Etude de propriétés d'apprentissage supervisé et non supervisé par des méthodes de Physique Statistique

Arnaud Buhot

Thèse

présentée par

Arnaud BUHOT

pour obtenir le titre de

Docteur

de l'UNIVERSITÉ GRENOBLE I - JOSEPH FOURIER

Arrêtés ministériels
du 5 juillet 1984 et du 30 mars 1992

Discipline : Physique Théorique

Étude de propriétés d'apprentissage supervisé et non supervisé par des méthodes de Physique Statistique

soutenue le 17 mai 1999 devant la Commission d'Examen :

JURY

MM.	**Bertrand FOURCADE**	**Président**
	Jean-Pierre NADAL	**Rapporteur**
	David SHERRINGTON	**Rapporteur**
Mme	**Elisabeth DUBOIS-VIOLETTE**	**Examinatrice**
M.	**Chris VAN DEN BROECK**	**Examinateur**
Mme	**Mirta B. GORDON**	**Directrice de Thèse**

Thèse préparée au sein du CEA Grenoble
Département de Recherche Fondamentale sur la Matière Condensée
Service de Physique Statistique, Magnétisme et Supraconductivité

À mes parents

Remerciements

Cette thèse s'est déroulée dans le Groupe Théorie du Service de Physique Statistique, Magnétisme et Supraconductivité au Département de Recherche Fondamentale sur la Matière Condensée au CEA Grenoble. Je tiens à remercier les directeurs du Service Jacques Flouquet et du Département Daniel Beysens pour les moyens qui ont été mis à ma disposition mais aussi pour l'excellent niveau scientifique de ce haut-lieu de la Physique.

Cette thèse n'aurait pu voir le jour sans ma directrice de thèse Mirta Gordon. Je tiens ici à la remercier pour ses qualités de recherche, d'écoute, de disponibilité mais aussi pour sa gentillesse. Il est connu que le responsable de thèse a autant d'importance que le sujet abordé. J'ai pu m'en rendre compte au cours de cette thèse grâce aux discussions journalières que nous avons pu avoir. Ces discussions parfois houleuses se sont toujours avérées instructives.

Je tiens à remercier le professeur David Sherrington de m'avoir fait l'honneur d'être rapporteur de cette thèse. Je suis d'ailleurs très enthousiaste à l'idée de pouvoir poursuivre mon apprentissage de la recherche dans son laboratoire.

Mes remerciements vont aussi à Jean-Pierre Nadal qui non seulement a accepté d'être rapporteur de cette thèse mais qui m'a aussi beaucoup apporté au cours des nombreuses discussions que nous avons pu avoir particulièrement lors de mon année passée à Paris.

Je suis aussi très reconnaissant à Elisabeth Dubois-Violette et à Chris Van den Broeck d'avoir participé à mon jury de thèse. Je remercie également Bertrand Fourcade qui non seulement a accepté d'être membre du jury mais qui a pris en charge la présidence de ce jury.

Mes plus sincères remerciements s'adressent aux chercheurs du SPSMS et plus particulièrement à Mireille Lavagna, à Daniel Grempel et à Jacques Villain. Grâce aux nombreuses discussions que nous avons pu avoir ma culture générale en Physique s'en est trouvée grandement élargie.

J'associe à ces personnes les membres du Laboratoire de Physique Statistique de l'Ecole Normale Supérieure à Paris où j'ai eu la chance de passer une année de Scientifique du Contingent. Cette année d'interruption de la thèse m'a permis de découvrir un autre laboratoire, d'autre sujets de recherche mais surtout de collaborer avec Werner Krauth; qu'il en soit ici chaleureusement remercié.

Je tiens aussi à exprimer ma gratitude aux personnes du groupe informatique du DRFMC ainsi qu'à celles du Centre Grenoblois de Calcul Vectoriel et plus particulièrement Laurent Colombet qui m'ont facilité l'utilisation des moyens informatiques. Mes remerciements vont également à Michèle Dominiak et à Marielle Perrier pour leur aide lors des tâches administratives.

Mes pensées vont à toutes les personnes qui m'ont aidé à obtenir ce titre de Docteur. Je pense notamment aux professeurs qui m'ont donné le goût de la Physique et

de la recherche. Je pense aussi aux amis sans qui la vie ne serait rien. Voici quelques prénoms, les personnes concernées se reconnaîtront, et que ceux qui auraient été oubliés ne se formalisent pas: Christian, Jean-Luc, Sylvie, Sophie, Sylvain, Yannick, Audrey, Jérôme, Marina, Yves, Flora, Stéphane...

Enfin, je remercie toute ma famille pour son soutien et plus particulièrement mes parents pour m'avoir donné toute la liberté et tous les moyens nécessaires pour réaliser mes études.

Grenoble, le 20 Mai 1999.

