer, ce qui peut aider à accélérer le processus d'apprentissage. Deuxièmement, sa dérivée est également simple : elle est de 1 pour les entrées positives et de 0 pour les entrées négatives, ce qui rend la rétropropagation plus efficace.

La ReLU est robuste au vanishing gradient, ce problème se produit lorsque les gradients deviennent très petits au fur et à mesure qu'ils sont rétropropagés à travers le réseau, de sorte que les poids de certaines couches du réseau ne sont pratiquement pas mis à jour pendant l'apprentissage. Comme la dérivée de ReLU est toujours 1 pour les entrées positives, elle permet de propager efficacement ces gradients et d'éviter ce problème.

Cependant, il faut noter que ReLU n'est pas sans défauts. L'un des problèmes majeurs est le « dying ReLU » : si un neurone donne une sortie négative, la dérivée de ReLU pour ce neurone sera 0, ce qui signifie que ce neurone n'apprendra plus rien pendant la rétropropagation. En d'autres termes, le neurone « meurt ». Plusieurs variantes de ReLU ont été proposées pour résoudre ce problème, comme Leaky ReLU, Parametric ReLU et GELU existent, mais il y en a plein d'autre.

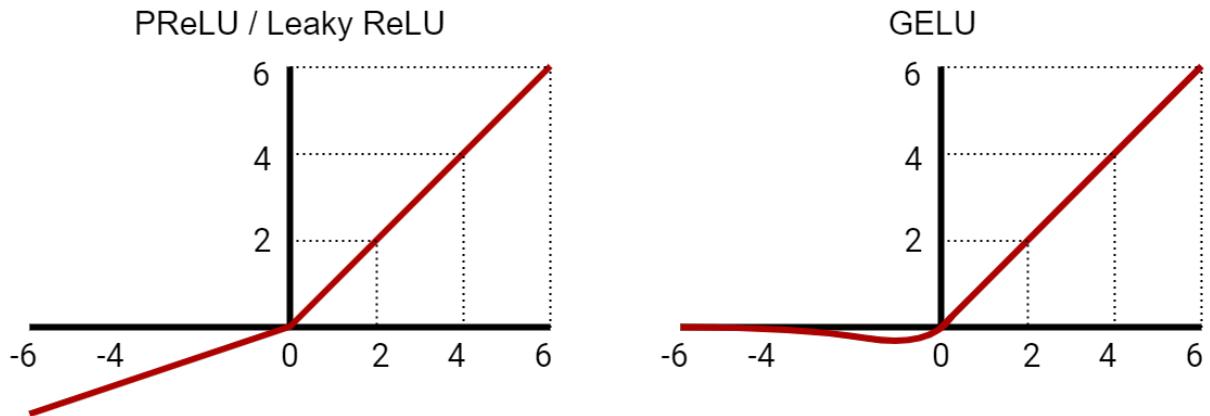


Fig. 17. – Courbe de la PReLU / Leaky ReLU et de la GELU.

Description du diagramme: La PReLU et la Leaky ReLU ont la même forme selon le paramètre qu'on leur inclus en code

Exemple d'implémentation de ces fonctions d'activations en PyTorch :

```
prelu = nn.PReLU(num_parameters=1, init=0.25)
leaky_relu = nn.LeakyReLU(negative_slope=0.01)
gelu = nn.GELU()
```

3.2.6 Fonction d'activation Softmax

La fonction d'activation softmax est fondamentale en deep learning, particulièrement dans la couche de sortie des tâches de classification multiclasse. Elle représente une généralisation de la fonction logistique. Le softmax « compresse » un vecteur z de dimension K , composé de valeurs réelles, en un autre vecteur de même dimension dont les valeurs, comprises entre 0 et 1, s'additionnent pour donner 1. Autrement dit, la fonction softmax permet de convertir des scores bruts en probabilités, facilitant ainsi l'interprétation des prédictions du modèle.

La formule mathématique de la fonction softmax est la suivante

$$s(z)_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}} \quad \text{pour } j = 1, \dots, K$$

```
def softmax(z):
    z_exp = np.exp(z)
    z_sum = np.sum(z_exp, axis=0, keepdims=True)
    return z_exp / z_sum
```

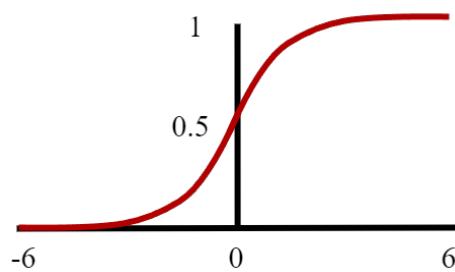


Fig. 18. – Courbe de la fonction d'activation softmax.

Dans cette formule plus sophistiquée que les précédentes, z est un vecteur de scores de dimension K . Pour chaque score z_j du vecteur z (j est un itérateur allant de 1 à K), $s(z)_j$ est la probabilité correspondante obtenue par la fonction softmax. La partie supérieure de la fraction, e^{z_j} est l'exponentiel du score z_j . La partie inférieure de la fraction, $\sum_{k=1}^K e^{z_k}$, est la somme des exponentielles de tous les scores dans le vecteur z . Tous les scores sont normalisés par la somme des exponentielles pour garantir que la somme des probabilités soit égale à 1.

Imaginons que nous avons un réseau neuronal qui est utilisé pour classer une image dans l'une de ces quatre catégories : chat, chien, oiseau ou poisson. Après avoir passé l'image par le réseau neuronal

$$z = \begin{pmatrix} 2.0 \\ 1.0 \\ -1.0 \\ 3.0 \end{pmatrix} \text{ où chaque score correspond à chaque catégorie dans cet ordre}$$

La fonction softmax transforme ces scores en probabilités. Pour le premier score (pour « chat »), la probabilité correspondante serait :

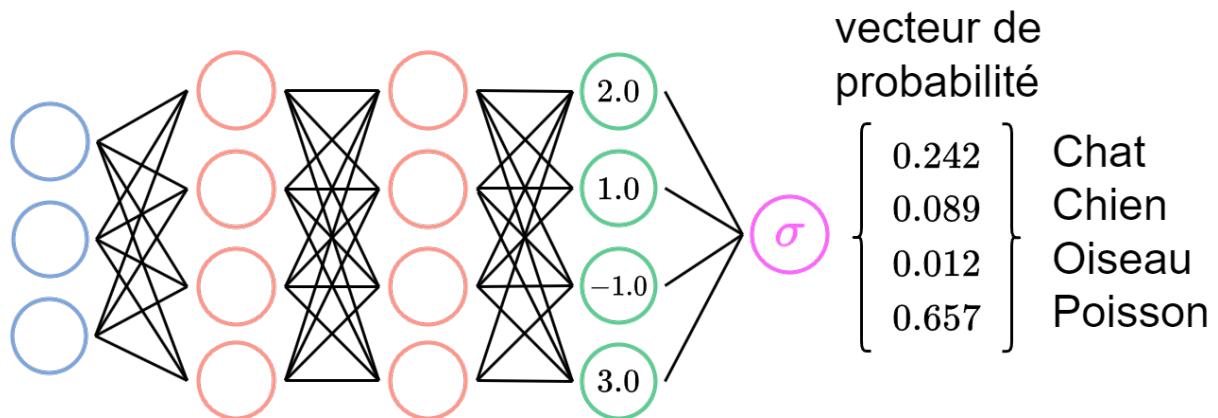


Fig. 19. – Réseau de neurones qui donne sa confiance avec une softmax comme fonction d'activation.

Description du diagramme: Le réseau de neurone donne un indicateur de sa confiance envers plusieurs catégories la somme des probabilités est de 1, ici le réseau de neurone pense à 65,7% que l'image est une image de poisson.

```
import math
# Notre vecteur de scores (logits)
z = [2.0, 1.0, -1.0, 3.0]
# Calculons les exponentielles de tous les scores
z_exp = [math.exp(i) for i in z]
# Somme de toutes les exponentielles
sum_z_exp = sum(z_exp)
# Application de la fonction softmax pour obtenir les probabilités
softmax = [round(i / sum_z_exp, 3) for i in z_exp]
print(softmax)
```

L'utilisation de la fonction softmax est recommandée principalement pour les problèmes de classification multiclass où chaque instance peut appartenir une seule classe parmi plusieurs.

Elle est utile dans ces cas, car elle donne une mesure de la confiance du modèle pour chaque classe possible pour une instance donnée.

$$\textcolor{violet}{s}(z)_j = \frac{e^{z_{\text{chat}}}}{\sum_{k=1}^K e^{z_k}}$$

Où $e^{z_{\text{chat}}}$ est l'exponentielle du score pour « chat », et $\sum_{k=1}^K e^{z_k}$ est la somme des exponentielles de tous les scores. En subsistant les valeurs concrètes dans cette formule, nous obtenons :

$$\textcolor{violet}{s}(z)_{\text{chat}} = \frac{e^{2.0}}{e^{2.0} + e^{1.0} + e^{-1.0} + e^{3.0}}$$

$$\textcolor{violet}{s}(z)_{\text{chat}} = \frac{7.389}{7.389 + 2.718 + 0.368 + 20.086}$$

$$\textcolor{violet}{s}(z)_{\text{chat}} = \frac{7.389}{30.561}$$

$$\textcolor{violet}{s}(z)_{\text{chat}} = 0.242$$

Cela signifie que, par exemple, la probabilité associée à la première valeur qui serait un chat de notre vecteur z est de 24,2%. De même, la probabilité associée à la deuxième valeur est 0.089 ou 8.9% et ainsi de suite. Toutes ces probabilités s'additionnent pour donner 1, ce qui est une propriété des probabilités.

3.2.7 Normalisation de la Sortie

Enfin, les fonctions d'activation peuvent aider à normaliser la sortie d'un neurone. Par exemple, la fonction d'activation softmax, souvent utilisée dans la couche de sortie des réseaux de neurones pour la classification multi-classes, transforme la sortie pondérée de chaque neurone en une probabilité entre 0 et 1. La somme de toutes ces probabilités est 1, ce qui permet d'interpréter chaque probabilité comme la confiance du réseau dans chaque classe possible.

3.3 Le Biais dans un Perceptron

Le biais est un terme ajouté dans une équation linéaire qui permet de décaler la sortie du modèle par une constante (une simple addition). En termes simples, le biais est intercepté dans une équation linéaire, c'est-à-dire le point où la ligne touche l'axe des ordonnées lorsque toutes les entrées sont à zéro.

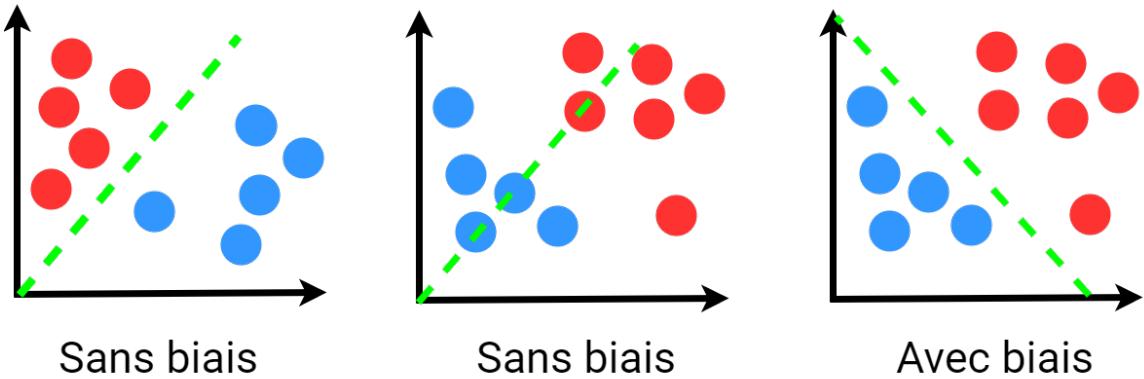


Fig. 20. – Diagramme de classification d'un perceptron avec et sans biais.

Description du diagramme: Dans le premier plot sans biais, le perceptron démarre à l'origine (le point où $x = 0$ et $y = 0$) arrive à bien séparé la classe rouge et la classe bleue. Dans le deuxième, il n'y arrive plus du tout, car la distribution ne lui permet pas à partir de l'origine, sur le troisième, il y arrive grâce au biais.

Le biais permet de donner de la flexibilité à notre perceptron et donc à notre modèle pour s'adapter aux données. Il permet de décaler la fonction d'activation. C'est essentiel parce qu'il donne au perceptron la capacité de produire une variété de sorties, même pour les mêmes entrées. Il permet à la fonction d'activation, par exemple une fonction sigmoïde, de se déplacer vers la gauche ou la droite, ce qui aide le modèle à s'adapter à différentes gammes de valeurs d'entrée. Sachez-que ce paramètre est activé par défaut sur PyTorch, vous n'aurez pas à le dire explicitement de l'inclure.

Cependant, il est intéressant de noter que dans certaines rares situations, les ingénieurs d'architecture de réseaux de neurones peuvent choisir de ne pas inclure de biais dans certaines couches cachées. Généralement des architectures de type CNN, les architectures qui servent à lire des images. Cela se produit souvent dans des situations où les données d'entrée sont centrées autour de zéro, et donc le biais n'apporte pas beaucoup de bénéfice.

Jusqu'à maintenant, nous calculons la sortie d'un perceptron $w^T x$ si nous ajoutons le biais cela donne $w^T x + b$ ce qui transforme notre perceptron de fonction linéaire à une fonction affine.

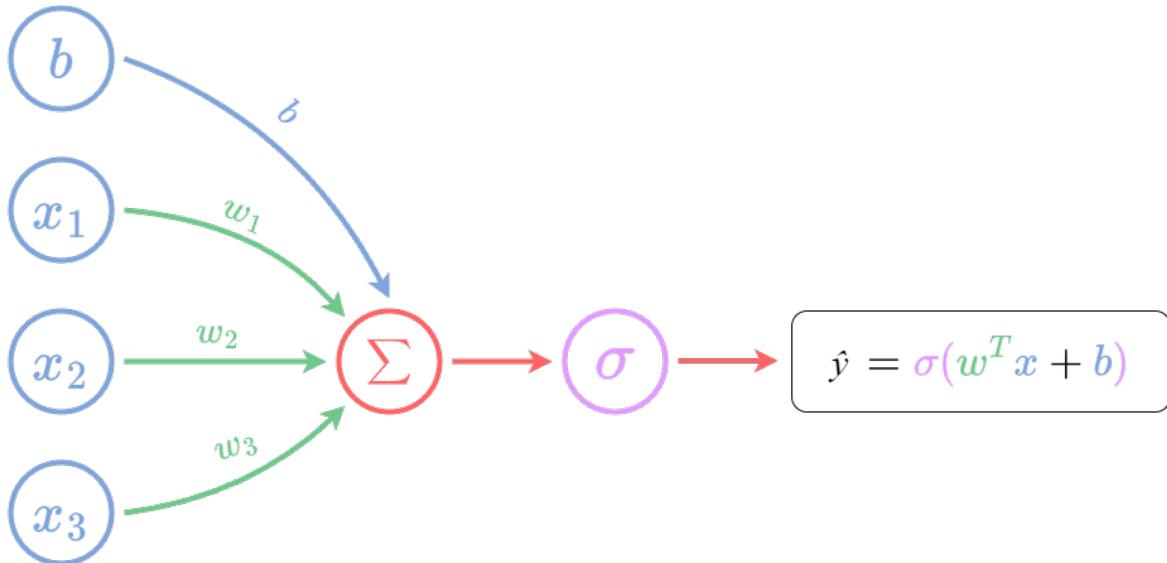


Fig. 21. – Ajout d'un biais au perceptron

3.4 Le Perceptron multicouches (MLP)

Un Perceptron multicouche, ou MLP en anglais pour « Multi-Layer Perceptron », correspond à ce que j'ai désigné par « réseau de neurones » précédemment. Il est composé d'une couche d'entrée, de couches cachées et d'une couche de sortie. En somme, il s'agit d'une architecture de réseau de neurones constituée de nombreux perceptrons. Le terme « MLP » est parfois utilisé pour évoquer une ou plusieurs couches de perceptrons.

3.4.1 Propagation avant (feed-forward)

Un réseau de neurones feed-forward, est un réseau où l'information (les sorties des perceptrons) se déplace uniquement dans une seule direction, de l'entrée vers la sortie sans boucle ni cycle, il existe des architectures de réseaux de neurones où l'information peut circuler dans les deux sens comme une architecture de type RNN pour (Recurrent Neural Network), ce genre d'architecture est utilisé pour traiter des données textuelles.

Une caractéristique des réseaux de neurones feed-forward est l'absence de connexion entre les neurones au sein de la même couche. Chaque neurone d'une couche est connecté à tous les neurones de la couche suivante.

Chaque neurone reçoit des valeurs de la couche précédente et somme ces valeurs, à cela est ajouté le biais. La fonction d'activation est ensuite appliquée à cette somme pour produire la sortie du neurone. Cette sortie est transmise à tous les neurones de la couche suivante. Cela est répété pour chaque couche jusqu'à ce que la couche de sortie soit atteinte.

Nous allons analyser le fonctionnement d'un réseau feed-forward qui souhaiterait prédire la taille d'une personne à partir de deux variables, son poids et son âge, calculer à la main serait fastidieux alors faisons le en python avec numpy, pour simplifier les calculs les fonctions d'activation seront des ReLU.

```
import numpy as np

def relu(x):
    return np.maximum(0, x)
```

```

# Initialiser les poids et les biais du réseau
weights = {
    'hidden_1': np.array([[0.1, 0.2, 0.3], [0.4, 0.5, 0.6]]),
    'hidden_2': np.array([[0.1, 0.2, 0.3], [0.4, 0.5, 0.6], [0.7, 0.8, 0.9]]),
    'output': np.array([[0.1], [0.2], [0.3]])
}
biases = {
    'hidden_1': np.array([0.1, 0.2, 0.3]),
    'hidden_2': np.array([0.1, 0.2, 0.3]),
    'output': np.array([0.1])
}

def feed_forward(inputs):
    # Calculer l'activation de la première couche cachée
    hidden_sum_1 = np.dot(inputs, weights['hidden_1']) + biases['hidden_1']
    hidden_activation_1 = relu(hidden_sum_1)
    print(f"Valeur de la somme pondérée dans la première couche cachée: {hidden_sum_1}")
    print(f"Valeur de l'activation dans la première couche cachée: {hidden_activation_1}")

    # Calculer l'activation de la deuxième couche cachée
    hidden_sum_2 = np.dot(hidden_activation_1, weights['hidden_2']) + biases['hidden_2']
    hidden_activation_2 = relu(hidden_sum_2)
    print(f"Valeur de la somme pondérée dans la deuxième couche cachée: {hidden_sum_2}")
    print(f"Valeur de l'activation dans la deuxième couche cachée: {hidden_activation_2}")

    # Calculer l'activation de la couche de sortie
    output_sum = np.dot(hidden_activation_2, weights['output']) + biases['output']
    output_activation = relu(output_sum)
    print(f"Valeur de la somme pondérée dans la couche de sortie: {output_sum}")
    print(f"Valeur de l'activation dans la couche de sortie: {output_activation}")

# Les données d'entrée pour l'exemple
# Poids = 70 kg, Âge = 25 ans
inputs = np.array([70, 25])
feed_forward(inputs)

# output :
>>> Valeur de la somme pondérée dans la première couche cachée: [17.1 26.7 36.3]
>>> Valeur de l'activation dans la première couche cachée: [17.1 26.7 36.3]
>>> Valeur de la somme pondérée dans la deuxième couche cachée: [37.9 46.01 54.12]

```

```

>>> Valeur de l`activation dans la deuxième couche cachée: [37.9 46.01
54.12]
>>> Valeur de la somme pondérée dans la couche de sortie: [29.328]
>>> Valeur de l`activation dans la couche de sortie: [29.328]

```

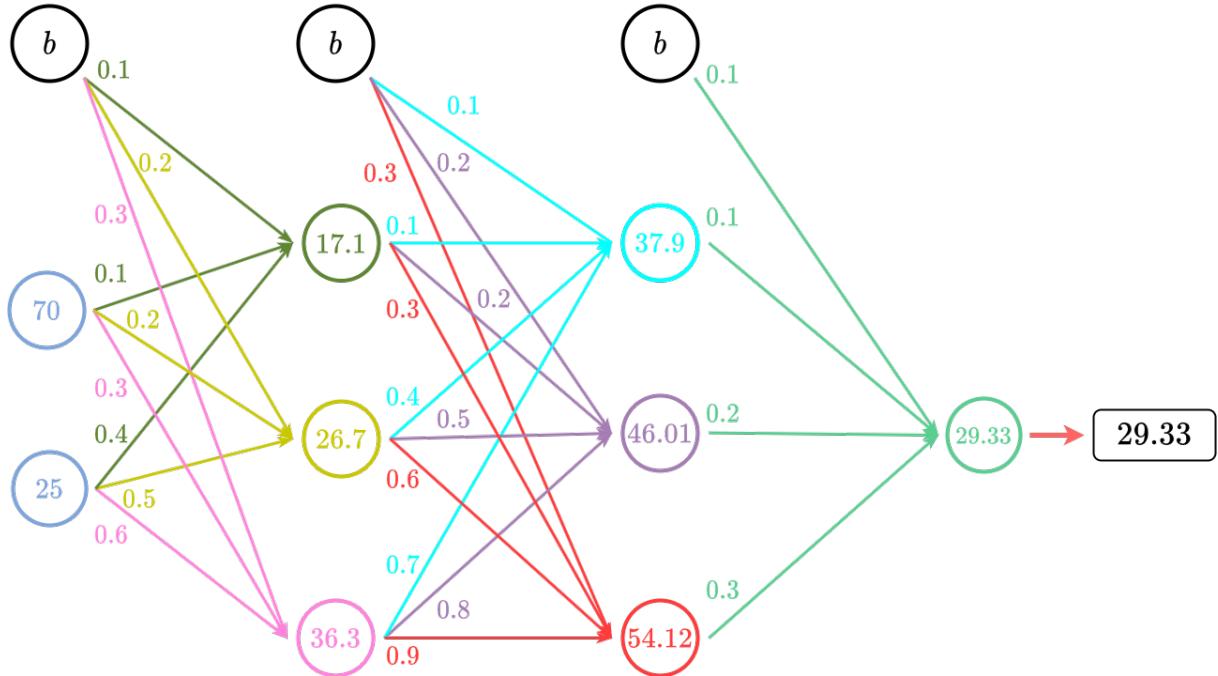


Fig. 22. – Réseau à perceptron multicouche feed-forward

Ça fait beaucoup de valeur, j'ai essayé de jouer sur les couleurs pour vous aider à y voir plus clair. Je ne vais pas faire à la main tous les calculs, mais quelques un pour que vous y voyiez quand même la logique sous-jacente. Les biais sont représentés du même design que les perceptrons, mais ils n'ont pas de valeurs d'entrée, le biais n'est qu'une addition, je vous rappelle la formule mathématique du calcul d'un perceptron $w^T x + b$ où w et x sont des vecteurs, cette formule multiplie les entrées par leur poids et additionne leur biais propre à chaque perceptron. Chaque perceptron contient une fonction d'activation ReLU, si la valeur de sortie du perceptron est négative alors la valeur sera à nulle (0).

Faisons le calcul pour un perceptron de la première couche de neurone cachée:

On multiplie chaque entrée par son poids : $70 \times 0.1 + 25 \times 0.4 = 7 + 10 = 17$ Ensuite, on ajoute le biais (0.1) pour obtenir la somme pondérée : $17 + 0.1 = 17.1$

Faisons le calcul pour 37.9 de la deuxième couche cachée.

Les valeurs obtenues de la première couche cachées sont $\begin{pmatrix} 17.1 \\ 26.7 \\ 36.3 \end{pmatrix}$ les poids pour la deuxième couche sont $\begin{pmatrix} 0.1 & 0.2 & 0.3 \\ 0.4 & 0.5 & 0.6 \\ 0.7 & 0.8 & 0.9 \end{pmatrix}$ et les biais sont $\begin{pmatrix} 0.1 \\ 0.2 \\ 0.3 \end{pmatrix}$.

On multiplie chaque entrée par son poids correspondant : $17.1 \times 0.1 + 26.7 \times 0.4 + 36.3 \times 0.7 = 1.71 + 10.68 + 25.41 = 37.8$. Ensuite, on ajoute le biais (0.1) pour obtenir la somme pondérée : $37.8 + 0.1 = 37.9$ après avoir obtenu la somme pondérée de 37.9 si la somme

pondérée obtenue est supérieure à zéro, alors la sortie est égale à la somme pondérée elle-même, sinon, la sortie est fixée à zéro. Dans notre cas, puisque la somme pondérée est égale à 37.9, qui est supérieure à zéro, la sortie de la fonction ReLU sera également égale à 37.9.

Ensuite pour obtenir la valeur de la couche de sortie 29.328 nous refaisons pareil, on utilise les valeurs des couches précédentes, $\begin{pmatrix} 37.9 \\ 46.01 \\ 54.12 \end{pmatrix}$ ainsi que les poids $\begin{pmatrix} 0.1 \\ 0.2 \\ 0.3 \end{pmatrix}$ et le biais pour la couche de sortie qui sera de 0.1

On multiplie chaque entrée par son poids correspondant : $37.9 \times 0.1 + 46.01 \times 0.2 + 54.12 \times 0.3 = 3.79 + 9.202 + 16.236 = 29.228$. Ensuite on ajoute le biais (0.1) pour obtenir la somme pondérée : $29.228 + 0.1 = 29.328$ puis nous utilisons toujours la fonction d'activation ReLU, la somme pondérée est positive alors la valeur ne change pas.

Maintenant, cette sortie doit-elle signifier quelque chose ? Pas vraiment, c'était la première feed-forward pass; les poids et les biais ont été définis arbitrairement sans se baser sur les données réelles. Dans une situation réelle, ces paramètres seraient ajustés en utilisant une méthode d'apprentissage comme la descente de gradient, que nous allons voir dans le prochain chapitre.

3.5 Le théorème d'approximation universelle

Le théorème d'approximation universelle est un théorème mathématique, nous ne nous attarderons pas sur sa preuve détaillée ici. Si vous êtes curieux de la preuve mathématique, je vous suggère de consulter le livre Deep Math [2] rédigé par Arnaud Bodin et François Recher.

En termes simples, ce théorème affirme qu'un réseau de neurones feed-forward, à condition d'avoir un nombre suffisant de couches cachées et de largeur adéquate, a la capacité d'approximer n'importe quelle fonction continue sur des ensembles fermés et bornés avec une précision quasi parfaite.

Si vous voulez modéliser une fonction mathématique complexe. Il pourrait s'agir de la modélisation de la trajectoire d'une fusée, de la prédiction de la météo, de la prédiction du prix des actions sur le marché boursier, ou de tout autre problème complexe que vous pouvez imaginer. Le théorème d'approximation universelle vous assure qu'il existe un réseau de neurones capable d'apprendre cette fonction à une précision satisfaisante.

Néanmoins, bien que le théorème d'approximation universelle offre une garantie théorique de la capacité des réseaux de neurones à apprendre une variété de fonctions, il ne donne pas de garanties pratiques. Il ne spécifie pas combien de neurones pourraient être nécessaires, comment ils doivent être organisés, ou comment les paramètres du réseau (les poids) peuvent être appris. De plus, le théorème ne traite pas de l'optimisation des paramètres du réseau, qui est une question majeure dans la pratique de l'apprentissage profond.

En dépit de ces limites, le théorème d'approximation universelle reste une pierre angulaire de la théorie du deep learning. Il fournit une justification théorique pour l'usage généralisé des réseaux de neurones pour une large gamme de tâches d'approximation de fonctions, et

il donne une indication que l'augmentation de la taille ou de la profondeur d'un réseau de neurones peut permettre de mieux approximer certaines fonctions.

3.6 Résumé

Nous avons vu ce qu'était un perceptron, qu'il est doté d'un ensemble d'entrée x et que chaque entrée a une pondération par des poids spécifiques, qui déterminent l'importance relative de chaque entrée. Un biais est également ajouté pour ajuster la sortie indépendamment des entrées.

L'importance des poids est illustrée par des exemples de calculs. Une entrée avec un poids élevé influence davantage le perceptron, ce qui est crucial pour le « feature engineering » en deep learning, où il est important de sélectionner les variables les plus pertinentes pour le problème à résoudre.

Ensuite, l'importance du biais dans un perceptron est discutée. Le biais est un terme constant qui permet de décaler la sortie du modèle, ce qui donne au perceptron la flexibilité nécessaire pour s'adapter à différentes données. Il permet de déplacer la fonction d'activation vers la gauche ou la droite, aidant ainsi le modèle à s'adapter à différentes gammes de valeurs d'entrée.

Nous avons exploré l'importance des fonctions d'activation pour les perceptrons. Ces fonctions transforment la sortie pondérée de la somme des entrées d'un perceptron introduisant des non-linéarités dans le modèle, ce qui permet au réseau d'apprendre des relations complexes.

Nous avons vu 4 fonctions d'activation :

- La fonction sigmoïde, une fonction couramment utilisée qui produit une sortie entre 0 et 1. Cependant, elle peut causer des problèmes de saturation pour les valeurs d'entrée extrêmes, ralentissant l'apprentissage.
- La fonction tangente hyperbolique (tanh), qui produit une sortie entre -1 et 1. Cette fonction est également sujette au problème de vanishing gradient lors de l'apprentissage.
- La fonction d'activation de Rectification (ReLU), qui est largement utilisée en raison de sa simplicité de calcul et de sa robustesse au vanishing gradient. Néanmoins, elle a le défaut du « dying ReLU » qui désactive certains neurones.
- La fonction d'activation Softmax, qui est principalement utilisée dans la couche de sortie pour les tâches de classification multiclass. Elle transforme les scores bruts en probabilités, facilitant l'interprétation des résultats.

Ensuite, nous avons vu le processus de propagation avant (feed-forward) dans les réseaux de neurones. Dans un réseau feed-forward, l'information (les sorties des perceptrons) se déplace uniquement dans une seule direction, de l'entrée vers la sortie, sans boucle ni cycle. Chaque neurone d'une couche est connecté à tous les neurones de la couche suivante, et il n'y a pas de connexion entre les neurones au sein de la même couche.

Le chapitre se conclut par une discussion sur le théorème d'approximation universelle, qui stipule que, théoriquement, un réseau de neurones feed-forward peut approximer n'importe quelle fonction continue à une précision arbitrairement petite, tant qu'il est suffisamment large et a suffisamment de couches cachées. Cependant, bien que ce théorème offre une garantie théorique de la capacité des réseaux de neurones, il ne donne pas de garanties pratiques.

3.7 Question

1. Qu'est-ce qu'un perceptron multicouche (MLP) ?

- a. Un type d'algorithmes de machine learning utilisé pour la classification binaire.
- b. Un type de réseau de neurones constitué d'une couche d'entrée, d'une ou plusieurs couches cachées, et d'une couche de sortie.
- c. Une fonction d'activation utilisée dans les réseaux de neurones.
- d. Un type de réseau de neurones où l'information se déplace dans les deux sens.

2. Qu'est-ce qu'un réseau de neurones feed-forward ?

- a. Un réseau de neurones où l'information peut circuler dans les deux sens.
- b. Un réseau de neurones où l'information se déplace uniquement dans une seule direction, de l'entrée à la sortie.
- c. Un type de fonction d'activation utilisée dans les réseaux de neurones.
- d. Un type de réseau de neurones utilisé pour le traitement du texte.

3. Comment sont connectés les neurones dans un réseau de neurones feed-forward ?

- a. Les neurones de la même couche sont connectés entre eux.
- b. Les neurones de chaque couche sont connectés à tous les neurones de la couche suivante.
- c. Les neurones de chaque couche sont connectés à tous les neurones de la couche précédente.
- d. Les neurones de chaque couche sont connectés à tous les neurones des autres couches.

4. Pourquoi avons-nous besoin de biais dans un réseau de neurones?

- A. Pour garantir que chaque neurone ait une sortie différente de zéro
- B. Pour ajuster la sortie du neurone indépendamment de ses entrées
- C. Pour garantir que les poids soient toujours positifs
- D. Pour augmenter la capacité du réseau à mémoriser les données d'entrée

5. Quelle est l'expression pour le calcul de la somme pondérée dans une couche cachée d'un réseau de neurones ?

- a. $\text{hidden_sum} = X \cdot W_{\text{hidden}} + b_{\text{hidden}}$
- b. $\text{hidden_sum} = W_{\text{hidden}} \cdot X + b_{\text{hidden}}$
- c. $\text{hidden_sum} = X + W_{\text{hidden}} + b_{\text{hidden}}$
- d. $\text{hidden_sum} = \frac{X}{W_{\text{hidden}}} + b_{\text{hidden}}$

6. A quoi sert une fonction d'activation ?

- a. Elle sert à introduire de la non-linéarité dans le modèle, permettant ainsi d'apprendre des relations complexes entre les entrées et les sorties.
- b. Elle sert à normaliser les entrées du modèle, en les transformant en valeurs entre 0 et 1.
- c. Elle sert à augmenter la vitesse de l'apprentissage du modèle, en accélérant la convergence de l'algorithme d'optimisation.
- d. Elle sert à réduire le nombre de paramètres du modèle, en introduisant des contraintes de parcimonie.

7. A quoi sert la non-linéarité en deep learning ?

- a. Elle permet de gérer les problèmes de vanishing et exploding gradients.

- b. Elle permet au modèle de capturer des relations complexes et non-linéaires entre les entrées et les sorties.
- c. Elle est utile pour accélérer la convergence de l'algorithme d'optimisation.
- d. Elle aide à réduire le surapprentissage en introduisant de l'irrégularité dans le modèle.

8. Quel est l'un des principaux avantages de la fonction d'activation ReLU ?

- a. Elle est complexe à calculer
- b. Elle est facile à calculer et sa dérivée est simple
- c. Elle produit une sortie constante
- d. Elle produit une sortie négative

Réponse : b

9. Quelle est la formule mathématique de la fonction d'activation ReLU ?

- a. $\sigma(x) = \min(0,x)$
- b. $\sigma(x) = \max(0,x)$
- c. $\sigma(x) = x^2$
- d. $\sigma(x) = x/2$

10. Quelle est la fonction principale de la fonction d'activation softmax ?

- a. Elle produit une sortie négative
- b. Elle produit une sortie constante
- c. Elle transforme les scores bruts en probabilités
- d. Elle produit une sortie binaire

11. Qu'est-ce que la « normalisation de sortie » d'une fonction d'activation ?

- a. Transformer la sortie d'un neurone en une probabilité entre 0 et 1
- b. Transformer la sortie d'un neurone en une valeur constante
- c. Transformer la sortie d'un neurone en une valeur négative
- d. Transformer la sortie d'un neurone en une valeur positive

12. Pourquoi la fonction d'activation Tanh est-elle parfois préférée à la fonction Sigmoide dans les couches cachées d'un réseau de neurones ?

- a. Parce que la Tanh a une sortie plus grande
- b. Parce que la Tanh est plus simple à calculer
- c. Parce que la Tanh produit des sorties centrées à zéro

13. Imaginez que vous avez un perceptron simple avec les poids [0.2, -0.5] et le biais est 0.1. Les entrées sont [2, 3]. La fonction d'activation est ReLU. Comment calculez-vous la sortie de ce perceptron ?

- a. $(0.2 \times 2 + -0.5 \times 3) + 0.1 = -0.9$, puis appliquez ReLU, donc la sortie est 0
- b. $(0.2 \times 2 + -0.5 \times 3) \times 0.1 = -0.09$, puis appliquez ReLU, donc la sortie est 0
- c. $(0.2 \times 2 - 0.5 \times 3) \times 0.1 = -0.09$, puis appliquez ReLU, donc la sortie est 0.09
- d. $(0.2 \times 2 + -0.5 \times 3) + 0.1 = -0.9$, puis appliquez ReLU, donc la sortie est -0.9

14. Lors l'exemple d'un réseau feed-forward de perceptron du chapitre 5.4 la tâche était de définir la taille d'une personne à partir de son poids et de son âge, ceci était une tâche de

- a. classification

- b. regression
 - c. les deux
15. **À quoi sert la fonction softmax dans la couche de sortie d'un réseau neuronal ?**
- a. Pour la classification binaire
 - b. Pour la classification multiclassée
 - c. Pour la régression
 - d. Pour l'apprentissage non supervisé

3.7.1 Correction

1. Qu'est-ce qu'un perceptron multicouche (MLP) ?

Réponse: b Un type de réseau de neurones constitué d'une couche d'entrée, d'une ou plusieurs couches cachées, et d'une couche de sortie.

2. Qu'est-ce qu'un réseau de neurones feed-forward ?

Réponse: b Un réseau de neurones où l'information se déplace uniquement dans une seule direction, de l'entrée à la sortie.

3. Comment sont connectés les neurones dans un réseau de neurones feed-forward ?

Réponse: b Un réseau de neurones où l'information se déplace uniquement dans une seule direction, de l'entrée à la sortie.

4. Pourquoi avons-nous besoin d'un biais dans un réseau de neurones?

Réponse: b Pour ajuster la sortie du neurone indépendamment de ses entrées

5. Quelle est l'expression pour le calcul de la somme pondérée dans une couche cachée d'un réseau de neurones ?

Réponse: a $\text{hidden_sum} = X \cdot W_{\text{hidden}} + b_{\text{hidden}}$

6. A quoi sert une fonction d'activation ?

Réponse: a Elle sert à introduire de la non-linéarité dans le modèle, permettant ainsi d'apprendre des relations complexes entre les entrées et les sorties.

7. A quoi sert la non-linéarité en deep learning ?

Réponse: b Elle permet au modèle de capturer des relations complexes et non-linéaires entre les entrées et les sorties

8. Quel est l'un des principaux avantages de la fonction d'activation ReLU ?

Réponse: b Elle est facile à calculer et sa dérivée est simple

9. Quelle est la formule mathématique de la fonction d'activation ReLU ?

Réponse: b $\sigma(x) = \max(0, x)$

10. Quelle est la fonction principale de la fonction d'activation softmax ?

Réponse: c Elle transforme les scores bruts en probabilités

11. Qu'est-ce que la « normalisation de sortie » d'une fonction d'activation ?

Réponse: a Transformer la sortie d'un neurone en une probabilité entre 0 et 1

12. Pourquoi la fonction d'activation Tanh est-elle parfois préférée à la fonction Sigmoïde dans les couches cachées d'un réseau de neurones ?

Réponse: c Parce que la Tanh produit des sorties centrées à zéro

13. Imaginez que vous avez un perceptron simple avec les poids [0.2, -0.5] et le biais est 0.1. Les entrées sont [2, 3]. La fonction d'activation est ReLU. Comment calculez-vous la sortie de ce perceptron ?

Réponse: a $(0.2 \times 2 + -0.5 \times 3) + 0.1 = -0.9$, puis appliquez ReLU, donc la sortie est 0

14. Lors l'exemple d'un réseau feed-forward de perceptron du chapitre 5.4 la tâche était de définir la taille d'une personne à partir de son poids et de son âge, ceci

étaient une tâche de

Réponse: b Regression

15. **À quoi sert la fonction softmax dans la couche de sortie d'un réseau neuronal ?**

Réponse: b Pour la classification multiclassse

4 Processus d'apprentissage avec la rétropropagation (Back-propagation)

Dans le chapitre précédent, nous avons appris ce qu'est un réseau de neurones, et surtout comment il fait ses prédictions à partir de données. Cela soulève une question, comment le modèle ajuste un ensemble de paramètres pour faire des prédictions exact ? La réponse à cette question est l'objet principal de ce chapitre, l'apprentissage.

Les meilleurs algorithmes de deep learning utilise l'apprentissage supervisé (supervised learning) comme paradigme où le modèle apprend à prédire une sortie à partir d'exemples d'entrée et de sortie labellisée. Dans ce chapitre, nous ne parlerons pas des paradigmes de l'apprentissage par renforcement (reinforcement learning) et de l'apprentissage non supervisé (unsupervised learning).

Comprendre le processus d'apprentissage d'un modèle de deep learning vous enlèvera une grosse partie « blackbox » sur le fonctionnement du deep learning. Ce chapitre a pour but de vous donner une bonne intuition de l'apprentissage et de ses problématiques. Nous allons explorer en détail les composantes clés de ce processus, notamment la fonction de perte, la descente de gradient et la backpropagation. Ces composantes permettent l'ajustement des paramètres de notre modèle et améliorent sa capacité à faire des prédictions précises à partir des données d'entraînement.

4.1 Composantes clés du processus d'apprentissage

Imaginons l'apprentissage d'un modèle de deep learning comme une traversée en voiture dans une ville inconnue à la recherche d'une destination précise – disons, un parking. Dans ce contexte, quatre éléments essentiels nous aideront à naviguer : les données, la fonction de coût, la descente de gradient et la rétropropagation.

Pour commencer, les données constituent le paysage urbain que nous explorons. Elles décrivent les routes, les immeubles, les sens de circulation, tous ces détails qui façonnent notre environnement de navigation. Les données alimentent notre véhicule, en fournissant à la fois le point de départ pour notre voyage et les repères nous permettant d'évaluer notre progression vers la destination.

C'est ici que la fonction de coût entre en jeu, agissant comme un senseur de distance qui nous indique à quel point nous sommes proches ou loin de notre parking. Chaque fois que nous faisons un mouvement – ou, dans le contexte du modèle, une prédition – la fonction de coût évalue la précision de ce mouvement en comparant notre position actuelle à la position du parking.

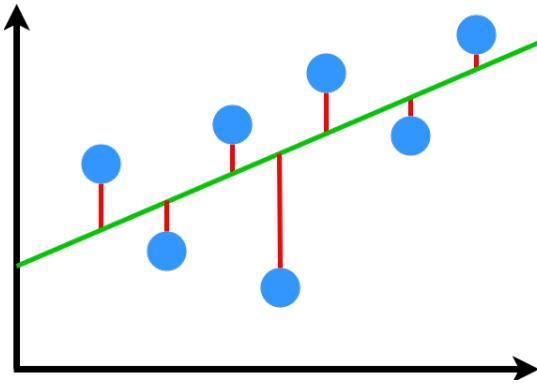


Fig. 23. – « Fonction de coût mesurant la « distance » (représentée en rouge) de chaque point par rapport à la prédition du modèle. »

Figure num_figure: Ce diagramme illustre la façon dont la fonction de coût mesure la « distance » (indiquée en rouge) entre notre position actuelle et la position du parking. Plus nous sommes loin du parking, plus la « distance » est grande.

Maintenant, alors que la fonction de coût nous indique si nous nous rapprochons ou nous éloignons du parking, elle ne nous dit pas quel chemin prendre. C'est là qu'interviennent la descente de gradient et la rétropropagation, qui travaillent ensemble comme un GPS sophistiqué.

La rétropropagation est la technologie sous-jacente de notre GPS, capable de retracer nos mouvements pour comprendre quelles décisions ont influencé notre distance au parking. Une fois que la fonction de coût a évalué notre « distance », la rétropropagation analyse cette « distance » et la distribue en arrière à travers notre itinéraire, pour comprendre comment chaque décision a contribué à l'écart actuel.

La descente de gradient, quant à elle, utilise les informations de la rétropropagation pour déterminer notre prochain mouvement. Elle examine tous les chemins possibles à chaque carrefour et choisit la direction qui semble réduire le plus notre « distance » au parking.

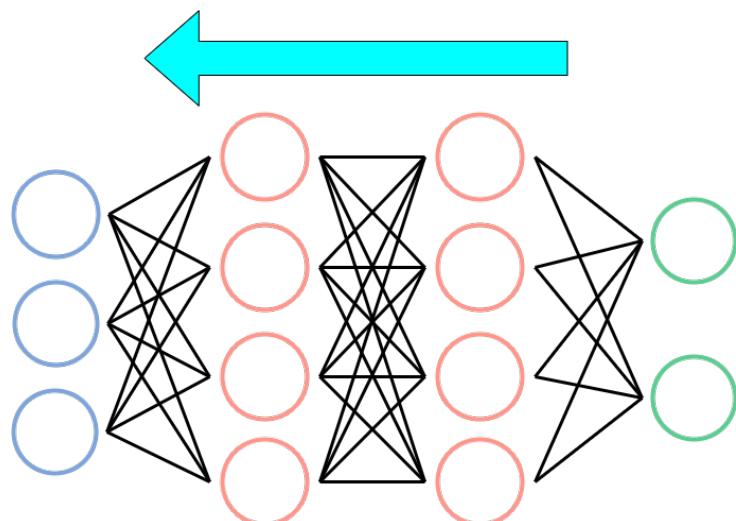


Fig. 24. – « Rétropropagation. Elle retrace le chemin de notre voyage, de notre position actuelle jusqu'à notre point de départ. »

En résumé, nos quatre composantes clés sont les données qui décrivent notre environnement, la fonction de coût qui mesure notre « distance » au parking, la rétropropagation qui nous aide à comprendre nos erreurs passées, et la descente de gradient qui nous guide vers notre destination. Tous ces éléments travaillent en symbiose pour minimiser notre « distance » au parking, nous rapprochant ainsi de notre objectif. Dans les sections suivantes, nous approfondirons chacune de ces composantes et leur rôle dans l'apprentissage d'un modèle de deep learning.

4.2 La fonction de perte (loss function)

4.2.1 Introduction à la fonction de perte (loss function)

La fonction de perte (loss function), également appelée fonction d'erreur (error function), est un élément indispensable du machine learning. Vous entendrez souvent les gens utiliser ces deux noms de fonctions de manières interchangeables. Elle est essentielle pour entraîner un modèle, car elle fournit une mesure quantitative de la performance du modèle. En termes simples, elle quantifie la différence entre la prédiction d'un modèle et la valeur réelle. La fonction de perte calcule donc l'erreur pour une seule instance d'apprentissage.

La fonction de coût est la moyenne des pertes pour les instances d'apprentissage. Elle résume en quelque sorte la performance globale du modèle sur l'ensemble des données d'entraînement. Cependant, le nom de fonction coût, fonction perte et fonction d'erreur sont fréquemment utilisés de manière interchangeable, le terme de fonction objectif apparaît parfois.

Au cœur de l'apprentissage en machine, l'objectif fondamental est de minimiser la fonction de perte. Cette minimisation est réalisée en trouvant un ensemble de poids (ou paramètres) w qui réduit au maximum l'erreur calculée par la fonction de perte.

Cette fonction de perte est essentielle, car elle compare les prédictions de l'algorithme aux valeurs réelles que nous cherchons à prédire. Autrement dit, elle mesure la divergence entre la sortie du modèle et la vérité terrain.

Pour illustrer, supposons que vous entraînez un algorithme pour prédire les températures futures à partir de données météorologiques. Pour chaque jour, l'algorithme produit une prédiction de la température, et la fonction de perte compare cette prédiction à la température réelle enregistrée ce jour-là. Plus la prédiction est précise, plus la valeur de la perte est faible, et vice versa.

Ce qu'il faut comprendre, c'est que la fonction de perte n'est pas une mesure absolue de l'exactitude ou de la précision. Elle est plutôt relative : sa valeur est utilisée pour comparer différentes prédictions et différents modèles entre eux. Une valeur de perte plus faible signifie simplement qu'un modèle ou une prédiction est meilleur(e) que d'autres selon le critère spécifique défini par la fonction de perte.

Le but de l'entraînement est de trouver un ensemble de paramètres qui minimise la valeur de la fonction de coût sur l'ensemble des données d'entraînement, en espérant que cela se généralisera à d'autres données inconnues. L'importance de la fonction de perte sert à guider notre algorithme d'apprentissage. Nous ne saurions pas quels paramètres produisent de bonnes prédictions et quels paramètres produisent de mauvaises prédictions.

4.2.2 Les différents types de fonction perte

Trouver la bonne fonction de perte se choisit selon si notre problématique est une régression ou une classification, ici sera présenté deux fonctions de perte pour des problématiques de régression et deux pour de la classification. Il en existe bien d'autre, choisir la bonne fonction perte est crucial pour le bon apprentissage de votre algorithme et peut-être souvent une fonction composée de plusieurs fonctions perte avec chacune une pondération.

4.2.2.1 L'Erreur Absolue Moyenne (MAE)

L'Erreur Absolue Moyenne, Mean Absolute Error (MAE) en anglais, ou encore L1 loss, est une métrique d'erreur. Cette mesure permet de quantifier l'erreur générée par un modèle de prédiction en calculant la moyenne des valeurs absolues des différences entre les prédictions du modèle et les valeurs réelles.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Où :

- n est le nombre total d'obersations ou de points de données
- y_i est la vérité terrain pour l'observation i
- \hat{y}_i est la valeur prédictée par le modèle pour l'observation i
- $|y_i - \hat{y}_i|$ est l'erreur absolue pour l'observation i

```
def MAE(y_true, y_pred):  
    return np.mean(np.abs(y_true - y_pred))
```

Chaque différence entre la valeur réelle y_i et la valeur prédictée \hat{y}_i est prise en valeur absolue, ce qui garantit que toutes les erreurs sont traitées de manière égale, qu'elles soient positives ou négatives. Ensuite, nous prenons la moyenne de ces erreurs absolues.

La MAE est particulièrement utile car elle peut être interprétée directement dans les unités de la variable que vous essayez de prédire. Par exemple, si vous prédisez les températures en degrés Celsius et que votre MAE est de 2, cela signifie que vos prédictions sont en moyenne à 2 degrés de la véritable température. Exemple le modèle prédit 16 degrés quand il en fait 18 en réalité.

La MAE traite toutes les valeurs de manière égale, qu'elles soient petites ou grandes à la différence de la MSE qui est au carré qui est sensible aux erreurs extrêmes.

4.2.2.2 La Fonction de perte Quadratique Moyenne (MSE)

L'erreur Quadratique Moyenne, Mean Squared Error (MSE) en anglais, parfois appelé L2 loss), est une autre métrique d'erreur, elle est une alternative à la MAE, elle s'applique, elle aussi, à des problématiques de régression, c'est-à-dire des tâches qui ne classifient pas un chien d'un chat par exemple. Elle mesure la moyenne des carrés des erreurs, c'est-à-dire la moyenne des différences au carré entre les valeurs prédictées et les valeurs réelles.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Où :

- n est le nombre total d'obersations ou de points de données
- y_i est la vérité terrain pour l'observation i
- \hat{y}_i est la valeur prédictée par le modèle pour l'observation i
- $(y_i - \hat{y}_i)^2$ est l'erreur au carré pour l'observation i

```
def MSE(y_true, y_pred):
    return np.mean((y_true - y_pred) ** 2)
```

Dans cette formule, chaque erreur est élevée au carré. Cela signifie que la MSE pénalise plus lourdement les grandes erreurs que les petites. Par conséquent, un modèle avec une MSE plus faible est un modèle qui a réussi à minimiser les grandes erreurs.

La MSE est employée comme fonction de perte en régression, grâce à sa propriété de différentiabilité. Pour comprendre ce que cela signifie, il faut d'abord comprendre ce qu'est une fonction différentiable. Une fonction est dite différentiable lorsqu'elle est lisse, sans cassures ni pointes, ce qui signifie qu'à tout point de cette fonction, nous pouvons tracer une tangente². En mathématiques, cette tangente correspond au taux de variation de la fonction à ce point précis et est représentée par la dérivée de la fonction.

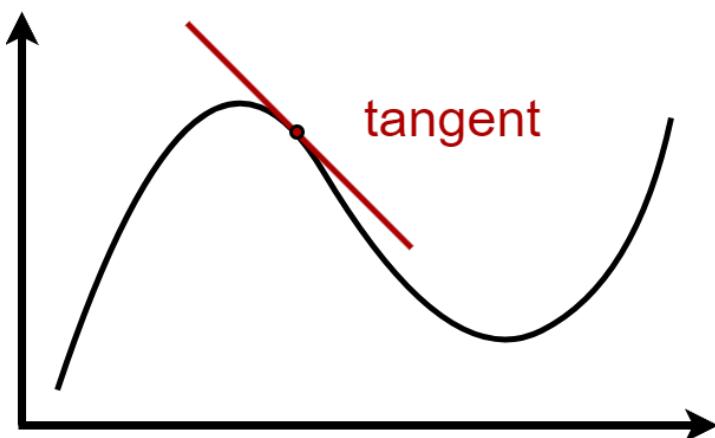


Fig. 25. – fonction différentiable en tout point avec une tangent.

Cette propriété est cruciale dans le contexte du deep learning. Les algorithmes d'optimisation tels que la descente de gradient stochastique³utilisent cette dérivée pour déterminer la direction à suivre afin de minimiser l'erreur de prédiction. Ils utilisent le principe que la dérivée est positive quand la fonction augmente, et négative quand elle diminue. En suivant cette direction, ils sont capables d'ajuster les paramètres du modèle pour réduire l'erreur, ce qui améliore la précision des prédictions.

La MSE a une particularité importante : elle donne un poids plus important aux erreurs importantes. Cela signifie que la MSE est plus sensible aux valeurs aberrantes (ou « outliers ») que d'autres métriques d'erreur comme la MAE.

Exemple : prenons le cas où nous avons un modèle de prédiction de température, qui a fait les prédictions suivantes pour cinq jours donnés :

- Prédiction : [15, 18, 20, 22, 25]

²J'ajouterais un schéma avec un exemple de fonction différentiable et un exemple de fonction non différentiable. Ex:fonction x^2 et $|x|$

³Attention !! Tu parles de la SGD trop tôt

Et voici les températures réelles qui ont été enregistrées ces jours-là :

- Vraies valeurs : [17, 20, 18, 21, 20]

Nous allons calculer la MAE et la MSE pour ces prédictions.

Calculons les erreurs absolues :

- Jour 1 : $|15 - 17| = 2$
- Jour 2 : $|18 - 20| = 2$
- Jour 3 : $|20 - 18| = 2$
- Jour 4 : $|22 - 21| = 1$
- Jour 5 : $|25 - 20| = 5$

Maintenant, prenons la moyenne de ces valeurs : $\text{MAE} = \frac{2+2+2+1+5}{5} = 2.4$

Calculons les erreurs au carré :

- Jour 1 : $(15 - 17)^2 = 4$
- Jour 2 : $(18 - 20)^2 = 4$
- Jour 3 : $(20 - 18)^2 = 4$
- Jour 4 : $(22 - 21)^2 = 1$
- Jour 5 : $(25 - 20)^2 = 25$

Maintenant, prenons la moyenne de ces valeurs $\text{MSE} = \frac{4+4+4+1+25}{5} = 7.6$

Si nous calculons la MAE et la MSE pour ces prédictions, nous constatons que :

- La MAE est de 2.4, ce qui signifie que les prédictions du modèle s'écartent en moyenne de 2.4 degrés de la réalité
- La MSE est de 7.6, une valeur supérieure à celle de la MAE, ce qui reflète la pénalisation plus sévère des grandes erreurs par la MSE par rapport à la MAE.

Ces fonctions de perte permettent d'évaluer la performance d'un modèle de prédiction et de quantifier l'importance des erreurs qu'il génère, chacune avec ses spécificités et ses avantages.⁴

4.2.2.3 Choisir entre la MSE et la MAE

Le choix entre ces deux fonctions de perte est important selon ce que l'on souhaite pénaliser. La MSE donne un poids plus important aux erreurs plus grandes, en les élevant au carré, tandis que la MAE traite toutes les erreurs de manière uniforme, en prenant simplement leur valeur absolue.

Par exemple, si notre modèle prédit que le temps de trajet d'un taxi, disons 15 minutes alors que trajet réel prend 45 minutes ! Cette erreur pourrait entraîner un client très mécontent. Il est adapté d'utiliser une MSE dans ce cas pour aider l'algorithme à ne pas faire de grosse erreur, et focaliser les efforts de changement des paramètres de notre modèle pour réduire la fréquence et l'ampleur de ces grandes erreurs.

4.2.2.4 Cross-Entropy Loss

La Cross-Entropy Loss, également connue sous le nom de Log Loss, est une fonction de perte utilisée pour les problèmes de classification. Ces termes sont souvent utilisés de manière

⁴Remettre une couche sur le fait que les métriques ont le même minimum mais compare juste deux prédictions de manière différente

interchangeable, bien que le terme Log Loss soit généralement employé pour les problèmes de classification binaire, tandis que la Cross-Entropy Loss englobe à la fois la classification binaire et multiclasses. Dans le contexte de la Cross-Entropy Loss, on utilise les termes « Binary Cross-Entropy » pour spécifier une classification binaire et « Categorical Cross-Entropy » pour une classification multiclasse.

Originaire de la théorie de l'information, la Cross-Entropy est une mesure essentielle utilisée pour entraîner les modèles de deep learning. Elle mesure la dissimilarité entre la distribution de probabilité prédite par le modèle et la vérité terrain. La Cross-Entropy punit sévèrement les prédictions confiantes mais incorrectes. Elle est donc préférée pour les problèmes de classification où une mauvaise classification confiante peut avoir un coût élevé comme la MSE face à la MAE pour les problèmes de régression. La Cross-Entropy a deux formules mathématiques dépendant du type de classification utilisé, à savoir la classification binaire et la classification multi-classes.

4.2.2.4.1 Binary Cross-Entropy Loss

Pour la classification binaire, où nous avons deux classes possibles (0 et 1), la formule est la suivante :

$$L_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

5

Dans cette formule, L_{BCE} est la loss de la la Binary Cross-Entropy, N est le nombre total d'observations, y_i est la « target » soit le vrai label à prédire pour chaque observation i , et \hat{y}_i est la probabilité prédite pour la classe 1 pour l'observation i ⁵. La somme est caculée sur toutes les observations.

Analysons chaque partie de la formule avant de l'utiliser :

- L_{BCE} : C'est la Binary cross entropy loss que nous essayons de calculer. C'est une mesure de la différence entre les probabilités prédites par le modèle \hat{y}_i et les étiquettes réelles y_i . Une valeur plus faible indique les prédictions du modèle sont proches des valeurs réelles.
- N est le nombre d'observations utilisé pour calculer la loss, appelé « batch size » comme les jeux de données sont trop volumineux, la loss calcule sur plusieurs fragments du jeu de données étape par étape. Dans la formule, elle est le diviseur pour permettre d'obtenir la moyenne de la loss sur tous les échantillons.
- $\sum_{i=1}^N$: c'est une somme sur tous les échantillons du « batch size » (une petite partie du jeu de données). Pour chaque échantillon i , nous calculons la contribution de cet échantillon à la loss totale et nous ajoutons toutes ces contributions pour obtenir la loss totale.
- y_i : C'est l'étiquette réelle de l'échantillon i . Pour une classification binaire, y_i est soit 0, soit 1.
- \hat{y}_i : C'est la probabilité prédite par le modèle que l'échantillon i appartienne à la classe positive (par exemple, que l'image soit celle d'un chat).

⁵Séparer la somme en deux (une somme sur les $y_i=0$ et une somme sur les $y_i=1$) Cela permet de visualiser plus facilement comment la fonction coût réagit aux erreurs de classification

⁶À intégrer que c'est une valeur comprise entre 0 et 1

- $\log(\hat{y}_i)$ et $\log(1 - \hat{y}_i)$: Le logarithme est utilisé pour amplifier l'effet des différences entre les probabilités prédites et les étiquettes réelles. Si la probabilité prédite est proche de l'étiquette réelle, la valeur du logarithme est proche de 0, ce qui contribue peu à la loss totale. En revanche, si la probabilité prédite est loin de l'étiquette réelle, la valeur du logarithme est un grand nombre négatif, ce qui contribue beaucoup à la loss totale.
- $y_i \cdot \log(\hat{y}_i)$: Ce terme mesure la contribution à la loss de la prédiction pour la classe négative. Si $y_i = 0$ (c'est-à-dire, si l'échantillon i n'appartient pas à la classe positive), alors cette contribution est $\log(1 - \hat{y}_i)$, qui est faible si \hat{y}_i est proche de 0 et grand si \hat{y}_i est proche de 1. Si $y_i = 1$ (autrement dit, si l'échantillon i appartient à la classe positive), alors cette contribution est 0. Car $1 - 1 = 0$

En résumé, la formule de la cross entropy loss mesure à quel point les probabilités prédites par le modèle sont proches des étiquettes réelles. Elle pénalise fortement les prédictions qui sont loin des étiquettes réelles, ce qui encourage le modèle à faire des prédictions précises. Le logarithme est utilisé pour amplifier l'effet des prédictions incorrectes, et les termes $y_i \cdot \log(\hat{y}_i)$ et $(1 - y_i) \cdot \log(1 - \hat{y}_i)$ permettent de calculer séparément la contribution à la loss des prédictions pour les classes positive et négative.

Maintenant, avec un exemple, prenons une situation où vous avez un modèle qui prédit si une image montre un chat (1) ou un chien (0). Supposons que nous avons trois images. Les étiquettes réelles (la réalité) et les probabilités prédites par le modèle sont les suivantes :

Image	Étiquette réelle y_i	Probabilité prédite d'être un chat \hat{y}_i
1	1 (c'est un chat)	0.9
2	0 (c'est un chien)	0.2
3	1 (c'est un chat)	0.6

Tableau 1. – Valeur du jeu de données avec prédiction du modèle

Reprendons notre formule de notre Binary Cross-Entropy loss plus haut :

Où :

- N est le nombre total d'échantillons (dans ce cas, $N = 3$),
- y_i est l'étiquette réelle de l'échantillon i (1 si c'est un chat, 0 si ce n'est pas un chat),
- \hat{y}_i est la probabilité prédite que l'échantillon i soit un chat.

Ici nous avons trois échantillons, donc nous allons calculer la cross entropy loss pour chaque échantillon et ensuite prendre la moyenne :

1. Pour l'échantillon 1:
 - $y_1 = 1$ (c'est un chat)
 - $\hat{y}_i = 0.9$ (la probabilité prédite d'être un chat)
 - Donc, la loss pour cet échantillon est : $1 \cdot \log(0.9) + (1 - 1) \cdot \log(1 - 0.9) = -0.105$
2. Pour l'échantillon 2:
 - $y_2 = 0$ (ce n'est pas un chat, autrement dit c'est un chien)
 - $\hat{y}_i = 0.2$ (la probabilité prédite d'être un chat)
 - Donc, la loss pour cet échantillon est : $0 \cdot \log(0.2) + (1 - 0) \cdot \log(1 - 0.2) = -0.223$
3. Pour l'échantillon 3 :

- $y_3 = 1$ (c'est un chat)
- $\hat{y}_3 = 0.6$ (la probabilité prédictive d'être un chat)
- Donc, la loss pour cet échantillon est : $1 \cdot \log(0.6) + (1 - 1) \cdot \log(1 - 0.6) = -0.511$

Ensuite, nous additionnons ces trois valeurs et nous les divisons par 3 (le nombre total d'échantillons) pour obtenir la cross entropy loss moyenne. C'est-à-dire nous allons calculer :

$$L_{\text{BCE}} = -\frac{1}{3} \times (-0.105 + -0.233 + -0.511)$$

$$L_{\text{BCE}} = 0.280$$

```
import numpy as np
# Les étiquettes réelles
y = np.array([1, 0, 1])
# Les probabilités prédictives
y_hat = np.array([0.9, 0.2, 0.6])

# Calcul de la cross entropy loss pour chaque échantillon
loss = -1 * (y * np.log(y_hat) + (1 - y) * np.log(1 - y_hat))

print(f"Loss pour l'échantillon 1 : {loss[0]:.3f}")
print(f"Loss pour l'échantillon 2 : {loss[1]:.3f}")
print(f"Loss pour l'échantillon 3 : {loss[2]:.3f}")

# Calcul de la cross entropy loss moyenne
mean_loss = np.mean(loss)

print(f"Loss moyenne : {mean_loss:.3f}")

# output
>>> Loss pour l'échantillon 1 : 0.105
>>> Loss pour l'échantillon 2 : 0.223
>>> Loss pour l'échantillon 3 : 0.511
>>> Loss moyenne : 0.280
```

Donc après avoir calculé la moyenne, nous obtenons la binary cross entropy loss L_{BCE} pour ces données est d'environ 0.280. Plus la valeur est faible, plus le modèle est proche des valeurs réelles.

Le calcul peut être effectué en un seul calcul qui donnerait :

$$L_{\text{BCE}} = -\frac{1}{3} \times [(1 \times \log(0.9) + (1 - 1) \times \log(1 - 0.9)) + (0 \times \log(0.2) + (1 - 0) \times \log(1 - 0.2)) + (1 \times \log(0.6) + (1 - 1) \times \log(1 - 0.6))]$$

4.2.2.4.2 Categorical Cross-Entropy Loss

Pour la classification multi-classes:

$$L_{\text{CCE}} = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^C y_{ik} \times \log(\hat{y}_{ik})$$

Comme pour la Binary Cross-Entropy je vais expliquer cette formule avant de l'utiliser.

- L_{CCE} : C'est la perte de l'entropie croisée catégorielle que nous essayons de minimiser
- $-\frac{1}{N}$: Nous prenons la moyenne des pertes calculées pour chaque échantillon dans le batch. N est le nombre d'échantillons dans le batch.
- $\sum_{i=1}^N$: Nous sommes sur tous les échantillons dans le lot
- C est le nombre de classes Pour chaque classe k , nous calculons la contribution de cette classe à la loss pour chaque échantillon.
- $\sum_{k=1}^C$ Pour chaque échantillon, nous sommes sur toutes les classes.
- y_{ik} : C'est l'étiquette réelle de l'échantillon i pour la classe k . Pour une classification multi-classe, y_{ik} est 1 si l'échantillon i appartient à la classe k et 0 sinon.
- \hat{y}_{ik} : C'est la probabilité prédictive par le modèle que l'échantillon i appartienne à la classe k .
- $\log(\hat{y}_{ik})$: Le logarithme est utilisé pour amplifier l'effet des différences entre les probabilités prédictives et les étiquettes réelles. Si la probabilité prédictive est proche de l'étiquette réelle, la valeur du logarithme est proche de 0, ce qui contribue peu à la loss totale. En revanche, si la probabilité prédictive est loin de l'étiquette réelle, la valeur du logarithme est un grand nombre négatif, ce qui contribue beaucoup à la loss totale.
- $y_{ik} \times \log(\hat{y}_{ik})$: Ce terme mesure la contribution à la loss de la prédition pour la classe k . Si $y_{ik} = 1$ (c'est-à-dire, si l'échantillon i appartient à la classe k), alors cette contribution est $\log(\hat{y}_{ik})$, qui est faible si \hat{y}_{ik} est proche de 1 et grande si \hat{y}_{ik} est proche de 0. Si $y_{ik} = 0$ (c'est-à-dire, si l'échantillon i n'appartient pas à la classe k), alors cette contribution est 0.

Voici un exemple d'utilisation de la CCE from scratch.

```
import numpy as np
# Définir les étiquettes réelles avec un one-hot-encoder
y = np.array([
    [1, 0, 0, 0], # Le premier échantillon est un chien
    [0, 1, 0, 0], # Le deuxième échantillon est un chat
    [0, 0, 0, 1] # Le troisième échantillon est un poisson
])

# Définir les probabilités prédictives par le modèle
y_hat = np.array([
    [0.7, 0.1, 0.1, 0.1], # Probabilités prédictives pour le premier échantillon
    [0.1, 0.7, 0.1, 0.1], # Probabilités prédictives pour le deuxième
    échantillon
    [0.1, 0.1, 0.1, 0.7] # Probabilités prédictives pour le troisième
    échantillon
])

# Calculer la categorical cross entropy loss avec une boucle
loss = 0.0
N = y.shape[0]
for i in range(N):
    for k in range(y.shape[1]):
        loss += y[i, k] * np.log(y_hat[i, k])

loss = -loss / N
print(f"Categorical Cross Entropy Loss: {loss:.3f}")
```

```
# output  
>>> Categorical Cross Entropy Loss : 0.357
```

Dans ce code, y contient les étiquettes réelles pour chaque échantillon et chaque classe sous forme de vecteurs one-hot. Le one-hot encoding est une représentation où chaque étiquette est un vecteur dont tous les éléments sont 0, sauf pour l'indice correspondant à la classe de l'échantillon, où la valeur est 1. Ainsi, chaque ligne de y correspond à un échantillon et est un vecteur one-hot, et chaque colonne correspond à une classe. Si un échantillon appartient à une classe, alors la valeur correspondante dans y est 1, sinon elle est 0.

y_{hat} contient les probabilités prédites par le modèle pour chaque échantillon et chaque classe. Chaque ligne de y_{hat} correspond à un échantillon, et chaque colonne correspond à une classe. Les valeurs dans y_{hat} sont des probabilités, donc elles sont comprises entre 0 et 1, et la somme des probabilités pour chaque échantillon est 1.

J'ai utilisé des boucles pour imiter une somme, puisque qu'une somme n'est en fait qu'une boucle qui parcourt chaque élément pour faire une grosse addition.

4.3 Descente de gradient

La descente de gradient est un algorithme d'optimisation utilisé pour le machine learning. Son objectif est de minimiser une fonction de coût, petit à petit, en modifiant les paramètres du modèle dans le but que notre fonction coût tende vers 0. Sans la descente de gradient, il serait difficile, voire impossible, de trouver ces valeurs de manière efficace. La descente de gradient n'est pas utilisée sous la forme que vous allez voir, mais la comprendre vous est indispensable pour comprendre la Stochastic Gradient Descent (SGD) et ses évolutions qui sont utilisées en pratique.

Prenons l'exemple où l'on se retrouve en voiture, à la recherche d'un parking dans une ville inconnue. On se retrouve à tourner en rond, à la recherche d'un emplacement pour stationner. Le défi réside dans le fait que l'on ne sait pas où se situent les parkings, ni lequel est le plus proche ou le moins cher.

C'est là que votre « GPS » entre en jeu, représentant ici la fonction de perte. Votre GPS ne peut pas vous dire directement où se trouve le meilleur parking, mais il peut vous donner une indication de la distance entre vous et le parking le plus proche. Plus vous êtes loin d'un parking, plus la valeur de cette fonction de perte est élevée.

À chaque carrefour, vous avez le choix entre plusieurs directions. Pour choisir, vous consultez votre GPS. Vous allez dans la direction qui semble réduire la distance au parking le plus proche. Parfois, vous pourriez vous tromper et vous éloigner de l'endroit où vous vouliez aller, augmentant ainsi votre « fonction de perte ». Mais grâce au feedback de votre GPS, vous pouvez corriger votre trajectoire et essayer une nouvelle direction.

Avec le temps, en utilisant cette approche de « descente de gradient », vous finirez par trouver un parking. De la même manière, dans un problème de machine learning, vous ajustez vos paramètres pas à pas pour minimiser votre fonction de perte, ce qui vous rapproche de la solution optimale.

Au vu l'importance de cet algorithme, heureusement qu'il n'est pas compliqué, pour les notions mathématiques, nous allons utiliser les dérivées et un peu d'algèbre.

4.3.1 Descente de Gradient en 1D

La descente de gradient en une dimension est la forme la plus simple de cette méthode d'optimisation. Dans ce contexte, nous disposons d'une seule variable — que nous appellerons « poids » (représenté par W pour « weights » en anglais) - à ajuster pour minimiser notre fonction de perte.

Sur le graphique suivant, l'axe des ordonnées représente le coût ou la perte. L'axe des abscisses, quant à lui, représente notre poids W . Notre objectif est d'atteindre le point minimal de la fonction de perte, représenté ici en vert. Les flèches rouges indiquent la direction de la descente de gradient — elles pointent vers l'opposé du gradient, ce qui est logique puisque nous cherchons à minimiser la fonction de perte, pas à la maximiser.

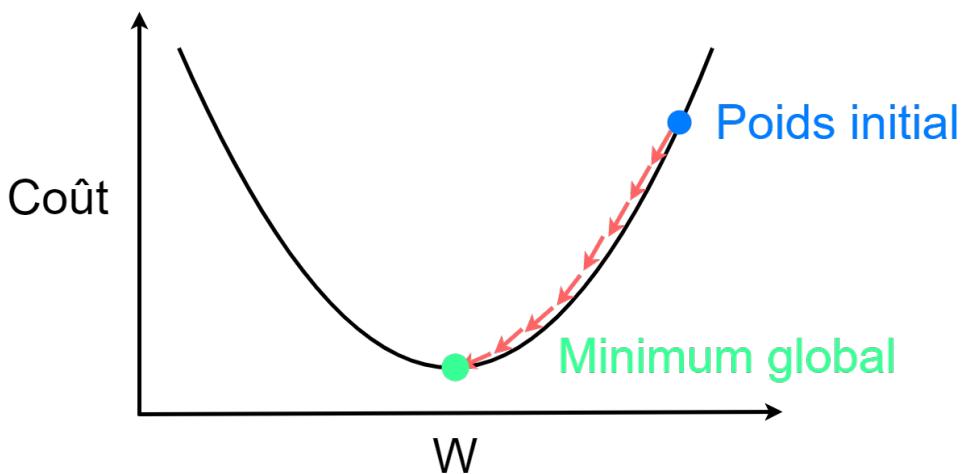


Fig. 26. – « Descente de gradient: le point bleu représente le poids initial (choisi aléatoirement). Les flèches rouges indiquent les étapes de descente de gradient (aussi appelées « pas », ou « learning rate » en anglais) qui nous rapprochent progressivement du minimum global, représenté par le point vert. »

4.3.2 Impact de la Taille du Pas (Learning Rate)

Le « pas » ou « learning rate » est un « hyperparamètre », un nouveau terme important à connaître, un paramètre est appris par le modèle et est optimisé par l'algorithme, par exemple les biais et les poids s'adapte lors de l'entraînement, eux sont des paramètres, le learning rate ne sera pas optimisé par l'algorithme lui-même, mais doit être défini par le développeur, c'est un « hyperparamètre ».

Lors d'une descente de gradient, le pas (learning rate) détermine la taille de la flèche sur les diagrammes, la taille de la progression afin que notre modèle converge vers le minimum. Plus précisément, il correspond au facteur multiplicatif appliqué à la dérivée de la fonction de perte dans la formule de mise à jour des poids, il est représenté par α :

$$W := W - \alpha \cdot \nabla J(W)$$

Où :

- W représente les paramètres (weights, les poids) du modèle, elle est généralement représentée par la lettre θ (theta) par convention, mais je trouve plus approprié d'utiliser la lettre W qui est aussi parfois utilisé.
- L'opérateur $:=$ est une mise à jour d'une valeur, en programmation, on utilise $=$.
- α est le taux d'apprentissage (learning rate) qui contrôle la taille des pas effectués lors de la mise à jour des paramètres, elle est présentée comme une flèche rouge dans les diagrammes.
- $\nabla J(W)$ le triangle à l'envers est « nabla » en grec mais c'est surtout le « gradient » dans le contexte du machine learning. Ici $\nabla J(W)$ est le gradient de la fonction coût J (une MSE par exemple). Le gradient représente la direction de la plus forte augmentation de la fonction coût, c'est pour cela que α est au négatif, afin de non pas de maximiser la fonction coût, mais de la minimiser, à cette étape le gradient dérivera la fonction coût, vous verrez ça en pratique plus loin.
- La fonction coût est représentée par la lettre J en référence à la matrice Jacobienne utilisée. Nous l'utiliserons plus tard pour simplifier les calculs.

```
def descente_gradient(theta, alpha, gradient):
    theta = theta - alpha * gradient
    return theta
```

Un taux d'apprentissage (learning rate) élevé entraîner peut aider à notre modèle à rapidement converger vers le minimum de la fonction de perte. Mais un taux d'apprentissage trop élevé ne convergera jamais, les mises à jour de poids deviennent si importantes que l'algorithme risque de « sauter » par-dessus le minimum recherché et peut même diverger. Sur une simple parabole cela peut aller, mais il faut imaginer que les problèmes d'optimisations sont très complexes.

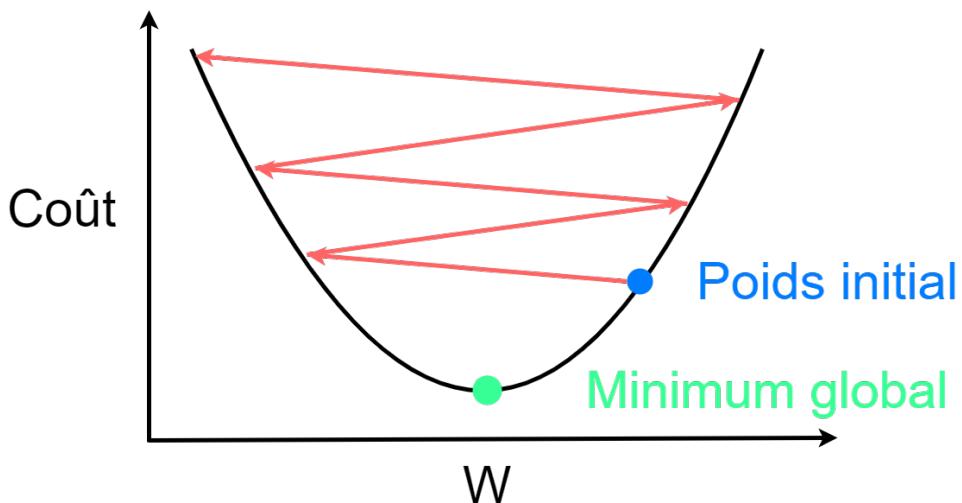


Fig. 27. – « Descente de gradient avec un learning rate trop élevé: le point bleu représente le poids initial (choisi aléatoirement). Les flèches rouges indiquent les étapes de descente de gradient avec un learning rate beaucoup trop élevé, qui entraînent des oscillations importantes et empêchent totalement la convergence vers le minimum global, représenté par le point vert. »

Inversement, un taux d'apprentissage faible conduit à des mises à jour plus petites, ce qui peut assurer une convergence plus stable, mais à un rythme beaucoup plus lent. Il y a donc

un risque de ne pas atteindre le minimum dans un délai raisonnable, surtout si l'on part d'un point initial éloigné du minimum.

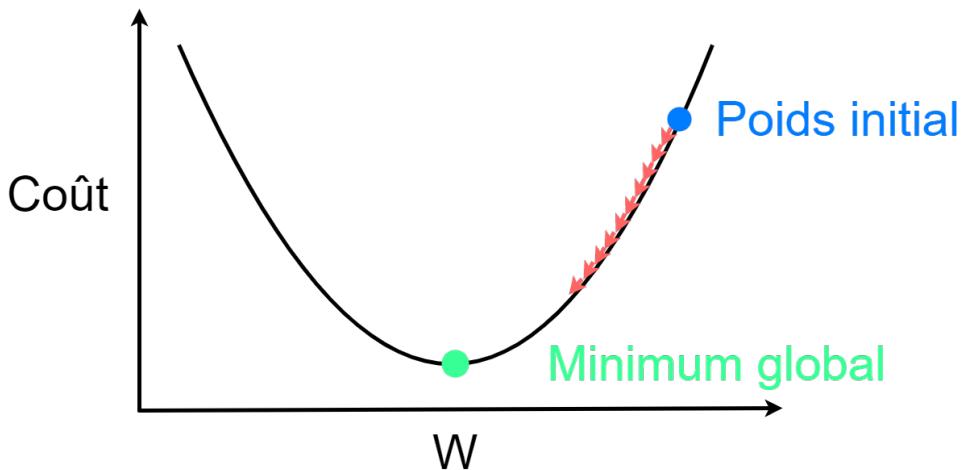


Fig. 28. – Descente de gradient avec un learning rate trop faible. Les flèches rouges indiquent les étapes de descente de gradient avec un learning rate beaucoup trop faible, l'entraînement fait des pas trop faible.

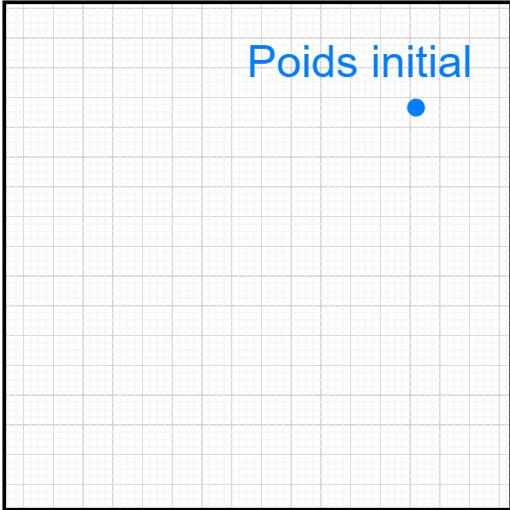
Le choix du learning rate est un exercice d'équilibrisme : il doit être assez grand pour permettre une convergence rapide (voir une convergence tout court), mais pas trop grand pour ne pas diverger et ni trop petit pour l'atteindre dans un délai raisonnable et ne pas rester coincé dans un minimum local. Dans la pratique, le taux d'apprentissage est souvent déterminé par essai et erreur, il n'y a pas de valeur universelle, ça dépend des architectures, mais ça se situe souvent entre 0.01 et 0.0001.

4.3.3 Descente de gradient en 2D

Dans le diagramme précédent, en forme de parabole d'une dimension, avait uniquement un paramètre à optimiser, nous pouvons imaginer qu'ici, il y aurait deux paramètres, le poids et son biais.

La descente de gradient est comme un explorateur perdu dans une vallée montagneuse par une nuit sans lune, il ne peut pas voir le paysage autour de lui. Tout ce qu'il sait, c'est où il se trouve (les valeurs actuelles du poids et de son biais), et il peut estimer la pente de la montagne sous ses pieds (avec le gradient son GPS).

Ce que la descente de gradient voit



La réalité

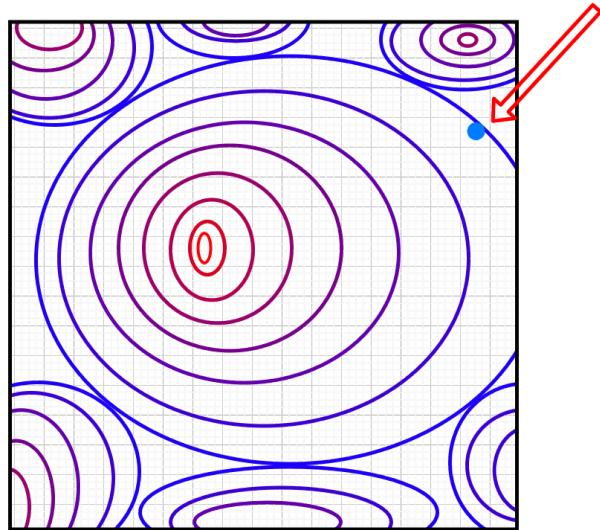


Fig. 29. – Descente de gradient d'une vision en deux dimensions avec les couleurs pour jouer sur la « profondeur » comme une carte topographique. La fonction coût qui doit tendre vers le minimum global, plus la couleur tend vers le bleu, plus notre fonction coût est élevée, plus nous tendons vers le rouge, plus notre fonction coût tend à se minimiser.

Notre explorateur ne voit pas le relief complet, il est dans un brouillard complet et doit se fier uniquement au gradient (son GPS) pour décider de sa prochaine étape qui est où la pente est la plus raide. L'explorateur est totalement aveugle à la topographie générale de la vallée et doit faire confiance au gradient pour le guider.

La descente de gradient est une méthode de recherche à l'aveugle qui utilise seulement l'information locale du gradient pour naviguer dans l'espace des paramètres. C'est probablement sa grande faiblesse, car elle peut la conduire à des solutions sous-optimales ou requérir un grand nombre d'itérations pour atteindre une solution acceptable.

4.3.4 Exemple mathématique de la descente de gradient

Vous n'allez jamais implémenter mathématiquement une descente de gradient « from scratch » en Python, mais l'utilisation des mathématiques avec des exemples très simples et concrets (une addition de 1, ici) permet de démysterifier les concepts du deep learning. Que vous ayez une base en mathématiques sur les calculs différentiels ou non, les calculs sont assez simples à comprendre intuitivement si vous savez que la dérivée de x^2 est $2x$. Je ne m'attends pas à ce que vous preniez une feuille blanche et un crayon à papier pour commencer à calculer les itérations de la descente de gradient, mais que vous renforciez votre intuition de son fonctionnement. D'abord avec un seul paramètre, puis avec deux paramètres en utilisant une matrice jacobienne pour donner un exemple ayant plus d'un paramètre. Ce n'est pas grave si vous ne comprenez pas réellement les mathématiques suivantes tant que vous avez bien saisies pourquoi notre résultat finit par converger.