Table des matières

Remerciements **5**

Introduction générale **9**

I Cas uniaxial : apprentissage optimal **15**

Introduction **17**

1 Apprentissage optimal **19**
1.1 Présentation du problème 19
1.2 Fonction de coût 20
1.3 Qualité de l'apprentissage 21
1.4 Intérêt de la Physique Statistique 22
1.4.1 Méthode des répliques 24
1.4.2 Hypothèse de symétrie des répliques 27
1.4.3 Interprétation des paramètres d'ordre 31
1.4.4 Limite de température nulle 31
1.5 Détermination du potentiel optimal 32
1.6 Cohérence de l'hypothèse de symétrie 35

2 Apprentissage Supervisé **37**
2.1 Présentation du problème 37
2.2 Notions d'erreur de classification 39
2.3 Reformulation du problème supervisé 40
2.4 Erreur de généralisation minimale 42
2.5 Caractéristiques du potentiel optimal 44
2.6 Distribution des stabilités 45
2.7 Simulations numériques 48
2.7.1 Description de l'algorithme utilisé 48
2.7.2 Erreur de généralisation de l'élève optimal 50
2.7.3 Distribution des stabilités 53

3 Apprentissage non supervisé **57**
3.1 Présentation du problème 57
3.2 Courbes d'apprentissage optimal 59
3.3 Approche bayesienne 65
3.3.1 Apprentissage de Gibbs 65
3.3.2 Performance de l'estimateur bayesien 66
3.4 Controverse 69
3.4.1 Résultats analytiques 70
3.4.2 Résultats numériques 71

3.4.3 Simulations numériques 72
3.4.4 Conclusions actuelles sur la controverse 75
3.5 Interprétation des 3 phases d'apprentissage 75
3.5.1 Apprentissage de la composante principale 76
3.5.2 Scénario supervisé 76
3.5.3 Discussion des phases de l'apprentissage optimal 77
3.5.4 Comportement du potentiel optimal 79

II Au-delà du cas uniaxial : étude de deux approches 83

Introduction 85

1 Machine à exemples supports 87
1.1 Présentation du problème 87
1.2 Calcul des propriétés d'apprentissage 89
1.2.1 Méthode des répliques 90
1.2.2 Tâche linéairement séparable 91
1.2.3 Tâche aléatoire 94
1.3 Discussion des propriétés d'apprentissage 94
1.3.1 Tâche linéairement séparable 94
1.3.2 Tâche aléatoire 98
1.3.3 Remarques générales 100
1.4 Erreur de reconnaissance 101
1.5 Tentative de borne pour l'erreur de généralisation 103
1.5.1 Présentation de la borne 103
1.5.2 Calcul de la borne pour les tâches étudiées 104
1.5.3 Ouvertures possibles 105

2 Machine de parité : Algorithme incrémental 107
2.1 Présentation de l'algorithme incrémental 108
2.2 Propriétés de l'algorithme incrémental 110
2.2.1 Approche par la Physique Statistique 110
2.2.2 Conditions de convergence 111
2.3 Détermination du nombre de perceptrons 112
2.4 Erreur d'apprentissage d'un perceptron simple 115
2.5 Règles d'apprentissage particulières 116
2.5.1 Minimisation de l'erreur d'apprentissage 117
2.5.2 Fonction de coût quadratique 121
2.6 Conclusion 124

Conclusion générale 127

Introduction générale

Le titre de cette thèse comporte des termes du langage commun comme *apprentissage* ou *supervisé*. Ceux-ci sont employés, au cours de la thèse, avec un sens précis et souvent réducteur par rapport à leur définition générale. Je vais profiter de cette introduction pour préciser les termes utilisés par la suite.

Par exemple, il est possible d'envisager plusieurs techniques d'apprentissage différentes. L'une d'elles, couramment utilisée dans l'enseignement scolaire, consiste à donner la règle générale (règle de grammaire, loi de la Physique) ainsi que les exceptions possibles à cette règle. L'élève doit alors apprendre la règle pour pouvoir ensuite l'appliquer à des situations particulières. Une deuxième technique possible consiste à présenter des exemples particuliers en nombre suffisant pour pouvoir en extraire la règle générale. Cette technique se rapproche de la démarche du chercheur qui, à partir des expériences, essaie d'extraire une loi générale, et aussi, de celle de l'enfant apprenant à parler. Cette technique d'apprentissage est celle que nous allons étudier dans cette thèse. Elle est généralement appelée *apprentissage par l'exemple*. L'apprentissage s'effectue à l'aide d'un *ensemble d'apprentissage*, constitué d'*exemples*, duquel on essaie d'extraire des règles générales.

Apprentissage supervisé et non supervisé

Nous allons considérer, au cours de la thèse, deux types différents d'apprentissage : *supervisé* et *non supervisé*. Dans chacun des deux cas, les exemples sont constitués de points d'un espace de *données*. Cet espace est généralement de grande dimension. Par exemple, dans les problèmes de la reconnaissance de forme [112], la dimension de l'espace des données correspond au nombre de pixels de chaque image. Chaque coordonnée d'un point associé à une image représente le niveau de gris du pixel correspondant. Pour le problème du diagnostic médical [115], chaque coordonnée de l'espace des données correspond à une donnée médicale comme le poids, la taille du patient ou des données plus spécifiques. La différence essentielle entre l'apprentissage supervisé et l'apprentissage non supervisé est que dans le premier, à chaque point est associé une classe. De manière imagée, on dit qu'un *professeur* associe une classe à chacun des points de l'espace des données.