4.3.4.1 Avec un seul paramètre

Dans cet exemple, nous utiliserons un « toy dataset », un jeu de données simple créé uniquement à des fins d'illustration. La fonction de perte utilisée sera la MSE, Le tableau ci-dessous présente les valeurs de notre toy dataset, où x est la valeur d'entrée et y est le label, c'est-à-dire la valeur à prédire.

x	y
1	2
2	3
3	4
4	5

Tableau 2. – Valeur du jeu de notre toy dataset, x et la valeur d'entrée, y le label soit la valeur à prédire, le modèle devra tendre à additionner 1 à chaque valeur d'entrée

Pour commencer, nous initialisons notre paramètre W , à une valeur arbitraire, en pratique les poids sont initiés aléatoirement. Pour cet exemple, choisissons $W = 0$. Nous devons également choisir une valeur pour le taux d'apprentissage α , pourquoi pas $\alpha = 0.1$.

Ensuite, nous entrons dans une boucle d'itérations. Rappel de la formule :

$$W := W - \alpha \cdot \nabla J(W)$$

Première itération :

1. **Calcul des prédictions** : Avec $W = 0$, nos prédictions sont toutes 0, car $Wx = 0$ pour toutes les valeurs de x . Cela signifie que, peu importe ce que nous multiplions par 0, le résultat sera toujours 0.
2. **Calcul de l'erreur** : L'erreur pour chaque paire (x, y) est $(y - Wx)^2 = y^2$, car $Wx = 0$. Donc, pour nos données, les erreurs sont $(2^2, 3^2, 4^2, 5^2) = (4, 9, 16, 25)$.
3. **Calcul de la fonction de coût** : La fonction de coût est la moyenne des erreurs, donc $J(W) = \frac{1}{4} \times (4 + 9 + 16 + 25) = 13.5$. Ici, nous divisons par 4 pour faire la moyenne.
4. **Calcul du gradient** : Le gradient est déterminé par la dérivée de la fonction coût par rapport à notre paramètre W , ici nous dérivons la fonction MSE. Le gradient est une mesure de la pente de la fonction coût, c'est-à-dire à quel point la fonction coût change lorsque nous changeons W . La dérivée de $(y_i - Wx_i)^2$ par rapport à W est $-2x_i(y_i - Wx_i)$,
 - en dérivant, le « 2 » provient de la dérivation de la fonction carrée (à droite de la parenthèse de la MSE)
 - x_i est l'entrée correspondante à la prédition actuelle
 - $(y_i - Wx_i)$ est simplement notre erreur actuelle pour cette entrée

Donc la dérivée de $J(W)$ par rapport à W est : $\nabla J(W) = \frac{1}{n} \sum_{i=1}^n -2x_i(y_i - Wx_i)$
Pour nos données, cela donne : $\nabla J(W) = \frac{1}{4}((-2 \times 1 \times (2 - 0)) + (-2 \times 2 \times (3 - 0)) + (-2 \times 3 \times (4 - 0)) + (-2 \times 4 \times (5 - 0))) = \frac{1}{4} \times -80 = -20$.

5. **Mise à jour de W** : Nous utilisons maintenant notre formule de mise à jour pour obtenir le nouveau W :

- $W := W - \alpha \times \nabla J(W) = 0 - 0.1 \times -20 = 2.0$. C'est notre ancienne valeur de W moins notre taux d'apprentissage fois le gradient. Cela nous donne notre nouvelle valeur de W qui, espérons-le, a une fonction de coût plus faible.

```
import numpy as np

x = np.array([1, 2, 3, 4])
y = np.array([2, 3, 4, 5])

W = 0
alpha = 0.1

for i in range(6):
    predictions = W * x # notre prédiction aussi appelé y_hat parfois
    errors = (y - predictions)**2
    cost = errors.mean()
    gradient = -2 * ((y - predictions) * x).mean()
    W = W - alpha * gradient
    print(f"Iteration {i+1}, J(W) = {cost}, W = {W}")

# Output
>>> Iteration 1, J(W) = 13.5, W = 2.0
>>> Iteration 2, J(W) = 3.5, W = 1.0
>>> Iteration 3, J(W) = 1.0, W = 1.5
>>> Iteration 4, J(W) = 0.375, W = 1.25
>>> Iteration 5, J(W) = 0.2188, W = 1.375
>>> Iteration 6, J(W) = 0.1796, W = 1.3125
>>> Iteration 10, J(W) = 0.1667, W = 1.3320
>>> Iteration 100, J(W) = 0.1667, W = 1.3333
```

Nous ne calculerons pas les itérations suivantes, la logique reste la même, avec encore les mêmes étapes. Ce qui était à comprendre ici, c'est que nous utilisons l'algorithme de la descente de gradient pour avoir une fonction coût qui tend vers 0 en ajustant un paramètre, Chaque paramètre a son propre gradient, qui guide comment il doit être ajusté pour minimiser la fonction de coût. En pratique, dans un cas réel, votre ordinateur optimisera des millions de gradients, avec une descente de gradient pour optimiser des millions de paramètres afin d'ajuster simultanément tous les poids du réseau.

Nous pouvons vérifier intuitivement que notre paramètre initialement à 0 qui ne pouvait que donner de mauvais résultats puisque qu'il prédisait uniquement 0 devient 1.33 au bout de 6 itérations et que par exemple pour la quatrième valeur de x qui serait $x_4 = 4$ coupler avec notre paramètre optimisé : $1.33 \times 4 = 5.32$, notre perceptron prédit 5.32 quand il fallait prédire 5, c'est plutôt pas mal, mais vous pouvez voir tout de suite la limite d'utiliser un seul perceptron, notre modèle est trop simpliste et nécessiterait plusieurs perceptron pour être un meilleur algorithme de calculatrice.

4.3.4.2 Avec deux paramètres, le biais en plus.

Précédemment, nous avons utilisé seulement le paramètre w_1 pour simplifier les calculs et éviter d'utiliser les dérivées partielles et une matrice Jacobienne.

Cette fois, il y a deux paramètres à prendre en compte : w_1 et son biais b_1 . Cela implique que nous allons devoir utiliser les dérivées partielles et la matrice Jacobienne pour ajuster ces paramètres.

En machine learning, nos modèles ont entre des milliers ou des millions de paramètres. Pour comprendre l'effet qu'à chaque paramètre sur la fonction de coût, nous utilisons une dérivée partielle

Une dérivée partielle est la dérivée d'une fonction par rapport à l'une de ses variables, en gardant toutes les autres constantes. Par exemple, $\frac{\partial J(W)}{\partial w_1}$ mesure comment la fonction de coût $J(W)$ change lorsque nous changeons seulement w_1 , en gardant b_1 constant. De même, $\frac{\partial J(W)}{\partial b_1}$ mesure comment $J(W)$ change lorsque nous changeons uniquement b_1 , en gardant w_1 constant.

La matrice Jacobienne est un outil qui nous permet de rassembler toutes ces dérivées partielles en une seule entité. En d'autres termes, elle est une généralisation de la dérivée pour les fonctions multivariées. Chaque élément de la matrice Jacobienne est une dérivée partielle de la fonction par rapport à l'une de ses variables.

Dans notre cas, la matrice Jacobienne est définie comme suit :

$$\nabla J(W) = \left[\frac{\partial J(W)}{\partial w_1}, \frac{\partial J(W)}{\partial b_1} \right]$$

Ainsi, $\nabla J(W)$ est un vecteur dont les composantes sont les dérivées partielles de $J(W)$ par rapport à w_1 et b_1 . En utilisant ce vecteur, nous pouvons mettre à jour simultanément w_1 et b_1 de manière à minimiser la fonction de coût.

L'idée de base de la descente de gradient est de modifier les paramètres dans la direction qui réduit le plus la fonction de coût. Les dérivées partielles nous indiquent dans quelle direction la fonction de coût change le plus rapidement, c'est pourquoi nous les utilisons pour mettre à jour nos paramètres. En d'autres termes, elles nous donnent la direction de la pente la plus raide que nous pouvons descendre pour réduire la fonction de coût.

Très bien, voyons maintenant comment cette approche s'adapte lorsqu'il y a deux paramètres dans notre modèle. Pour simplifier, nous supposerons que nous avons maintenant une fonction affine de la forme $y = w_1x + b_1$, où w_1 est le poids et b_1 est le biais.

Pour ce cas, nous initialisons nos paramètres w_1 et b_1 à des valeurs arbitraires. Prenons $w_1 = 0$ et $b_1 = 0$. Et nous prenons toujours $\alpha = 0.1$ qui sera notre taux d'apprentissage.

Le calcul de la mise à jour des poids devient maintenant :

$$[w_1, b_1] := [w_1, b_1] - \alpha \cdot \nabla J(W)$$

Où $\nabla J(W)$ est maintenant le gradient de J par rapport à W et est calculé comme suit :

$$\nabla J(W) = \left[\frac{\partial J(W)}{\partial w_1}, \frac{\partial J(W)}{\partial b_1} \right]$$

C'est la matrice Jacobienne de J par rapport à W .

Chaque $\frac{\partial J(W)}{\partial W_i}$ ou $\frac{\partial J(W)}{\partial b_i}$ est une dérivée partielle de J par rapport à W_i ou b_i qui est calculée en prenant la moyenne de $-2x_i(y_i - (w_1x_i + b_1))$ pour $\frac{\partial J(W)}{\partial w_1}$ et de $-2(y_i - (w_1x_i + b_1))$ pour $\frac{\partial J(W)}{\partial b_1}$ sur toutes les paires de données (x, y) dans notre jeu de données.

Ainsi, chaque paramètre est mis à jour individuellement en fonction de sa contribution à l'erreur totale, comme indiqué par sa dérivée partielle respective. C'est l'essence de l'utilisation des dérivées partielles et de la matrice Jacobienne dans l'algorithme de descente de gradient.

À titre d'exemple, nous calculerons la première itération de ce processus pour le jeu de données précédent.

Première itération :

1. Calcul des prédictions : Avec $w_1 = 0$ et $b_1 = 0$, nos prédictions sont toutes 0, car $w_1x + b_1 = 0$ pour toutes les valeurs de x .
2. Calcul de l'erreur : L'erreur pour chaque paire (x, y) est $(y - (w_1x + b_1))^2$, donc pour nos données, les erreurs sont les mêmes que dans le cas précédent.
3. Calcul de la fonction de coût : La fonction de coût reste la même, donc $J(W) = 13.5$.
4. Calcul du gradient : Le gradient est maintenant un vecteur de dérivées partielles :

- Pour w_1 : $\frac{\partial J(W)}{\partial w_1}$
 $= \frac{1}{4}(-2 \times 1 \times (2 - 0) + -2 \times 2 \times (3 - 0) + -2 \times 3 \times (4 - 0) + -2 \times 4 \times (5 - 0))$
 $= \frac{1}{4} \times -80 = -20.$
- Pour b_1 : $\frac{\partial J(W)}{\partial b_1}$
 $= \frac{1}{4}(-2 \times (2 - 0) + -2 \times (3 - 0) + -2 \times (4 - 0) + -2 \times (5 - 0))$
 $= \frac{1}{4} \times -28 = -7.$

Donc $\nabla J(W) = [-20, -7]$.

5. Mise à jour de W : Nous utilisons maintenant notre formule de mise à jour pour obtenir les nouvelles valeurs w_1 et b_1 :

- $w_1 := w_1 - \alpha \times \frac{\partial J(W)}{\partial w_1} = 0 - 0.1 \times -20 = 2.0$
- $b_1 := b_1 - \alpha \times \frac{\partial J(W)}{\partial b_1} = 0 - 0.1 \times -7 = 0.7$

Reste des calculs effectuer à l'ordinateur avec python.

```
import numpy as np
x = np.array([1, 2, 3, 4])
y = np.array([2, 3, 4, 5])
```

```
w1 = 0
b1 = 0
```

```

alpha = 0.1

for i in range(20):
    predictions = W1 * x + b1 # notre prédiction aussi appelé y_hat parfois
    errors = (y - predictions)**2
    cost = errors.mean()
    dW1 = -2 * ((y - predictions) * x).mean() # dérivée partielle par rapport
à W1
    db1 = -2 * (y - predictions).mean() # dérivée partielle par rapport à b1
    W1 = W1 - alpha * dW1
    b1 = b1 - alpha * db1
    print(f"Iteration {i+1}, J(W) = {round(cost, 3)}, W1 = {round(W1, 3)}, b1
= {round(b1, 3)}")

# ouput
>>> Iteration 1, J(W) = 13.5, W1 = 2.0, b1 = 0.7
>>> Iteration 2, J(W) = 6.09, W1 = 0.65, b1 = 0.26
>>> Iteration 3, J(W) = 2.761, W1 = 1.545, b1 = 0.583
>>> Iteration 4, J(W) = 1.265, W1 = 0.936, b1 = 0.394
>>> Iteration 5, J(W) = 0.592, W1 = 1.335, b1 = 0.547
>>> Iteration 6, J(W) = 0.288, W1 = 1.059, b1 = 0.47
>>> Iteration 10, J(W) = 0.044, W1 = 1.127, b1 = 0.556
>>> Iteration 15, J(W) = 0.025, W1 = 1.13, b1 = 0.626
>>> Iteration 20, J(W) = 0.018, W1 = 1.109, b1 = 0.678
>>> Iteration 233, J(W) = 0.0, W1 = 1.0, b1 = 1.0

```

Ainsi, même avec plusieurs paramètres, le concept reste le même : nous ajustons chaque paramètre en fonction de son influence sur l'erreur globale. C'est ce qui fait l'efficacité de la descente de gradient.

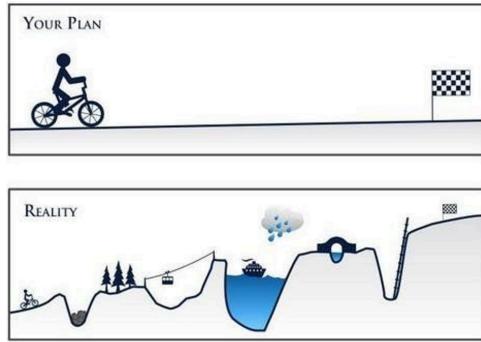
Dans ce deuxième exemple, nous avons ajouté un biais à notre modèle, ce qui le rend plus sophistiqué et plus flexible, et lui permet de s'adapter à nos données. Le biais permet au modèle de ne pas être contraint de passer par l'origine, ce qui peut améliorer la précision de nos prédictions.

Après 233 itérations, notre poids W_1 est de 1.0 et notre biais b_1 est également de 1.0. Cela signifie que notre modèle prédit $y = 1.0x + 1.0$. Par exemple, pour la quatrième valeur de x qui est $x_4 = 4$, notre modèle prédit $1.0 \times 4 + 1.0 = 5.0$, ce qui est exactement la valeur réelle de 5.

Cela montre que l'ajout d'un biais à notre modèle a permis d'améliorer considérablement la précision de nos prédictions. En fait, après suffisamment d'itérations, notre modèle est capable de prédire parfaitement les valeurs de y pour les données d'entrée données.

4.3.5 Qu'est-ce qu'un minimum local ?

Les descentes de gradient ne sont pas de simple parabole en forme de « U » vers laquelle on pourrait simplement tendre vers zéro sans rencontrer d'obstacle sur le chemin. Parmi ces obstacles, les minimums locaux sont l'un des problèmes que nous rencontrons fréquemment lors d'une descente de gradient.



En réalité, le paysage de la descente de gradient peut être parsemé de minimums locaux et globaux, d'interstices, de crêtes, et de plateaux, compliquant ainsi la convergence vers le minimum de notre fonction de coût. Tous ces problèmes viennent avec la chance que nous pouvons avoir lors de l'initialisation aléatoire des poids du modèle au démarrage de l'entraînement. Par exemple, si l'initialisation commence à gauche de la figure, nous pourrions nous retrouver dans un très mauvais minimum local. En revanche, si le poids d'initialisation est au centre, nous convergerons plus facilement vers le minimum global. Enfin, si le poids commence à droite, nous nous retrouverons également dans un minimum local, mais il s'agit d'un minimum local relativement acceptable.

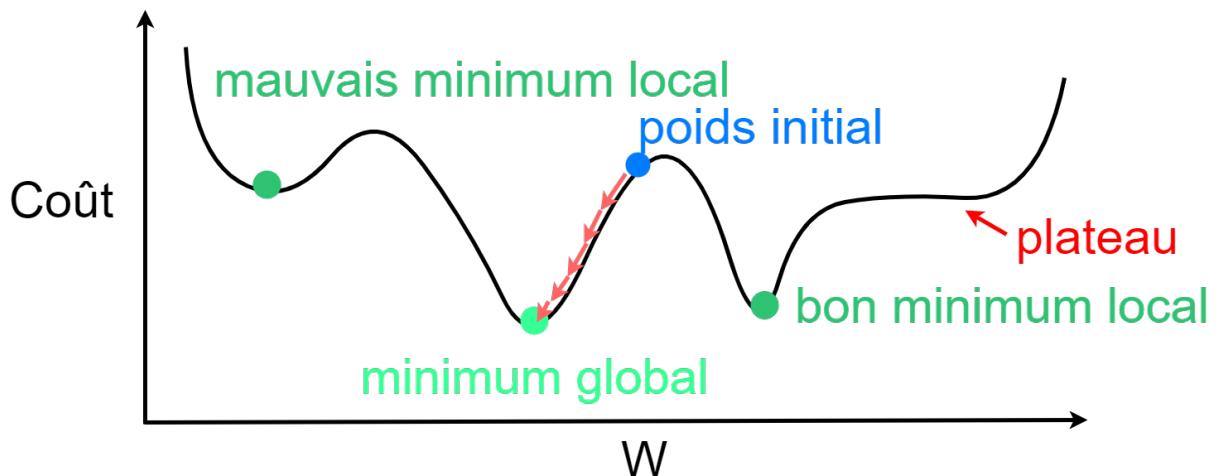


Fig. 31. – « Descente de gradient : le point bleu représente le poids initial (choisi aléatoirement). Les flèches rouges indiquent les étapes de descente de gradient (aussi appelées ‹ pas ›, ou ‹ learning rate › en anglais) qui nous rapprochent progressivement du minimum global, représenté par le point vert. La flèche rouge nous montre un plateau problématique pour la convergence, c'est le phénomène du vanishing gradient que nous verrons vers la fin du chapitre. »

Ces minimums locaux nous apprennent que notre algorithme de la descente de gradient ne nous promet pas de nous donner le paramètre optimal par rapport à la fonction coût s'il tombe dans un minimum local. La nuance est que tous les minimums locaux ne sont pas mauvais. Certains peuvent être relativement acceptables, comme le minimum de droite dans la figure précédente.

Lorsque nous effectuons une descente de gradient, nous n'avons pas une vue d'ensemble de la fonction de coût. Nous ne voyons que la pente locale à l'endroit où notre paramètre se trouve.

Avec un algorithme de descente de gradient un peu plus sophistiqué, on pourrait imaginer qui aurait une « mémoire » des pas précédents et éviter de retomber dans le même minimum local.

Il existe aussi des méthodes d'initialisation des poids plus intelligente qu'une simple initialisation purement aléatoire pour éviter de commencer l'entraînement du modèle dans une région ayant plein de mauvais minimums locaux. Mais même avec les meilleures techniques d'initialisation de poids et d'algorithme de la descente de gradient les plus avancées, nous n'avons quand même pas la garantie de trouver le minimum global et il est même impossible d'avoir la certitude que notre algorithme n'est pas tombé dans un minimum local plutôt que global. Dans des cas d'application réelle (plus compliqué que la simple descente de gradient 1D présenté plus haut), le paysage d'une fonction coût multidimensionnelle, l'espace des possibilités que notre paramètre puisse prendre est gigantesque et souvent irrégulier. Le sujet de l'optimisation dans le deep learning est tout sauf un sujet trivial.

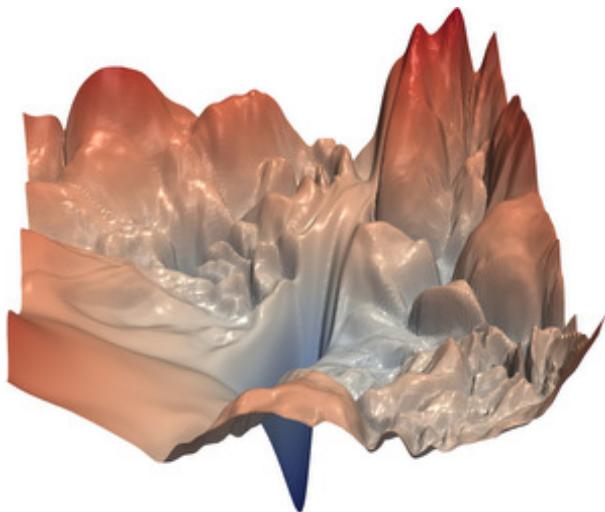


Fig. 32. – « Descente de gradient sophistiquée d'un réseau »

Cette visualisation a été rendue possible via plusieurs méthodes pour rendre une descente de gradient multidimensionnelle à une visualisation en 3D [3].

4.3.6 Les points de selle (saddle points)

Le point de selle, ou « saddle point » en anglais, tire son nom de sa ressemblance avec une selle de cheval. Ce terme provient de la géométrie et se réfère à un point où la courbure de la surface change de signe. Visualisez une selle de cheval. Vous pouvez y observer une courbure ascendante dans une direction (comme le dos d'un cheval) et une courbure descendante dans l'autre (comme les côtés de la selle sur lesquels les jambes du cavalier reposent).

En termes mathématiques, dans un espace à deux dimensions, un point de selle est l'endroit où la courbe est à la fois concave et convexe. Ce phénomène se traduit par la formation d'un creux dans une direction et d'un pic dans l'autre. Dans le contexte de la descente de gradient, c'est un endroit où le gradient de la fonction coût est nul. C'est là que réside la distinction cruciale entre un point de selle et un minimum local. Un minimum local représente un creux dans toutes les directions, contrairement à un point de selle.

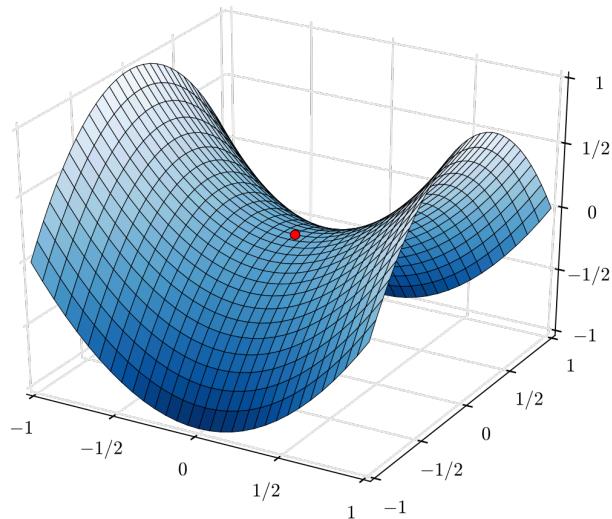


Fig. 33. – « Une représentation d'un point de selle en rouge dans une fonction à deux variables. On observe une courbe ascendante dans une direction et une courbe descendante dans l'autre, formant un point de selle. »

Le problème de la descente de gradient... c'est qu'elle se repose sur le gradient pour déterminer la direction à prendre pour continuer à minimiser la fonction. Si le gradient est nul, le gradient « ne sait pas » dans quelle direction aller. Le point de selle est à la frontière entre une zone où la fonction est en train de diminuer et une zone où elle est en train d'augmenter.

Si visuellement, il est facile pour nous de voir que certaines directions sont ascendantes et d'autres descendantes, mais l'algorithme n'a pas cette vue d'ensemble. Comme dans la métaphore de la voiture au début du chapitre, le gradient est le « GPS » il est limité par l'information que le gradient lui fournit. Sur un point de selle la surface est mathématiquement plate, l'algorithme n'a aucune raison de bouger.

Contrairement à ce que notre intuition peut nous laisser penser, il est beaucoup plus probable de tomber dans une point de selle que de tomber dans un minimum local. Nos sens nous trompent du que nous n'avons pas une bonne intuition sur des espaces de milliers de dimensions, où il y a beaucoup plus de direction dans lesquelles la courbure peut changer.

Ce problème trompe les algorithmes d'optimisations classiques en laissant penser qu'ils ont trouvé le minimum. Comme son gradient est nul, l'algorithme s'arrête prématurément et créera un modèle sous-optimal qui n'aura pas atteint le plein potentiel de l'architecture du modèle de deep learning

Ces questions d'optimisation sont des sujets de recherche encore actifs, optimiser la vitesse d'apprentissage ce sont des entraînements plus court et plus économique. Pour résoudre ce problème les ingénieurs d'architecture n'utilise pas de descente de gradient « vanillia » mais des versions sophistiquées comme la descente de gradient stochastique que nous verrons en fin de ce chapitre

Une des choses à retenir de ce sous-chapitre assez théorique et que nos intuitions et compréhensions sur de faible dimension ne se généraliseront pas forcément dans cas de plusieurs milliers de dimensions.

4.3.7 La descente de gradient stochastique, l'apprentissage par lot

La descente de gradient classique n'est pas utilisée en pratique, des variantes comme la descente de gradient stochastique (SGD) et ses variantes résolvent plusieurs problèmes d'optimisation cités précédemment.

L'idée de la SGD est la même que la descente de gradient classique, nous cherchons à minimiser une fonction coût. La différence est que la SGD calcule différemment les mises à jour à effectuer.

Auparavant, nous avons calculé le gradient sur l'ensemble du jeu de données (batch) qui était de 4 échantillons avant de faire chaque mise à jour des poids. Ce processus n'est pas vraiment réalisable avec de gros jeu de données en plus d'être très coûteux en puissance de calculs. La SGD met à jour les poids pour chaque mini-lot (mini-batch) du jeu de données, une petite partie portion du jeu de données.

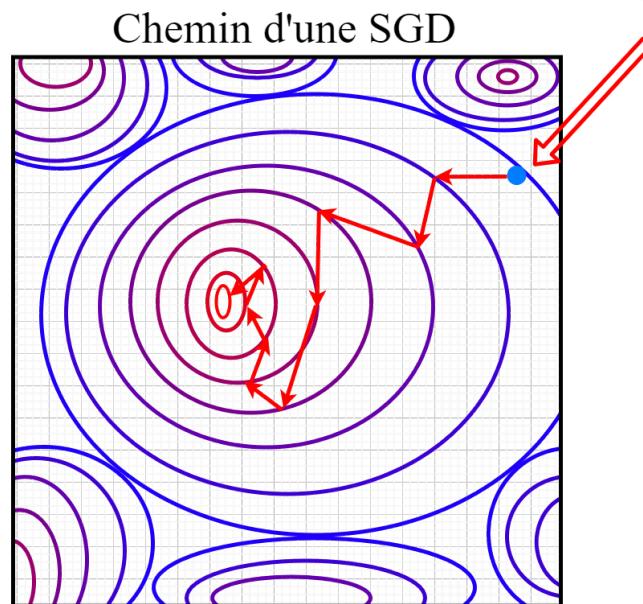


Fig. 34. – « Visualisation de la Descente de Gradient Stochastique : La mise à jour des poids se fait après chaque mini-lot (batch) plutôt qu'après avoir traversé l'ensemble de l'ensemble d'apprentissage. »

L'utilisation de mini-lot (mini-batch) permet de mettre plus régulièrement à jour les poids, l'algorithme peut converger plus rapidement, il n'a pas à calculer le gradient pour tout le jeu de données en entier, mais uniquement son mini-lot. En pratique, nous utilisons de grand ensemble de données qui rend l'utilisation d'une descente de gradient classique ni réellement réalisable ni optimisé, il faudrait des cartes graphiques ayant de gigantesque mémoire vive.

L'approche stochastique de la mise à jour des poids introduit cependant une certaine quantité de bruit dans le processus d'optimisation. Alors que dans la descente de gradient traditionnelle, chaque mise à jour est calculée en utilisant la totalité des données, garantissant ainsi que chaque mise à jour va dans la direction optimale, la SGD, en utilisant un échantillon ou un petit lot à la fois, peut effectuer des mises à jour qui ne vont pas exactement dans la direction optimale (vous pouvez le voir sur le diagramme d'ailleurs, il va même en arrière). Cela peut être une bonne chose, car cela peut aider l'algorithme à éviter de rester coincé dans les minimums locaux, mais cela peut aussi rendre le processus de convergence plus chaotique.

La taille de ces mini-lots (batch size) deviennent alors un nouveau hyperparamètre à ajuster au développeur comme l'est le learning rate. Si nous avons un batch size = 1, la SGD effectuera une mise à jour de ses paramètres après chaque échantillon qui serait la SGD « pure ». Si nous utilisons un batch size de la taille du jeu de données, on revient à une descente de gradient classique. Comme le taux d'apprentissage, le choix de la taille des batchs (mini-lot) est un exercice d'équilibrisme entre la vitesse convergence et la stabilité de l'algorithme. Il est bon de savoir que de trop petits batchs size augmentent la probabilité d'être coincé dans un minimum local. En pratique, nous essayons d'avoir un batch size le plus gros possible selon notre matérielle informatique (la mémoire vive de la carte graphique) afin de ne pas avoir à effectuer la rétropropagation trop régulièrement qui demande beaucoup de ressource computationnelle.

4.4 La rétropropagation (Backpropagation)

La rétropropagation (< backpropagation > ou simplement < backprop >) est au cœur de l'apprentissage du deep learning, c'est la descente de gradient appliquée aux réseaux de neurones. Sans la backpropagation, nous ne pourrions pas apprendre efficacement les poids dans un réseau de neurones. Il deviendrait vite extrêmement coûteux en puissance de calculs d'entraîner un modèle de deep learning à mesure que la profondeur du réseau augmente.

Il s'agit là du chapitre le plus complexe de ce livre. Vous n'aurez probablement jamais de votre vie besoin d'implémenter vous-même une backpropagation, car tous les frameworks modernes de deep learning possède une implémentation optimisée de la backpropagation. Néanmoins, il est tout de même important de comprendre cet algorithme auquel le deep learning en dépendant pour son apprentissage.

La backpropagation a été introduit par des pionniers du domaine, David E. Rumelhart, Geoffrey E. Hinton et Ronald J. Williams en 1986[4]. Elle permet de quantifier l'erreur de chaque neurone par rapport à la sortie attendue, et de répartir ensuite cette erreur à travers le réseau pour mettre à jour les poids de chaque couche caché les unes après les autres.

Geoff Hinton after writing the paper on backprop in 1986



Fig. 35. – « G. Hinton après avoir écrit l'article sur la rétropropagation en 1986 au début du deuxième hiver de l'IA. « *Je suppose que vous n'êtes pas encore prêts pour cela, mais vos enfants vont adorer.* » Meme tiré du film *Retour vers le futur*. »

4.4.1 La backpropagation dans le processus d'apprentissage

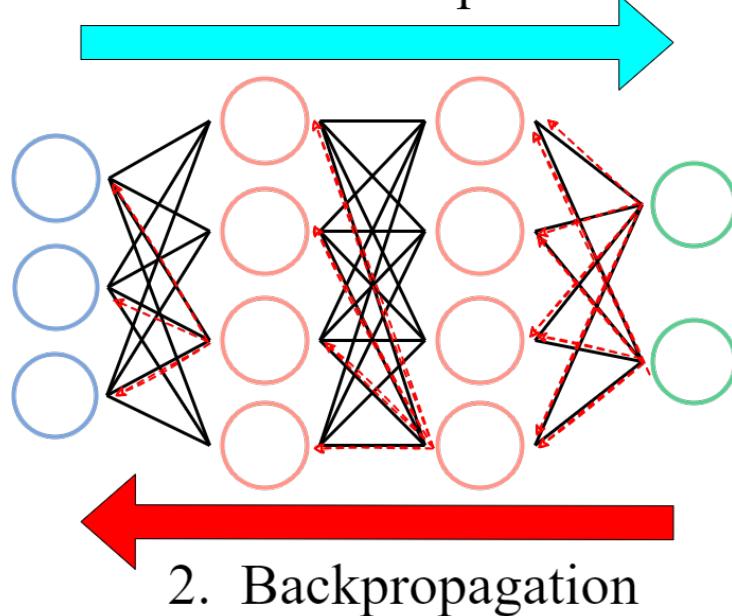
La première étape dans le processus d'apprentissage du modèle est la propagation en avant (forward propagation). Durant cette phase, le modèle fait une prédiction basée sur les données d'entrée (forward pass), en traitant l'information à travers les différentes couches de neurones, de la couche d'entrée à la couche de sortie. C'est lors de cette phase que le réseau quantifie l'erreur qu'il fait par rapport à la valeur cible, noté \hat{y} (y chapeau) cette erreur et calculée selon la fonction coût choisie par le développeur (MSE, BCE (Binary Cross Entropy) etc).

La deuxième étape est la rétropropagation (backpropagation). Avec l'erreur calculée lors de la forward pass, la backpropagation va prendre le chemin inverse de la propagation en avant en remontant l'erreur de la couche de sortie vers la couche d'entrée.

La backpropagation calcule le gradient de la fonction de coût par rapport à chaque paramètre du réseau, c'est-à-dire chaque poids et biais. Je rappelle que le gradient est votre GPS qui indique à l'algorithme dans quelle direction converger pour minimiser l'erreur. Si le gradient d'un poids est grand en valeur absolue, cela signifie qu'une petite modification de ce poids pourrait entraîner une grande réduction de l'erreur.

L'utilisation des gradients permet d'ajuster chaque paramètre du modèle pour réduire l'erreur avec la descente de gradient, la pondération de l'ajustement des paramètres se fait selon le taux d'apprentissage (learning rate). En répétant plusieurs fois ce processus de prédiction, calculer l'erreur, ajuster les paramètres du réseau et l'on répète jusqu'à que le modèle soit suffisamment entraîné.

1. Fordward pass



2. Backpropagation

Fig. 36. – « 1. La forward pass, la première étape, effectue des prédictions sur un lot de données (la taille du lot, ou < batch size >, est choisie par le développeur). 2. Avec l'erreur obtenue, la backpropagation calcule le gradient pour chaque paramètre (illustré en rouge) et effectue un ajustement proportionnel à ce gradient. Cet ajustement est fait dans le sens qui minimise la fonction de coût. »

4.4.2 Les Principes Fondamentaux de la Backpropagation

La backpropagation utilise les calculs différentiels, notamment les dérivées partielles et la règle de la chaîne. Nous avons déjà utilisé les dérivées partielles quand nous avions effectué une descente de gradient ayant 2 paramètres.

4.4.2.1 Le Rôle des Dérivées Partielles

Les dérivées partielles permettent de quantifier la manière dont un changement infime d'un poids ou d'un biais affecte la fonction de coût. Elles donnent une mesure de l'importance de chaque poids et biais dans la détermination de la sortie du réseau.

Plus la valeur absolue d'une dérivée partielle est grande, plus cela signifie qu'un petit changement du poids ou du biais (paramètre) correspondant entraînera une grande modification de la fonction de coût. À l'inverse, plus la dérivée partielle est proche de zéro, plus la fonction de coût sera insensible à des modifications de ce paramètre.

4.4.2.2 La Dérivée en Chaîne dans la Backpropagation

La règle de la chaîne est utilisée pour calculer la backpropagation, elle permet de calculer la dérivée d'une fonction composée. La backpropagation propage les erreurs de la couche de sortie vers la couche d'entrée.

Supposons que nous avons une fonction coût J (« J », comme « jacobien » de matrice jacobien utilisé lors de la descente de gradient), qui est une fonction de sortie d'un neurone ayant une fonction d'activation a (« a » comme « activation ») dans le réseau, c'est-à-dire $J = J(a)$ où a est lui-même une fonction des entrées pondérées à ce neurone, $a = \sigma(z)$, et z est une fonction

des poids (weight) w et des biais b , c'est-à-dire $z = w \cdot x + b$ la sortie du perceptron avant la fonction d'activation. Ici σ est la fonction d'activation du neurone.

Notre objectif est de comprendre comment un petit changement des poids w ou des biais b affectent la fonction coût J . Cela revient à calculer les dérivées partielles $\frac{\partial J}{\partial w}$ et $\frac{\partial J}{\partial b}$. Pour calculer cela, on applique une dérivée en chaîne :

$$\frac{\partial J}{\partial w} = \frac{\partial J}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial w}$$

De la même manière pour $\frac{\partial J}{\partial b}$ si nous cherchons à calculer l'effet du biais b à la place d'un poids w .

$$\frac{\partial J}{\partial b} = \frac{\partial J}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial b}$$

Dans ces deux formules permettent de calculer le gradient de J par rapport à tous les poids et biais du réseau, en une seule fois à travers le réseau.

4.4.3 L'apprentissage d'un réseau avec la Backpropagation : pas à pas

L'apprentissage du réseau est un processus en quatre étapes : feedforward, calcul de l'erreur, calcul du gradient et mise à jour des poids.

Dans cet exemple, nous allons examiner un réseau de neurones très simple avec une seule entrée, deux couches cachées contenant chacune un seul perceptron, et une seule sortie.

Pour comprendre comment fonctionne la backpropagation, nous allons utiliser l'exemple d'un réseau ayant un seul perceptron avec son biais par couche. Ce réseau aura une couche d'entrée, deux couches cachées et une couche de sortie.

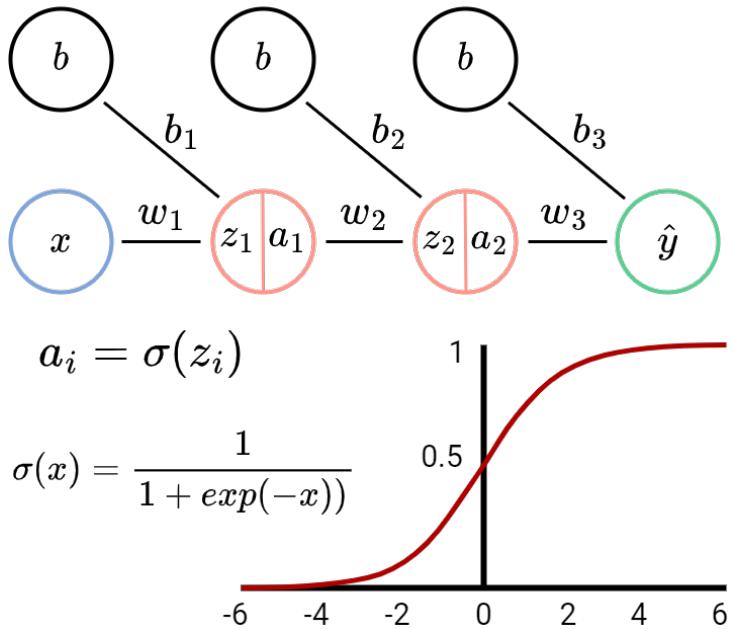


Fig. 37. – Notre petit réseau de neurones. Il se compose d'une entrée, de deux couches cachées (avec un perceptron et son biais dans chacune), et d'une couche de sortie. Le terme z_i représente la somme pondérée des entrées d'un perceptron avant l'application de la fonction d'activation σ , qui est une fonction sigmoïde dans notre cas.

4.4.3.1 Feedforward: la première étape

La première étape est quand un réseau de neurones fait une prédiction (forward pass). Dans cette phase, le réseau de neurones fait des prédictions en utilisant ses poids et biais actuels.

Prenons un réseau de neurones avec deux couches cachées, chacune ayant un seul perceptron, et une couche de sortie avec une sortie. Chaque perceptron utilise la fonction d'activation sigmoïde, et l'erreur est calculée à l'aide de l'erreur quadratique moyenne (MSE).

Pour simplifier, disons que notre réseau ne contient qu'un seul échantillon d'entrée, x , et une sortie attendue, y .

Tout d'abord, l'entrée x est multipliée par le poids w du premier perceptron, puis un biais b est ajouté. C'est ce qu'on appelle une combinaison linéaire.

$$z_1 = w_1 x + b_1$$

Ce résultat z_1 est ensuite passé à travers une fonction d'activation (la fonction sigmoïde dans notre cas) pour donner le résultat a_1 (pour « activation »). La fonction sigmoïde est définie par :

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Le processus est ensuite répété pour la seconde couche cachée et la couche de sortie.

$$a^1 = \sigma(z_1)$$

La deuxième couche cachée, fait de même avec la sortie de la première couche cachée, a_1 , pour produire une autre sortie, a_2 .

$$a^2 = \sigma(w^2 a^1 + b^2)$$

Voilà notre prédiction est faite, la prédiction est nommée \hat{y} . Mais, comment savons-nous à quel point nous avons bien (ou mal) fait ?

4.4.3.2 Calcul de l'erreur : la deuxième étape

C'est là qu'intervient notre fonction coût pour calculer l'erreur de notre modèle, la MSE ici. Nous prenons notre prédiction \hat{y} et la soustrayons à la valeur réelle y , on met au carré le résultat et nous avons l'erreur.

$$J = \frac{1}{2}(y - \hat{y})^2$$

Ici c'est l'erreur pour une seule prédiction pour faire simple. Pour un lot de données, nous ferions cela chaque prédiction, puis nous prendrions la moyenne de toutes ces erreurs. Pour le moment, restons-en à une seule prédiction.

Le facteur $\frac{1}{2}$ est une commodité mathématique pour la dérivation, ça simplifie les calculs que vous verrez plus tard.

Maintenant que nous savons à quel point nous avons bien (ou mal?) fait, comment utilisons-nous cette information pour améliorer notre réseau de neurones ?

4.4.3.3 Calcul du gradient: la troisième étape

C'est là qu'intervient la backpropagation. Nous allons prendre notre erreur et la « propager en arrière » à travers notre réseau pour trouver à quel point chaque poids et chaque biais a contribué à cette erreur.

La backpropagation calcule le gradient de chaque paramètre, c'est-à-dire les poids et les biais, chacun a son propre gradient. Notre réseau a deux couches cachées, chaque couche ayant un seul perceptron (un seul poids et un seul biais), plus un poids et un biais pour la couche de sortie, nous avons six gradients à calculer.

Les voici :

1. Le gradient du poids w^1 de la première couche cachée
2. Le gradient du biais b^1 de la première couche cachée
3. Le gradient du poids w^2 de la deuxième couche cachée
4. Le gradient du biais b^2 de la deuxième couche cachée
5. Le gradient du poids w^3 de la couche de sortie
6. Le gradient du biais b^3 de la couche de sortie

Cependant, pour simplifier notre explication, nous allons uniquement calculer le gradient pour le poids w^2 de la deuxième couche cachée. Ces calculs peuvent être appliqués de la même manière pour les autres paramètres du réseau.

Nous allons calculer le gradient de la fonction de coût par rapport à notre paramètre w^2 (le poids de la deuxième couche cachée) affecte la fonction coût J . Le gradient donnera une indication de la façon dont nous devons ajuster w^2 pour minimiser l'erreur.

On commence par la fonction de coût MSE que l'on a défini précédemment :

$$J = \frac{1}{2}(y - \hat{y})^2$$

Pour cela, on va utiliser la règle de la chaîne qui dit que la dérivée d'une fonction composée est le produit des dérivées. Donc, on va décomposer $\frac{\partial J}{\partial w^2}$ en trois parties.

1. La dérivée de J par rapport à \hat{y} (la prédiction), que l'on notera $\frac{\partial J}{\partial \hat{y}}$
2. La dérivée de \hat{y} par rapport à z^2 (la sortie du neurone avant la couche d'activation), que l'on notera $\frac{\partial \hat{y}}{\partial z^2}$.
3. La dérivée de z^2 par rapport à w^2 (le poids), que l'on notera $\frac{\partial z^2}{\partial w^2}$.

On a donc :

$$\frac{\partial J}{\partial w^2} = \frac{\partial J}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z^2} \cdot \frac{\partial z^2}{\partial w^2}$$

Allons-y étape par étape :

1. La dérivée de J par rapport à \hat{y} est simplement :

$$\frac{\partial J}{\partial \hat{y}} = y - \hat{y}$$

2. La dérivée de \hat{y} par rapport à z^2 est la dérivée de la fonction d'activation sigmoïde. Si on note $\sigma(z)$ la fonction sigmoïde, alors sa dérivée est $\sigma(z) \cdot (1 - \sigma(z))$. On a donc :

$$\frac{\partial \hat{y}}{\partial z^2} = \hat{y} \cdot (1 - \hat{y})$$

3. La dérivée de z^2 par rapport à w^2 est simplement la valeur de l'entrée a^1 , car $z^2 = w^2 \cdot a^1 + b^2$. Donc on a :

$$\frac{\partial z^2}{\partial w^2} = a^1$$

Si on empile tout ensemble on obtient :

$$\frac{\partial J}{\partial w^2} = (y - \hat{y}) \cdot \hat{y} \cdot (1 - \hat{y}) \cdot a^1$$

On a notre gradient du poids w^2 ! En utilisant cette formule, nous pouvons calculer la direction dans laquelle nous devons ajuster le poids w^2 pour minimiser la fonction de coût J .

4.4.3.4 Mise à jour des poids: la quatrième étape

Maintenant, supposons que nous avons nos six gradients, nous pouvons faire ce pour quoi nous sommes ici : mettre à jour nos poids et nos biais. Pour cela, nous prenons chaque poids et chaque biais et le déplaçons un petit peu dans la direction opposée à son gradient. C'est la descente de gradient que nous avons déjà vu.

Pour chaque poids :

$$w := w - \alpha \frac{\partial J}{\partial w}$$

Pour chaque biais :

$$b := b - \alpha \frac{\partial J}{\partial b}$$

Où α est le taux d'apprentissage (learning rate) et $\frac{\partial J}{\partial w}$ ou $\frac{\partial J}{\partial b}$ est le gradient de la fonction coût rapport à w ou b .

Voilà, enfin, une itération complète de la backpropagation ! C'est le processus que notre réseau de neurones parcourt à chaque fois qu'il apprend à partir d'un lot (batch) de nos données.

4.4.3.5 Exemple en code

Voici un exemple de code en PyTorch pour illustrer ces étapes

```
import torch
from torch import nn

# Définition de notre réseau de neurones simple
class SimpleNetwork(nn.Module):
    def __init__(self):
        super(SimpleNetwork, self).__init__()
        self.layer1 = nn.Linear(1, 1)
        self.layer2 = nn.Linear(1, 1)
        self.layer3 = nn.Linear(1, 1)

    def forward(self, x):
        x = torch.sigmoid(self.layer1(x))
        x = torch.sigmoid(self.layer2(x))
        return self.layer3(x)

# Instanciation de notre réseau
net = SimpleNetwork()

# Définition de notre fonction de coût MSE
criterion = nn.MSELoss()

# Définition de notre optimiseur (SGD)
optimizer = torch.optim.SGD(net.parameters(), lr=0.01) # lr est notre taux
d'apprentissage

# Exemple d'entrée et de sortie
x = torch.tensor([2.0])
y = torch.tensor([1.0])

# Étape 1 : Feedforward
y_pred = net(x)

# Étape 2 : Calcul de l'erreur
loss = criterion(y_pred, y)

# Étape 3 : Backpropagation
loss.backward()
```

```

# Étape 4 : Mise à jour des poids
optimizer.step()

# Remise à zéro des gradients pour la prochaine étape
optimizer.zero_grad()

```

Dans ce code, `loss.backward()` effectue la backpropagation et calcule les gradients pour tous les paramètres du réseau. L'appel à `optimizer.step()` effectue ensuite la mise à jour des poids en fonction des gradients calculés. La dernière ligne, `optimizer.zero_grad()`, remet les gradients à zéro pour préparer la prochaine itération.

4.4.4 Les problèmes du Vanishing Gradient et de l'Exploding Gradient

Le vanishing gradient (gradient qui disparaît) et l'exploding gradient (gradient qui explose) sont des problèmes que la backpropagation rencontre avec des réseaux profonds, c'est-à-dire des réseaux avec de nombreuses couches cachées. Ils sont liés à la manière dont les gradients de la fonction coût sont calculés et propagés dans le réseau lors de la backpropagation. Ils se produisent lorsqu'un réseau de neurones est suffisamment profond, c'est-à-dire un réseau avec beaucoup de couche cachée, et rendent l'entraînement de tels réseaux difficile, voire impossible à entraîner.

Le problème du vanishing gradient se produit lorsque les gradients des couches profondes deviennent très petits, presque zéro. Cela rend le réseau de neurones très lent à apprendre, voire incapable d'apprendre. Ces petits gradients entraînent des modifications minuscules des poids, rendant l'apprentissage extrêmement lent ou stagnant. À l'inverse, le problème du gradient explosif se produit quand les gradients deviennent trop grands. Ce qui entraîne une mise à jour trop agressive des paramètres et provoque une instabilité et faire diverger l'apprentissage.

Mais pourquoi ces problèmes se produisent-ils ? Pour répondre à cette question, nous devons regarder de plus près le processus de backpropagation et la manière dont les gradients sont calculés.

Lors du calcul du gradient de la fonction de coût par rapport aux poids et aux biais, la règle de la chaîne est utilisée pour propager l'erreur de la couche de sortie à la couche d'entrée. Ce produit est ensuite transmis à la couche précédente. Si la dérivée de la fonction d'activation est très petite (proche de zéro), le produit sera également très petit, ce qui peut conduire à un vanishing gradient si ce processus est répété à travers de nombreuses couches. À l'inverse, si la dérivée de la fonction d'activation ou les poids du réseau sont trop grands, le produit sera également trop grand, ce qui peut entraîner un exploding gradient.

Certaines fonctions d'activation, comme la sigmoïde et la tangente hyperbolique (tanh), ont tendance à saturer, c'est-à-dire que leurs sorties tendent vers leurs limites pour les grandes valeurs d'entrée. Pour la sigmoïde, ces limites sont 0 et 1, tandis que pour la tanh, elles sont -1 et 1. Cela signifie que leurs dérivées deviennent très petites, proches de zéro, pour ces grandes entrées. C'est pourquoi ces fonctions d'activation ne sont pas utilisées pour des réseaux profonds, car elles causent du vanishing gradient.

Mathématiquement, la dérivée de la fonction sigmoïde est donnée par :

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

où $\sigma(x)$ est la fonction sigmoïde. De même, la dérivée de la tanh est :

$$\tanh'(x) = 1 - \tanh^2(x)$$

On peut voir que dans les deux cas, lorsque x est grand en valeur absolue, la dérivée tend vers 0. C'est pourquoi ces fonctions d'activation causent du vanishing gradient dans les réseaux profonds

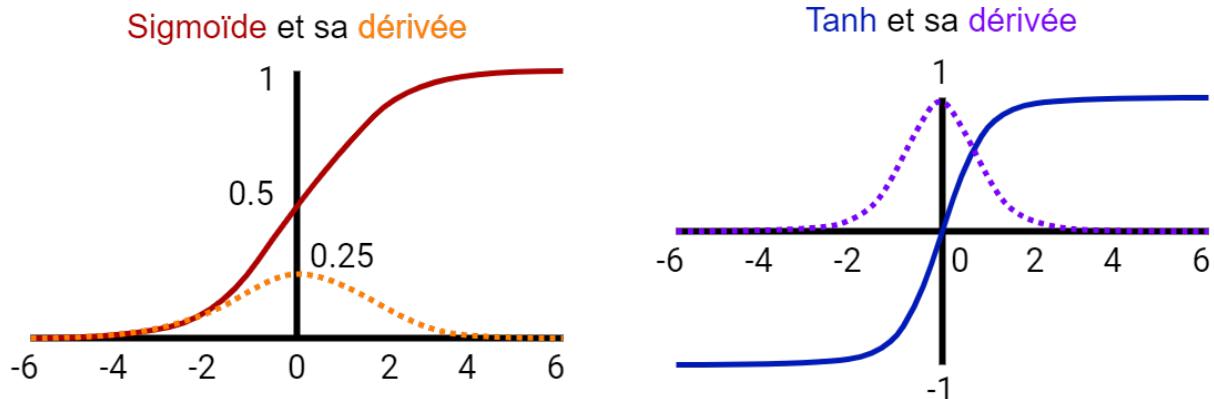


Fig. 38. – Fonction d'activation sigmoïde et Tanh avec leur dérivée.

Pour illustrer le vanishing gradient, prenons l'exemple d'un réseau de neurones avec 5 couches cachées, et supposons que le gradient de la fonction d'activation est de 0,1 pour chaque neurone. Le gradient pour la première couche cachée serait alors de $0,1^5 = 0,001$. Cela rendrait la mise à jour des poids de cette couche presque insignifiante, ralentissant considérablement l'apprentissage, voire le rendant impossible. Même si mathématiquement le réseau finira par converger au bout d'un long moment, en informatique la précision n'étant pas infinie, le modèle pourrait ne jamais converger.

Pour éviter le vanishing gradient, une solution est d'utiliser des fonctions d'activation qui ne saturent pas, comme la fonction ReLU (Unité Linéaire Rectifiée). La ReLU est définie comme :

$$f(x) = \max(0, x)$$

Sa dérivée est donc :

$$f'(x) = \begin{cases} 1 & \text{si } x > 0 \\ 0 & \text{sinon} \end{cases}$$

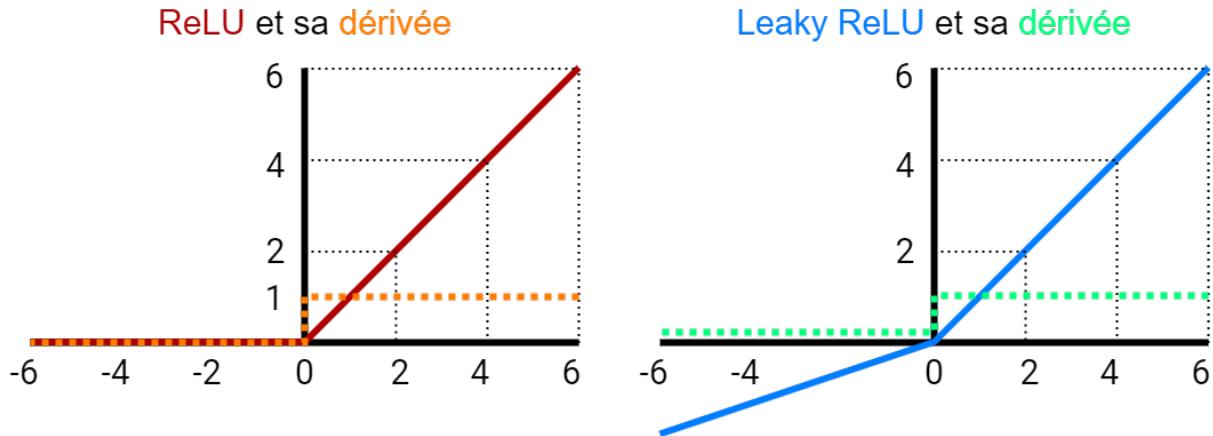


Fig. 39. – Fonction d’activation ReLU et Leaky ReLU avec leur dérivée.

Ainsi, la ReLU ne sature pas pour les entrées positives, évitant le problème du vanishing gradient. Néanmoins, la ReLU n’est pas sans défaut. Si les neurones sont « cloués » à 0, alors les poids auront un gradient nul, ce qui entraînera le problème du « dying ReLU ». C’est pourquoi des variantes comme la Leaky ReLU ont été proposées, avec une petite pente non nulle pour les entrées négatives :

$$f'(x) = \begin{cases} 1 & \text{si } x > 0 \\ \alpha & \text{sinon (avec } \alpha \text{ une petite constante, par exemple 0,01)} \end{cases}$$

Quant à l’exploding gradient, il se produit lorsqu’un gradient devient trop grand, entraînant une mise à jour excessive des poids lors de la descente de gradient, comme si le taux d’apprentissage était anormalement élevé. Le modèle peut alors osciller autour du minimum global ou même diverger complètement. Ce phénomène est plus fréquent dans les tâches où les sorties sont sensibles aux valeurs d’entrée plus anciennes, comme en traitement du langage naturel.

Reprendons l’exemple précédent, mais supposons maintenant que le gradient de la fonction d’activation est de 10 pour chaque neurone. Le gradient pour la première couche cachée serait alors de $10^5 = 100000$, ce qui entraînerait une mise à jour massive des poids de cette couche, déstabilisant pour le modèle.

Des poids initiaux élevés dans le réseau peuvent contribuer à l’exploding gradient. S’ils sont initialisés avec de grandes valeurs, cela engendrera des valeurs d’activation élevées, qui seront multipliées lors de la rétropropagation pour donner des gradients encore plus importants, amplifiant le problème. Pour éviter cela, des techniques d’initialisation comme celle de Xavier[5]. sont utilisées afin de calibrer correctement les poids initiaux en fonction de la taille des couches adjacentes.

4.4.5 Avenir de la Backpropagation dans le Deep Learning

La recherche sur la manière dont nos algorithmes apprennent est un sujet de recherche très actif, de meilleurs algorithmes d’apprentissage résoudraient d’énorme problème, comme avoir besoin de moins de données (de bonnes qualités notamment) et moins de puissance de calculs s’ils apprennent plus vite.

À l'heure actuelle, les alternatives à la backpropagation classique restent moins performantes classique[6], [7], [8], [9]. Cependant, nul ne peut prédire avec certitude ce que nous réserve l'avenir. Il est tout à fait envisageable que dans quelques années, la backpropagation soit supplante par de nouvelles méthodes plus avancées.

Geoffrey Hinton, l'un des pères de la backpropagation, travaille encore activement sur de nouveaux algorithmes en quête d'une alternative plus efficace à la backpropagation[10]. G. Hinton et Y. Bengio, tous deux lauréats du prix Turing (l'équivalent du prix Nobel dans le domaine de l'informatique) pour leur travail sur le deep learning, admettent que la backpropagation n'est probablement pas l'approche optimale pour l'apprentissage en deep learning. Ces chercheurs étudient la manière de créer des modèles de deep learning « intelligent », et il semble improbable que notre cerveau apprenne avec un algorithme similaire à la backpropagation.

De nombreux chercheurs dans le domaine pensent que la clé d'une meilleure méthode d'apprentissage pourrait résider dans la biologie et le fonctionnement de notre cerveau[11]. Ces recherches, bien que prometteuses, sont encore à leurs débuts. Il est trop tôt pour déterminer si les futures alternatives à la backpropagation seront inspirées par la biologie ou suivront une voie radicalement différente.

L'exploration de ces méthodes alternatives s'accompagne souvent d'une réflexion sur les mécanismes d'apprentissage dans le cerveau. Les neurones biologiques communiquent par des spikes, des impulsions électriques brèves et intenses, très différentes des activations continues utilisées dans les réseaux de neurones artificiels. De plus, la plasticité synaptique, le mécanisme par lequel les connexions entre neurones sont renforcées ou affaiblies, est bien plus complexe que la simple règle de mise à jour des poids utilisée dans la backpropagation.

Des modèles comme les réseaux de neurones à spikes (spiking neural networks) cherchent à se rapprocher de la biologie en utilisant des neurones à spikes et des règles d'apprentissage inspirées de la plasticité synaptique[12]. Bien que prometteuses, ces approches sont encore loin d'égaler les performances de la backpropagation sur les tâches classiques de deep learning et est kplutot un domaine de recherche pour des modèles dans des systèmes embarqués puisque ce genre d'architectures utilisent moins de puissance de calculs.

4.5 Résumé

La fonction de perte, aussi appelée fonction d'erreur, est un élément central du deep learning. Elle mesure la différence entre les prédictions d'un modèle \hat{y} et les valeurs réelles y . La fonction de coût est la moyenne des pertes pour l'ensemble d'un batch, la taille de ce batch, le batch size est défini par le développeur. L'objectif est de minimiser la fonction de perte en trouvant l'ensemble de paramètres qui réduit au maximum l'erreur calculée par la fonction de perte.

Il existe différents types de fonctions de perte selon que le problème est de régression ou de classification. Pour les problèmes de régression, deux fonctions de perte communes sont l'Erreur Absolue Moyenne (MAE) et l'Erreur Quadratique Moyenne (MSE).

- La MAE est la moyenne des valeurs absolues des différences entre les prédictions du modèle et les valeurs réelles. Elle est utile car elle peut être interprétée directement dans les unités de la variable que vous essayez de prédire. Par exemple, si vous prédissez les températures en

degrés Celsius et que votre MAE est de 2, cela signifie que vos prédictions sont en moyenne à 2 degrés de la véritable température.

- La MSE est une alternative à la MAE, elle mesure la moyenne des carrés des erreurs, c'est-à-dire la moyenne des différences au carré entre les valeurs prédictes et les valeurs réelles. La MSE pénalise plus lourdement les grandes erreurs que les petites, ce qui la rend plus sensible aux valeurs aberrantes que d'autres métriques d'erreur comme la MAE.

Pour les problèmes de classification, la Cross-Entropy Loss est la plus utilisée, elle mesure la dissimilarité entre la distribution de probabilité prédictée par le modèle et la vérité terrain. Il existe deux types de Cross-Entropy Loss : Binary Cross-Entropy pour les problèmes de classification binaire, et Categorical Cross-Entropy pour les problèmes de classification multilabels.

- La Binary Cross-Entropy est utilisée lorsque nous avons deux classes possibles (0 et 1). Elle pénalise fortement les prédictions qui sont loin des étiquettes réelles, ce qui encourage le modèle à faire des prédictions précises.
- La Categorical Cross-Entropy est utilisée pour les problèmes de classification multilabels. Elle est similaire à la Binary Cross-Entropy mais adaptée pour plus de deux classes.

Nous avons vu que la descente de gradient est une méthode d'optimisation, elle optimise nos paramètres pour minimiser une fonction de coût (fonction erreur). Elle procède par itérations, modifiant progressivement selon le taux d'apprentissage (learning rate) les paramètres du modèle afin de minimiser la fonction de coût.

Dans le cas d'un modèle à un seul paramètre, la mise à jour des paramètres est effectuée en utilisant la règle de mise à jour de la descente de gradient.

La notion de minimum local est introduite, qui est un point où la fonction de coût atteint une valeur basse, mais pas nécessairement la plus basse possible. L'algorithme de descente de gradient peut se retrouver coincé dans ces minimums locaux.

Les points de selle sont également abordés. Ce sont des points où la courbe est à la fois concave et convexe, formant un creux dans une direction et un pic dans l'autre. Pour la descente de gradient, un point de selle est un endroit où le gradient de la fonction coût est nul.

Enfin, la descente de gradient stochastique (SGD) est mentionnée. C'est une variante de la descente de gradient qui résout plusieurs problèmes d'optimisation. La SGD met à jour les poids pour chaque mini-lot du jeu de données, ce qui la rend plus réalisable et moins coûteuse en puissance de calcul pour de grands ensembles de données.

Nous avons vu la backpropagation qui est l'application de la descente de gradient pour les réseaux de neurones. Cet algorithme a été introduit par David E. Rumelhart, Geoffrey E. Hinton et Ronald J. Williams en 1986, il permet de quantifier l'erreur de chaque neurone par rapport à la sortie attendue et de la répartir à travers le réseau pour mettre à jour les poids de chaque couche cachée.

L'apprentissage par backpropagation comprend deux étapes principales : la forward pass et la backpropagation. La forward pass prédit la sortie basée sur les données d'entrée, quantifie

l'erreur par rapport à la valeur à prédire y et la calcule selon la fonction de coût choisie. La backpropagation prend ensuite cette erreur et la remonte à travers le réseau, en calculant le gradient de la fonction de coût par rapport à chaque paramètre du réseau.

La backpropagation utilise des calculs différentiels, notamment les dérivées partielles et la règle de la chaîne, pour calculer le gradient de la fonction de coût par rapport à tous les poids et biais du réseau en une seule fois.

Cependant, la backpropagation peut rencontrer des problèmes, tels que le vanishing gradient ou exploding gradient. Ces problèmes sont liés à la manière dont les gradients de la fonction de coût sont calculés et propagés dans le réseau. De nombreuses recherches sont menées pour trouver des alternatives à la backpropagation, mais à l'heure actuelle, aucune n'a encore réussi à surpasser la backpropagation classique.

4.6 Questions

1. **Qu'est-ce que la descente de gradient ?**
 - a. Une méthode d'optimisation pour trouver le minimum d'une fonction
 - b. Un algorithme pour trouver le maximum d'une fonction
 - c. Une méthode pour calculer la dérivée d'une fonction
 - d. Une technique pour trouver le zéro d'une fonction
2. **Quelle est la grande faiblesse de la descente de gradient ?**
 - a. Elle nécessite beaucoup de mémoire pour stocker toutes les données
 - b. Elle peut conduire à des solutions sous-optimales ou nécessiter un grand nombre d'itérations pour atteindre une solution acceptable
 - c. Elle ne peut pas être utilisée avec des fonctions non linéaires
 - d. Elle nécessite que toutes les variables soient indépendantes
3. **Qu'est-ce que le gradient dans le contexte de la descente de gradient ?**
 - a. La direction dans laquelle se trouve le minimum de la fonction de coût
 - b. La vitesse à laquelle la fonction de coût change
 - c. La direction dans laquelle la fonction de coût augmente le plus rapidement
 - d. La valeur maximale que la fonction de coût peut atteindre
4. **Qu'est-ce qu'un minimum local dans le contexte de la descente de gradient ?**
 - a. Le point le plus bas de la fonction de coût
 - b. Un point où la fonction de coût est plus faible que dans les points voisins, mais pas nécessairement le plus bas de tous
 - c. Un point où la fonction de coût est plus élevée que dans les points voisins
 - d. Le point le plus élevé de la fonction de coût
5. **Qu'est-ce que le taux d'apprentissage dans l'algorithme de descente de gradient ?**
 - a. Le nombre d'itérations que l'algorithme va effectuer
 - b. La taille des pas que l'algorithme va faire à chaque itération
 - c. Le nombre de variables que l'algorithme va optimiser simultanément
 - d. La vitesse à laquelle l'algorithme va converger vers le minimum
6. **Pourquoi la descente de gradient stochastique est-elle souvent préférée à la descente de gradient classique ?**
 - a. Elle permet de traiter des ensembles de données plus importants
 - b. Elle est moins susceptible de tomber dans des minimums locaux
 - c. Elle est plus rapide à calculer
 - d. Toutes les réponses précédentes sont correctes
7. **Dans le contexte de la descente de gradient, qu'est-ce qu'un point de selle ?**
 - a. Un point où le gradient de la fonction de coût est nul
 - b. Un point où la fonction de coût atteint son maximum
 - c. Un point où la fonction de coût atteint son minimum
 - d. Un point où la fonction de coût change de signe
8. **Qu'est-ce qu'une fonction de perte (loss function) dans le contexte de l'apprentissage machine ?**
 - a. Elle quantifie la différence entre la prédiction d'un modèle et la valeur réelle.

- b. Elle mesure la précision de la prédiction d'un modèle.
- c. Elle calcule le temps nécessaire pour former un modèle.
- d. Elle évalue la complexité d'un modèle.

9. Pourquoi la MSE est-elle plus sensible aux valeurs aberrantes que la MAE ?

- a. Parce que chaque erreur est élevée au carré dans la MSE.
- b. Parce que chaque erreur est prise en valeur absolue dans la MSE.
- c. Parce que la MSE calcule la somme des erreurs, pas leur moyenne.
- d. Parce que la MSE utilise la racine carrée des erreurs.

Réponse : a

10. Pour un problème de classification multiclass j'utilise quelle loss ?

- a. MSE
- b. MAE
- c. Binary Cross-Entropy
- d. Categorical Cross-Entropy

11. Qu'est-ce que la backpropagation en deep learning ?

- a. Un moyen d'optimiser les poids dans un réseau de neurones
- b. Un algorithme pour générer de nouvelles données
- c. Un outil pour visualiser l'apprentissage d'un modèle
- d. Une méthode pour créer des architectures de réseaux de neurones

Réponse : a

12. Quelle est l'ordre du processus d'apprentissage d'un modèle ?

- a. Calcul de l'erreur, feedforward, calcul du gradient, mise à jour des poids.
- b. Feedforward, calcul du gradient, calcul de l'erreur, mise à jour des poids.
- c. Feedforward, calcul de l'erreur, calcul du gradient, mise à jour des poids.
- d. Calcul du gradient, feedforward, mise à jour des poids, calcul de l'erreur.

13. Quelle est la fonction de la backpropagation ?

- a. Elle calcule l'erreur du modèle
- b. Elle fait une prédiction basée sur les données d'entrée
- c. Elle calcule le gradient de la fonction de coût par rapport à chaque paramètre du réseau
- d. Elle calcule le taux d'apprentissage

14. Qu'est-ce que la < forward propagation >?

- a. Une méthode pour initialiser les poids dans un réseau de neurones.
- b. Le processus d'ajustement des poids dans un réseau de neurones.
- c. Le processus par lequel le modèle fait une prédiction basée sur les données d'entrée.
- d. Une mesure de l'erreur d'un réseau de neurones.

15. Qu'est-ce que le gradient dans le contexte de la backpropagation ?

- a. Une mesure de la vitesse d'apprentissage du modèle.
- b. Une mesure du temps de calcul nécessaire pour former le modèle.
- c. Une mesure de l'importance de chaque poids et biais dans la détermination de la sortie du réseau.

- d. Une mesure de la complexité du modèle.

16. **Qu'est-ce que le problème du vanishing gradient?**

- a. Un problème où les gradients deviennent trop grands, provoquant une mise à jour trop agressive des paramètres.
- b. Un problème où les gradients deviennent très petits, rendant le réseau de neurones très lent à apprendre.
- c. Un problème où le réseau de neurones ne peut pas apprendre de nouvelles caractéristiques.
- d. Un problème où le réseau de neurones oublie les caractéristiques qu'il a apprises précédemment.

4.7 Réponse

1. Qu'est-ce que la descente de gradient ?

Réponse: A - Une méthode d'optimisation pour trouver le minimum d'une fonction

2. Quelle est la grande faiblesse de la descente de gradient ?

Réponse: B - Elle peut conduire à des solutions sous-optimales ou nécessiter un grand nombre d'itérations pour atteindre une solution acceptable

3. Qu'est-ce que le gradient dans le contexte de la descente de gradient ?

Réponse: A - La direction dans laquelle se trouve le minimum de la fonction de coût

4. Qu'est-ce qu'un minimum local dans le contexte de la descente de gradient ?

Réponse: B - Un point où la fonction de coût est plus faible que dans les points voisins, mais pas nécessairement le plus bas de tous

5. Qu'est-ce que le taux d'apprentissage dans l'algorithme de descente de gradient ?

Réponse: B - La taille des pas que l'algorithme va faire à chaque itération

6. Pourquoi la descente de gradient stochastique est-elle souvent préférée à la descente de gradient classique ?

Réponse: D - Toutes les réponses précédentes sont correctes

7. Dans le contexte de la descente de gradient, qu'est-ce qu'un point de selle ?

Réponse: A - Un point où le gradient de la fonction de coût est nul

8. Qu'est-ce qu'une fonction de perte (loss function) dans le contexte de l'apprentissage machine ?

Réponse: A - Elle quantifie la différence entre la prédiction d'un modèle et la valeur réelle

9. Pourquoi la MSE est-elle plus sensible aux valeurs aberrantes que la MAE ?

Réponse: A - Parce que chaque erreur est élevée au carré dans la MSE

10. Pour un problème de classification multiclass j'utilise quelle loss ?

Réponse: D - La Categorical Cross-Entropy

11. Qu'est-ce que la backpropagation en deep learning ?

Réponse: A Un moyen d'optimiser les poids dans un réseau de neurones

12. Quelle est l'ordre du processus d'apprentissage d'un modèle ?

Réponse: C Feedforward, calcul de l'erreur, calcul du gradient, mise à jour des poids

13. Quelle est la fonction de la backpropagation ? Réponse: C

Elle calcule le gradient de la fonction de coût par rapport à chaque paramètre du réseau

14. Qu'est-ce que la forward propagation ?

Réponse: C Le processus par lequel le modèle fait une prédiction basée sur les données d'entrée.

15. Qu'est-ce que le gradient dans le contexte de la backpropagation ?

Réponse: c Une mesure de l'importance de chaque poids et biais dans la détermination de la sortie du réseau.

16. Qu'est-ce que le problème du vanishing gradient ?

Réponse: b Un problème où les gradients deviennent très petits, rendant le réseau de neurones très lent à apprendre.

5 Projet: Projet avec des MLP et mesurer ses performances

Dans ce chapitre, vous allez suivre et réaliser deux mini-projets guidés que nous mènerons de bout en bout à partir des connaissances théoriques acquises. Vous disposerez d'un cahier des charges détaillant la procédure à appliquer. Le premier mini-projet, nommé « Calculatrice », porte sur un problème de régression : il s'agira de créer un modèle capable d'effectuer une addition. Le second mini-projet, que nous appellerons « HandwrittenDigits », consistera à développer un modèle capable de reconnaître un chiffre écrit à la main dans une image.

5.1 Projet 1: Créer une calculatrice

Dans ce projet, nous allons résoudre un problème de régression : prédire le résultat d'une opération mathématique (addition ou multiplication). Nous utiliserons un jeu de données que nous allons générer nous-mêmes, composé de paires de nombres et de leurs résultats. Vous serez amené à construire un Modèle de Perceptron Multicouche (MLP) pour résoudre ce problème.

Dans ce chapitre, vous apprendrez à construire des réseaux de neurones avec le framework PyTorch, les projets seront relativement faciles à réaliser, l'idée n'est pas de réaliser un fantastique projet, mais de prendre en main un framework de deep learning.

5.1.1 Cahier des charges

1. **Générer un dataset pour une opération mathématique de deux nombres avec leurs résultats correspondants.** Pour cela, vous pourriez générer des paires de nombres entiers aléatoires dans une certaine plage, disons, de 0 à 100, mais je vous invite à jouer avec ces valeurs et calculer leur somme. Assurez-vous de générer suffisamment de données pour entraîner votre modèle, 1000 échantillons me semble correct, mais pouvez très bien essayer d'en générer 100 ou 10000 pour voir l'effet sur l'entraînement.
2. **Crée un jeu de données pour la phrase d'entraînement et la phase de test** ou alors générer un jeu de données que vous allez diviser pour les phases d'entraînement et de test, vous pouvez utiliser la fonction `train_test_split` de scikit-learn.
3. **Construire un modèle de MLP pour prédire le résultat de l'opération.** Le réseau aura deux neurones en entrée (pour les deux nombres à additionner) et un neurone en sortie (pour le résultat de l'addition). Vous pouvez commencer par un modèle simple avec une seule couche cachée, puis essayer d'ajouter plus de couches et voir l'influence du nombre de couches cachées sur les résultats. Je vous invite à tester des réseaux avec peu de couches cachées, mais avec beaucoup de perceptron par couche cachée et les comparer.
4. **Entraîner le modèle et optimiser les hyperparamètres.** Vous pouvez utiliser une stratégie d'optimisation comme la descente de gradient stochastique (SGD) ou un optimiseur plus sophistiqué comme Adam, j'expliquerai le fonctionnement de l'optimiseur Adam dans un autre chapitre. Assurez-vous de suivre attentivement la perte d'entraînement et la perte de validation pendant l'entraînement pour éviter le « sur-ajustement » (overfitting).
5. **Évaluer le modèle sur l'ensemble de test.** Une fois que vous êtes satisfait de votre modèle, évaluez-le sur l'ensemble de test pour voir comment il se comporte sur des données qu'il n'a jamais vues auparavant.

6. **Analyser les résultats.** Le réseau est-il capable de prédire correctement les sommes ? Comment la performance varie-t-elle avec différentes plages de nombres ? Le réseau est-il capable de généraliser à des nombres qu'il n'a jamais vus précédemment ?
7. **Évaluer le modèle en utilisant différentes métriques de performance et fonction de perte :**
 - Mean Squared Error (MSE)
 - Mean Absolute Error (MAE)
 - R-squared (R^2)

5.1.2 importer vos bibliothèques python et configurer votre matériel.

Dans notre mini-projet nous utiliser diverses bibliothèques Python très utilisées pour le deep learning.

```
import numpy as np # manipulation de tableaux
import matplotlib.pyplot as plt # visualisation de données
import torch # notre framework de deep learning
from torch import nn # définition d'architectures de réseaux de neurones (nn comme Neural Network)
from torch.utils.data import DataLoader, TensorDataset # gestion et la manipulation des ensembles de données
from tqdm import tqdm # affiche une barre de progression pendant l'entraînement du modèle
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score # métrique d'évaluation des modèles
import random # génération de nombres aléatoires
from torchsummary import summary # affiche le nombre de paramètres du modèle
```

Après avoir importé ces bibliothèques, la question du matériel se pose. Le deep learning est notoirement gourmand en ressources de calcul, et choisir entre un CPU et un GPU peut faire la différence entre un modèle qui prend des heures ou des jours à s'entraîner. Pour cela, nous allons utiliser PyTorch pour identifier le meilleur matériel disponible :

```
device = torch.device("cuda:0" if torch.cuda.is_available() else "cpu")
```

Cette ligne détecte si un GPU compatible avec CUDA est disponible, et dans ce cas, configure device pour utiliser ce GPU. Sinon, le CPU sera utilisé.

Pour une reproductibilité des résultats j'ai ajouté un seed qui vous permettra d'avoir exactement les mêmes résultats que moi si vous choisissez d'utiliser comme valeur 42, c'est une valeur aléatoire vous pouvez utiliser 1 ou toute autre valeur, et à chaque fois que lanceriez un entraînement, vous aurez toujours le même « hasard » produit par l'ordinateur.

```
def set_seed(seed_value=42):
    random.seed(seed_value)
    np.random.seed(seed_value)
    torch.manual_seed(seed_value)
    torch.cuda.manual_seed_all(seed_value)
    torch.backends.cudnn.deterministic = True
    torch.backends.cudnn.benchmark = False
```

Pour finaliser la configuration, nous exécutons ces lignes :

```

set_seed(42)

if torch.cuda.is_available():
    print("Le GPU est utilisé")
else:
    print("Le GPU n'est PAS utilisé, le CPU est utilisé")

```

Ces dernières lignes donnent un retour sur le matériel qui sera utilisé.

5.1.3 Le jeu de données

L'objectif de notre modèle de deep learning est de prédire le résultat d'opérations mathématiques comme les additions **ou** les multiplications sur deux nombres entier. Pour ce faire, nous devons générer un ensemble de données synthétiques qui représente bien le problème en question.

La fonction `create_calculator_dataset` est un générateur de données conçu pour ce but précis. Il génère une matrice où chaque ligne est une paire de nombre entiers aléatoires. Ensuite nous créons l'étiquettes, c'est-à-dire la réponse, ce que notre modèle doit prédire, le résultat d'une addition ou d'une multiplication, ces étiquettes seront stockés dans le vecteur `y`. Les modèles de deep learning sont bien plus performant quand les données d'entraînement sont normalisés, alors nous divisons `X` et `y` par des constantes qui garantit que les valeurs restent dans une plage plus maniable pour les modèles de machine learning, entre 0 et 1.

```

def create_calculator_dataset(num_samples, min_value, max_value, operation):
    X = np.random.randint(min_value, max_value+1, (num_samples, 2))
    if operation == 'add':
        y = X[:, 0] + X[:, 1]
    elif operation == 'multiply':
        y = X[:, 0] * X[:, 1]
    else:
        raise ValueError("Operation not recognized. Use 'add' or 'multiply'")

    X = X / max_value
    if operation == 'add':
        y = y / (2 * max_value)
    elif operation == 'multiply':
        y = y / (max_value * max_value)
    return X, y

```

- `num_samples` définit le nombre total d'échantillons que vous souhaitez générer. Plus vous aurez d'échantillons plus votre modèle sera reboste, cela enrichie de variété de combinaison possible votre dataset, ce qui aidera votre modèle à être plus performant sur une plus grande variété de données.
- `min_value` et `max_value` fixent la plage de valeurs pour les opérandes, nous irons de 0 à 100
- `operation` indique le type d'opération à effectuer.

Nous utiliserons notre fonction plus tard pour générer notre jeu de données.

5.1.4 Architecture du réseau

La prochaine étape consiste à définir l'architecture du réseau de neurones qui fera l'objet de notre étude. Nous allons utiliser PyTorch. Il existe deux approches pour définir une architec-

ture de deep learning avec PyTorch: l'une à l'aide Sequential de PyTorch et l'autre via une classe personnalisée qui hérite de nn.Module.

5.1.4.1 Modèle séquentiel

La première approche exploite la capacité de Sequential pour définir une séquence linéaire de couches. C'est une méthode simple et efficace pour des architectures avec un enchaînement ordonné de couche. Je vous la présente ici, mais je ne l'utiliserais peu, elle sera parfois utilisée dans des classes personnalisées héritant de nn.module pour définir une partie d'une architecture sophistiquée que nous verrons dans ce livre.

```
sequential_model = nn.Sequential(  
    nn.Linear(2, 40, bias=True), # Couche d'entrée  
    nn.ReLU(),  
    nn.Linear(40, 40, bias=True), # Première couche cachée  
    nn.ReLU(),  
    nn.Linear(40, 40, bias=True), # Deuxième couche cachée  
    nn.ReLU(),  
    nn.Linear(40, 40, bias=True), # Troisième couche cachée  
    nn.ReLU(),  
    nn.Linear(40, 40, bias=True), # Quatrième couche cachée  
    nn.ReLU(),  
    nn.Linear(40, 1, bias=True) # Couche de sortie  
)
```

Dans cette architecture, chaque nn.Linear représente un perceptron, il est nommé comme cela puisque qu'un perceptron effectue des calculs linéaires qui effectuent une somme pondérée des entrées. Dans le code le paramètre bias=True est le réglage par défaut pour les perceptrons dans PyTorch, indiquant que chaque perceptron à son propre biais individuel qui s'ajoute à sa sortie. Le paramètre bias est par défaut à True, j'aurai pu ne pas le préciser qu'il en serait de même pour la définition de l'architecture du réseau, je l'ai explicitement mentionné ici pour en clarifier sa présence, si vous ne voyez pas bias=False, c'est que le biais est inclus.

- **nn.Linear(2, 40)**: représente la couche d'entrée du réseau. Elle prend un vecteur d'entrée de taille 2 et le transforme en un vecteur de dimension 40. Ce nombre (40) est arbitraire et pourrait être modifié selon les besoins de l'application.
- **nn.ReLU()**: La fonction d'activation ReLU (Rectified Linear Unit) est utilisée après chaque couche linéaire. Elle introduit une non-linéarité dans le modèle, ce qui permet au réseau d'apprendre des fonctions non linéaires.
- **nn.Linear(40, 40)**: Ces couches représentent les couches cachées du réseau. Chacune prend un vecteur de taille 40 en entrée et produit un vecteur de taille 40 en sortie. Encore une fois, le choix de 40 est arbitraire. L'idée est souvent de conserver la même dimensionnalité pour toutes les couches cachées pour simplifier l'architecture, mais ce n'est pas une règle stricte. Il faut quand même éviter de réduire vers la fin du réseau, cela compresserait l'information qui avait été développé par les couches précédentes, soit on conserve la même largeur des couches tout au long, soit nous élargissons le réseau vers les dernières couches.
- **nn.Linear(40, 1)**: Cette dernière couche linéaire transforme le vecteur de taille 40 en un vecteur de taille 1, qui correspond à la sortie du réseau.