Le but de l'apprentissage supervisé consiste à trouver un *élève* capable de classer correctement, c'est-à-dire, de manière identique au professeur, les exemples de l'ensemble d'apprentissage mais aussi, si possible, des exemples nouveaux ne faisant pas partie de cet ensemble. Cette dernière propriété correspond à la généralisation à partir d'exemples, décrite précédemment.

Dans le cas de l'apprentissage non supervisé, l'ensemble d'apprentissage est uniquement constitué d'un ensemble de points d'un espace de grande dimension. Une des possibilités d'apprentissage consiste alors à déterminer des propriétés générales de la densité de probabilité des exemples qui a permis d'obtenir l'ensemble d'apprentissage. Ces propriétés sont, par exemple, la détermination des composantes principales, la détermination du nombre d'amas de la densité de probabilité, leur

localisation, etc.

La question qui se pose alors est la suivante : qu'elle est l'approche à utiliser pour déterminer les caractéristiques du professeur ou les propriétés de la densité de probabilité des exemples, ou encore, quelle approche utiliser pour l'apprentissage ? Une des réponses possibles à cette question a été apportée par l'étude des *réseaux de neurones*.

Les réseaux de neurones

La théorie de l'apprentissage est fortement influencée par la modélisation du cerveau. En effet, sous certains aspects, le cerveau peut être considéré comme une *machine* capable d'apprendre des tâches souvent très complexes. Il constitue une source d'inspiration pour la compréhension de l'apprentissage.

En 1943, McCulloch et Pitts [75] ont modélisé les briques élémentaires du cerveau, les *neurones*, par des variables binaires $\sigma_i = \pm 1$. La valeur $+1$ correspond à un état excité du neurone alors que la valeur -1 correspond à l'état au repos. Les *synapses* reliant les neurones entre eux sont représentées par des couplages J_{ij} entre ces neurones. Un couplage positif correspond à un caractère excitateur du neurone i vers le neurone j et un couplage négatif à un caractère inhibiteur. Le couplage est d'autant plus efficace que la valeur absolue de J_{ij} est importante. La dynamique d'un tel *réseau de neurones* s'exprime de manière simple. Le neurone i calcule la somme pondérée des activités synaptiques des autres neurones à l'instant t :

$$h_i = \sum_{j \neq i} \sigma_j J_{ji}.$$

L'état du neurone i à l'instant $t + \delta t$ est excité si le *champ* h_i est supérieur à un *seuil* θ_i, il reste inactif sinon. D'un point de vue neurobiologique, cette modélisation est caricaturale. Toutefois, elle constitue un point de départ intéressant pour une étude théorique de l'apprentissage [59, 5].

En 1961, Rosenblatt [104, 105] a proposé un réseau de neurones particulier, appelé le *perceptron*. Celui-ci est constitué de N neurones d'*entrée* reliés à un unique neurone de *sortie* par des couplages J_i avec $i = 1, \cdots, N$. Cette architecture en arbre ne possède plus réellement de dynamique. En effet, à chaque configuration $\boldsymbol{\xi} = \{\xi_1, \cdots, \xi_N\}$ des neurones d'entrée, qui peut être constituée de variables ξ_i réelles ou binaires, le réseau associe de manière déterministe, une classe σ correspondant à l'état du neurone de sortie grâce à la règle suivante :

$$\sigma = \operatorname{sign}\left(\sum_{i=1}^{N} J_i \xi_i - \theta \right),$$

où θ est le seuil du neurone de sortie. La sortie du perceptron peut être généralisé aisément à des valeurs réelles en remplaçant la fonction signe par une fonction d'activation g continue.

Le perceptron permet ainsi de classer des points $\boldsymbol{\xi} = \{\xi_1, \cdots, \xi_N\}$ d'un espace de *données* de dimension N. La classification faite par chaque perceptron dépend des valeurs des couplages J_i et du seuil θ. Elle sépare en deux l'espace des données avec un hyperplan normal au vecteur $\mathbf{J} = \{J_1, \cdots, J_N\}$, placé à une distance θ de l'origine. Les points d'un côté de l'hyperplan sont classés $+1$ tandis que les autres sont classés -1.

Apprentissage

Chaque perceptron simple, défini par l'ensemble de ses couplages J_i et de son seuil θ, constitue un élève possible pour effectuer la tâche définie par le professeur. Le fait d'adapter les couplages et le seuil afin de classer les exemples de l'ensemble d'apprentissage de la même façon que le professeur constitue la phase d'apprentissage du réseau de neurones. Une des possibilités pour adapter les paramètres du perceptron simple est de minimiser le nombre d'erreurs de classification de l'élève pour les exemples de l'ensemble d'apprentissage. Cette approche semble réservée à l'apprentissage supervisé, mais en fait, il n'en est rien. En effet, il est possible de remplacer le nombre d'erreurs par une *fonction de coût* qui dépend à la fois des paramètres du perceptron et des exemples de l'ensemble d'apprentissage. La fonction doit être adaptée à la quantité que l'on veut apprendre. Dans le cas de l'apprentissage non supervisé, les couplages J_i pourront, par exemple, représenter une direction d'anisotropie de la densité de probabilité des exemples.

Pour des tâches très complexes, il est possible d'utiliser plusieurs perceptrons simples sous forme de réseaux en couches [58], ce qui permet d'augmenter le nombre de paramètres à déterminer, augmentant ainsi la complexité des tâches susceptibles d'être apprises.

La phase d'apprentissage consiste à obtenir les paramètres du perceptron simple qui minimisent la fonction de coût. Ce n'est pas le problème algorithmique de cette phase d'apprentissage qui va nous intéresser au cours de cette thèse, mais plutôt l'étude des *propriétés* de l'apprentissage. Plus spécifiquement, nous étudierons le comportement des propriétés du perceptron obtenu après la phase d'apprentissage en fonction du nombre d'exemples de l'ensemble d'apprentissage, de la fonction de coût utilisée, etc.

Une approche très fructueuse pour cette étude consiste à utiliser les méthodes de la Physique Statistique [3, 14, 42, 52, 58, 93, 126]. L'idée est de considérer la fonction de coût comme une énergie. Les propriétés du minimum de cette fonction de coût seront alors obtenues en considérant le système physique correspondant à température nulle (température pour laquelle l'énergie d'un système physique est minimale). Toutefois, l'existence de paramètres *gelés*, représentés par l'ensemble d'apprentissage, nécessite l'introduction des méthodes utilisées pour les systèmes désordonnés comme les verres de spins. En particulier, nous utiliserons de manière intensive la méthode des répliques [81].

Plan de la thèse

Nous nous sommes intéressés, au cours de cette thèse, à l'étude des propriétés d'apprentissage supervisé et non supervisé.

Dans la première partie, nous avons essayé de déterminer les performances optimales ainsi qu'une procédure, un algorithme, permettant d'obtenir ces performances optimales. Nous avons pu résoudre deux problèmes différents en utilisant la même approche pour les deux : une approche variationnelle, permettant de déterminer le potentiel *optimal*. Les paramètres minimisant la fonction de coût correspondante possèdent des performances de généralisation optimales. Cette méthode, développée indépendamment par nous-mêmes et par Kinouchi et Caticha [62] pour le problème d'apprentissage supervisé, a été généralisé par Reimann et Van den Broeck [101, 120]. Nous présentons dans le premier chapitre, cette approche généralisée utilisant la méthode des répliques dans la limite thermodynamique.

Dans le deuxième chapitre, nous développons les résultats que nous avons obtenus dans le cas où la tâche à apprendre est linéairement séparable, c'est-à-dire, pour l'apprentissage des couplages d'un perceptron simple, lorsque le professeur est

lui-même un perceptron simple. Les résultats analytiques permettent de déduire un algorithme simple permettant de trouver l'élève avec les performances optimales. Celui-ci est obtenu par minimisation de la fonction de coût optimale, déterminée par notre approche variationnelle. Ces résultats étant obtenus dans la limite thermodynamique, où la dimension de l'espace des données est infinie, avec certaines approximations, nous avons effectué des simulations numériques pour essayer de confirmer les résultats théoriques et déterminer les effets de taille finie. Ces effets seront discutés.

Le dernier chapitre de cette première partie est dédié à un problème d'apprentissage non supervisé : la détection des deux amas d'une densité de probabilité de points à partir d'un ensemble d'exemples distribués selon cette densité. Ce problème a été très étudié par le passé en raison des nombreuses applications possibles (reconnaissance de formes...). Nous avons élargi la recherche des performances optimales au cas où l'écart et la largeur des amas sont quelconques. Les résultats obtenus en fonction de ces paramètres mettent en évidence, pour la première fois pour ce type d'apprentissage, l'existence de transitions de phases du premier ordre dans les performances optimales en fonction de la taille de l'ensemble d'apprentissage. Nos résultats ont soulevé une controverse, car ils suggèrent que la branche métastable de haute performance serait accessible, en contradiction avec les résultats d'une autre approche de l'apprentissage optimal, l'approche bayesienne. La conclusion de cette controverse n'a pas encore pu être donnée mais des pistes pour expliquer les raisons de cette contradiction seront apportées. La fin du chapitre est dédiée à la discussion des différentes phases d'apprentissage. En fonction de la taille de l'ensemble d'apprentissage, ces phases sont soit interprétées en termes de l'apprentissage d'une composante principale ou soit reliées à l'apprentissage d'une formulation supervisée du problème. Nous discutons aussi de la forme du potentiel optimal en fonction des phases d'apprentissage ainsi que de son évolution en fonction de la taille de l'ensemble d'apprentissage.

La deuxième partie de la thèse aborde le problème de l'apprentissage de tâches complexes, nécessitant des réseaux de neurones plus élaborés (ou un nombre de paramètres plus important) que lors de la première partie de la thèse. Nous nous sommes limités, dans cette deuxième partie, à l'étude de tâches supervisées.