L'architecture du réseau est illustrée ci-dessous :

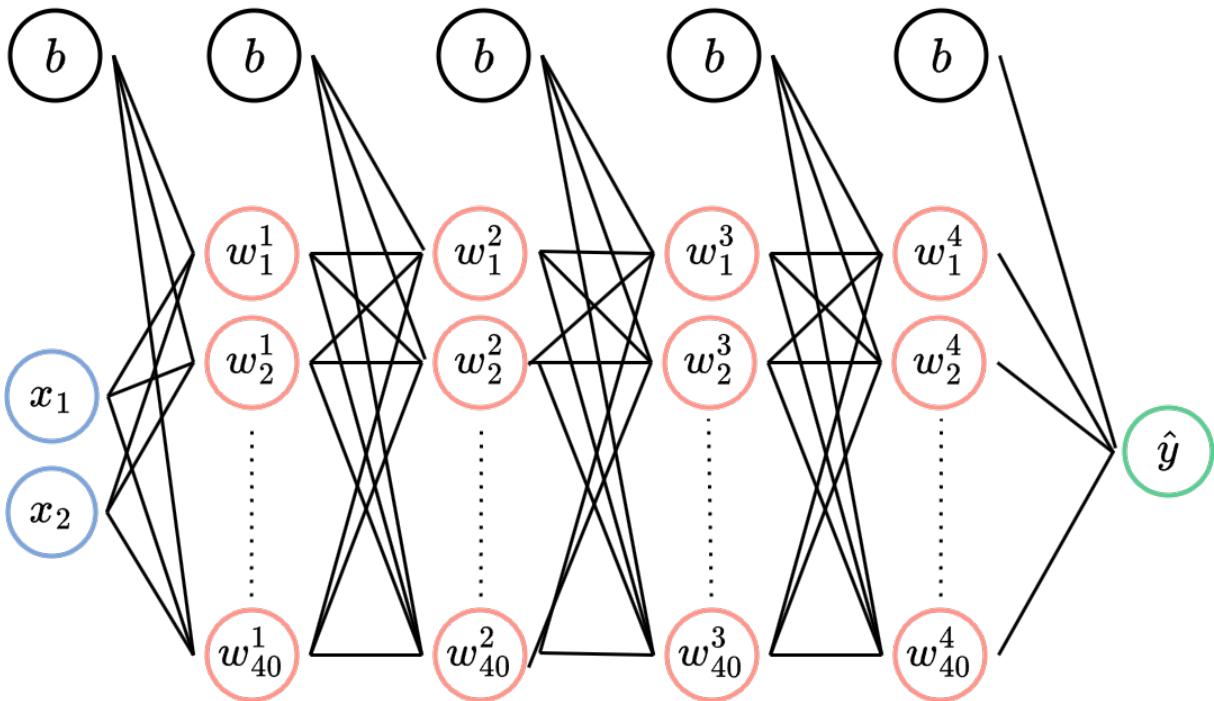


Fig. 40. – Architecture d'un réseau ayant 2 entrées, 4 couches cachées ayant chacune 40 perceptrons par couche cachée et 1 perceptron dans la couche de sortie, chaque perceptron à son propre biais.

5.1.4.2 Pourquoi 40 ?

Le choix du nombre « 40 » est une valeur complètement aléatoire. En général, des valeurs correspondant à une puissance de 2 sont généralement choisies (2, 4, 8, 16, 32, 64...). Il est possible de choisir une architecture plus large, mais cela risque de ne pas améliorer les performances et augmente le risque de sur-apprentissage (overfitting).

5.1.4.3 Quelle architecture choisir ?

Le choix de l'architecture de votre modèle de deep learning est davantage un art qu'une science exacte. Il n'existe pas de théorème mathématique pour déterminer quelle doit être votre architecture pour votre problème et avec vos données. La définition de votre architecture est un processus itératif d'expérimentation et d'ajustement. Vous apprendrez rapidement à choisir les bonnes architectures de deep learning en expérimentant. C'est pour cela que je vous conseille vivement d'essayer par vous-même les codes du livre qui se trouvent sur le répertoire [GitHub](https://github.com/MatteoEleouet/DeepLearningEnFrancais) du livre. URL : <https://github.com/MatteoEleouet/DeepLearningEnFrancais>

5.1.4.3.1 Complexité du Problème

Le premier facteur à considérer est la complexité intrinsèque du problème à résoudre. Si on décide de créer un jeu de donnée ayant une plage de 0 à 100 000 à la place de 0 à 100 comme actuellement, la tâche devient plus sophistiquée, et nécessitera une architecture plus sophistiquée pour résoudre un problème plus sophistiqué.

5.1.4.3.2 Contraintes Computationnelles

La disponibilité des ressources computationnelles peut également influencer le choix de l'architecture. Des architectures plus grandes et plus complexes nécessiteront plus de mémoire et de temps de calcul, ce qui peut être problématique pour des applications en temps réel ou des dispositifs avec des capacités limitées.

Des entraînements sur des réseaux de taille moyenne (en million de paramètres) peut nécessiter plusieurs jours d'entraînement selon votre GPU.

5.1.4.4 Modèle Basé sur une Classe Personnalisée

Gagner en flexibilité vous permettra de créer de meilleures architectures de deep learning. Cette seconde méthode consiste à définir une classe qui hérite de `nn.Module`. Elle permet une personnalisation plus poussée de l'architecture, en particulier pour des structures de réseau plus complexes. Par conséquent, c'est la méthode utilisée par les chercheurs et ingénieurs du domaine et celle que nous utiliserons tout le long du livre.

Lorsque vous définissez une classe personnalisée en PyTorch, deux méthodes essentielles doivent être implémentées :

- **`__init__`**: Cette méthode est appelée lors de l'initialisation de votre objet. Ici, vous définirez tous les composants nécessaires à votre réseau, tels que les couches linéaires, les fonctions d'activation et autres techniques de deep learning.
- **`forward`**: La méthode `forward` définit la logique de la propagation avant, autrement dit le chemin emprunté par le signal à travers le réseau lors de l'apprentissage ou de l'évaluation.

Maintenant que ceci est compris créons une classe personnalisée reprenant l'architecture conçue avec `nn.Sequential`.

```
class MLP(nn.Module):  
    def __init__(self):  
        super(CustomModel, self).__init__()  
        self.layer1 = nn.Linear(2, 40)  
        self.layer2 = nn.Linear(40, 40)  
        self.layer3 = nn.Linear(40, 40)  
        self.layer4 = nn.Linear(40, 40)  
        self.layer5 = nn.Linear(40, 40)  
        self.layer_out = nn.Linear(40, 1)  
        self.relu = nn.ReLU()  
  
    def forward(self, x):  
        x = self.layer1(x)  
        x = self.relu(x)  
        x = self.layer2(x)  
        x = self.relu(x)  
        x = self.layer3(x)  
        x = self.relu(x)  
        x = self.layer4(x)  
        x = self.relu(x)  
        x = self.layer5(x)  
        x = self.relu(x)
```

```

        x = self.layer_out(x)
    return x

```

La fonction `super(MLP, self).__init__()` permet d'hériter de toutes les fonctionnalités de la classe parente `nn.Module`. Les couches sont définies dans le constructeur (`__init__`) et la logique de la propagation vers l'avant (la forward pass) est mise en place dans la méthode `forward`. Si ceci n'est pas très clair pour vous, je vous invite à comprendre les concepts de Programmation Orienté Objet (POO) en Python.

5.1.4.4.1 Classe Évolutive

Comme dit précédemment, pour découvrir la bonne architecture il faut expérimenter alors nous allons la définir notre class de cette manière :

```

class MLP(nn.Module):
    def __init__(self, input_size=2, hidden_size=40, output_size=1,
num_hidden_layers=4, activation_fn=nn.ReLU()):
        super(MLP, self).__init__()
        self.layers = nn.ModuleList([nn.Linear(input_size, hidden_size)])
        self.layers.extend([nn.Linear(hidden_size, hidden_size) for _ in
range(num_hidden_layers-1)])
        self.output_layer = nn.Linear(hidden_size, output_size)
        self.activation_fn = activation_fn

    def forward(self, x):
        for layer in self.layers:
            x = self.activation_fn(layer(x))
        x = self.output_layer(x)
        return x

```

Dans cette version de notre classe `MLP` (*Multi Layer Perceptron*), nous avons introduit plusieurs arguments dans le constructeur pour permettre une flexibilité, en utilisant `nn.ModuleList`, une classe conteneur spécialement conçue pour stocker une liste de couches. Nous utilisons une boucle pour générer dynamiquement le nombre désiré de couches cachées, tel que spécifié par le paramètre `num_hidden_layers`. Cet artifice rend l'architecture du réseau facilement modulable, simplifiant ainsi les expérimentations avec différentes profondeurs de réseau. Vous pourrez l'influence d'ajouter plus ou moins de profondeur à votre réseau. Vous pourrez aussi ajouter plus ou moins de largeur (nombre d'unité par couche) à votre réseau avec le paramètre `hidden_size`.

La méthode `forward` représente la procédure de propagation avant (ou forward pass) du réseau. Lorsqu'un ensemble de données d'entrée (ou un batch) est introduit dans le réseau, il passe successivement à travers chaque couche cachée, subissant à chaque étape une transformation linéaire suivie d'une activation non-linéaire (ici, ReLU). Finalement, les données transformées sont acheminées vers la couche de sortie à travers `self.output_layer`.

Je vous invite à expérimenter avec les différentes fonctions d'activation. Chacune a ses propres avantages et inconvénients, et le choix peut affecter significativement les performances de votre modèle. Ce n'est qu'en expérimentant que vous pourrez véritablement comprendre leur impact et leur utilité dans des contextes spécifiques.

5.1.4.5 L'importance de la profondeur et la largeur d'un modèle

La profondeur du modèle se réfère au nombre de couches cachées, augmenter la profondeur du modèle signifie ajouter des couches cachées, un réseau large serait le nombre de perceptron par couche.

La profondeur de votre réseau est la capacité à capturer des représentations hiérarchiques. Chaque couche cachée supplémentaire permet au modèle d'effectuer une transformation non linéaire additionnelle sur les données entrantes. Si vous n'avez pas assez de couche cachée vous n'aurez peut-être pas assez traité votre donnée d'entrée pour en extraire les caractéristiques nécessaires pour que la fin de votre réseau (tête de classification) puisse avoir des caractéristiques pertinentes pour classifier votre donnée.

Néanmoins, à partir d'un nombre de couches, ajouter des couches cachées n'améliore plus la performance, voir même la réduire, car un réseau plus profond est un réseau plus difficile à entraîner. Effet que l'on comprend quand on calcule une backpropagation d'un réseau très profond, au fur et à mesure l'instabilité devient forte.

La largeur de votre réseau est la capacité est complémentaire à la profondeur. Une couche large possède davantage d'unités neuronales capables de traiter simultanément différentes caractéristiques pertinentes à un même niveau hiérarchique. Les modèles larges sont particulièrement adaptés lorsque les données possèdent une grande quantité d'informations simultanées ou indépendantes, comme c'est le cas pour des images haute résolution ou des signaux multidimensionnels complexes. En augmentant la largeur, on évite ainsi le phénomène de goulot d'étranglement informationnel qui pourrait se produire si une couche trop étroite était utilisée. Néanmoins, comme pour la profondeur, augmenter indéfiniment la largeur ne conduit pas nécessairement à une amélioration des performances. Au-delà d'un certain seuil, ajouter davantage d'unités neuronales par couche peut entraîner une augmentation inutile du nombre total de paramètres sans bénéfice en termes de performance prédictive, voir même l'inverse, car les réseaux trop gros ont la capacité de trop bien retenir les informations et donc ne pas pouvoir se généraliser sur une autre distribution que la distribution du jeu d'entraînement.

Comment savoir à l'avance quelle profondeur et largeur doit avoir votre future modèle ?

Vous ne pouvez pas vraiment savoir à l'avance, avec l'expérience vous allez développer des intuitions du genre d'architecture qui fonctionne pour tel ou tel problème à résoudre, mais vous devrez toujours expérimenter plusieurs architectures pour en être sûr, le Deep Learning demeure fondamentalement une discipline expérimentale où plusieurs configurations doivent être testées afin d'identifier celle qui offre les meilleures performances pour un problème donné.

Sachez qu'il existe un nombre théorique infini d'architecture possible, mais certaines heuristiques peuvent guider ce choix :

- Pour traiter des données complexes nécessitant plusieurs niveaux hiérarchiques successifs (par exemple la reconnaissance visuelle), privilégier initialement une architecture profonde.
- Pour traiter des données riches en informations simultanées mais sans hiérarchie évidente (par exemple signaux multidimensionnels comme la voix), privilégier initialement une architecture large.

Vous devrez toujours tester empiriquement plusieurs combinaisons profondeur-largeur afin d'affiner progressivement vers l'architecture optimale. Un bon point de départ est de consulter la littérature scientifique récente pour identifier quelles architectures ont déjà fait leurs preuves sur des problèmes similaires. Vous verrez que la communauté scientifique a progressivement convergé vers un ensemble relativement restreint d'architectures de référence qui se sont révélées particulièrement efficaces pour certaines classes de problèmes. Typiquement, les architectures contenant des blocs de transformateurs domine le paysage.

5.1.5 Fonctions d'entraînement et de préparation

Maintenant que nous avons posé les bases de notre infrastructure — la création de données, la constitution du modèle et la configuration du matériel — nous pouvons aborder la quintessence de l'apprentissage machine : le processus d'entraînement. Maintenant nous aller jusqu'à la mise à jour des poids du modèle

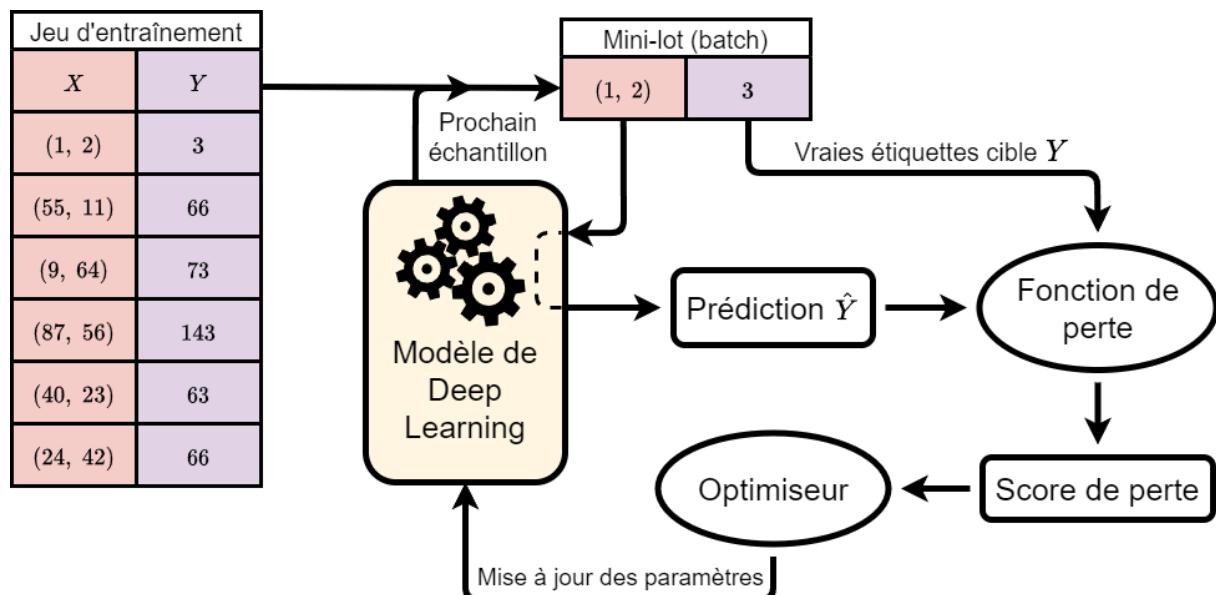


Fig. 41. – Représentation du processus d'entraînement d'un modèle de deep learning à partir d'un ensemble de données. Ici le modèle utilise un batch size de taille de 1.

5.1.5.1 Dataloader: Diviser pour mieux régner

Nous avons déjà vu comment générer un jeu de données synthétiques. Le DataLoader intervient à ce stade pour diviser ces données en lots (batches), ce qui permet une mise à jour plus fréquente des poids du réseau, accélérant ainsi la convergence. Si vous n'utilisez pas de DataLoader, votre réseau fera une prédiction sur le jeu de données en entier, soit une « époque », avant de faire la moindre optimisation. Par ailleurs, il est impératif de séparer les données en ensembles d'entraînement et de validation. L'ensemble de validation sert à estimer la performance du modèle sur des données non vues pendant l'entraînement, et offre une vision sur un surajustement potentiel, et à partir de combien d'époque.

```
from sklearn.model_selection import train_test_split

def create_calculator_dataloaders(num_samples=1000, min_value=0,
max_value=10,
                                operation='add', batch_size=64,
```

```

val_split=0.2):
    X, y = create_calculator_dataset(num_samples, min_value, max_value,
operation)
    X_train, X_val, y_train, y_val = train_test_split(X, y,
test_size=val_split, random_state=42)

    X_train_tensor = torch.tensor(X_train, dtype=torch.float32).to(device)
    y_train_tensor = torch.tensor(y_train, dtype=torch.float32).to(device)
    X_val_tensor = torch.tensor(X_val, dtype=torch.float32).to(device)
    y_val_tensor = torch.tensor(y_val, dtype=torch.float32).to(device)

    train_dataset = TensorDataset(X_train_tensor, y_train_tensor)
    train_dataloader = DataLoader(train_dataset, batch_size=batch_size,
shuffle=True)
    val_dataset = TensorDataset(X_val_tensor, y_val_tensor)
    val_dataloader = DataLoader(val_dataset, batch_size=batch_size)

    return train_dataloader, val_dataloader

```

5.1.5.2 L'Optimiseur: Le Maître d'Œuvre

L'optimiseur c'est le type d'algorithme de descente de gradient choisi pour minimiser la fonction perte. Deux des plus populaires sont la Descente de Gradient Stochastique (SGD) et Adam. Adam n'a pas encore été présenté, c'est une évolution de la SGD, je vous invite à essayer les deux pour voir s'il y a ou non des différences. L'algorithme Adam sera présenté dans un autre chapitre.

```

def get_optimizer(model, optimizer_name="sgd", lr=0.1):
    if optimizer_name.lower() == "adam":
        return torch.optim.Adam(model.parameters(), lr=lr)
    elif optimizer_name.lower() == "sgd":
        return torch.optim.SGD(model.parameters(), lr=lr)
    else:
        raise ValueError("Optimizer not recognized. Use 'adam' or 'sgd'.")

```

5.1.5.3 Le Processus d'Entraînement: Une Symphonie en Plusieurs Actes

L'entraînement d'un modèle de deep learning est un processus itératif qui se déroule en plusieurs étapes clés :

- Propagation avant, la prédiction (Forward Pass):** Les données d'entrée traversent le réseau de neurones pour générer une prédiction.
- Calcul de la Perte:** La prédiction est comparée à la vérité terrain à l'aide d'une fonction de perte, généralement une mesure d'erreur.
- Rétropropagation (Backpropagation):** Le gradient de cette perte est calculé par rapport à chaque paramètre du modèle.
- Optimisation:** Les poids du modèle sont mis à jour dans la direction qui minimise la perte.

Ce cycle est répété pour chaque lot d'échantillons (batch) de l'ensemble d'entraînement jusqu'à ce que le modèle ait vu toutes les données une fois. Chaque passage complet à travers l'ensemble de données est appelé une époque.

```

def train_model(model, train_dataloader, val_dataloader, epochs=10,
optimizer_name="sgd", lr=0.1):
    model.to(device)
    optimizer = get_optimizer(model, optimizer_name, lr)
    loss_fn = nn.MSELoss()
    train_loss_history, val_loss_history = [], []

    for epoch in tqdm(range(epochs), desc="Training Progress"):
        model.train()
        running_train_loss = 0.0
        for X_batch, y_batch in train_dataloader:
            optimizer.zero_grad()
            outputs = model(X_batch)
            loss = loss_fn(outputs.view(-1), y_batch)
            loss.backward()
            optimizer.step()
            running_train_loss += loss.item()

        epoch_train_loss = running_train_loss / len(train_dataloader)
        train_loss_history.append(epoch_train_loss)

        model.eval()
        running_val_loss = 0.0
        with torch.no_grad():
            for X_batch, y_batch in val_dataloader:
                outputs = model(X_batch)
                loss = loss_fn(outputs.view(-1), y_batch)
                running_val_loss += loss.item()

        epoch_val_loss = running_val_loss / len(val_dataloader)
        val_loss_history.append(epoch_val_loss)

    return train_loss_history, val_loss_history

```

5.1.6 Entraînement et visualisation des résultats

Il est temps de passer à la phase de l'entraîner notre de modèle. L'entraînement est le processus par lequel le modèle ajuste ses paramètres internes (ou « poids ») pour minimiser une fonction de perte. Cette fonction mesure l'écart entre les prédictions du modèle et les vérités de terrain. L'objectif est de parvenir à un modèle qui généralise bien, c'est-à-dire qui performe efficacement sur de nouvelles données jamais vues auparavant.

5.1.6.1 Initialisation du modèle

Nous débutons par la définition des paramètres (appelé « hyper-paramètre ») qui vont régir la création de notre jeu de données ainsi que la configuration de l'architecture de notre modèle.

```

# Paramètres pour la création du jeu de données
num_samples = 10000      # échantillons générer. Ex: 5000, 10000, etc.
min_value = 0             # Valeur minimale des nombres. Ex: 0, 10, -10, etc.
max_value = 100           # Valeur maximale des nombres. Ex: 100, 200 etc.
operation = 'add'         # Opération à effectuer. 'add' ou 'multiply'.
batch_size = 64            # Taille des lots pour l'entraînement. Ex: 32, 64,

```

```

etc.

learning_rate = 0.1           # Taux d'apprentissage. "adam", 0.001 ou "sgd",
0.1
optimizer = "sgd"            # Optimiseur. "adam" ou "sgd"

# Paramètres pour le modèle MLP
input_size = 2                # Taille de l'entrée. Pour notre cas, c'est toujours
2.
hidden_size = 16               # Nombre de perceptron dans les couches cachées. Ex: 8,
16, etc.
output_size = 1                # Taille de la sortie. Pour notre cas, c'est toujours
1.
num_hidden_layers = 3          # Nombre de couches cachées. Ex: 1, 2, 3, 4, etc.
activation_fn = nn.ReLU()     # Fonction d'activation. Ex: nn.ReLU(),
nn.Sigmoid() nn.Tanh() et nn.LeakyReLU().

# Paramètres pour l'entraînement
epochs = 10                   # Nombre d'époques pour l'entraînement. Ex: 5, 10, 20, etc.

```

Chaque hyperparamètre a un rôle précis :

- **num_samples** influence la quantité de données sur laquelle le modèle va s'entraîner. Un plus grand nombre d'échantillons peut améliorer la généralisation du modèle, mais augmente également le temps d'entraînement, car votre modèle exécutera plus d'itération par époque.
- **min_value** et **max_value** déterminent la plage des données et peuvent être ajustés pour tester la capacité du modèle à généraliser sur différentes plages de nombres.
- **operation** définit l'opération arithmétique que le modèle doit apprendre à résoudre. Changer cette opération peut grandement influencer la complexité de la tâche d'apprentissage, les multiplications sont plus difficiles à < approximer > qu'une addition.
- **batch_size** a un impact sur la stabilité et la vitesse de l'entraînement. Des lots plus petits peuvent conduire à une convergence plus rapide mais peuvent aussi causer une instabilité pendant l'entraînement.
- **input_size**, **hidden_size**, **output_size**, et **num_hidden_layers** définissent l'architecture du réseau. Varier ces valeurs peut avoir des conséquences significatives sur la capacité du modèle à apprendre des représentations complexes.
- **activation_fn** est cruciale pour introduire des non-linéarités dans le modèle, permettant d'apprendre des fonctions plus complexes que de simples combinaisons linéaires des entrées.
- **epochs** détermine le nombre de fois que le modèle verra l'ensemble du jeu de données. Un nombre insuffisant d'époques peut conduire à un sous-apprentissage, tandis qu'un nombre trop élevé peut causer du surapprentissage.
- **learning_rate** : Ce paramètre contrôle la taille des pas faits dans la direction du gradient. Un taux d'apprentissage trop élevé peut conduire à des sauts trop grands, manquant potentiellement le minimum, tandis qu'un taux trop bas peut ralentir l'entraînement et se coincer dans un minimum local.
- **optimizer** : Le choix de l'optimiseur peut influencer la vitesse et la qualité de la convergence du modèle. -SGD (Stochastic Gradient Descent) est un choix classique, mais des

optimiseurs plus avancés comme Adam peuvent accélérer l'entraînement et parfois aboutir à de meilleures performances.

L'expérimentation avec diverses configurations d'hyperparamètres est essentielle pour maîtriser l'art du deep learning. Je vous incite à manipuler le taux d'apprentissage, à osciller entre des valeurs élevées pour une convergence rapide et des valeurs faibles pour une plus grande finesse dans la recherche du minimum de la fonction de perte. Interrogez-vous sur l'optimiseur le plus adapté à votre problème : serait-ce Adam avec ses corrections adaptatives qui facilitent la navigation dans des paysages de perte complexes, ou SGD, dont la simplicité et la constance pourraient se montrer avantageuses dans certains contextes ?

Ne négligez pas l'impact potentiel des fonctions d'activation : chaque choix, qu'il s'agisse de ReLU, de Tanh, ou de LeakyReLU, apporte une dynamique différente à la propagation de l'activation dans votre réseau. Le deep learning est une discipline empirique où l'intuition se forge à travers la pratique et l'expérimentation. En testant une pléthore de combinaisons, vous affûterez non seulement la performance de votre modèle, mais aussi votre compréhension conceptuelle. Chaque essai est une étape vers une expertise accrue et chaque erreur, une leçon précieuse. C'est en explorant activement cette vaste étendue de possibilités que vous progresserez dans le domaine fascinant du deep learning.

Pour ceux qui visent l'excellence, je vous encourage à incorporer dans mon code des techniques d'optimisation des hyperparamètres, comme la recherche par grille (grid search) ou la recherche aléatoire (random search). Ces méthodes systématiques permettent d'explorer l'espace des hyperparamètres de manière structurée et peuvent être particulièrement utiles lorsqu'on travaille avec des jeux de données de petite ou moyenne taille. Cependant, dans des contextes professionnels où l'on doit composer avec des modèles de grande envergure et des contraintes temporelles strictes, ces méthodes peuvent se révéler prohibitives en termes de ressources et de temps.

Pour pallier cela, des techniques plus avancées telles que l'optimisation bayésienne offre un compromis entre exhaustivité et efficience, en se basant sur des principes probabilistes pour guider la recherche vers les régions les plus prometteuses de l'espace des hyperparamètres.

N'oubliez pas que l'entraînement de modèles de deep learning est souvent une entreprise de longue haleine. Les compétences acquises à travers l'expérimentation sur des modèles plus simples vous seront inestimables lorsque vous aborderez des problématiques plus ardues. Chaque expérimentation vous rapproche de la maîtrise du deep learning, et chaque échec est une occasion d'apprentissage. C'est en osant explorer et en apprenant de chaque essai que vous progresserez dans cet art.

5.1.6.2 Préparation des données

Les données sont préparées à l'aide de la fonction `create_calculator_dataloaders`, qui génère des jeux de données pour l'entraînement et la validation et les encapsule dans des DataLoaders. Cela permet de les découper en lots et d'optimiser le processus d'entraînement.

```
train_dataloader, val_dataloader = create_calculator_dataloaders(  
    num_samples=num_samples, min_value=min_value, max_value=max_value,  
    operation=operation, batch_size=batch_size, val_split=0.2  
)
```

5.1.6.3 Processus d'entraînement

Maintenant appelons notre fonction `train_model()` avec les hyper-paramètres choisis pour lancer le cycle d'apprentissage, orchestrant la descente de gradient par l'intermédiaire de l'optimiseur sélectionné (SGD ou Adam), ajustant les poids de notre modèle à travers les époques, tout en gardant trice de l'historique de perte pour l'analyse future.

```
train_loss_history, val_loss_history = train_model(  
    model, train_dataloader, val_dataloader, epochs=epochs,  
    optimizer_name=optimizer, lr=learning_rate  
)
```

5.1.6.4 Affichage des courbes de perte

Nous observons la représentation graphique des courbes de perte d'entraînement et de validation d'un modèle au fil des époques. L'axe des abscisses indique les époques, qui sont les cycles complets de passage de l'ensemble des données d'entraînement à travers le modèle. L'axe des ordonnées montre la perte, calculée par une fonction de coût, qui mesure l'écart entre les prédictions du modèle et les valeurs réelles. L'échelle logarithmique utilisée sur cet axe met en évidence les variations de la perte, même quand ces variations sont très petites, permettant une meilleure distinction entre les phases d'amélioration rapide et plus lente de la performance du modèle. La courbe bleue représente la perte d'entraînement qui diminue avec le temps, reflétant l'apprentissage du modèle, tandis que la courbe orange montre la perte de validation, qui évalue la performance du modèle sur un ensemble de données non vu pendant l'entraînement, un indicateur clé de la capacité du modèle à généraliser.

```
plt.figure(figsize=(12, 6))  
plt.plot(train_loss_history, label="Training Loss")  
plt.plot(val_loss_history, label="Validation Loss")  
plt.title("Training and Validation Loss")  
plt.xlabel("Epoch")  
plt.ylabel("Loss")  
plt.yscale('log') # Appliquer une échelle logarithmique  
plt.legend()  
plt.show()
```

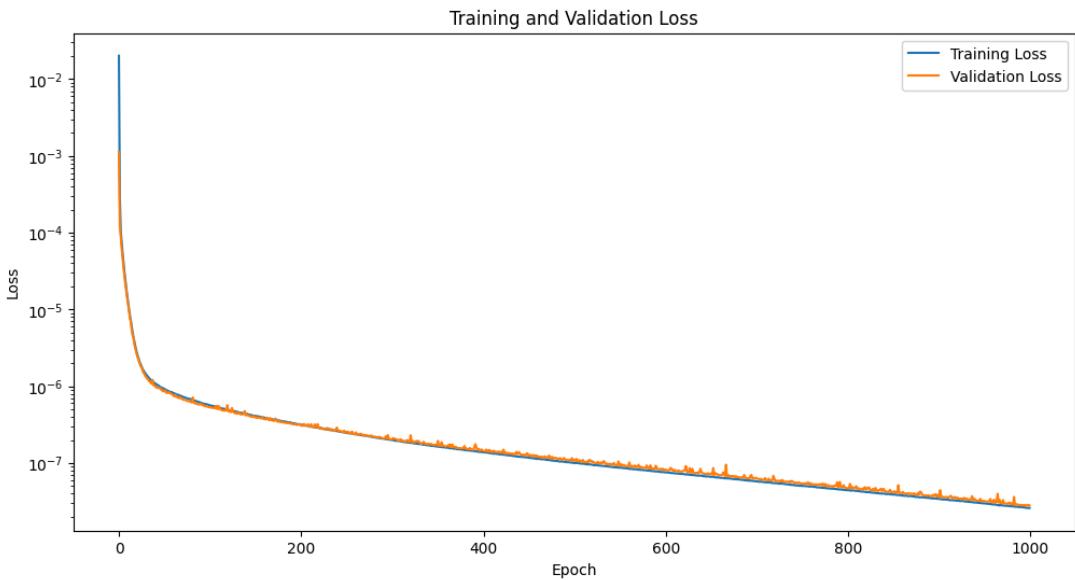


Fig. 42. – Courbes de perte d’entraînement (bleu) et de validation (orange) en fonction des époques, sur une échelle logarithmique. Ces tendances illustrent comment le modèle apprend et s’ajuste pour minimiser l’erreur à travers les cycles d’entraînement

5.1.6.5 Résumé de notre architecture

La fonction `summary` a été utilisée pour fournir un aperçu concis de l’architecture de notre modèle, révélant la configuration des couches, leurs formes de sortie et le décompte des paramètres. L’architecture se compose de quatre couches linéaires alternées avec des activations ReLU, chaque couche linéaire ayant respectivement 48, 272, 272 et 17 paramètres.

```
summary(model, input_size=(1, input_size))
```

Layer (type)	Output Shape	Param #
<hr/>		
Linear-1	[-1, 1, 16]	48
ReLU-2	[-1, 1, 16]	0
Linear-3	[-1, 1, 16]	272
ReLU-4	[-1, 1, 16]	0
Linear-5	[-1, 1, 16]	272
ReLU-6	[-1, 1, 16]	0
Linear-7	[-1, 1, 1]	17
<hr/>		
Total params:	609	
Trainable params:	609	
Non-trainable params:	0	
<hr/>		

5.1.7 Test et Visualisation: Révéler l’Âme du Modèle

Après avoir entraîné un modèle, l’étape suivante consiste à évaluer son efficacité et à interpréter ses performances. Cette phase ne se limite pas à l’obtention d’une métrique de performance. Il s’agit aussi de comprendre ce que le modèle a appris et comment il prend ses décisions.

```

# Fonction de test du modèle
def test_model(model, num_samples, min_value, max_value, operation):
    X_test, y_test = create_calculator_dataset(num_samples, min_value,
max_value, operation)
    X_test_tensor = torch.tensor(X_test, dtype=torch.float32).to(device)
    with torch.no_grad():
        y_pred = model(X_test_tensor)
        y_pred_denormalized = (y_pred * (2 *
max_value)).squeeze().cpu().numpy()
    return y_pred_denormalized, y_test * (2 * max_value)

# Fonction de tracé des prédictions
def plot_predictions(y_true, y_pred):
    plt.figure(figsize=(12, 8))
    plt.scatter(range(len(y_true)), y_true, color='green', label="True
values")
    plt.scatter(range(len(y_pred)), y_pred, color='blue',
alpha=0.5, label="Predictions")
    for i, (true, pred) in enumerate(zip(y_true, y_pred)):
        difference = pred-true
        if abs(difference) > 1:
            plt.annotate("", xy=(i, pred), xytext=(i, true),
arrowprops=dict(arrowstyle="->", color='red'))
    plt.title("True values vs. Predictions")
    plt.legend()
    plt.show()

```

Dans le code ci-dessus, `test_model` est une fonction qui prend en entrée un modèle entraîné, un nombre d'échantillons, une plage de valeurs et une opération mathématique. Elle génère un ensemble de données de test (`X_test` et `y_test`) et utilise le modèle pour faire des prédictions. L'utilisation de `torch.no_grad()` est utilisé lorsqu'on fait une inférence, ce qu'il se passe, c'est qu'on indique à PyTorch de ne pas stocker les gradients du modèle, d'où le « no grad ». Lors de l'entraînement, non seulement les poids du modèle sont chargés en mémoire, mais également les gradients associés à ces poids. Ces gradients, nécessaires pour la mise à jour des poids durant l'apprentissage, doublent pratiquement la quantité de mémoire requise. En revanche, lors de l'inférence, les gradients ne sont pas nécessaires car on n'entraîne pas le modèle, donc pas besoin des gradients et on économise beaucoup de RAM à notre GPU.

La fonction `plot_predictions` visualise les valeurs réelles (`y_true`) et les prédictions du modèle (`y_pred`). Les annotations rouges indiquent où les prédictions diffèrent significativement des valeurs réelles, ce qui aide à identifier les points où le modèle est moins précis.

Le code suivant montre comment ces fonctions sont utilisées pour tester et évaluer un modèle :

```

## Test et visualisation des prédictions
num_samples = 2000

y_pred, y_true = test_model(model, num_samples, min_value, max_value,
operation)
plot_predictions(y_true, y_pred)

```

```

#? Évaluation des performances à l'aide de différentes métriques
mse = mean_squared_error(y_true, y_pred)
mae = mean_absolute_error(y_true, y_pred)
r2 = r2_score(y_true, y_pred)

print(f"Mean Squared Error (MSE): {mse:.4f}")
print(f"Mean Absolute Error (MAE): {mae:.4f}")
print(f"R-squared (R2): {r2:.4f}")

>>> Mean Squared Error (MSE): 1.6901
>>> Mean Absolute Error (MAE): 1.0354
>>> R-squared (R2): 0.9990

```

Ce segment de code évalue le modèle en utilisant des métriques standard comme l'erreur quadratique moyenne (MSE), l'erreur absolue moyenne (MAE), et le coefficient de détermination R^2 , qui mesure la proportion de la variance des données expliquée par le modèle. Ces mesures fournissent une vue quantitative de la performance du modèle.

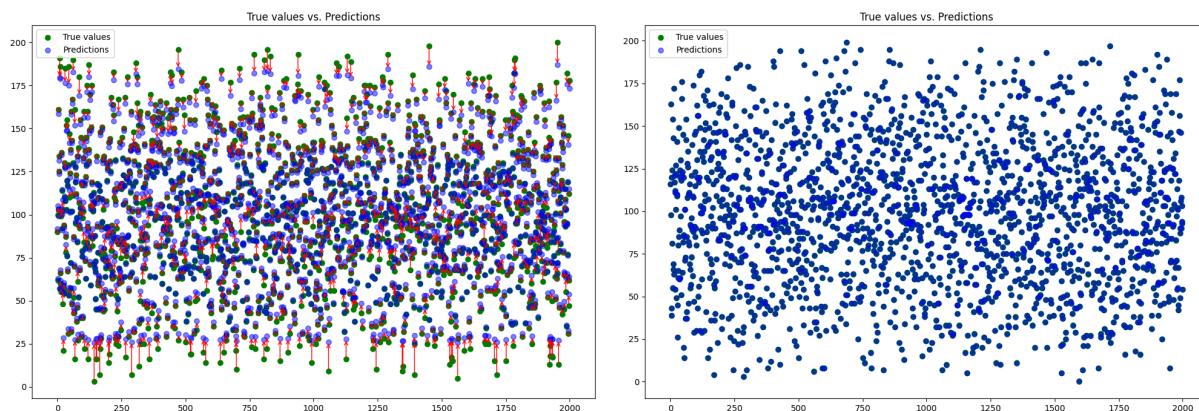


Fig. 43. – Modèle sous-entraîné à gauche, modèle bien entraîné à droite. La différence de l'erreur est indiquée par une flèche rouge.

Il est important de tester vos modèles et si possible de rendre ça visuel pour le développement de vos modèles. Cette phase permet non seulement de mesurer la performance, mais aussi de comprendre où et comment le modèle peut être amélioré.

5.1.7.1 Faire des inférences à la main

Pour ceux qui souhaitent tester le modèle avec leurs propres valeurs, la fonction `predict_sum` est conçue à cet effet. Elle permet de faire des prédictions sur des paires de nombres de votre choix. Voici comment elle fonctionne :

```

def predict_sum(model, num1, num2, max_value=100):
    # Normalisation des entrées
    X_input = torch.tensor([[num1/max_value, num2/max_value]],
    dtype=torch.float32).to(device)

    # Inférence
    with torch.no_grad():
        prediction = model(X_input)

    # Retour à l'échelle originale pour obtenir le résultat

```

```

predicted_sum = prediction * (2 * max_value)
return predicted_sum.item()

# Exemple d'utilisation
result = predict_sum(model, 50, 50)
print(f"La somme prédictée de 50 et 50 est : {result:.2f}")
print(f"La somme réelle de 50 et 50 si on arrondit au entier est :
{np.round(result):.0f}")

# Test de la fonction avec d'autre valeur
result = predict_sum(model, 6, 6)
print(f"La somme prédictée de 6 et 6 est : {result:.2f}")
print(f"La somme réelle de 6 et 6 si on arrondit au entier est :
{np.round(result):.0f}")

>>> La somme prédictée de 50 et 50 est : 100.02
>>> La somme réelle de 50 et 50 si on arrondit au entier est : 100

>>> La somme prédictée de 6 et 6 est : 11.89
>>> La somme réelle de 6 et 6 si on arrondit au entier est : 12

```

Cette fonction est un outil pratique pour expérimenter avec le modèle et observer comment il gère différentes entrées numériques.

5.2 Projet 2: Reconnaissance d'écriture manuscrits

Dans ce projet numéro deux, notre objectif est de construire et d'entraîner un modèle de deep learning capable de reconnaître et de prédire la classe d'images représentant des chiffres manuscrits. Nous allons utiliser le célèbre jeu de donnée MNIST rendu célèbre par le Français Yann Le Cun, MNIST regroupe 60 000 images d'apprentissages et 10 000 images de test. Ce sont des images de chiffre allant de 0 à 10 en noir et blanc, d'une résolution de 28 pixels par 28, de toutes petites images. C'est toujours un projet-guidé, mais vous pouvez décider de réaliser le projet de votre côté à partir du cahier des charges.

5.2.1 Cahier des charges

1. **Importer et explorer le dataset :** Téléchargez le jeu de données MNIST en utilisant la bibliothèque `torchvision`, qui contient des images en noir et blanc de chiffres écrits à la main. Appliquez les transformations nécessaires pour convertir les images en tenseurs et normalisez-les. Séparez ensuite les données d'entraînement en ensembles d'entraînement et de validation. Avant de plonger dans la construction du modèle, familiarisez-vous avec le jeu de données. Sélectionnez un échantillon d'images du lot d'entraînement et visualisez-les.
2. **Conception du modèle MLP :** Définissez la structure du modèle de Perceptron Multicouche. Ce modèle doit inclure une couche d'entrée adaptée à la taille des images MNIST, plusieurs couches cachées et une couche de sortie avec une fonction d'activation softmax pour la classification des chiffres.
3. **Entraînement du modèle :** Procédez à l'entraînement du modèle sur l'ensemble d'entraînement. Utilisez la fonction de perte `CrossEntropyLoss` et un optimiseur comme SGD ou Adam. Assurez-vous de suivre la perte d'entraînement et de validation pour surveiller les performances du modèle et ajustez les hyperparamètres si nécessaire.
4. **Validation du modèle :** Après chaque époque, évaluez les performances du modèle sur l'ensemble de validation. Cela aidera à détecter et à éviter le surajustement.
5. **Évaluation des performances :** Une fois l'entraînement terminé, testez le modèle sur l'ensemble de test MNIST pour évaluer sa capacité à généraliser sur de nouvelles données. Calculez des métriques telles que l'exactitude, la précision, le recall et le F1 score pour quantifier les performances du modèle.
6. **Visualisation des résultats de prédiction :** Visualisez les prédictions du modèle sur un ensemble d'images test pour une analyse qualitative. Cela permet de comprendre les capacités et les limites du modèle dans la classification des chiffres manuscrits.

5.2.2 Importation des bibliothèques Python

Comme dans chaque projet, nous importons les diverses bibliothèques Python

```
import torch
import torch.nn as nn
import torch.optim as optim
import torchvision
import torchvision.transforms as transforms
from torch.utils.data import DataLoader, random_split
```

```

from tqdm import tqdm # pour la barre de progression
import matplotlib.pyplot as plt # pour le tracé graphique
from PIL import Image # pour la lecture d'images
import os
import random
from sklearn.metrics import confusion_matrix
import seaborn as sns

```

5.2.3 Configuration du Matériel de Calcul

Nous configurons ensuite l'environnement matériel pour l'exécution de notre modèle. L'utilisation d'un GPU (Graphics Processing Unit) est préférable à un CPU (Central Processing Unit) pour l'entraînement de modèles de deep learning, en raison de la capacité supérieure du GPU à effectuer des calculs parallèles, accélérant ainsi significativement le processus d'apprentissage.

```

device = torch.device("cuda:0" if torch.cuda.is_available() else "cpu")

if torch.cuda.is_available():
    print("Le GPU est utilisé")
else:
    print("Le GPU n'est PAS utilisé, le CPU est utilisé")

```

Ce code détecte automatiquement la disponibilité d'un GPU compatible CUDA (une technologie de NVIDIA). Si un GPU est disponible, le modèle s'exécutera sur celui-ci, sinon il reviendra à l'utilisation du CPU. Il est crucial de s'assurer que l'environnement de travail, qu'il s'agisse d'un ordinateur local avec un GPU NVIDIA ou d'un environnement cloud comme Google Colab, est correctement configuré pour utiliser le GPU.

5.2.4 Importation et Préparation des Données

Notre projet se concentre sur la reconnaissance de chiffres manuscrits, un défi classique dans le domaine du deep learning. Nous utilisons le jeu de données MNIST, largement reconnu et employé comme point de référence dans l'apprentissage automatique pour les tâches de classification d'images.

La préparation des données commence par la définition des transformations à appliquer aux images du jeu de données :

```

batch_size = 32

transform = transforms.Compose([
    transforms.ToTensor(),
    transforms.Normalize((0.5,), (0.5,))
])

```

Dans cette configuration :

- **batch_size = 32** : Ce paramètre définit le nombre d'exemples traités en un seul lot. Un batch size de 32 est un équilibre entre efficacité de calcul et capacité à généraliser, ce nombre de batchs (lot en français), doit être suffisant pour permettre à notre descente de gradient de faire une approximation suffisante du gradient sur l'ensemble des données tout en gardant la variabilité nécessaire pour une bonne convergence.

- **transforms.Compose([...])** : Cette fonction combine plusieurs transformations en une seule opération. Dans notre cas, deux transformations sont appliquées :
 - **transforms.ToTensor()** : Convertit les images de leur format d'origine (généralement des tableaux numpy ou Image de la bibliothèque PIL) en tenseur PyTorch, cette conversion est nécessaire pour que notre modèle PyTorch traitent les données dans un format qu'il connaît.
 - **transforms.Normalize((0.5,), (0.5,))** : Normalise les images, les valeurs **(0.5,)** pour la moyenne et l'écart-type sont utilisées pour centrer et redimensionner les pixels des images. Cela aide l'algorithme à converger, autrement dit à mieux s'entraîner.

Nous téléchargeons les données via la bibliothèque torchvision de la manière suivante :

```
train_dataset = torchvision.datasets.MNIST(root='./data', train=True,
transform=transform, download=True)
test_dataset = torchvision.datasets.MNIST(root='./data', train=False,
transform=transform, download=True)
```

Nous chargeons les jeux de données d'entraînement et de test MNIST. Le paramètre `train` indique si nous chargeons le jeu de données d'entraînement `True` ou de test `False`. Le `transform` appliqué assure les images sont correctement transformées en tenseurs et normalisées.

Nous divisons ensuite le jeu d'entraînement en sous-ensembles d'entraînement et de validation :

```
train_size = int(0.8 * len(train_dataset))
val_size = len(train_dataset) - train_size
train_dataset, val_dataset = random_split(train_dataset, [train_size,
val_size])
```

Dans ce processus, nous allouons 80% du jeu de données d'entraînement pour l'entraînement proprement dit et les 20% restants pour la validation. Le sous-ensemble de validation joue un rôle fondamental : il nous permet d'évaluer la performance du modèle sur des données qu'il n'a pas rencontrées lors de l'entraînement, offrant ainsi une mesure objective de sa capacité de généralisation. Bien que le ratio de 80/20 soit couramment utilisé, il est important de souligner que ce n'est pas une règle fixe. En effet, dans le cadre de jeux de données volumineux, une fraction plus petite de l'ensemble global, parfois aussi peu que 1%, peut être suffisante pour la validation, à condition que cette portion soit représentative de l'ensemble des données. Cette flexibilité dans la répartition des données est essentielle pour adapter la configuration à la spécificité et à la complexité du projet en cours.

Poursuivant notre configuration, nous créons trois `DataLoaders`

```
train_loader = DataLoader(dataset=train_dataset, batch_size=batch_size,
shuffle=True)
val_loader = DataLoader(dataset=val_dataset, batch_size=batch_size,
shuffle=True)
test_loader = DataLoader(dataset=test_dataset, batch_size=batch_size,
shuffle=False)
```

Les `DataLoaders` de PyTorch fournissent une interface pour charger les données par lots (batch), permettant un traitement efficace et parallèle. Le paramètre `shuffle=True` pour les

jeux d'entraînement et de validation garantit que les données sont mélangées à chaque époque, ce qui est essentiel pour éviter un apprentissage biaisé et pour améliorer la généralisation. Pour l'ensemble de test, `shuffle` est défini sur `False` car l'ordre des données de test n'affecte pas l'évaluation du modèle.

5.2.4.1 Pourquoi est-il important de normaliser ses données ?

Premièrement, la normalisation des données aide à améliorer la stabilité numérique de notre modèle pendant entraînement. Le deep learning utilise la descente de gradient, le gradient de l'erreur par rapport aux poids du réseau est calculé pour mettre à jour ces derniers. Si les données d'entrée présentent des échelles très différentes, les gradients peuvent également varier considérablement, ce qui demanderait à nos poids de savoir gérer des données avec de grosse variabilité, une tâche plus difficile. En normalisant les données, on s'assure que les gradients restent dans des plages gérables, facilitant ainsi la mise à jour des poids et améliorant la stabilité générale de l'apprentissage.

Supposons un modèle linéaire très simple avec deux caractéristiques x_1 et x_2 et deux poids w_1 et w_2 . La sortie du modèle est donnée par :

$$y = w_1x_1 + w_2x_2$$

Imaginons que x_1 prenne des valeurs autour de 1, tandis que x_2 varie autour de 10 000. Alors, même une petite modification du poids w_2 aura un effet considérable sur la sortie du modèle comparé à une modification similaire sur w_1 . Le gradient associé au poids w_2 sera donc beaucoup plus grand que celui associé à w_1 , ce qui peut conduire à des mises à jour disproportionnées :

$$\frac{\partial y}{\partial w_2} = x_2 \approx 10\,000$$

le poids associé à la caractéristique dominante risque d'évoluer trop rapidement ou trop lentement par rapport aux autres caractéristiques.

La figure ci-dessous démontre visuellement ces principes à travers les distributions de deux caractéristiques synthétiques avant et après la normalisation

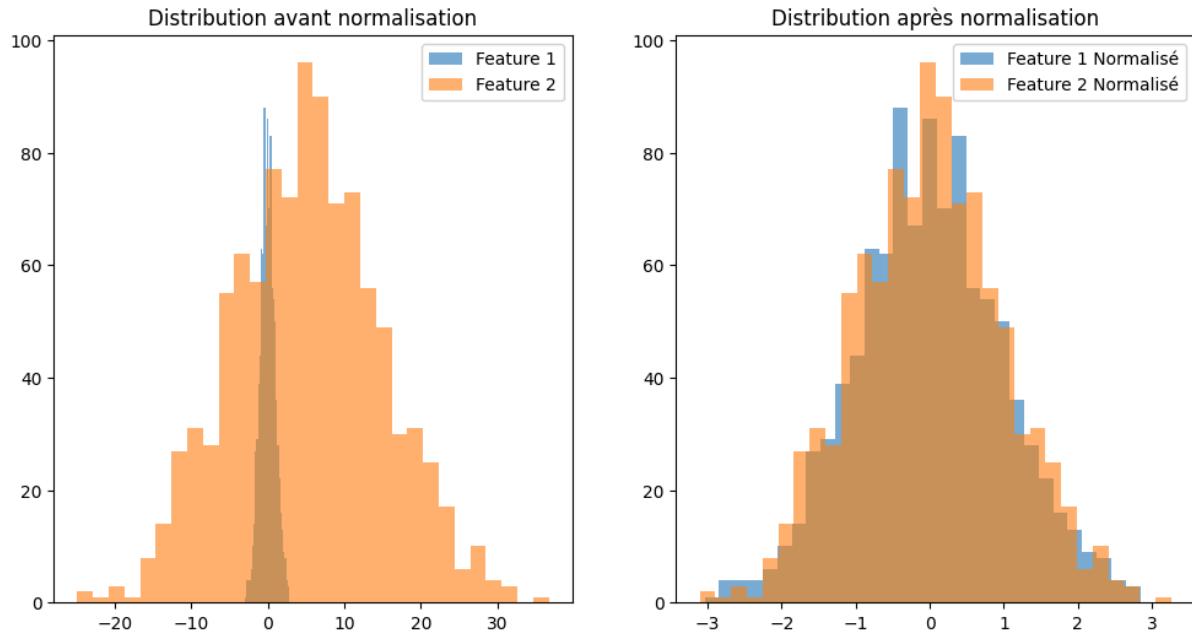


Fig. 44. – Impact de la normalisation sur les distributions de caractéristiques. À gauche, les distributions initiales des « Feature 1 » et « Feature 2 » montrent des écarts significatifs en termes de moyenne et de variance. À droite, après normalisation, les distributions sont ajustées pour avoir des moyennes de zéro et des variances unitaires, ce qui démontre une homogénéité facilitant les processus d'apprentissage. Cette transformation normalise les caractéristiques pour qu'elles contribuent également à l'apprentissage du modèle sans qu'une caractéristique domine en raison de son échelle.

Deuxièmement, la normalisation peut aider à accélérer l'entraînement. Lorsque les données sont normalisées, chaque dimension a approximativement la même échelle, ce qui rend la surface d'erreur (le paysage que la descente de gradient explore pour trouver le minimum) plus uniforme. Ceci est particulièrement important dans les réseaux de neurones profonds, où des différences d'échelle peuvent entraîner des chemins d'apprentissage inefficaces et des temps d'entraînement prolongés.

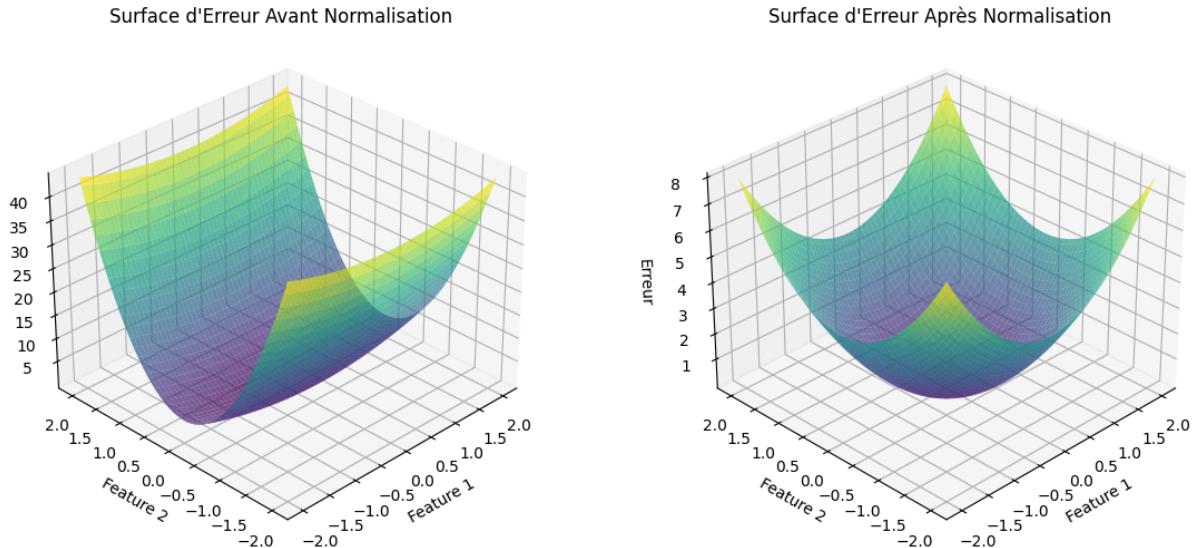


Fig. 45. – Comparaison des surfaces d’erreur avant et après normalisation des caractéristiques. La surface d’erreur asymétrique avant normalisation illustre les défis de l’optimisation face à des caractéristiques de différentes échelles. Après normalisation, la surface devient uniforme et symétrique, ce qui simplifie la descente de gradient et favorise une convergence rapide et stable.

Par exemple, considérons un jeu de données médical avec deux caractéristiques : « âge » (variant entre 0 et 100 ans) et « nombre total de globules blancs » (variant entre 4 000 et 11 000 cellules par microlitre). Sans normalisation, la seconde caractéristique dominera systématiquement le calcul du modèle simplement parce qu’elle présente des valeurs numériques beaucoup plus grandes. Pourtant, il est possible que l’âge soit tout aussi pertinent pour prédire certaines maladies. Normaliser ces deux caractéristiques permet donc au modèle de considérer équitablement leur importance relative sans biais lié à leur échelle numérique.

Troisièmement, la normalisation des données aide à éviter le problème du « vanishing/exploding gradient ». Dans les réseaux de neurones profonds, le gradient peut devenir extrêmement petit (vanishing) ou grand (exploding) à mesure qu’il se propage à travers les couches. Si les données d’entrée ne sont pas normalisées, cette tendance est exacerbée, car les poids peuvent recevoir des mises à jour trop importantes ou négligeables, rendant l’apprentissage inefficace ou même impossible.

5.2.4.2 Comment cela s’explique mathématiquement ?

Mathématiquement parlant, deux méthodes principales existent pour réaliser cette opération

Z-normalization (« Z-scoring ») qui consiste à centrer chaque caractéristique autour de zéro puis diviser par son écart-type :

$$Z = \frac{x - \mu}{\sigma}$$

où : x est la valeur originale, μ est la moyenne et σ l’écart-type. La soustraction centre les données, et la division les réduit.

Prenons un exemple avec les tailles : 150, 160, 170, 180, 190 cm. La moyenne μ est :

$$\mu = \frac{150, 160, 170, 180, 190}{5} = 170 \text{ cm}$$

L'écart-type σ se calcule via la variance. Écarts au carré : 400, 100, 0, 100, 400. Variance = $\frac{1000}{5} = 200$. Ainsi :

$$\sigma = \sqrt{200} \approx 14.14$$

cm

Pour 180 cm :

$$Z = \frac{180 - 170}{14.14} \approx 0.71$$

La valeur 180 cm devient 0.71, soit 0.71 écart-type au-dessus de la moyenne. Cette méthode est utile en deep learning pour stabiliser la forward pass et la backpropagation, notamment avec des poids initialisés selon une loi normale.

Le Min-Max Scaling : transforme les données dans un intervalle, souvent [0, 1]. La formule est la suivante :

$$x_{\text{scaled}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}$$

où : x est la valeur originale, x_{min} est la valeur minimale de la caractéristique et x_{max} est la valeur maximale.

Reprendons notre exemple de tailles : 150, 160, 170, 180, 190 cm. Ici, $x_{\text{min}} = 150$ cm et $x_{\text{max}} = 190$ cm.

Pour la valeur 180 cm :

$$x_{\text{scaled}} = \frac{180 - 150}{190 - 150} = \frac{30}{40} = 0.75$$

Ainsi, 180 cm est transformé en 0.75 sur une échelle de 0 à 1. Cette méthode est particulièrement utile pour les algorithmes qui nécessitent des entrées dans un intervalle spécifique, comme les réseaux de neurones avec des fonctions d'activation sigmoid ou tanh.

5.2.4.3 Pourquoi avoir plusieurs ensembles de données ?

L'essence même du deep learning est l'aptitude d'un modèle à interpréter et à tirer des conclusions à partir de données qui lui préalablement inconnues. Cette compétence, qui démarque une intelligence artificielle performante d'une simple réplique algorithmique, est le fruit d'un processus d'entraînement et de validation méthodique. Dans ce contexte, le recours à des ensembles de données multiples n'est pas une simple formalité, mais un passage obligé vers l'établissement d'une performance authentique et fiable.

Le meilleur moyen que quiconque ait trouvé pour déterminer les performances d'un modèle face à de nouvelles données inconnues consiste à le tester sur des données non vues et de voir comment il réagit. Aucune méthode alternative ne peut remplacer cette vérification expéri-

mentale cruciale. C'est pourquoi nous avons découpé notre jeu de données en trois parties, **jeu d'entraînement, jeu de validation et jeu de test**.

L'ensemble d'entraînement : Les Exercices Quotidiens

Prenons l'analogie de l'étudiant. L'ensemble d'entraînement est le manuel d'exercices utilisé au quotidien. À travers des sessions répétées, l'étudiant (le modèle) s'efforce de comprendre et de résoudre des problèmes (les données) dans un cadre sécurisé (l'environnement d'apprentissage). Cependant, si nous devions évaluer cet étudiant, il serait peu judicieux de le tester sur des exercices qu'il a déjà pratiqués. Il les connaîtrait par cœur, certes, mais cela ne prouverait pas qu'il a assimilé la matière. De même, évaluer un modèle sur l'ensemble d'entraînement ne nous apprend rien sur sa capacité à généraliser au-delà des exemples appris.

L'Ensemble de Validation : Les Tests Blancs

Ici intervient l'ensemble de validation, équivalent des tests blancs qui préparent l'étudiant aux conditions réelles de l'examen. Ce sont des exercices qu'il n'a pas encore résolus, mais qui sont conçus pour être représentatifs de sa progression actuelle. Le modèle est ajusté et affiné en fonction de ses performances sur l'ensemble de validation, similaire à l'étudiant qui apprend de ses erreurs sur les tests blancs et améliore sa compréhension des sujets. La validation permet de calibrer le modèle sans compromettre la neutralité de l'examen final, l'ensemble de test.

L'Ensemble de Test : L'Examen Final

L'ensemble de test est l'ultime baromètre de l'efficacité d'apprentissage du modèle. Tout comme les questions d'un examen final sont inconnues de l'étudiant jusqu'au jour J, l'ensemble de test est gardé secret et ne révèle ses données au modèle qu'au terme de son entraînement. Cette approche garantit une évaluation objective de la capacité du modèle à traiter de nouvelles informations, en dehors de toute familiarité préalable. Si l'étudiant réussit son examen final, nous pouvons affirmer avec assurance qu'il a bien intégré les leçons. De la même manière, si un modèle parvient à interpréter correctement l'ensemble de test, nous pouvons être confiants dans sa capacité à fonctionner dans des conditions réelles et imprévisibles.

La fuite de données : Un écueil à éviter

Le phénomène de fuite de données se produit lorsque l'information des données de test influence le processus d'apprentissage. C'est une forme de tricherie involontaire qui peut mener à une surévaluation des capacités réelles du modèle. Pour l'éviter, il est essentiel de pratiquer une stricte hygiène des données en maintenant une séparation rigoureuse entre les ensembles d'entraînement, de validation et de test tout au long du processus d'apprentissage.

L'intégrité de l'évaluation : La Clé de la Vérité

Il est tentant de vouloir utiliser les données de test pour améliorer la performance du modèle, en particulier si les résultats initiaux sont décevants. Cependant, cela équivaudrait à donner à l'étudiant les réponses de l'examen à l'avance. Le modèle peut alors paraître performant, mais sa réussite serait illusoire.

5.2.4.4 Comment fonctionne un DataLoader ?

Le *Dataloader* est une interface élaborée destinée à la manipulation et à la préparation des données, orchestrant leur chargement, leur transformation et leur distribution aux modèles de deep learning. Ce module est doté d'une architecture capable de rationaliser le traitement des données, allant de leur stockage initial jusqu'à leur consommation algorithmique.

Lors de l’instanciation d’un *DataLoader* au sein de l’écosystème PyTorch, nous déployons une suite de commandes orchestrant le chargement des données, leur mise en forme selon les spécifications du modèle, suivie de leur agencement en lots structurés. La représentation graphique offerte par le schéma suivant décompose visuellement ce processus en étapes distinctes.

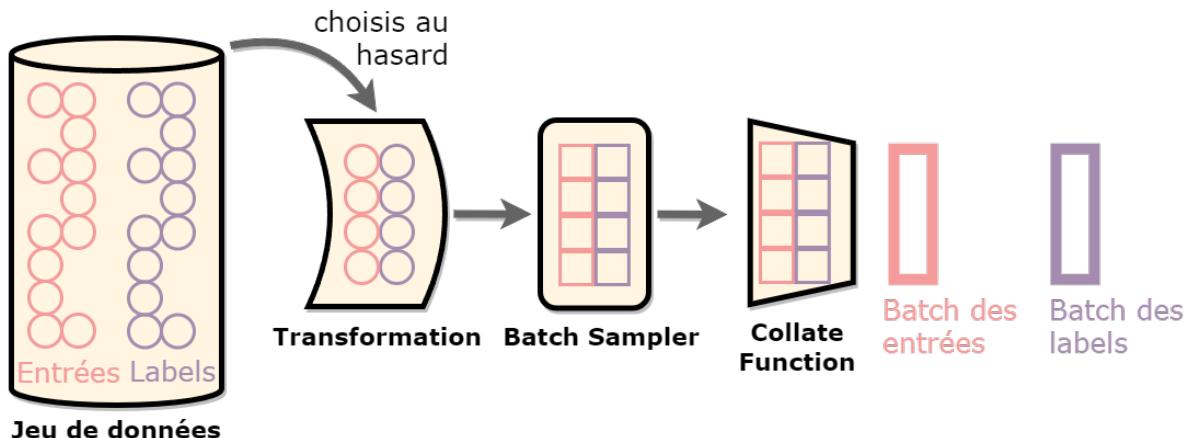


Fig. 46. – Illustration schématique du flux opérationnel d’un *DataLoader* en PyTorch, de la sélection aléatoire des données, leur transformation, le regroupement en lots, jusqu’à la séparation en entrées et labels prêts pour l’entraînement.

La figure ci-dessus illustre le flux de données à travers un *DataLoader*. Le processus débute par la sélection aléatoire d’échantillons à partir du jeu de données, ce qui est essentiel pour introduire de la stochasticité dans l’entraînement et empêcher le modèle de simplement mémoriser l’ordre des observations. Les échantillons sont ensuite soumis à une transformation qui sert à la normalisation des données ou autre transformation comme l’augmentation des données que nous utiliserons dans un autre projet.

Ensuite, le « *Batch Sampler* » entre en scène pour regrouper les données transformées en lots de taille prédéfinie. Le regroupement des données en lots est une étape qui exploite efficacement la puissance de calcul parallèle moderne, permettant ainsi un entraînement plus rapide et plus efficace du modèle.

Après la formation des lots, la « *Collate Function* » est appelée à structurer ces lots de manière à ce qu’ils puissent être traités de manière cohérente par le modèle. Dans cette phase, les ajustements nécessaires, comme le remplissage ou la mise en forme des tenseurs, sont effectués pour s’assurer que chaque lot a une forme et une taille uniformes.

Le résultat final est une séparation des données en deux groupes distincts : le « *Batch des entrées* », qui contiendra les caractéristiques ou les données d’entrée pour le modèle, et le « *Batch des labels* », qui contiendra les étiquettes ou les cibles que le modèle apprendra à prédire. Ce partitionnement assure que les données sont prêtes à être ingérées par le modèle pour la propagation avant et la mise à jour des poids lors de la propagation arrière.

5.2.5 Exploration des données

L’exploration des données est une phase du développement important dans tout projet de deep learning, elle permet de développer une compréhension des caractéristiques des données

à disposition. Cette phase permet de détecter des tendances, des anomalies, des schémas et des relations intrinsèques dans le jeu de données, fournissant ainsi une base solide pour la conception et l'optimisation de modèles plus efficaces.

Pour initier notre exploration, commençons par visualiser un sous-ensemble des images de notre jeu de données. Cette visualisation nous permet de comprendre la variabilité et la qualité des images, ainsi que les défis potentiels auxquels notre modèle pourrait être confronté, tels que la variabilité de l'écriture manuscrite ou la présence de bruit dans les images.

```
print("Forme des données d'entraînement:", train_dataset[0][0].shape)
print("Forme des données de validation:", val_dataset[0][0].shape)
print("Forme des données de test:", test_dataset[0][0].shape)

>>> Forme des données d'entraînement: torch.Size([1, 28, 28])
>>> Forme des données de validation: torch.Size([1, 28, 28])
>>> Forme des données de test: torch.Size([1, 28, 28])
```

Les sorties disent que chaque image, qu'elle soit dans l'ensemble d'entraînement, de validation ou de test, a la forme [1, 28, 28]. Chaque image est un tableau de 28 par 28 pixels, avec un seul canal de couleur (niveaux de gris), si la forme était de [3, 28, 28] il s'agirait d'une image avec 3 canaux de couleur, une image RGB (Red, Green, Blue).

Le code suivant illustre cette étape initiale d'exploration :

```
dataiter = iter(train_loader)
images, labels = next(dataiter)

batch_size = len(images)

random_indices = random.sample(range(batch_size), 10)

plt.figure(figsize=(15, 6)) # largeur, hauteur

for i, idx in enumerate(random_indices):
    image = images[idx].squeeze(0).numpy()
    label = labels[idx].item()

    plt.subplot(2, 5, i + 1)
    plt.imshow(image, cmap='gray')
    plt.title(f"Label: {label}")
    plt.axis('off')

plt.subplots_adjust(wspace=0.1)
plt.show()
```

Ce code démarre par la création d'un itérateur pour notre `DataLoader` (`train_loader`), qui est une collection d'images et d'étiquettes du jeu de données MNIST. Un batch d'images et d'étiquettes est ensuite extrait de cet itérateur. La sélection aléatoire de 10 images de ce batch est effectuée pour assurer une diversité dans les exemples visualisés. Chaque image est alors affichée avec son étiquette correspondante.

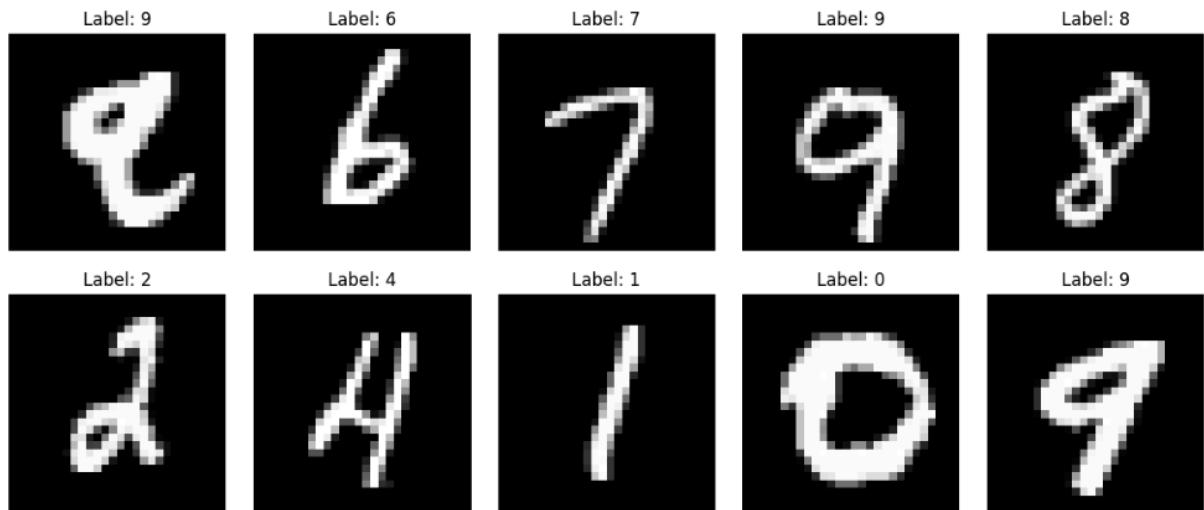


Fig. 47. – Sélection aléatoire de 10 exemples issus du jeu de données MNIST. Chaque sous-image représente un chiffre unique, comme indiqué par son étiquette correspondante.

Dans notre projet de deep learning, nous avons déjà examiné la forme des images du jeu de données MNIST et visualisé un échantillon de ces images. Maintenant, approfondissons notre exploration pour obtenir une compréhension plus complète des données.

Une autre analyse importante est la distribution des étiquettes dans le jeu de données. Le code suivant génère un histogramme montrant la fréquence de chaque chiffre dans l'ensemble d'entraînement :

```
labels = [label for _, labels in train_loader for label in labels]
plt.figure(figsize=(10, 6))
plt.hist(labels, bins=np.arange(-0.5, 10.5, 1), rwidth=0.8)
plt.title("Distribution des étiquettes dans MNIST")
plt.xlabel("Chiffre")
plt.ylabel("Nombre d'exemples")
plt.xticks(np.arange(0, 10, 1))
plt.show()
```

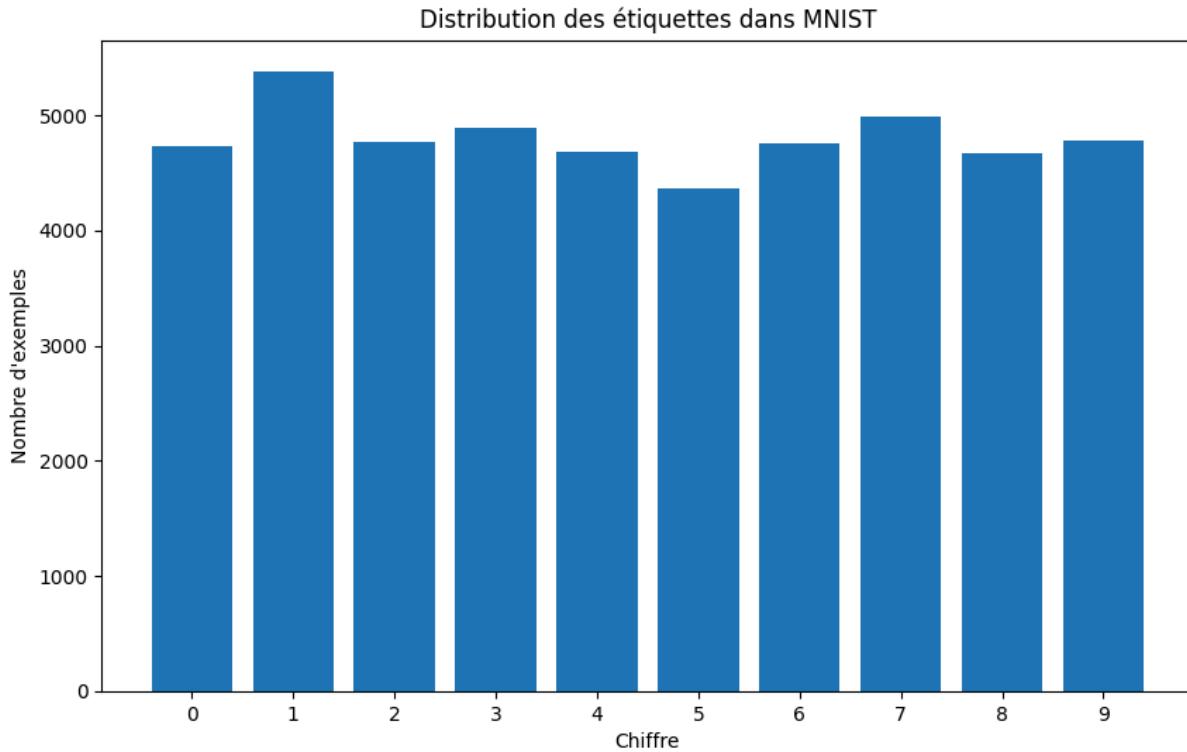


Fig. 48. – Histogramme par nombre d'exemples par classe du jeu MNIST. Cette visualisation montre la fréquence de chaque chiffre dans l'ensemble d'entraînement, fournissant des indices sur l'équilibre ou le déséquilibre des classes dans le jeu de données.

Cet histogramme nous montre la répartition des différentes classes (chiffres) dans le jeu de données. Une répartition équilibrée des classes est idéal pour s'assurer que le modèle s'entraîne de manière équitable sur chaque chiffre. Un déséquilibre marqué pourrait entraîner un biais dans le modèle, le rendant plus performant pour reconnaître certains chiffres que d'autres. Cependant, il convient de noter que cette situation idéale est plutôt exceptionnelle dans les applications réelles.

Par exemple pour la détection de transactions frauduleuses, cela représente un faible pourcentage des données, si vous avez disons 1% de données frauduleuse, un modèle naïve qui répond toujours « transaction non-frauduleuse » aura raison 99% du temps, et au même fortement réduit sa fonction objectif (loss function) ! Il y a plein de techniques pour remédier à ces problèmes que vous verrez dans les projets suivants.

5.2.6 Définition de l'architecture du réseau

Nous définissons notre réseau de manière très similaire au mini-projet calculatrice, sauf que cette fois nous traitons des images et des données structurelles. Les images du MNIST sont des tableaux de 28x28 pixels, ce qui donne une dimension totale de 784 pixels par image. Cependant, un réseau MLP classique ne peut pas traiter directement ces images sous forme de tableaux 2D. Il nécessite des données d'entrée sous forme de vecteurs 1D. D'où la nécessité de transformer chaque image 28x28 en un vecteur linéaire de 784 éléments. C'est ce qu'on appelle « mettre à plat » les images (flattening).

Le code suivant illustre cette transformation :

```

class MLP(nn.Module):
    def __init__(self, input_size=784, hidden_size=40, output_size=10,
num_hidden_layers=4):
        super(MLP, self).__init__()
        self.layers = nn.ModuleList()
        self.layers.append(nn.Linear(input_size, hidden_size, bias=True))
        for i in range(num_hidden_layers):
            self.layers.append(nn.Linear(hidden_size, hidden_size, bias=True))
        self.layers.append(nn.Linear(hidden_size, output_size, bias=True))

    def forward(self, x):
        x = x.view(x.size(0), -1) # Mettre à plat les images
        for layer in self.layers[:-1]:
            x = torch.relu(layer(x))
        x = self.layers[-1](x)
        return x

```

Dans la méthode `forward`, `x = x.view(x.size(0), -1)` modifie la forme des images entrantes. Chaque image, initialement un tableau 2D, est transformée en un vecteur 1D. Cela permet au réseau de traiter chaque pixel de l'image comme une caractéristique individuelle.

Dans la construction de `MLP`, `input_size=784` spécifie que le réseau attend des vecteurs de 784 éléments en entrée, correspondant aux images mises à plat. Les couches linéaires `nn.Linear` successives traitent ces données, avec des fonctions d'activation `ReLU` introduites entre les couches pour apporter la non-linéarité nécessaire à l'apprentissage de représentations complexes.

5.2.7 Entraînement du modèle

Après avoir défini l'architecture de notre réseau de neurones multicouches (`MLP`) et exploré notre jeu de données, nous arrivons à une étape cruciale : l'entraînement du modèle.

Le code suivant illustre le processus d'entraînement de notre modèle `MLP` sur le jeu de données `MNIST` :

La première étape est la création du modèle en instanciant la classe `MLP`. Le modèle est ensuite chargé sur le GPU si disponible, ce qui accélère significativement les calculs

```
model = MLP().to(device) # Le charger sur le GPU si disponible
```

La fonction de perte (`criterion`) mesure à quel point les prédictions du modèle s'écartent des étiquettes réelles. Ici, `CrossEntropyLoss` est utilisé. L'optimiseur (`optimizer`), tel que la Descente de Gradient Stochastique (`SGD`), met à jour les poids du modèle en fonction du gradient de la fonction de perte. Le taux d'apprentissage (`lr`) contrôle la taille des pas de l'optimisation.

```
criterion = nn.CrossEntropyLoss() # fonction de perte
optimizer = optim.SGD(model.parameters(), lr=0.01) # lr = learning rate
```

L'entraînement se déroule sur plusieurs époques. À chaque époque, le modèle est entraîné sur l'ensemble d'entraînement et validé sur l'ensemble de validation.

- **Phase d'entraînement** : Cette phase, le modèle apprend à partir de l'ensemble d'entraînement. Pour chaque lot (batch) d'images et d'étiquettes, le modèle fait une

prédition `outputs`, calcule la perte `loss`, effectue une rétropropagation du gradient `loss.backward()` et met à jour les poids `optimizer.step()`.

- **Phase de validation** : La validation permet d'évaluer la performance du modèle sur des données non vues pendant l'entraînement. Elle aide à détecter le surajustement *overfitting*. Pendant la validation, les gradients ne sont pas calculés `torch.no_grad()`, car le modèle n'est pas mis à jour, nous n'entraînons pas notre modèle sur notre jeu de validation, nous vérifions pendant l'entraînement comment le modèle se comporte sur des données qu'il n'a jamais vues. Durant la phase de validation, le modèle évalue simplement les données sans ajuster ses paramètres. L'utilisation de `torch.no_grad()` indique que les gradients n'ont pas besoin d'être calculés, pas besoin de calculer les gradients de la fonction perte par rapport à chaque paramètre du modèle à travers la rétropropagation, mais également pas de mise à jour de ses paramètres.

```
num_epochs = 20
train_losses = []
val_losses = []

for epoch in range(num_epochs):
    model.train()
    train_loss = 0.0

    # Entraînement
    for images, labels in tqdm(train_loader, desc=f"Epoch {epoch+1}"):
        images, labels = images.to(device), labels.to(device)
        optimizer.zero_grad()
        outputs = model(images)
        loss = criterion(outputs, labels)
        loss.backward()
        optimizer.step()
        train_loss += loss.item()

    train_losses.append(train_loss/len(train_loader))

    # Validation
    model.eval()
    val_loss = 0.0
    with torch.no_grad():
        for images, labels in val_loader:
            images, labels = images.to(device), labels.to(device)
            outputs = model(images)
            loss = criterion(outputs, labels)
            val_loss += loss.item()

    val_losses.append(val_loss/len(val_loader))
```

Les variables `train_losses` et `val_losses` sont utilisées pour stocker les pertes enregistrées pendant les phases d'entraînement et de validation respectivement.

5.2.7.1 Evaluation de l'entraînement

Pour suivre les performances de notre modèle, nous visualisons la perte (loss) au cours de l'entraînement pour les ensembles d'entraînement et de validation.

```
# Affichage des pertes d'entraînement et de validation
plt.figure(figsize=(12, 6))
plt.plot(train_losses, label='Train Loss')
plt.plot(val_losses, label='Validation Loss')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.legend()
plt.show()

print('Entraînement terminé')
```

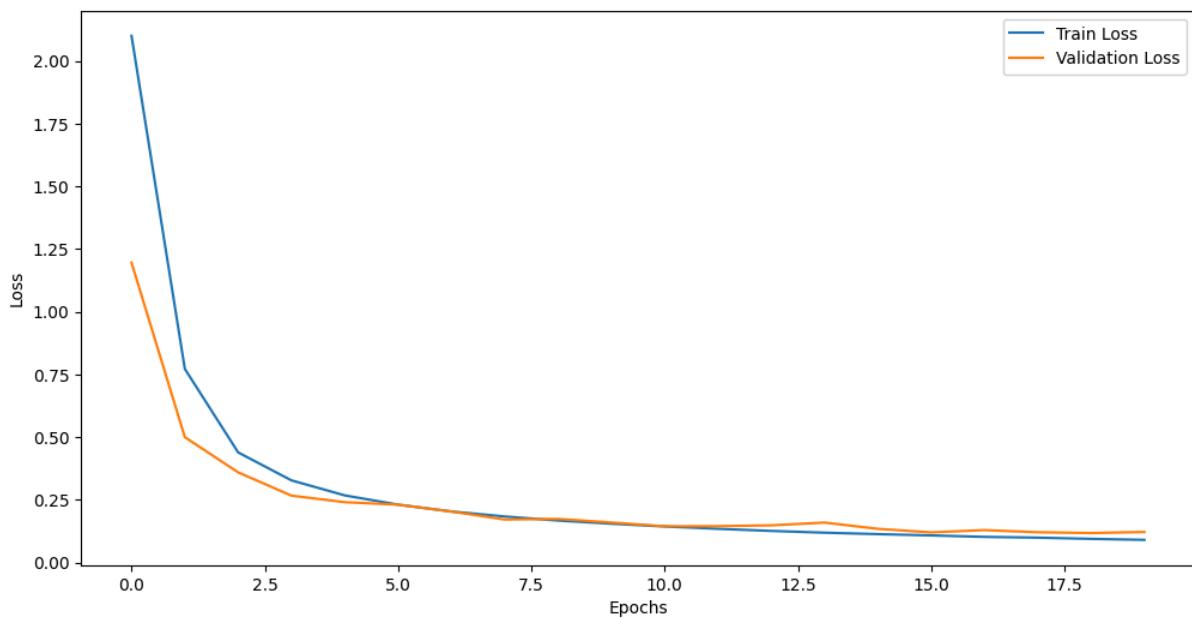


Fig. 49. – Courbes de perte d'entraînement et de validation sur 20 époques.

La perte d'entraînement (*Train Loss*) montre comment le modèle s'améliore sur les données qu'il a déjà vues, tandis que la perte de validation (*Validation Loss*) indique la performance du modèle sur des nouvelles données, qu'il n'a pas utilisées pour l'apprentissage. En analysant la courbe de perte, nous pouvons tirer plusieurs conclusions :

1. **Diminution rapide des pertes au début** : Cela suggère que le modèle apprenne efficacement les caractéristiques fondamentales des données.
2. **Convergence des pertes d'entraînement et de validation** : Si les deux pertes diminuent à un rythme similaire et convergent vers une valeur basse, cela indique que le modèle généralise bien et qu'il n'y a pas de surajustement significatif.
3. **Plateau des courbes de perte** : Vers la fin des époques, les courbes se stabilisent, indiquant que le modèle a atteint ses capacités d'apprentissage avec les paramètres et la structure actuels.