Le premier chapitre est dédié à l'étude d'une nouvelle approche, celle des machines à exemples supports [122, 123], par des méthodes de Physique Statistique. Contrairement aux méthodes qui complexifient les réseaux élèves, cette approche propose d'étendre l'espace des données en un espace des représentations. Ceci permet de réduire la phase d'apprentissage à la détermination d'un perceptron simple mais dans ce nouvel espace. L'extension de l'espace des données augmente le nombre de paramètres à déterminer lors de l'apprentissage et permet d'envisager l'apprentissage de tâches complexes. Nous avons utilisé la méthode des répliques pour déterminer les performances d'apprentissage d'une famille de transformations particulières de l'espace des données pour deux tâches extrêmes : la première est l'apprentissage d'un problème linéairement séparable dans l'espace des données et la deuxième une tâche où les classes des exemples sont choisies aléatoirement. Les performances pour ces deux tâches et pour cette famille de transformations nous ont permis de déduire des remarques générales sur le comportement des machines à exemples supports. Nous avons, de plus, introduit une notion d'erreur de reconnaissance, dans le but de borner l'erreur de généralisation et de comprendre les raisons des bonnes performances observées lors des applications réalistes.

Dans le deuxième chapitre, nous présentons l'étude d'un algorithme incrémental [13] qui permet de créer un réseau de neurones à une couche cachée. Cet algorithme construit itérativement des représentations internes binaires des exemples. Ces représentations internes correspondent à des classes, associées aux exemples par des perceptrons simples. Leurs couplages sont appris successivement au cours

de la phase d'apprentissage, afin de corriger les erreurs de classification des perceptrons précédents et cela jusqu'à ce que les exemples de l'ensemble d'apprentissage soient tous correctement classés. La classe d'un exemple est le produit des variables binaires correspondant à sa représentation interne. Une machine effectuant ainsi la classification est appelée machine de parité. Nous nous sommes intéressés à la détermination du nombre de perceptrons simples nécessaires à la convergence de l'algorithme incrémental en fonction de la taille de l'ensemble d'apprentissage. Les résultats obtenus nous permettent d'en déduire la capacité de la machine de parité pour cet algorithme particulier, la capacité étant le nombre maximal d'exemples de classes aléatoires qu'il est possible d'apprendre avec une architecture (nombre de perceptrons simples) donnée. Nous obtenons le résultat surprenant que, pourvu que les représentations internes soient apprises avec un algorithme efficace, cette capacité est proche de la capacité de la machine de parité, obtenue indépendamment de l'algorithme utilisé pour l'apprentissage.

Les questions soulevées par certains des résultats présentés, au cours de cette thèse, ouvrent des perspectives de recherche qui seront discutées dans la conclusion générale.

Première partie

Cas uniaxial : apprentissage optimal

Introduction

Cette première partie est consacrée à l'étude de l'apprentissage d'une direction privilégiée dans un espace de très grande dimension. L'information nécessaire à un tel apprentissage est donnée par l'intermédiaire d'un ensemble d'apprentissage constitué d'exemples. La plupart des algorithmes d'apprentissage permettant d'obtenir une direction proche de la direction privilégiée sont basés sur la minimisation d'une fonction de coût qui dépend de l'ensemble d'apprentissage.

Dans cette partie, nous nous intéressons, non seulement, au calcul des performances de tels algorithmes, mais aussi à la recherche de la fonction de coût pour laquelle les performances sont optimales.

Dans le premier Chapitre, après une présentation générale du problème due à Reimann et Van den Broeck [101, 120], les performances optimales d'un tel apprentissage sont obtenues à l'aide des outils de la Physique Statistique et plus particulièrement de la méthode des répliques.

Deux applications du problème général sont étudiées dans les Chapitres 2 et 3. La première application concerne la classification de données en deux classes distinctes dans le cas particulier où la séparation de ces deux classes peut s'effectuer par un hyperplan séparateur. La direction normale à l'hyperplan est la direction privilégiée que l'on cherche à déterminer. Les exemples sont constitués par des points de l'espace des données, distribués aléatoirement, et de leurs classes respectives. Les propriétés optimales obtenues par les méthodes de la Physique Statistique sont confirmées par des simulations numériques.

La deuxième application consiste à détecter une structure en deux amas d'un ensemble de points dans un espace de grande dimension. La direction privilégiée, dans ce cas, correspond à la direction qui relie les centres des deux amas. Suivant l'écart et la largeur des amas, différentes phases d'apprentissage optimal sont observées et sont analysées. Ces différentes phases sont, généralement, séparées par des transitions du premier ou du second ordre. Dans ce problème, nous avons suscité une controverse entre deux approches différentes de l'apprentissage optimal, dont les prédictions ne coïncident pas pour certaines valeurs des paramètres du problème. Cette controverse qui n'a pas encore pu être résolue, sera discutée.

Chapitre 1

Apprentissage optimal

Cette partie du mémoire est dédiée à l'étude de l'apprentissage optimal d'une direction privilégiée à partir d'exemples, dans un espace de très grande dimension. J'ai choisi de présenter au Chapitre 1 le calcul des propriétés de l'apprentissage optimal, dans le cas le plus général, afin de présenter ensuite de manière simple les résultats que nous avons obtenus pour deux problèmes d'apprentissage classiques, qui seront développés dans les deux Chapitres suivants : un cas d'apprentissage supervisé (Chap.2) et un autre d'apprentissage non supervisé (Chap.3). Bien que la formulation générale ait été présentée par Reimann et Van den Broeck [101, 120], le cas particulier de l'apprentissage supervisé avait été résolu précédemment par nous-mêmes [21, 25, 26], ainsi que par Kinouchi et Caticha [61, 62], de manière indépendante. Nous avons démontré, de plus, que l'hypothèse de symétrie des répliques utilisée pour déterminer les propriétés optimales est cohérente dans le cas général [22].