La courbe de perte est un outil pour l'ajustement de la validation du modèle. Si la perte de validation commence à augmenter alors que la perte d'entraînement continue de diminuer,

cela peut être un signe de surajustement, signifiant que le modèle apprend par cœur les données d'entraînement au lieu de généraliser à partir d'elles.

5.2.7.2 Analyse des performances par classe

Lors de l'évaluation des performances d'un modèle de classification, il est courant d'examiner la précision globale ou la perte moyenne sur l'ensemble des classes. Cependant, cette approche peut masquer des disparités significatives dans les performances du modèle pour différentes classes. Une analyse plus approfondie, qui décompose la performance par classe, peut fournir des informations précieuses sur les forces et les faiblesses spécifiques du modèle.

Le code suivant illustre une méthode pour calculer la perte moyenne par classe lors de l'évaluation du modèle sur l'ensemble de test :

```
model.eval()
class_loss = {i: 0.0 for i in range(10)} # Initialisation de la perte par classe
class_counts = {i: 0 for i in range(10)} # Compteur pour chaque classe

criterion = nn.CrossEntropyLoss(reduction='none') # Utilisation de 'none' pour conserver la perte de chaque échantillon

with torch.no_grad():
    for images, labels in test_loader:
        images, labels = images.to(device), labels.to(device)
        outputs = model(images)
        losses = criterion(outputs, labels).detach() # Perte individuelle par échantillon

        _, predictions = torch.max(outputs, 1)
        for label, prediction, loss in zip(labels, predictions, losses):
            if label == prediction: # Si la prédiction est correcte
                class_loss[label.item()] += loss.item()
                class_counts[label.item()] += 1

# Calcul de la moyenne des pertes par classe et affichage
for cls in range(10):
    if class_counts[cls] > 0:
        class_loss[cls] /= class_counts[cls]
    print(f'Classe {cls}: Perte moyenne = {class_loss[cls]:.4f}')
```

L'exécution de ce code produit une sortie similaire à celle-ci :

```
>>> Classe 0: Perte moyenne = 0.0201
>>> Classe 1: Perte moyenne = 0.0183
>>> Classe 2: Perte moyenne = 0.0822
>>> Classe 3: Perte moyenne = 0.0125
>>> Classe 4: Perte moyenne = 0.0476
>>> Classe 5: Perte moyenne = 0.0933
>>> Classe 6: Perte moyenne = 0.0769
>>> Classe 7: Perte moyenne = 0.0403
>>> Classe 8: Perte moyenne = 0.1862
>>> Classe 9: Perte moyenne = 0.0830
```

Cette sortie révèle des variations notables dans la perte moyenne pour différentes classes. Par exemple, la classe 8 présente une perte moyenne nettement plus élevée (0.1862) que les autres classes, ce qui suggère que le modèle a plus de difficultés à classer correctement les images de cette classe.

En revanche, la classe 3 affiche la perte moyenne la plus faible (0.0125), indiquant que le modèle est particulièrement performant pour reconnaître les images de cette classe.

Ces informations peuvent orienter les efforts d'amélioration du modèle. Par exemple, si la classe 8 est identifiée comme problématique, on pourrait envisager de collecter davantage de données d'entraînement pour cette classe, d'appliquer des techniques d'augmentation de données spécifiques à cette classe, ou d'ajuster l'architecture du modèle pour mieux capturer les caractéristiques distinctives de cette classe.

L'analyse de la perte par classe offre également un aperçu de la difficulté intrinsèque de chaque classe. Certaines classes peuvent présenter des variations plus importantes ou des caractéristiques moins distinctives, les rendant plus difficiles à classer correctement, même pour un modèle bien entraîné.

5.2.7.3 Matrice de confusion : un outil d'analyse approfondie

L'analyse de la perte par classe a révélé des disparités notables dans les performances de notre modèle, en particulier pour la classe 8 qui présente une perte moyenne nettement plus élevée que les autres classes. Pour mieux comprendre ce qui se passe, nous allons maintenant examiner la matrice de confusion.

La matrice de confusion est un outil puissant pour visualiser les performances d'un modèle de classification. Elle compare les étiquettes réelles (True Label) aux étiquettes prédites (Predicted Label) pour chaque classe, permettant ainsi d'identifier non seulement les prédictions correctes (éléments diagonaux), mais aussi les types d'erreurs commises par le modèle (éléments non diagonaux).

Voici le code utilisé pour générer la matrice de confusion de notre modèle :

```
model.eval()
all_labels = []
all_predictions = []

with torch.no_grad():
    for images, labels in test_loader:
        images, labels = images.to(device), labels.to(device)
        outputs = model(images)
        _, predictions = torch.max(outputs, 1)
        all_labels.extend(labels.cpu().numpy())
        all_predictions.extend(predictions.cpu().numpy())

cm = confusion_matrix(all_labels, all_predictions)

plt.figure(figsize=(10, 10))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', cbar=False,
            xticklabels=list(range(10)), yticklabels=list(range(10)))
plt.xlabel('Predicted Label')
```

```

plt.ylabel('True Label')
plt.title('Confusion Matrix')
plt.show()

```

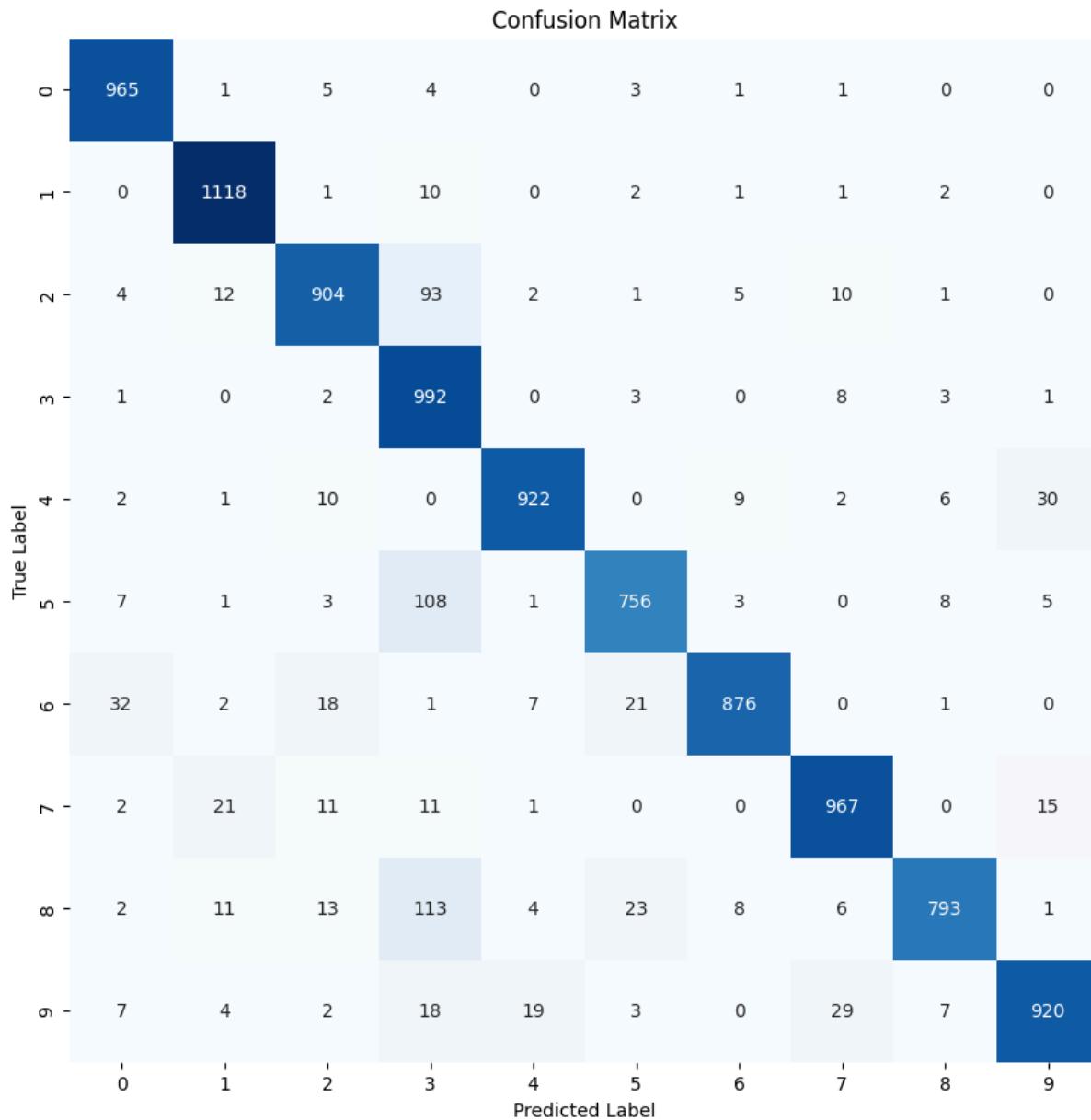


Fig. 50. – Matrice de confusion de notre modèle entraîné sur MNIST. Les éléments diagonaux montrent le nombre de prédictions correctes pour chaque classe, tandis que les éléments non diagonaux indiquent les erreurs de classification.

En examinant la matrice de confusion, nous pouvons tirer plusieurs enseignements :

- Prédictions correctes** : Les éléments diagonaux montrent que notre modèle classe correctement la plupart des images pour chaque chiffre. C'est un bon signe global.
- Classe 8 problématique** : Cependant, si nous nous concentrons sur la ligne correspondant à la classe 8, nous voyons qu'un nombre significatif d'images de 8 sont incorrectement classées comme des 3 ou 5. Cela explique pourquoi la perte moyenne était si élevée pour cette classe. Notre modèle semble avoir du mal à distinguer les 8 des chiffres visuellement similaires.

3. **Trop prédiction de la classe 3** : Notre modèle semble effectivement une très faible perte pour la classe 3, mais il prédit trop souvent la classe 3 et 5 qui donne les deux pires score de perte à ces classes, et intuitivement on peut imaginer qu'un 5 peut être confondu visuellement avec un 3 et un 8 peut être confondu avec un 3.
4. **Autres erreurs notables** : Nous pouvons également voir que notre modèle confond parfois les 4 avec les 9, et les 7 avec les 9. Bien que moins prononcées que pour la classe 8, ces erreurs indiquent d'autres zones d'amélioration potentielles.

Pour avoir ces résultats j'ai entraîné le modèle surseulement 10 époques, si vous entraînez le modèle sur plus d'époque, vous devriez avoir de meilleurs résultats.

5.2.8 Prédiction sur l'ensemble de test

Après avoir utilisé le jeu d'entraînement et de validation, place au jeu de test, l'examen final. Le code ci-dessous met le modèle en mode évaluation `model.eval()`, ce mode désactive certaines fonctionnalités spécifiques à l'entraînement, telles que la normalisation par lots (*batch normalization*) et la désactivation de neurones (*dropout*) vous n'avez vu aucune de ces méthodes encore, vous verrez de quoi il s'agit au chapitre suivant mais ils ont un impact sur modèle durant l'entraînement que nous souhaitons désactiver pendant la phase d'évaluation

```
all_labels, all_predictions = [], []
model.eval()
with torch.no_grad():
    for images, labels in test_loader:
        images, labels = images.to(device), labels.to(device)
        outputs = model(images)
        _, predicted = torch.max(outputs.data, 1)
        all_labels.extend(labels.cpu().numpy())
        all_predictions.extend(predicted.cpu().numpy())
```

Dans ce code, nous itérons sur l'ensemble de test `test_loader`. Pour chaque lot d'images, nous réalisons une prédiction `outputs = model(images)`, et nous utilisons `torch.max()` pour obtenir la classe prédite avec la plus haute probabilité. L'opérateur `==` est utilisé pour comparer les prédictions avec les véritables étiquettes, et nous accumulons le nombre total de prédictions correctes ainsi que le nombre total d'étiquettes.

Examiner de près les différentes métriques utilisées pour évaluer les performances de notre modèle de classification.

```
from sklearn.metrics import accuracy_score, precision_score, recall_score,
f1_score
accuracy = accuracy_score(all_labels, all_predictions)
precision = precision_score(all_labels, all_predictions, average='weighted')
recall = recall_score(all_labels, all_predictions, average='weighted')
f1 = f1_score(all_labels, all_predictions, average='weighted')

print(f"Exactitude sur l'ensemble de test : {accuracy:.2f}%")
print(f'Precision moyenne : {precision:.2f}')
print(f'Recall moyen : {recall:.2f}')
print(f'Score F1 moyen : {f1:.2f}')
```

```
>>> Exactitude sur l'ensemble de test : 96%
>>> Precision moyenne : 0.96
>>> Recall moyen : 0.96
>>> Score F1 moyen : 0.96
```

Une précision, un recall et une F1 de 96% sur l'ensemble de test est sont de très bons scores et suggère que notre modèle s'est très bien entraîné, notre modèle MLP est très efficace pour reconnaître les chiffres du MNIST.

5.2.8.1 Explications des différentes Métriques de classifications

L'exactitude (*accuracy en anglais*) est le ratio des prédictions correctes sur l'ensemble des prédictions effectuées. Mathématiquement, elle est définie comme :

$$\text{Accuracy} = \frac{\text{Nombre de prédictions correctes}}{\text{Nombre total de prédictions}} \times 100$$

Cette métrique est simple et intuitive. Elle offre une vue d'ensemble immédiate de la performance du modèle. Cependant, son utilité est limitée dans les cas où les classes sont déséquilibrées. Par exemple, dans un jeu de données où 90% des échantillons appartiennent à une classe, un modèle qui prédit toujours cette classe atteindra une précision de 90%, sans réellement apprendre à distinguer les caractéristiques pertinentes des différentes classes.

La précision (distincte de l'exactitude) mesure la proportion de prédictions correctes parmi les prédictions attribuées à une certaine classe. Elle est définie comme :

$$\text{Precision} = \frac{\text{True Positive (TP)}}{\text{TP} + \text{False Positive (FP)}}$$

Dans le contexte de la détection du COVID-19, les TP représentent les cas où le modèle identifie correctement un patient ayant le COVID-19, tandis que les FP sont les situations où le modèle indique à tort qu'une personne saine est infectée. Un faux positif dans ce scénario pourrait conduire à un isolement inutile, de l'anxiété et des coûts supplémentaires pour les individus et le système de santé.

Le Recall (Rappel Moyen), évalue la capacité du modèle à identifier toutes les instances réelles d'une classe spécifique. Il est calculé comme :

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{False Negative (FN)}}$$

où les FN sont les instances qui sont réellement positives mais n'ont pas été identifiées comme telles par le modèle. Cette métrique est importante quand il faut capturer tous les cas positifs. Par exemple, dans le diagnostic médical, un faux négatif (manquer un cas de maladie réelle) peut être beaucoup plus grave qu'un faux positif.

Le score F1 est une moyenne harmonique de la précision et du recall, offrant ainsi un équilibre entre ces deux métriques. Il est calculé comme suit :

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Le score F1 est particulièrement utile dans les cas où il est nécessaire de trouver un équilibre entre la précision et le recall. Un score F1 élevé indique que le modèle a une performance équilibrée en termes de précision et de recall, réduisant efficacement les faux positifs et les faux négatifs.

5.2.8.2 Pourquoi Utiliser Plusieurs Métriques ?

L'utilisation de plusieurs métriques est essentielle car aucune métrique n'est complète en soi. La précision seule peut donner une image trompeuse de la performance, en particulier dans des scénarios avec des classes déséquilibrées. La combinaison de plusieurs métriques permet une évaluation plus robuste et nuancée du modèle. Par exemple, si nous avions 90 % de classes < 9 >, et que notre modèle choisit de toujours choisir la classe < 9 >, il aurait 90 % de précision, pourtant ça ne ferait pas de lui un bon modèle, il aurait alors un *faible rappel*. Il n'existe jamais la métrique parfaite, il est important de connaître les points fort et faible de chacune et de les comprendre pour pouvoir bien interpréter les résultats.

5.2.9 Visualisation des prédictions

Comme dans le mini-projet précédent auquel nous avions pu tester de faire des additions sur les valeurs que nous lui donnons, ici nous allons lui donner une image en entrée du réseau, j'ai téléchargé 10 images. Pour cela, j'ai sélectionné 10 images que vous pouvez retrouver et télécharger depuis le répertoire [GitHub](#) du livre, situé dans le dossier *DeepLearningEnFrancais/Code/07_projet_mlp/data/MNIST/testing*. Vous pouvez soit télécharger individuellement ces images, soit cloner l'ensemble du dépôt pour votre commodité. Une fois téléchargées, assurez-vous de définir correctement le chemin vers ces images dans votre script.

```
image_directory = "./data/MNIST/testing"

def predict_image(image_path, model, transform):
    image = Image.open(image_path).convert('L')
    image_tensor = transform(image).unsqueeze(0).to(device)
    model.eval()
    with torch.no_grad():
        outputs = model(image_tensor)
        _, predicted = torch.max(outputs.data, 1)
    return predicted.item(), image # Retournez la prédiction et l'image
ouverte

# Listez tous les fichiers dans le répertoire
all_files = os.listdir(image_directory)

image_files = [f for f in all_files if f.endswith('.png')]

# Exécutez la prédiction sur toutes les images
plt.figure(figsize=(12, 6))
for i, image_file in enumerate(image_files[:10]):
    full_image_path = os.path.join(image_directory, image_file)
    prediction, image = predict_image(full_image_path, model, transform)

    plt.subplot(2, 5, i+1)
    plt.imshow(image, cmap='gray')
```

```

plt.title(f"Prédiction: {prediction}")
plt.axis('off')

plt.subplots_adjust(wspace=0.1) # Ajustement de l'espace entre les plots
plt.show()

```

Le code que nous allons utiliser illustre comment réaliser des prédictions sur des images individuelles à l'aide de notre modèle de réseau de neurones. Nous avons une fonction `predict_image`, qui ouvre une image donnée, la convertit en niveaux de gris (puisque notre modèle a été entraîné sur des images en niveaux de gris), et la transforme en un tenseur compatible avec notre modèle. Le modèle est ensuite utilisé pour prédire la classe de l'image. Cette prédition est retournée avec l'image pour une visualisation ultérieure.

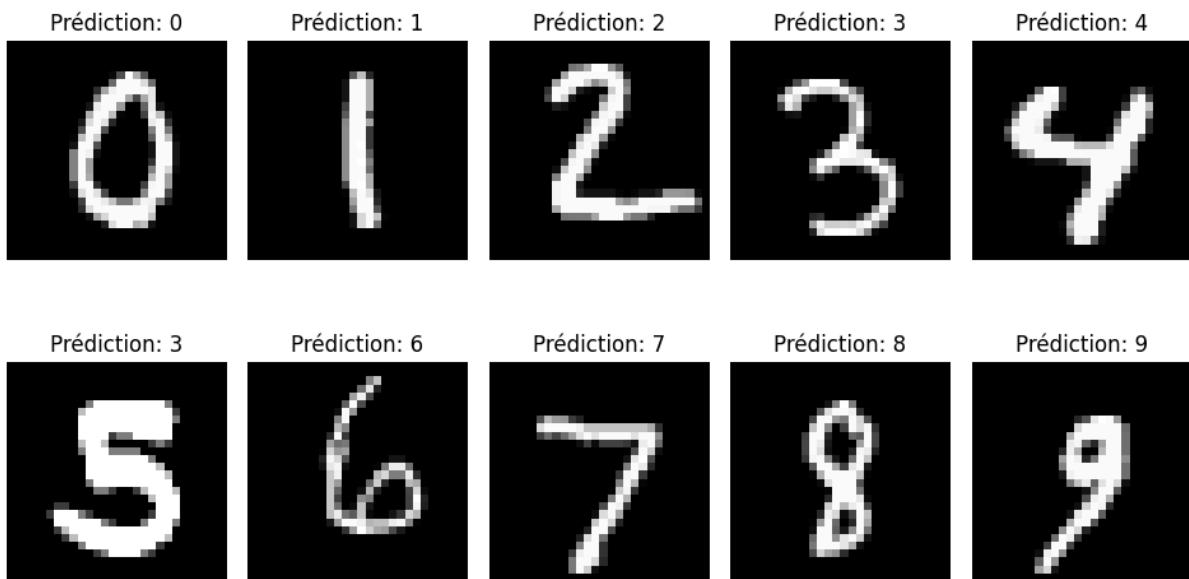


Fig. 51. – Résultats des inférences de notre modèle sur 10 données.

5.3 Résumé

Dans ce chapitre, nous avons réalisé deux mini-projets de deep learning, couvrants les deux problématiques du deep learning ; la régression et la classification. Le premier projet, intitulé « Calculatrice », visait à construire un modèle capable de prédire le résultat d'opérations mathématiques basiques, tandis que le second projet se concentrerait sur la reconnaissance de chiffres manuscrits à l'aide du jeu de données MNIST. Ces projets ont été soigneusement conçus pour vous fournir une expérience pratique et approfondie des différentes phases de développement d'un modèle de deep learning, depuis la génération de données jusqu'à l'évaluation des performances.

L'utilisation du framework PyTorch permet une immersion dans les techniques et les outils essentiels du deep learning, facilitant ainsi l'apprentissage et l'application des connaissances acquises, comme vous l'avez remarqué, en pratique, pas besoin de savoir implémenter les mathématiques de la rétropropagation du gradient pour entraîner notre modèle, PyTorch nous offre une facilité de programmation dit haut niveau et de gagner en productivité.

Dans le projet « calculatrice », vous avez abordé un problème de régression où le but était de prédire le résultat d'une opération mathématique. Vous avez commencé par générer un jeu de

données en créant des données synthétiques, c'est-à-dire, des données qui ne sont pas issues du monde réel, cette approche vous aura peut-être permis de jouer sur le nombre de données à fournir à notre modèle, si nous fournissons trop peu de données à notre modèle, notre modèle n'arrivait pas à se généraliser. Nous avons défini une architecture flexible qui nous permet de modifier sa largeur (le nombre d'unités, de perceptron ici, par couche cachée) et sa profondeur (nombre de couches cachées), cette flexibilité vous aura peut-être permis d'explorer plusieurs architectures différentes démontrant l'importance de l'expérimentation et de coder sa manière à pouvoir configurer notre manière de modifier notre approche de l'entraînement modèle au fur et à mesure des itérations. Vous avez sûrement remarqué l'optimiseur « Adam » fonctionne mieux que l'optimiseur « SGD » si vous avez bien surveillé la performance des modèles sur la courbe de la fonction de perte de votre modèle.

L'évaluation des modèles entraînés se fait sur des ensembles de test, permettant d'apprécier leur capacité à généraliser sur de nouvelles données. Cette étape est cruciale pour valider la fiabilité des modèles et leur applicabilité dans des situations réelles. Les résultats sont analysés sous différents angles, questionnant la capacité des modèles à prédire correctement les résultats et à généraliser sur de nouvelles données.

Les métriques de performance telles que l'erreur quadratique moyenne (MSE), l'erreur absolue moyenne (MAE) et le coefficient de détermination (R^2) sont utilisées pour une évaluation quantitative des performances. Ces métriques offrent des perspectives variées sur la performance des modèles, soulignant l'importance d'une évaluation multidimensionnelle.

La configuration matérielle, incluant la détection et l'utilisation d'un GPU compatible CUDA, est une partie pratique importante du projet. Elle influence directement la vitesse et l'efficacité de l'entraînement des modèles.

Les bibliothèques Python couramment utilisées dans le domaine du deep learning sont introduites, permettant aux apprenants de se familiariser avec les outils et techniques standards de l'industrie. La manipulation de tableaux multidimensionnels avec NumPy, la visualisation des données avec Matplotlib et la construction et l'entraînement des modèles avec PyTorch sont des compétences essentielles développées tout au long du projet.

La reproductibilité des résultats est un aspect clé du projet, soulignée par l'importance de fixer une graine aléatoire (seed). Cela garantit que les résultats peuvent être reproduits de manière fiable, un aspect crucial dans un domaine où les expériences doivent souvent être répétées.

Deux méthodes pour définir l'architecture des réseaux en PyTorch sont explorées : l'utilisation de `Sequential` pour des architectures simples et la création de classes personnalisées héritant de `nn.Module` pour des structures plus complexes. Cette flexibilité souligne la diversité des approches disponibles pour les praticiens du deep learning.

La segmentation des données en ensembles d'entraînement et de validation, ainsi que l'utilisation de `DataLoaders`, sont essentielles pour une gestion efficace des jeux de données et une évaluation équilibrée des modèles.

La visualisation des courbes de perte d'entraînement et de validation offre une perspective visuelle sur les performances du modèle au fil du temps. De plus, la fonction `summary` fournit

un aperçu détaillé de l'architecture du modèle, y compris le nombre de paramètres de chaque couche.

Le projet de reconnaissance de chiffres manuscrits avec le jeu de données MNIST est un exemple de classification d'images appliquée. Les images de 28x28 pixels représentant des chiffres de 0 à 9 sont converties en tenseurs et normalisées pour l'entraînement du modèle de MLP. Cette étape de préparation des données est essentielle pour la réussite du projet.

La normalisation des images et la division du jeu de données MNIST en sous-ensembles d'entraînement et de validation sont des étapes clés de la préparation des données. Trois DataLoaders sont créés pour faciliter le chargement et la manipulation des données lors de l'entraînement.

L'exploration des données, incluant la visualisation d'images, le calcul d'images moyennes et la création d'histogrammes, est une étape essentielle pour comprendre le jeu de données MNIST. Cette compréhension est cruciale pour l'adaptation et l'optimisation du modèle.

L'évaluation du modèle sur l'ensemble de test MNIST et le calcul de différentes métriques permettent de quantifier la performance du modèle. L'utilisation de plusieurs métriques offre une évaluation complète de la capacité du modèle à classer les chiffres manuscrits.

La visualisation des prédictions du modèle sur un ensemble d'images test offre une perspective qualitative sur les capacités du modèle. Cette phase permet de comprendre les forces et les limites du modèle dans la classification des chiffres manuscrits.

J'espère que cette exploration approfondie des projets « Calculatrice » et « Reconnaissance de chiffres manuscrits » a renforcé votre compréhension du domaine fascinant du deep learning. Ces projets sont conçus pour stimuler votre curiosité et développer vos compétences en résolution de problèmes, vous préparant ainsi à relever des défis plus complexes dans les chapitres à venir.

5.4 Questions

1. **Quelle est la différence entre créer un modèle Séquentiel et par l'héritage d'une classe Python avec nn.Module ?**
 - a. Le modèle Séquentiel est plus flexible
 - b. La classe nn.Module permet une personnalisation plus avancée
 - c. Le modèle Séquentiel est plus adapté aux architectures complexes
 - d. Il n'y a aucune différence significative
2. **Sur quels critères faut-il réfléchir pour définir une architecture ? (plusieurs bonnes réponses)**
 - a. La taille de l'ensemble des données
 - b. Le type de tâche à réaliser (classification, régression, etc.)
 - c. La capacité de calcul disponible
 - d. La couleur préférée de l'utilisateur
3. **Peut-on combiner le module Sequential de PyTorch avec une classe personnalisée héritant de nn.Module ?**
 - a. Oui, cela est possible et peut-être utile pour définir des blocs de notre architecture qui vont se répéter.
 - b. Non, cela n'est jamais possible
 - c. Oui, mais seulement dans certains cas spécifiques
 - d. Non, car ils servent des objectifs différents
4. **Quel est le meilleur optimiseur entre SGD et Adam ?**
 - a. SGD est habituellement meilleur
 - b. Adam est habituellement meilleur
 - c. Cela dépend de la situation spécifique et du problème
 - d. Aucun des deux n'est utile
5. **Quelle est la meilleure fonction d'activation entre ReLU et Sigmoid ? (plusieurs bonnes réponses)**
 - a. ReLU est habituellement la meilleure
 - b. Sigmoid est habituellement la meilleure
 - c. Cela dépend du type de réseau et de la couche
 - d. Aucune des deux n'est efficace
6. **Quelle est la différence entre le mode Eval et entraînement de PyTorch ? (plusieurs bonnes réponses)**
 - a. Le mode Eval utilise plus de mémoire pour accélérer l'inférence
 - b. Le mode Entraînement désactive certaines fonctions comme le dropout
 - c. Le mode Eval utilise moins de mémoire parce qu'il ne stocke pas les gradients
 - d. Le mode Eval désactive certaines fonctions comme le dropout
7. **Pourquoi utilise-t-on un GPU pour entraîner un modèle de deep learning ?**
 - a. Le GPU offre une plus grande capacité de stockage pour les données
 - b. Le GPU accélère les calculs parallèles nécessaires lors de l'entraînement
 - c. L'utilisation d'un GPU est plus économique en termes de consommation d'énergie
 - d. Le GPU est nécessaire pour visualiser les résultats de l'entraînement

8. C'est quoi normaliser ses données ?

- a. Changer toutes les données en zéros et uns
- b. Supprimer les données inutiles
- c. Ajuster les données pour qu'elles aient une moyenne de 0 et un écart-type de 1
- d. Convertir toutes les données en un format unique

9. À quoi sert un DataLoader PyTorch ?

- a. À charger des données depuis le disque dur
- b. À optimiser les modèles de deep learning
- c. À préparer et organiser les données pour l'entraînement
- d. À augmenter la taille des ensembles de données

10. Pourquoi faut-il utiliser un DataLoader PyTorch ?

- a. Pour accélérer le processus d'entraînement
- b. Pour permettre une meilleure répartition des données en lots
- c. Pour faciliter la manipulation des données
- d. Toutes les réponses sont correctes

11. Pourquoi est-il important de normaliser ses données ?

- a. Pour réduire la taille du fichier de données
- b. Pour accélérer le processus d'entraînement du modèle
- c. Pour s'assurer que les caractéristiques contribuent équitablement à l'apprentissage du modèle
- d. Pour augmenter la résolution des images

12. Pourquoi utilise-t-on des ensembles d'entraînement, de validation et de test dans le deep learning ?

- a. Pour augmenter la taille de l'ensemble de données global
- b. Pour évaluer la performance du modèle à différentes étapes et éviter le surajustement
- c. Pour entraîner des modèles différents sur chaque ensemble
- d. Uniquement pour respecter les normes du domaine

13. Quelle est la différence entre le jeu de validation et de test ?

- a. Le jeu de validation est utilisé pour l'entraînement, le test pour l'évaluation
- b. Il n'y a pas de différence
- c. Le jeu de test est utilisé pour l'entraînement, la validation pour l'évaluation
- d. Le jeu de validation est utilisé pour ajuster les hyperparamètres, le test pour évaluer la généralisation

14. Quelle est la meilleure métrique pour une tâche de classification ?

- a. L'exactitude (accuracy)
- b. La précision (precision)
- c. Le recall
- d. Il n'y a pas de meilleure métrique, il faut en utiliser plusieurs, elles ont toutes leur point fort et point faible

15. Dans le projet de reconnaissance de chiffres manuscrits, pourquoi utilise-t-on des métriques comme l'exactitude, la précision, le recall et le F1 score pour évaluer le modèle ?

- a. Pour montrer que le modèle fonctionne parfaitement
 - b. Pour obtenir une évaluation complète et équilibrée des performances du modèle
 - c. Car ce sont les seules métriques disponibles dans PyTorch
 - d. Pour respecter les normes réglementaires
16. **Comment la normalisation des données influence-t-elle l'entraînement d'un modèle de deep learning ?**
- a. Elle permet d'accélérer l'entraînement en réduisant la variabilité des données
 - b. Elle augmente la précision des prédictions en changeant la distribution des données
 - c. Elle n'a aucun impact sur l'entraînement ou la performance du modèle
 - d. Elle rend le modèle plus complexe et difficile à entraîner
17. **Quel est l'impact de l'utilisation de la fonction d'activation ReLU dans un réseau de neurones ?**
- a. Elle transforme toutes les entrées en sorties positives
 - b. Elle introduit une non-linéarité dans le modèle
 - c. Elle normalise les données d'entrée
 - d. Elle réduit la dimensionnalité des données

5.5 Réponses

1. **Quelle est la différence entre créer un modèle Séquentiel et par l'héritage d'une classe Python avec nn.Module ?**

Réponse: b La classe nn.Module permet une personnalisation plus avancée

2. **Sur quels critères faut-il réfléchir pour définir une architecture ?**

Réponses: a, b, c La taille de l'ensemble des données, le type de tâche à réaliser, et la capacité de calcul disponible

3. **Peut-on combiner le module Sequential de PyTorch avec une classe personnalisée héritant de nn.Module ?**

Réponse: a Oui, cela est possible et peut-être utile pour définir des blocs de notre architecture qui vont se répéter

4. **Quel est le meilleur optimiseur entre SGD et Adam ?**

Réponse: b Adam est habituellement meilleur

5. **Quelle est la meilleure fonction d'activation entre ReLU et Sigmoid ?**

Réponse: a, c ReLU est habituellement la meilleure, cela dépend du type de réseau et de la couche

6. **Quelle est la différence entre le mode Eval et entraînement de PyTorch ?**

Réponse: c, d Le mode Eval utilise moins de mémoire parce qu'il ne stocke pas les gradients, le mode Eval désactive certaines fonctions comme le dropout

7. **Pourquoi utilise-t-on un GPU pour entraîner un modèle de deep learning ?**

Réponse: b Le GPU accélère les calculs parallèles nécessaires lors de l'entraînement

8. **C'est quoi normaliser ses données ?**

Réponse: c Ajuster les données pour qu'elles aient une moyenne de 0 et un écart-type de 1

9. **À quoi sert un DataLoader PyTorch ?** Réponse: c À préparer et organiser les données pour l'entraînement

10. **Pourquoi faut-il utiliser un DataLoader PyTorch ?**

Réponse: d Toutes les réponses sont correctes

11. **Pourquoi est-il important de normaliser ses données ?**

Réponse: c Pour s'assurer que les caractéristiques contribuent équitablement à l'apprentissage du modèle

12. **Pourquoi utilise-t-on des ensembles d'entraînement, de validation et de test dans le deep learning ?**

Réponse: b Pour évaluer la performance du modèle à différentes étapes et éviter le surajustement

13. **Quelle est la différence entre le jeu de validation et de test ?**

Réponse: d Le jeu de validation est utilisé pour ajuster les hyperparamètres, le test pour évaluer la généralisation

14. Quelle est la meilleure métrique pour une tâche de classification ?

Réponse: d Il n'y a pas de meilleure métrique, il faut en utiliser plusieurs, elles ont toutes leur point fort et point faible

15. Dans le projet de reconnaissance de chiffres manuscrits, pourquoi utilise-t-on des métriques comme l'exactitude, la précision, le recall et le F1 score pour évaluer le modèle ?

Réponse: b Pour obtenir une évaluation complète et équilibrée des performances du modèle

16. Comment la normalisation des données influence-t-elle l'entraînement d'un modèle de deep learning ?

Réponse: a Elle permet d'accélérer l'entraînement en réduisant la variabilité des données

17. Quel est l'impact de l'utilisation de la fonction d'activation ReLU dans un réseau de neurones ?

Réponse: b Elle introduit une non-linéarité dans le modèle

6 La Généralisation des modèles de Deep Learning

La généralisation, ou la capacité d'un modèle à performer de manière fiable sur des données inédites, constitue l'essence même de l'IA. Après tout, quel intérêt y aurait-il à développer un modèle qui excelle sur les données d'entraînement mais échoue lamentablement face à de nouveaux exemples ? C'est précisément sur cette aptitude à généraliser que repose la valeur des modèles de deep learning.

La généralisation est loin d'être une tâche triviale. Les modèles de deep learning, avec leur capacité à modéliser des relations complexes et non linéaires, sont particulièrement vulnérables aux écueils du surajustement (overfitting) et du sous-ajustement (underfitting). Le premier se produit lorsqu'un modèle, trop complexe, commence à mémoriser les spécificités et le bruit des données d'entraînement au détriment de l'apprentissage des tendances générales. À l'inverse, le sous-ajustement survient quand un modèle, trop simple, ne parvient pas à saisir la richesse des relations présentes dans les données.

Trouver le juste équilibre entre ces deux extrêmes est au cœur du défi de la généralisation. C'est ici qu'intervient le fameux compromis biais-variance, un concept fondamental en apprentissage statistique. Le biais représente l'erreur due aux hypothèses simplificatrices du modèle, tandis que la variance reflète sa sensibilité aux fluctuations des données d'entraînement. Un bon modèle doit minimiser ces deux sources d'erreur pour atteindre une généralisation optimale.

Nous aborderons les stratégies de validation croisée (cross-validation) pour estimer de manière fiable les performances de généralisation, et nous discuterons de l'importance d'un ensemble de données diversifié et représentatif pour atténuer le biais de sélection. En maîtrisant ces notions, vous serez en mesure de développer des modèles robustes et polyvalents.

6.1 Présentation des concepts clés: généralisation, surajustement et sous-ajustement

6.1.1 Le surajustement (overfitting)

Imaginez que vous soyez un espion en herbe qui, pour sa première mission, doit identifier les super-héros parmi les invités d'une fête costumée sur le thème des bandes dessinées. Au début, vous utilisez des indices généraux pour identifier les super-héros : une cape, un masque, peut-être un peu de spandex. Cela fonctionne assez bien, un peu comme un modèle de deep learning qui apprend des caractéristiques générales à partir d'un ensemble de données.

Cependant, au fur et à mesure que la soirée avance (ou l'entraînement de votre modèle) et que vous gagnez en confiance, vous commencez à baser vos identifications sur des détails de plus en plus triviaux et absurdes. Vous voyez quelqu'un avec des chaussures rouges et concluez immédiatement que c'est Superman (parce que, évidemment, Superman adore les chaussures rouges !). Ou vous décidez que toute personne portant des lunettes est Clark Kent en déguisement. Ces associations deviennent de plus en plus ridicules, tout comme un modèle de deep learning commence à surajuster en mémorisant les détails spécifiques et non pertinents des données d'entraînement.

Lorsque de nouveaux invités arrivent, vos théories s'effondrent spectaculairement. La personne en chaussures rouges est en fait déguisée en Deadpool, et celle avec des lunettes est

juste un fan de Harry Potter. C'est l'overfitting en action : vous avez appris des associations tellement spécifiques qu'elles ne sont plus valables pour un groupe plus large. Vous avez basé vos suppositions sur des détails tellement idiosyncrasiques qu'ils ne s'appliquent pas au-delà de la poignée de personnes que vous avez initialement rencontrées.

Pourquoi cela s'est-il produit ? Dans votre enthousiasme d'espion débutant, vous avez pris des raccourcis comiques et complètement erronés, en associant des caractéristiques aléatoires à des identités de super-héros. En termes de deep learning, cela équivaut à un modèle qui s'attache à des détails non représentatifs des données d'entraînement.



Fig. 52. – Lorsque le modèle a été un peu trop entraîné sur une caractéristique du père, la généralisation devient un défi comique et surprenant.

6.1.2 Le sous-ajustement (underfitting)

Poursuivant notre analogie, considérons à présent le sous-ajustement comme le pendant du surajustement. Si, dans le cas précédent, vous étiez concentré sur des indices trop spécifiques pour identifier les super-héros, dans le scénario du sous-ajustement, c'est comme si vous aviez décidé de ne regarder que des éléments trop généraux ou évidents, au point de manquer des nuances importantes. Imaginons que vous basiez uniquement sur la présence d'une cape pour identifier un super-héros. Ce critère est si vague et répandu que vous finiriez par classer chaque personne portant une simple cape de fête comme un super-héros, en omettant de reconnaître les véritables héros parmi eux.

En termes de deep learning, le sous-ajustement survient lorsque le modèle est trop simpliste et ne parvient pas à saisir la complexité sous-jacente des données. Cela peut se produire pour diverses raisons : peut-être que le modèle n'a pas assez de couches ou de neurones pour traiter la complexité des données, ou il se pourrait que l'algorithme d'entraînement n'ait pas été entraîné suffisamment longtemps pour capturer les relations entre les caractéristiques.

Dans notre scénario de fête, cela se traduit par un manque de discernement qui vous empêche de distinguer entre un invité déguisé en magicien et un véritable maître de l'illusion. Si vous aviez observé plus attentivement, peut-être auriez-vous remarqué que le vrai magicien mani-

pule avec aisance un jeu de cartes ou fait disparaître des objets, tandis que le faux magicien ne fait que porter un grand chapeau noir. Mais votre méthode trop simpliste vous a privé de ces indices.

Le sous-ajustement est également préjudiciable : il se manifeste par une incapacité à effectuer des prédictions précises, même sur l'ensemble d'entraînement, ce qui entraîne une mauvaise performance du modèle en général. C'est comme si un modèle de reconnaissance d'images classait toutes les images avec du bleu et du vert comme des paysages, sans distinguer entre les images de forêts, de villes avec des parcs ou des peintures abstraites.

Pour combattre le sous-ajustement, il serait nécessaire d'augmenter la complexité du modèle ou d'allonger la durée de l'entraînement. En revenant à notre fête, cela équivaudrait à prendre le temps de noter des détails plus subtils, comme la façon dont une personne réagit lorsqu'elle est confrontée à un tour de magie ou si elle semble avoir une dextérité qui dépasse la moyenne d'un simple amateur.

Le sous-ajustement et le surajustement sont deux écueils dans le développement de modèles de deep learning performants. L'un néglige la complexité, tandis que l'autre s'y attarde de façon excessive. Le défi consiste à trouver l'équilibre optimal, où le modèle est suffisamment sophistiqué pour apprendre les subtilités des données, sans pour autant se perdre dans leurs spécificités. Cet équilibre est essentiel pour atteindre une bonne généralisation, permettant au modèle de fonctionner efficacement dans diverses situations et de s'adapter à de nouvelles données, tout comme un espion compétent qui doit reconnaître un super-héros dans toutes sortes de contextes, en dépassant les apparences pour découvrir la véritable essence de son caractère.