1.1 Présentation du problème

Comme nous l'avons vu dans l'introduction, les exemples sont supposés être des points d'un espace des données de dimension N, échantillonnés selon une densité de probabilité inconnue dont on cherche à déterminer les propriétés. On dispose pour cela d'un nombre P d'exemples $\boldsymbol{\xi}^\mu$ qui forment ce que l'on appelle généralement l'*ensemble d'apprentissage* :

$$\mathcal{L}_\alpha = \{\boldsymbol{\xi}^\mu\}_{\mu=1,\cdots,P}\,. \tag{1.1}$$

L'indice α reflète la taille réduite $\alpha \equiv P/N$ de l'ensemble d'apprentissage. Cet ensemble contient l'information dont on dispose pour déterminer les propriétés de la densité de probabilité des exemples. Par la suite, le terme densité de probabilité sera souvent remplacé par distribution.

La quantité d'information apportée par les exemples de l'ensemble d'apprentissage étant limitée, il est nécessaire de faire certaines hypothèses sur leur distribution afin de pouvoir apprendre à partir des exemples.

Dans cette partie, nous supposons premièrement que la distribution des exemples possède une symétrie axiale : elle est invariante par rapport à une rotation autour d'un axe. Cet axe, passant par l'origine, est caractérisé par sa direction $\mathbf{B}$. Par la suite, on supposera $\mathbf{B}$ normé : $\mathbf{B}\cdot\mathbf{B} = 1$. Cette direction *privilégiée* est supposée inconnue.

Nous supposons connue la distribution des exemples suivant les directions orthogonales à $\mathbf{B}$. Dans toute cette partie, nous allons considérer la plus simple des distri-

butions, la *gaussienne normale*. Cette distribution est une gaussienne de moyenne nulle et de variance unité.

Enfin, suivant la direction $\mathbf{B}$, la distribution est supposée être perturbée par rapport à la gaussienne normale. Il est alors possible d'écrire la distribution des exemples de la manière générale suivante :

$$P(\boldsymbol{\xi}|\mathbf{B}) = \frac{1}{(2\pi)^{N/2}} \exp\left(-\frac{\boldsymbol{\xi}\cdot\boldsymbol{\xi}}{2} - V^*\left(\boldsymbol{\xi}\cdot\mathbf{B}\right)\right) \tag{1.2}$$

où $\boldsymbol{\xi}$ représente un point de l'espace des données. La fonction V^* caractérise la *perturbation* par rapport à la gaussienne normale dans la direction $\mathbf{B}$. La distribution suivant cette direction privilégiée est donnée par :

$$P\left(\lambda \equiv \boldsymbol{\xi}\cdot\mathbf{B}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\lambda^2}{2} - V^*\left(\lambda\right)\right). \tag{1.3}$$

La condition de normalisation de la distribution des exemples impose la contrainte suivante sur la perturbation V^* :

$$1 = \int D\lambda \, \exp\left(-V^*\left(\lambda\right)\right) \tag{1.4}$$

avec $D\lambda$ la mesure gaussienne définie par :

$$D\lambda \equiv \frac{d\lambda}{\sqrt{2\pi}} \exp\left(-\frac{\lambda^2}{2}\right). \tag{1.5}$$

Le domaine d'intégration de cette intégrale et de toutes les intégrales qui apparaîtront par la suite sera l'espace entier sauf indication contraire. Dans le cas présent, l'intégrale s'étend donc de $-\infty$ à $+\infty$.

Dans la pratique, la direction privilégiée $\mathbf{B}$ ainsi que la fonction V^* sont *a priori* inconnues. Pour cette étude, nous supposons que la forme complète de cette fonction V^* nous est connue. La forme de la distribution des exemples étant supposée connue, il ne reste plus alors que la direction privilégiée à déterminer. Ce problème peut paraître simpliste. Toutefois, de nombreuses applications réalistes peuvent s'y ramener comme nous le verrons dans les deux Chapitres suivants. De plus, les solutions à ce problème montrent une telle richesse qu'il est intéressant de l'étudier et de bien le comprendre avant de passer à des problèmes plus complexes.

1.2 Fonction de coût

Le but de l'apprentissage est de trouver la direction privilégiée $\mathbf{B}$ inconnue. Pour déterminer cette direction, on dispose de l'ensemble d'apprentissage dont la taille P finie ne permet pas la détermination exacte de $\mathbf{B}$. L'apprentissage consiste alors à déterminer une direction qui soit la plus proche possible de la direction privilégiée. Afin de déterminer cette direction, il est possible de minimiser une *fonction de coût* qui dépend de l'ensemble d'apprentissage $\mathcal{L}_\alpha$ et d'une direction $\mathbf{J}$ par rapport à laquelle on effectue la minimisation.

Nous allons maintenant discuter du choix de la fonction de coût. Les exemples de l'ensemble d'apprentissage étant supposés distribués indépendamment les uns des autres, il est possible de se restreindre à une fonction de coût additive :

$$E(\mathbf{J};\mathcal{L}_\alpha) = \sum_{\mu=1}^{P} V(\mathbf{J}\cdot\boldsymbol{\xi}^\mu) \tag{1.6}$$

où la fonction V est appelée *potentiel*. Ce potentiel dépend de la direction $\mathbf{J}$ et de l'exemple $\boldsymbol{\xi}^\mu$ uniquement par l'intermédiaire de leur produit scalaire. Cette restriction supplémentaire dans le choix de la fonction de coût ne limite pas les possibilités d'apprentissage. En effet, ce produit scalaire est le seul scalaire pertinent que l'on puisse construire avec la direction $\mathbf{J}$ et l'exemple $\boldsymbol{\xi}^\mu$. Les deux autres scalaires que l'on puisse construire sont les normes des vecteurs. Or, la direction $\mathbf{J}$ est normée à un et la norme des exemples est indépendante de l'exemple considéré dans la limite qui nous intéresse par la suite (taille de l'espace des données infinie). Ces normes de vecteurs n'apportent donc aucune information pertinente pour l'apprentissage.

La fonction de coût doit être choisie de telle façon que la direction $\mathbf{J}^*$ qui la minimise soit aussi proche que possible de la direction $\mathbf{B}$. Dans la littérature, il y a un grand nombre de fonction de coût *ad hoc*, proposée pour l'apprentissage de tâches variées [48, 102, 126]. Dans ce Chapitre, nous montrerons que si la tâche est bien définie, c'est-à-dire, que la perturbation V^* est connue, il est alors possible de déterminer le potentiel optimal. Le minimum de la fonction de coût correspondante donne la direction $\mathbf{J}^*$ la plus proche possible de $\mathbf{B}$.

1.3 Qualité de l'apprentissage

La détermination de la direction privilégiée a été ramenée à la minimisation d'une fonction de coût. L'étude de l'apprentissage concerne deux aspects, l'un algorithmique, l'autre concernant l'étude des propriétés générales des algorithmes.

Le problème algorithmique est la détermination du minimum de la fonction de coût pour un ensemble d'apprentissage donné. Les algorithmes généralement utilisés, qu'ils soient basés sur des méthodes de descente de gradient ou de recuit simulé, devront garantir d'avoir obtenu le minimum absolu de la fonction de coût et non pas un minimum relatif (ou local). Dans le deuxième Chapitre, des simulations numériques de la fonction de coût optimale seront présentées. Dans le cas étudié, la fonction de coût ne possède qu'un seul minimum, ce qui simplifie les simulations. Ceci n'est pas le cas pour la deuxième application considérée au troisième Chapitre.

L'aspect le plus développé dans cette thèse concerne l'étude des propriétés du minimum de la fonction de coût. En effet, le choix de la fonction de coût est *a priori* arbitraire. Le but de l'apprentissage étant de trouver la direction privilégiée, il est important de trouver une fonction de coût dont le minimum en soit proche. Ce minimum dépend évidemment de la fonction de coût choisie mais aussi de la perturbation V^* de la distribution des exemples dans la direction $\mathbf{B}$ ainsi que de l'ensemble d'apprentissage $\mathcal{L}_\alpha$ considéré.

Une possibilité pour caractériser la *qualité* de l'apprentissage est de considérer le produit scalaire entre les deux directions $\mathbf{B}$ et $\mathbf{J}^*$. En ne considérant que des directions normées à un, ce produit scalaire $R \equiv \mathbf{B} \cdot \mathbf{J}^*$ est compris entre -1 et 1. Par la suite, ce paramètre R sera considéré comme une mesure de la qualité de l'apprentissage. Un apprentissage parfait correspond à $R = 1$ car dans ce cas les deux directions sont identiques. Comme nous l'avons déjà dit, puisqu'on ne dispose que d'un nombre fini d'exemples pour l'apprentissage, un apprentissage parfait n'est généralement pas envisageable.

Puisque la direction $\mathbf{J}^*$ minimise la fonction de coût pour un ensemble d'apprentissage donné, la qualité R de l'apprentissage dépend de cet ensemble. Par la suite, nous nous intéresserons plus particulièrement à la moyenne de cette qualité. Cette moyenne sera effectuée sur tous les ensembles d'apprentissage possédant le même nombre d'exemples, distribués selon la même densité de probabilité. Quelques remarques sur la moyenne de la qualité s'imposent à ce niveau. Par exemple, un choix aléatoire de la direction $\mathbf{J}^*$ correspond à une valeur moyenne de R nulle. Une fonction de coût ne sera donc intéressante qu'à la condition que la direction cor-

respondante à son minimum possède un produit scalaire strictement positif avec la direction privilégiée. Si la distribution des exemples est symétrique par rapport à l'origine, ce qui est le cas pour une fonction V^* paire, alors il faut considérer la moyenne de la valeur absolue de R puisque les directions $\mathbf{B}$ et $-\mathbf{B}$ sont équivalentes.