6.2 Surajustement et Sous-ajustement: Compréhension et Diagnostic

Supposons que notre modèle MLP entraîné sur le jeu de données MNIST du chapitre précédent soit très bon sur le jeu d'entraînement, disons une exactitude de 99%, mais ne parvienne qu'à une exactitude de 60% sur l'ensemble de test. Ce scénario est un exemple classique d'un surajustement ou « overfitting », ici le modèle a appris à reconnaître avec précision les chiffres dans l'ensemble d'entraînement, mais cette performance ne se généralise pas aux données inédites de l'ensemble de test.

Le surajustement est intrinsèquement lié à la complexité des modèles de deep learning. Un modèle de deep learning va s'ajuster aux données d'entraînement, si vous décidez de faire une architecture plus grosse, chaque paramètre supplémentaire offre au modèle une opportunité d'affiner davantage sa correspondance avec les données d'entraînement, parfois au détriment de la capacité à prédire des données non vues. Si votre modèle commence à même apprendre les détails spécifiques et le « bruit » qui se trouve dans les données d'entraînement, au lieu de découvrir des motifs et des tendances généralisables, l'entraînement sera raté, car votre modèle ne se généralisera pas. Simplifier un modèle peut se traduire par la réduction du nombre de couches ou de neurones.

L'entraînement prolongé est un facteur susceptible de causer le surajustement. Lorsqu'un modèle est entraîné pendant un nombre excessif d'itérations, il peut commencer à mémoriser spécifiquement les détails des données d'entraînement plutôt que d'apprendre des caractéristiques généralisables. Pour cela on utilisera une méthode qui se nomme l'*early-stopping*.

L'early stopping est une stratégie pour prévenir le surajustement. Cette technique consiste à arrêter l'entraînement d'un modèle lorsque la performance sur l'ensemble de validation commence à se dégrader, signe que le modèle ne parvient plus à généraliser et commence à apprendre par cœur les spécificités des données d'entraînement. En pratique, cela signifie surveiller la performance du modèle sur un ensemble de validation à chaque époque et arrêter l'entraînement lorsque l'erreur de validation ne diminue plus ou commence à augmenter. Le modèle est alors restauré à l'état où il a obtenu la meilleure performance sur l'ensemble de validation. Cette approche permet de capturer le modèle à un état auquel il est suffisamment entraîné pour reconnaître les patterns dans les données mais pas trop pour inclure le bruit aléatoire.

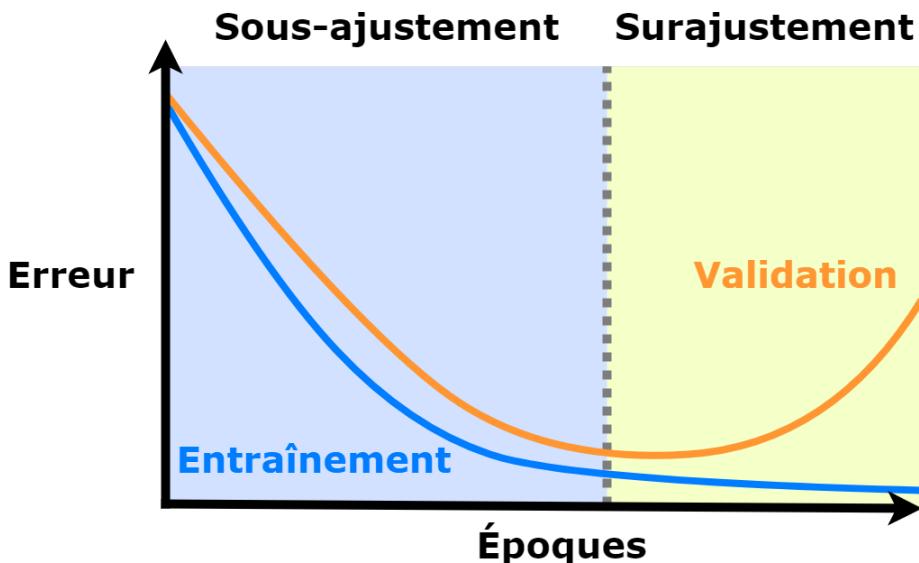


Fig. 53. – Représentation schématique de l'évolution de l'erreur d'entraînement (en bleu) et de l'erreur de validation (en orange) au fil des époques. La région à gauche, où les deux courbes baissent, caractérise une phase de sous-ajustement où le modèle ne capture pas encore toute la complexité des données. À droite, la divergence des courbes marque la transition vers le surajustement, où l'erreur de validation augmente, indiquant une perte de généralisation du modèle face à de nouvelles données. Au milieu en pointillé, le point optimal d'arrêt de l'entraînement avant que le modèle ne devienne trop spécialisé sur les données d'entraînement

Le sous-ajustement, en contraste, survient lorsque le modèle est trop simple pour capturer la structure et la complexité des données. Il se manifeste par une performance médiocre tant sur les données d'entraînement que de test. Il indique qu'un modèle est incapable de modéliser adéquatement les relations dans les données, souvent dû à un manque de profondeur ou de capacité dans l'architecture du réseau. Cela peut également résulter d'un entraînement insuffisant ou d'une initialisation inappropriée des poids du réseau.

6.2.1 Visualiser l'Overfitting et l'Underfitting

Pour illustrer concrètement les concepts d'overfitting et d'underfitting, considérons un exemple de régression simple dans un espace à deux dimensions. Supposons que nous cherchions à modéliser la relation entre une variable d'entrée x et une variable de sortie y . Notre objectif est d'entraîner un modèle capable de prédire la valeur de y pour une nouvelle valeur de x .

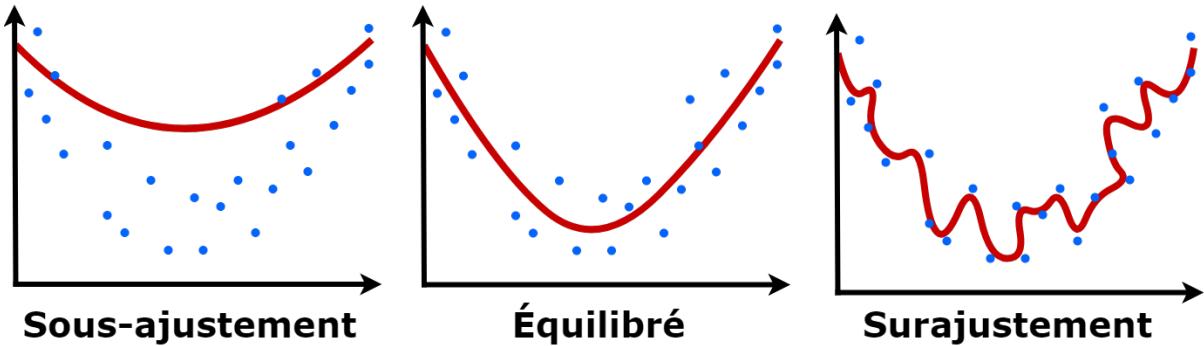


Fig. 54. – Pour illustrer concrètement les concepts d’overfitting et d’underfitting, considérons un exemple de régression simple dans un espace à deux dimensions. Supposons que nous cherchions à modéliser la relation entre une variable d’entrée x et une variable de sortie y . Notre objectif est d’entraîner un modèle capable de prédire la valeur de y pour une nouvelle valeur de x .

Dans le scénario d’underfitting (à gauche), le modèle est trop simple, représenté ici par une droite. Il ne parvient pas à capturer la relation non linéaire entre x et y . Bien qu’il ne soit pas perturbé par le bruit dans les données, ses prédictions resteront imprécises, car il ne peut pas représenter la complexité de la vraie relation sous-jacente.

À l’opposé, le scénario d’overfitting (à droite) montre un modèle qui est devenu trop complexe, ici un polynôme de degré élevé. Il s’ajuste parfaitement à chaque point d’entraînement, y compris le bruit et les variations aléatoires. Bien que ce modèle ait une erreur d’entraînement très faible, il est peu probable qu’il généralise bien à de nouvelles données. Il a essentiellement « mémorisé » les données d’entraînement plutôt que d’apprendre la véritable relation sous-jacente.

Le scénario du milieu illustre un bon équilibre. Le modèle, ici un polynôme de degré modéré, capture la tendance générale des données sans être excessivement influencé par des points individuels. Il réalise un bon compromis entre l’ajustement aux données d’entraînement et la capacité à généraliser à de nouvelles données.

En pratique, nous surveillons les performances du modèle sur les ensembles d’entraînement et de validation pour détecter ces phénomènes. Un signe d’overfitting est lorsque l’erreur sur l’ensemble d’entraînement continue de diminuer tandis que l’erreur sur l’ensemble de validation stagne ou commence à augmenter. Cela indique que le modèle commence à mémoriser le bruit dans les données d’entraînement. À l’inverse, des erreurs élevées à la fois sur les ensembles d’entraînement et de validation suggèrent un underfitting.

L’objectif est de trouver le juste équilibre, où le modèle capture les relations pertinentes dans les données d’entraînement, comme en témoigne une faible erreur d’entraînement, tout en conservant une bonne performance sur l’ensemble de validation, indiquant sa capacité à généraliser. Des techniques telles que la régularisation et l’early stopping sont essentielles pour atteindre cet équilibre optimal dans les modèles de deep learning.

6.2.2 Le biais de sélection

Le biais de sélection est un autre facteur qui affecte la généralisation des modèles de deep learning, ce biais survient lorsqu'il existe une discordance significative entre les données

sur lesquelles le modèle a été entraîné et celles sur lesquelles il est testé. Si les données d'entraînement sont trop homogènes ou si elles ne reflètent pas fidèlement la diversité des cas que le modèle rencontrera dans des situations réelles, il y a un risque élevé que le modèle ne parvienne pas à généraliser correctement.

Prenons l'exemple d'un modèle conçu pour la reconnaissance d'images d'animaux. Si le modèle est entraîné exclusivement sur des images prises par des appareils photo professionnels et testé sur des images prises par un smartphone, la différence dans la qualité d'image, l'éclairage et même la perspective peut entraîner une chute des performances. De même, un modèle entraîné uniquement sur des textes formels pourrait se retrouver démunie face à des textes informels ou des argots, car il n'aura pas appris la variabilité et les nuances de la langue naturelle en dehors d'un cadre formel.

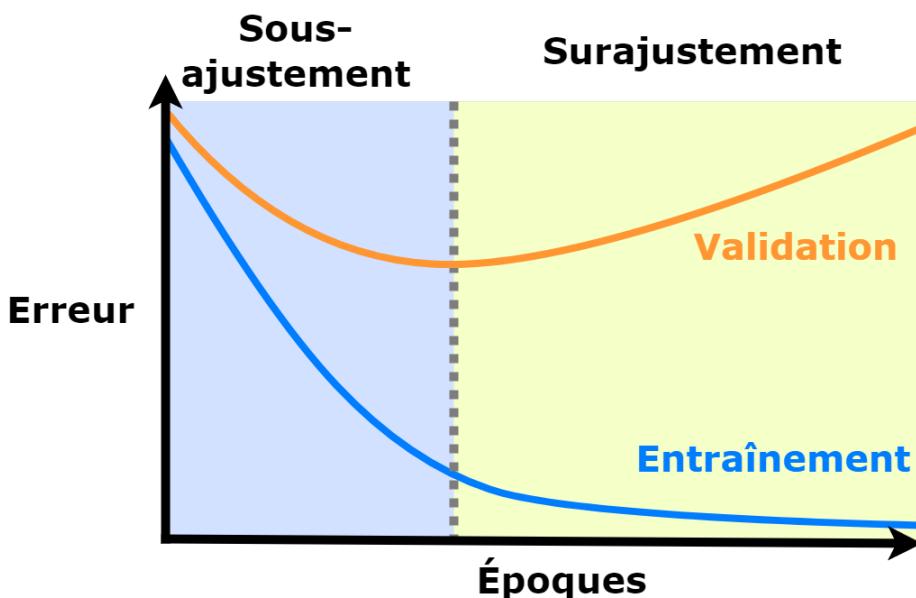


Fig. 55. – Cette courbe montre l'effet du biais de sélection sur la performance d'un modèle de deep learning. L'écart croissant entre l'erreur d'entraînement et l'erreur de validation indique un modèle surajusté aux données homogènes et non représentatives, révélant un manque de généralisation sur des données diversifiées.

D'autre part, l'ajout de données est souvent la clé pour améliorer la généralisation des modèles de deep learning. Un ensemble de données volumineux et diversifié expose le modèle à une variété de scénarios, l'obligeant à apprendre les caractéristiques fondamentales qui déterminent les sorties plutôt que les particularités spécifiques à un sous-ensemble de données. Plus il y a de données, plus le modèle peut affiner sa compréhension des caractéristiques générales qui sont vraiment prédictives. Cela est analogue à la façon dont une expérience humaine plus riche conduit à une meilleure capacité de généralisation. Par exemple, un médecin ayant vu des milliers de patients sera mieux à même de diagnostiquer une maladie rare qu'un novice. De la même manière, un modèle d'apprentissage profond entraîné sur une large gamme d'exemples apprendra à distinguer le signal du bruit, renforçant ainsi sa capacité de généralisation.

6.3 Principes et Méthodologies de la Cross-Validation

La cross-validation, aussi connue sous le nom de validation croisée, est une méthode pour évaluer la performance et la capacité de généralisation d'un modèle de machine learning ou

de deep learning. Cette méthode est utile pour limiter les risques de surajustement et de sous-ajustement, en fournissant une estimation plus fiable de la performance du modèle sur de nouvelles données. Cette méthode est utilisée lorsque vous avez peu de données, et pas assez pour constituer un ensemble de validation suffisamment fiable.

La validation croisée permet de surmonter ce dilemme en offrant une méthode pour estimer la performance d'un système sans nécessiter un ensemble de test dédié. Le principe sous-jacent est simple, il s'agit de diviser l'ensemble de données en plusieurs sous-ensembles, puis d'entraîner le modèle sur certains de ces sous-ensembles tout en le testant sur le sous-ensemble de validation. On divise l'ensemble de données en K sous-ensembles (ou « folds ») de taille égale. Le processus se déroule ensuite en K itérations, où à chaque itération, un des K folds est utilisé comme ensemble de validation, tandis que les $K - 1$ autres folds sont combinés pour former l'ensemble d'entraînement. Le modèle est alors entraîné sur l'ensemble d'entraînement et évalué sur l'ensemble de validation. Ce processus est répété K fois, de sorte que chaque fold serve exactement une fois d'ensemble de validation.

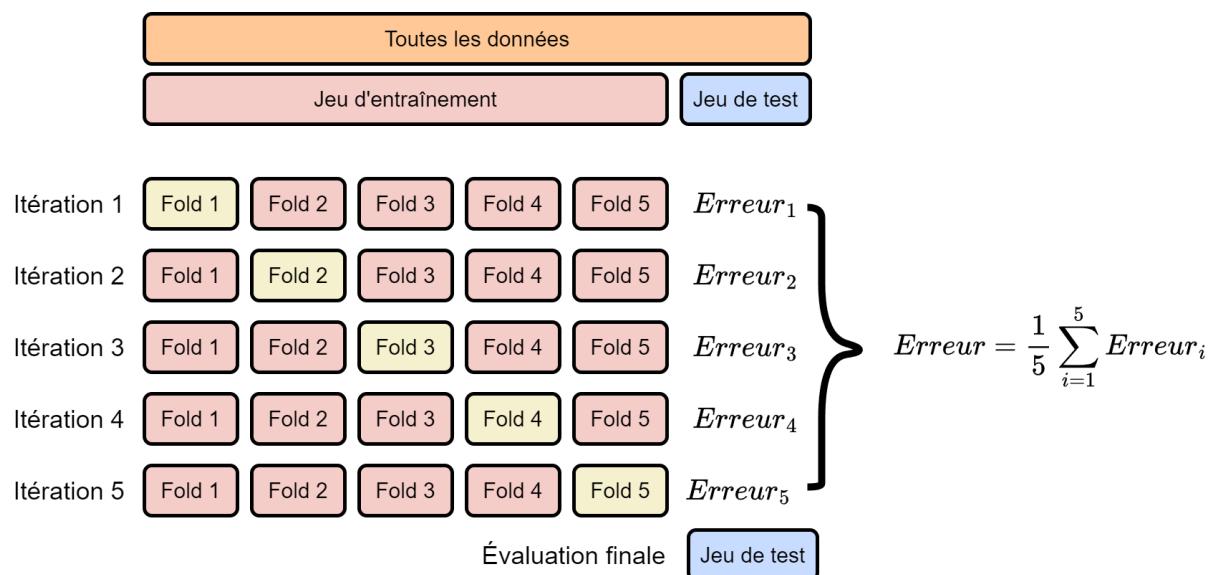


Fig. 56. – L'ensemble de données complet est divisé en cinq sous-ensembles, ou « folds », chaque rangée représente une itération de validation croisée où un fold différent est utilisé comme jeu de validation (en jaune) et les quatre autres comme jeu d'entraînement (en rose), pour finir par le jeu de test (en bleu).

À la fin des K itérations, nous obtenons K scores de performance (par exemple, l'exactitude ou l'erreur quadratique moyenne, MSE), un pour chaque fold. La performance globale du modèle est alors estimée en prenant la moyenne de ces K scores :

$$\text{Performance} = \frac{1}{K} \times \sum_{i=1}^K \text{Performance}_i$$

où Performance_i est le score de performance obtenu lorsque le i -ème fold est utilisé comme ensemble de validation.

Cette approche présente plusieurs avantages majeurs. Premièrement, elle permet d'utiliser toutes les données disponibles pour l'entraînement et la validation, ce qui est particulièrement

précieux lorsque les données sont limitées. Deuxièmement, en répétant le processus K fois avec différents ensembles de validation, elle fournit une estimation plus robuste et moins variable de la performance du modèle, en comparaison à une simple division en ensembles d'entraînement et de test. Cela réduit le risque que l'estimation de la performance soit biaisée par une division particulière des données.

Le choix de K , c'est-à-dire le nombre de folds, est un hyperparamètre dans la cross-validation. Une valeur couramment utilisée est $K = 5$, ce qui correspond à une validation croisée à 5 folds. Cependant, il faut répéter le processus d'entraînement K fois, ce qui peut être computationnellement coûteux, cette méthode existe, et est souvent appliquée sur des tâches de Machine Learning classique, mais est peu pratiqué pour le deep learning qui a des entraînements bien plus longs que le machine learning classique, lancé 5 fois un entraînement pour mieux estimer la généralisation du modèle n'est pas forcément intéressant par rapport à une simple division entre le jeu d'entraînement et de validation.

6.4 Comprendre le compromis biais-variance en deep learning

Le compromis biais-variance (« bias-variance trade-off ») est un concept en Machine Learning qui décrit la tension existante entre deux types d'erreurs commises par un modèle : le biais (« bias ») et la variance. Le biais mesure la tendance d'un modèle à apprendre systématiquement des représentations incorrectes ou simplifiées des données, tandis que la variance quantifie sa sensibilité aux fluctuations spécifiques de l'ensemble d'entraînement. Autrement dit, un modèle avec un biais élevé fait des hypothèses trop simplistes sur les données, alors qu'un modèle avec une variance élevée est excessivement sensible au bruit et aux particularités de l'échantillon d'apprentissage.

Ce compromis est directement lié aux notions de sous-ajustement (« underfitting ») et de sur-ajustement (« overfitting ») vu précédemment. Un modèle sous-ajusté présente un biais élevé : il ne capture pas suffisamment les tendances sous-jacentes des données. À l'inverse, un modèle sur-ajusté possède une variance élevée : il mémorise les détails spécifiques du jeu d'entraînement, y compris le bruit, ce qui nuit à sa capacité à généraliser à de nouvelles données.

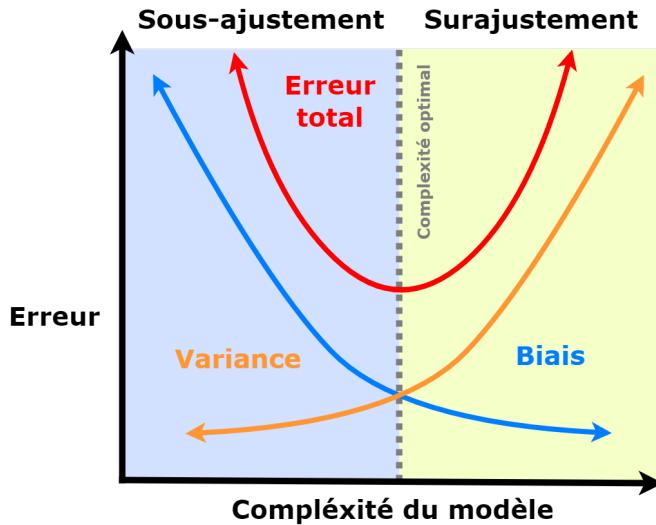


Fig. 57. – Représentation du compromis biais-variance : l’erreur totale (rouge) résulte de la somme du biais (bleu) et de la variance (orange). La complexité optimale du modèle minimise cette erreur.

Considérons l’exemple : une chercheuse en météorologie mesure quotidiennement la vitesse du vent au sommet d’une montagne. Elle suppose que ces mesures résultent de la somme d’une courbe idéale, stable d’une année sur l’autre, et d’un bruit aléatoire représentant les fluctuations imprévisibles quotidiennes.

Notre objectif est alors d’estimer cette courbe idéale à partir des données bruitées. Nous pouvons ajuster différents modèles à ces données :

- Si nous choisissons une famille de modèles simples (par exemple, des régressions linéaires ou polynomiales de faible degré), les courbes obtenues seront toutes très similaires entre elles, mais éloignées des points observés. Ces modèles possèdent un biais élevé car ils sont incapables de saisir la complexité réelle des données. Cependant, leur simplicité implique une faible sensibilité aux variations spécifiques des jeux de données utilisés : ils présentent donc une faible variance.
- À l’inverse, en choisissant une famille de modèles complexes (par exemple, des polynômes de degré élevé ou des réseaux neuronaux profonds sans régularisation), chaque courbe s’ajuste étroitement aux points observés. Chaque modèle est alors très différent selon l’ensemble précis utilisé pour son entraînement : cela traduit une forte variance. En revanche, leur flexibilité leur permet théoriquement de capturer toute forme complexe sous-jacente aux données, réduisant ainsi fortement leur biais intrinsèque.

Formalisation mathématique du compromis biais-variance

Formalisons maintenant rigoureusement ce concept. Supposons que nous disposions d’une distribution sous-jacente \mathcal{B} générant nos jeux de données \mathcal{D} :

$$\mathcal{D} \sim \mathcal{B}$$

Chaque point observé (x_i, y_i) provient d’une fonction inconnue $f(x)$ additionnée d’un bruit indépendant ε_i :

$$y_i = f(x_i) + \varepsilon_i$$

où $E[\varepsilon_i] = 0$, c'est-à-dire que le bruit a une espérance nulle.

Notre objectif est d'apprendre une fonction $\hat{f}(x)$ qui approxime au mieux la fonction inconnue $f(x)$. Pour cela, nous utilisons un algorithme d'apprentissage qui, à partir d'un jeu de données \mathcal{D} , produit une estimation $\hat{f}(x, \mathcal{D})$.

L'erreur quadratique moyenne (MSE) pour un point de test x se définit par :

$$\text{MSE}(x) = E\left[\left(y - \hat{f}(x, \mathcal{D})\right)^2\right]$$

où l'espérance est prise sur la distribution des y pour un x fixé et sur tous les jeux de données d'entraînement possibles \mathcal{D} .

Cette erreur peut être décomposée de manière élégante. Commençons par substituer $y = f(x) + \varepsilon$:

$$\text{MSE}(x) = E\left[\left(f(x) + \varepsilon - \hat{f}(x, \mathcal{D})\right)^2\right]$$

En développant cette expression et en utilisant le fait que $E[\varepsilon] = 0$ et que ε est indépendant de $\hat{f}(x, \mathcal{D})$, nous obtenons :

$$\text{MSE}(x) = E\left[\left(f(x) - \hat{f}(x, \mathcal{D})\right)^2\right] + E[\varepsilon^2]$$

Le terme $E[\varepsilon^2]$ représente la variance du bruit, que nous noterons σ^2 . C'est l'erreur irréductible, présente même avec un modèle parfait.

Pour décomposer davantage le premier terme, ajoutons et soustrayons $E[\hat{f}(x, \mathcal{D})]$:

$$E\left[\left(f(x) - \hat{f}(x, \mathcal{D})\right)^2\right] = E\left[\left(f(x) - E[\hat{f}(x, \mathcal{D})] + E[\hat{f}(x, \mathcal{D})] - \hat{f}(x, \mathcal{D})\right)^2\right]$$

En développant cette expression et en notant que les termes croisés s'annulent (car $E[\hat{f}(x, \mathcal{D}) - E[\hat{f}(x, \mathcal{D})]] = 0$), nous obtenons :

$$E\left[\left(f(x) - \hat{f}(x, \mathcal{D})\right)^2\right] = \left(f(x) - E[\hat{f}(x, \mathcal{D})]\right)^2 + E\left[\left(E[\hat{f}(x, \mathcal{D})] - \hat{f}(x, \mathcal{D})\right)^2\right]$$

Ainsi, l'erreur quadratique moyenne se décompose en trois termes :

$$\text{MSE}(x) = \underbrace{\left(f(x) - E[\hat{f}(x, \mathcal{D})]\right)^2}_{\text{Biais au carré}} + \underbrace{E\left[\left(E[\hat{f}(x, \mathcal{D})] - \hat{f}(x, \mathcal{D})\right)^2\right]}_{\text{Variance}} + \underbrace{\sigma^2}_{\text{Erreur irréductible}}$$

Cette décomposition nous permet de comprendre intuitivement les trois sources d'erreur dans notre modèle :

- Le **biais au carré** mesure l'écart entre la vraie fonction $f(x)$ et la moyenne des prédictions de notre modèle sur tous les jeux de données possibles. Un biais élevé indique que notre modèle fait systématiquement des erreurs, quelle que soit la donnée d'entraînement utilisée.

- La **variance** mesure la variabilité des prédictions de notre modèle pour différents jeux de données d'entraînement. Une variance élevée signifie que notre modèle est très sensible aux fluctuations dans les données d'entraînement.
- L'**erreur irréductible** est due au bruit inhérent aux données et ne peut être éliminée, même avec un modèle parfait.

Pour un ensemble test de taille t , l'erreur quadratique moyenne totale est simplement la moyenne des erreurs sur chaque point :

$$\text{MSE}_{\text{total}} = \frac{1}{t} \sum_{i=1}^t \text{MSE}(x_i)$$

Intuitivement, un modèle avec une variance élevée « mémorise » les données d'entraînement, y compris le bruit et les aberrations, plutôt que d'apprendre les tendances générales. En conséquence, il performera très bien sur les données d'entraînement, mais généralisera mal à de nouvelles données. La variance serait la variabilité de ces prédictions autour de leur moyenne. Si nos prédictions sont très différentes les unes des autres, notre modèle a une variance élevée.

Le biais serait alors la différence entre la moyenne de ces prédictions et la vraie valeur y . Si en moyenne nos prédictions sont éloignées de la vraie valeur, notre modèle est biaisé. Même si nous avions un modèle parfait qui prédit en moyenne la vraie valeur (biais nul) avec des prédictions consistantes (variance nulle), il y aurait toujours une erreur due au bruit dans les données. C'est ce qu'on appelle l'erreur irréductible.

Le compromis biais-variance découle du fait qu'il est utopique de minimiser à la fois le biais et la variance. Les modèles simples, comme les modèles ayant peu de paramètres, ont tendance à avoir un biais élevé, mais une faible variance. Ils font des hypothèses fortes sur la forme de la fonction cible, ce qui peut les empêcher de capturer les relations complexes dans les données. Cependant, ils sont moins sensibles aux fluctuations des données d'entraînement.

Le compromis biais-variance stipule qu'en général, les modèles plus complexes auront un biais plus faible, mais une variance plus élevée, tandis que les modèles plus simples auront un biais plus élevé, mais une variance plus faible.

Pour illustrer ces concepts, imaginons trois scénarios distincts :

1. **Sous-ajustement (biais élevé, variance faible)** : Considérons un modèle linéaire simple pour prédire les prix des maisons. Si la relation réelle entre la superficie et le prix est non linéaire, notre modèle linéaire ne pourra pas capturer cette complexité. Par conséquent, il aura un biais élevé, car il simplifie trop la relation. Cependant, comme il est simple, il ne sera pas très sensible aux variations dans les données d'entraînement, ce qui signifie qu'il aura une variance faible.
2. **Sur-ajustement (biais faible, variance élevée)** : Maintenant, imaginons que nous utilisons un modèle polynômial de degré très élevé pour prédire les prix des maisons. Ce modèle peut s'ajuster parfaitement aux données d'entraînement, y compris le bruit et les anomalies. Par conséquent, il aura un biais faible, car il peut capturer presque toutes les nuances des données d'entraînement. Cependant, il sera très sensible aux variations dans les données d'entraînement, ce qui signifie qu'il aura une variance élevée. Ce modèle performera bien sur les données d'entraînement, mais généralisera mal aux nouvelles données.

3. **Équilibre optimal (biais modéré, variance modérée)** : Enfin, considérons un modèle de complexité intermédiaire, tel qu'un modèle polynômial de degré modéré ou un réseau de neurones avec un nombre raisonnable de couches et de neurones. Ce modèle peut capturer les tendances générales dans les données sans être trop influencé par le bruit. Par conséquent, il aura un biais modéré et une variance modérée, atteignant un bon équilibre entre les deux.

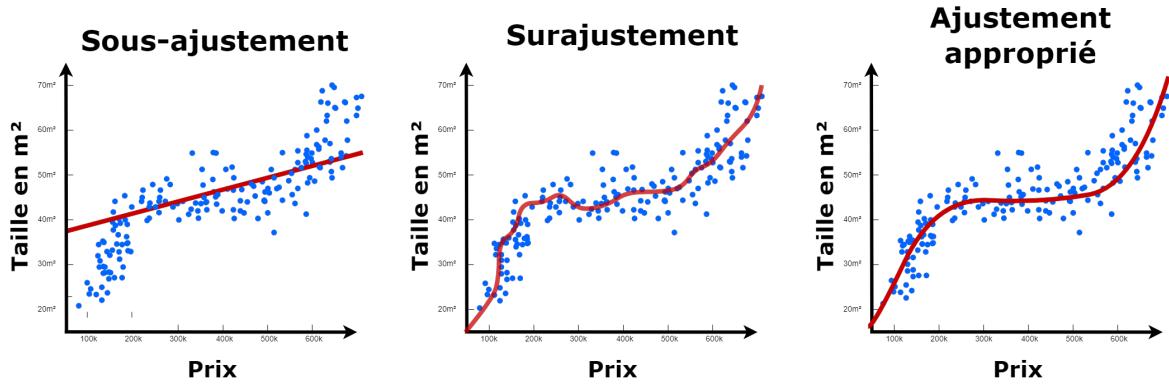


Fig. 58. – Illustration du compromis biais-variance : Les trois graphiques montrent des modèles de complexité croissante ajustés à un même ensemble de données. À gauche, un modèle linéaire simple présente un biais élevé et une faible variance, sous-ajustant les données. Au centre, un modèle polynomial de degré modéré équilibre biais et variance, offrant une meilleure approximation de la relation sous-jacente. À droite, un modèle polynomial de degré élevé présente un biais faible mais une variance élevée, sur-ajustant les données et capturant le bruit.

L'objectif est de développer un modèle suffisamment complexe pour saisir les relations pertinentes dans les données, ce qui se traduit par un faible biais. Cependant, il faut veiller à ce que le modèle ne soit pas trop complexe au point de commencer à modéliser le bruit, ce qui entraînerait une variance élevée.

Cet équilibre optimal dépend de plusieurs facteurs clés. Le premier est la complexité de la fonction sous-jacente que nous cherchons à apprendre, qui est directement liée au nombre de paramètres de notre modèle. Le deuxième facteur est la quantité de données d'entraînement à notre disposition. Enfin, le niveau de bruit présent dans ces données joue également un rôle crucial.

Lorsque nous disposons d'un volume important de données d'entraînement, nous pouvons nous permettre d'utiliser des modèles plus complexes. En effet, la variance sera réduite grâce à l'abondance d'exemples. Intuitivement, plus nous avons d'exemples, moins il est probable qu'un modèle complexe soit induit en erreur par des fluctuations spécifiques à l'ensemble d'entraînement.

Cependant, si les données sont fortement entachées de bruit, même un grand ensemble de données ne permettra pas d'éliminer complètement la variance. Dans ce cas, il peut être préférable d'opter pour un modèle plus simple, même s'il ne capture pas toutes les subtilités des données. Un modèle plus simple sera moins enclin à sur-ajuster le bruit, offrant ainsi une meilleure capacité de généralisation.

6.5 Résumé

La généralisation désigne la capacité d'un modèle de deep learning à maintenir de bonnes performances sur des données inédites, différentes de celles utilisées lors de son entraînement. Cette aptitude est essentielle, car un modèle performant uniquement sur les données d'entraînement n'a qu'une utilité très limitée en pratique.

Deux phénomènes nuisent à cette généralisation : le surajustement (overfitting) et le sous-ajustement (underfitting). Le surajustement survient lorsqu'un modèle trop complexe mémorise les particularités et le bruit des données d'entraînement, échouant à identifier les tendances générales. À l'inverse, le sous-ajustement apparaît quand un modèle trop simple ne parvient pas à saisir les relations réelles présentes dans les données, entraînant ainsi des performances médiocres même sur l'ensemble d'entraînement.

Ces deux phénomènes sont étroitement liés au compromis biais-variance. Le biais mesure l'erreur systématique due aux hypothèses simplificatrices du modèle, tandis que la variance quantifie la sensibilité du modèle aux fluctuations spécifiques des données d'entraînement. Un modèle sous-ajusté présente un biais élevé et une faible variance, alors qu'un modèle surajusté a un faible biais mais une forte variance. L'objectif est donc de trouver un équilibre optimal entre ces deux extrêmes afin d'obtenir une généralisation efficace.

Pour détecter et prévenir ces problèmes, plusieurs stratégies existent :

- **L'early stopping** consiste à arrêter l'entraînement dès que la performance sur l'ensemble de validation cesse de s'améliorer, évitant ainsi que le modèle ne mémorise le bruit des données.
- **La validation croisée (cross-validation)** permet une estimation robuste et fiable des performances du modèle en divisant les données en plusieurs sous-ensembles utilisés alternativement pour l'entraînement et la validation.
- **La diversification des données** permet de réduire le biais de sélection. Un ensemble d'entraînement varié et représentatif améliore significativement la capacité du modèle à généraliser à des situations nouvelles.

6.6 Questions

1. **Qu'est-ce que la généralisation en deep learning ?**
 - A. L'aptitude d'un modèle à encoder efficacement les données d'entraînement
 - B. La capacité d'un modèle à performer de manière fiable sur des données inédites
 - C. Le processus de simplification d'un modèle complexe
 - D. La technique pour accélérer l'apprentissage des réseaux de neurones
2. **Le surajustement (overfitting) se caractérise par :**
 - A. Une performance médiocre à la fois sur les données d'entraînement et de test
 - B. Une performance égale sur les données d'entraînement et de test
 - C. Une excellente performance sur les données d'entraînement mais médiocre sur les données de test
 - D. Une performance meilleure sur les données de test que sur les données d'entraînement
3. **Quel est le principal problème causé par le sous-ajustement (underfitting) ?**
 - A. Le modèle mémorise les données d'entraînement
 - B. Le modèle est trop complexe pour les données disponibles

- C. Le modèle est trop simple pour capturer la complexité des données souvent car il a reçu un entraînement qui a été arrêté de manière précoce
 - D. Le modèle devient trop spécifique aux exemples d'entraînement
4. L'**early stopping** est une technique qui consiste à :
- A. Arrêter l'entraînement avant que le modèle ne converge
 - B. Réduire le nombre de paramètres du modèle
 - C. Arrêter l'entraînement lorsque l'erreur sur l'ensemble de validation commence à augmenter
 - D. Éliminer les exemples d'entraînement difficiles
5. Un signe typique de surajustement lors de l'entraînement d'un modèle est :
- A. L'erreur d'entraînement et l'erreur de validation diminuent ensemble
 - B. L'erreur d'entraînement et l'erreur de validation augmentent ensemble
 - C. L'erreur d'entraînement continue de diminuer tandis que l'erreur de validation commence à augmenter
 - D. L'erreur d'entraînement stagne tandis que l'erreur de validation continue de diminuer
6. Dans la validation croisée à K-fold, quelle est la méthode employée ?
- A. Diviser les données en K ensembles, utiliser K-1 ensembles pour l'entraînement et 1 pour la validation, puis répéter en changeant l'ensemble de validation
 - B. Sélectionner aléatoirement K exemples pour la validation
 - C. Entraîner K modèles différents sur les mêmes données
 - D. Tester le modèle K fois sur le même ensemble de validation
7. Le biais de sélection en deep learning fait référence à :
- A. La préférence du modèle pour certaines classes de sortie
 - B. La discordance entre les données d'entraînement et les données réelles
 - C. L'erreur due aux hypothèses simplificatrices du modèle
 - D. La tendance à sélectionner des modèles trop complexes
8. Dans le compromis biais-variance, un modèle avec un biais élevé et une faible variance est généralement :
- A. Un modèle complexe qui mémorise les données d'entraînement
 - B. Un modèle simple qui ne capture pas toute la complexité des données
 - C. Un modèle parfaitement équilibré
 - D. Un modèle qui généralise bien à de nouvelles données
9. Dans la décomposition de l'erreur quadratique moyenne, l'erreur irréductible représente :
- A. L'erreur due aux hypothèses simplificatrices du modèle
 - B. L'erreur due à la variabilité des prédictions pour différents jeux de données
 - C. L'erreur due au bruit inhérent aux données, impossible à éliminer
 - D. L'erreur due au manque de données d'entraînement

6.7 Réponses

1. **Qu'est-ce que la généralisation en deep learning ?**

Réponse: B - La capacité d'un modèle à performer de manière fiable sur des données inédites

2. **Le surajustement (overfitting) se caractérise par :**

Réponse: C - Une excellente performance sur les données d'entraînement mais médiocre sur les données de test

3. **Quel est le principal problème causé par le sous-ajustement (underfitting) ?**

Réponse: C - Le modèle est trop simple pour capturer la complexité des données souvent car il a reçu un entraînement qui a été arrêté de manière précoce

4. **L'early stopping est une technique qui consiste à :**

Réponse: C - Arrêter l'entraînement lorsque l'erreur sur l'ensemble de validation commence à augmenter

5. **Un signe typique de surajustement lors de l'entraînement d'un modèle est :**

Réponse: C - L'erreur d'entraînement continue de diminuer tandis que l'erreur de validation commence à augmenter

6. **Dans la validation croisée à K-fold, quelle est la méthode employée ?**

Réponse: A - Diviser les données en K ensembles, utiliser $K-1$ ensembles pour l'entraînement et 1 pour la validation, puis répéter en changeant l'ensemble de validation

7. **Le biais de sélection en deep learning fait référence à :**

Réponse: B - La discordance entre les données d'entraînement et les données réelles

8. **Dans le compromis biais-variance, un modèle avec un biais élevé et une faible variance est généralement :**

Réponse: B - Un modèle simple qui ne capture pas toute la complexité des données

9. **Dans la décomposition de l'erreur quadratique moyenne, l'erreur irréductible représente :**

Réponse: C - L'erreur due au bruit inhérent aux données, impossible à éliminer

Bibliographie

- [1] H. Touvron *et al.*, « LLaMA: Open and Efficient Foundation Language Models », *arXiv preprint arXiv:2302.13971*, févr. 2023, [En ligne]. Disponible sur: <https://arxiv.org/pdf/2302.13971.pdf>
- [2] A. Bodin et F. Recher, *Deep Math Mathématiques (simples) des réseaux de neurones (pas trop compliqués)*. Exo7, 2021. [En ligne]. Disponible sur: <https://exo7math.github.io/deepmath-exo7/>
- [3] H. Li, Z. Xu, G. Taylor, C. Studer, et T. Goldstein, « Visualizing the Loss Landscape of Neural Nets », *arXiv preprint arXiv:1712.09913*, 2018.
- [4] D. E. Rumelhart, G. E. Hinton, et R. J. Williams, « Learning representations by back-propagating errors », *Nature*, vol. 323, n° 6088, p. 533-536, 1986, [En ligne]. Disponible sur: <http://www.cs.toronto.edu/~hinton/absps/naturebp.pdf>
- [5] X. Glorot et Y. Bengio, « Understanding the difficulty of training deep feedforward neural networks », in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, p. 249-256. [En ligne]. Disponible sur: <http://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf>
- [6] D.-H. Lee, S. Zhang, A. Fischer, et Y. Bengio, « Difference Target Propagation », in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2015, p. 498-515. doi: 10.48550/arXiv.1412.7525.
- [7] W.-D. K. Ma, J. Lewis, et W. B. Kleijn, « The HSIC Bottleneck: Deep Learning without Back-Propagation », *arXiv preprint arXiv:1908.01580*, 2019, doi: 10.48550/arXiv.1908.01580.
- [8] A. Choromanska *et al.*, « Beyond Backprop: Online Alternating Minimization with Auxiliary Variables », *arXiv preprint arXiv:1806.09077*, 2018, doi: 10.48550/arXiv.1806.09077.
- [9] M. Jaderberg *et al.*, « Decoupled Neural Interfaces using Synthetic Gradients », *arXiv preprint arXiv:1608.05343*, 2016, doi: 10.48550/arXiv.1608.05343.
- [10] G. Hinton, « The Forward-Forward Algorithm: Some Preliminary Investigations », *arXiv preprint arXiv:2212.13345*, 2022, doi: 10.48550/arXiv.2212.13345.
- [11] Y. Bengio, D.-H. Lee, J. Bornschein, T. Mesnard, et Z. Lin, « Towards Biologically Plausible Deep Learning », *arXiv preprint arXiv:1502.04156*, 2015, doi: 10.48550/arXiv.1502.04156.
- [12] A. Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier, et A. Maida, « Deep learning in spiking neural networks », *Neural Networks*, vol. 111, p. 47-63, mars 2019, doi: 10.1016/j.neunet.2018.12.002.