1.4 Intérêt de la Physique Statistique

La Physique Statistique permet l'étude de la qualité de l'apprentissage R, lorsque celui-ci se fait par minimisation d'une fonction de coût. En effet, un système physique à température nulle est dans l'état qui minimise l'énergie de celui-ci. Si l'on considère la fonction de coût comme une énergie, on peut obtenir les propriétés du minimum en considérant le système correspondant à température nulle. Les variables dynamiques de notre système correspondent à la direction $\mathbf{J}$. L'ensemble d'apprentissage sera considéré comme un ensemble de variables gelées. Ce problème se rapproche de celui des systèmes désordonnés tels que les verres de spins [111, 63, 81] où une partie des variables est dynamique et une autre gelée. Dans le cas des verres de spins, les couplages entre spins sont généralement considérés comme aléatoires et gelés tandis que les spins sont des variables dynamiques. Cette analogie entre ces problèmes se retrouve aussi dans la méthode utilisée pour les résoudre : la méthode des répliques [33, 15, 95, 96, 97, 98, 81].

Afin de pouvoir étudier les propriétés du minimum de la fonction de coût, nous allons la considérer comme une énergie. L'énergie d'un système physique est minimale à température nulle ; il est donc intéressant d'introduire une température fictive T que l'on fera tendre par la suite vers 0. Ceci revient à considérer un système dont les variables microscopiques sont les N composantes de la direction recherchée $\mathbf{J}$. La distribution de ces variables est donnée par la distribution de Gibbs :

$$P(\mathbf{J};\mathcal{L}_\alpha,\beta) = \frac{1}{Z}\exp\left(-\beta E\left(\mathbf{J};\mathcal{L}_\alpha\right)\right). \tag{1.7}$$

La fonction de partition du système correspondant à la fonction de coût (1.6) s'écrit alors :

$$Z\left(\beta;\mathcal{L}_\alpha\right) \equiv \int dP(\mathbf{J})\,\exp\left(-\beta E\left(\mathbf{J};\mathcal{L}_\alpha\right)\right) \tag{1.8}$$

où $\beta = 1/T$ et $dP(\mathbf{J})$ est la distribution uniforme sur l'ensemble des directions d'un espace de dimension N :

$$dP(\mathbf{J}) = \delta\left(\mathbf{J}\cdot\mathbf{J} - 1\right)\; d\mathbf{J}. \tag{1.9}$$

Cette distribution dite *a priori* est choisie arbitrairement. Le choix d'une distribution uniforme sur l'espace des phases de la variable dynamique est le choix le plus simple que l'on puisse envisager. Ce choix est, de plus, cohérent avec l'hypothèse sous-jacente que la direction privilégiée peut correspondre de manière équiprobable à n'importe quelle direction. Cette remarque est importante pour la définition de l'optimalité dont on discutera plus tard. L'hypothèse que toutes les directions peuvent correspondre à la direction privilégiée de manière équiprobable n'est pas sans conséquence. Imaginons que la direction privilégiée possède uniquement des coordonnées binaires ($B_i = \pm 1$). Il est assez évident de penser que l'apprentissage optimal devra tenir compte de cette information supplémentaire par l'intermédiaire d'une distribution *a priori* différente de (1.9), car cette information n'est pas contenue dans l'ensemble d'apprentissage. Cette information est d'ailleurs difficile à prendre en compte et de nombreux auteurs s'y sont intéressés [19, 64, 65, 67, 78, 79, 83, 108, 121].

L'énergie libre F du système pour un ensemble d'apprentissage donné $\mathcal{L}_\alpha$ s'écrit :

$$F(\beta;\mathcal{L}_\alpha) \equiv -\frac{1}{\beta}\ln Z(\beta;\mathcal{L}_\alpha)\,. \tag{1.10}$$

Une difficulté importante pour caractériser le minimum de F, et les propriétés de R, provient du fait que la fonction de partition et, donc, ses propriétés sont des variables aléatoires qui dépendent de l'ensemble d'apprentissage. Une possibilité pour s'affranchir de cette dépendance est de considérer la limite thermodynamique pour laquelle le nombre d'exemples P de l'ensemble d'apprentissage ainsi que la dimension N de l'espace des données divergent de telles façons que la *taille réduite* $\alpha \equiv P/N$ reste finie. Dans cette limite particulière, l'énergie libre du système devient généralement indépendante du choix particulier de l'ensemble d'apprentissage. Cette propriété appelée propriété d'automoyennage est alors équivalente à la relation suivante :

$$\lim_{\substack{N\to+\infty\\ P\to+\infty\\ \alpha=P/N}} \frac{1}{N}F(\beta;\mathcal{L}_\alpha) = \lim_{\substack{N\to+\infty\\ P\to+\infty\\ \alpha=P/N}} \frac{1}{N}\,\overline{F(\beta;\mathcal{L}_\alpha)}, \tag{1.11}$$

où la barre sur l'énergie libre symbolise la moyenne sur tous les ensembles d'apprentissage de taille réduite α. Cette relation signifie que, dans la limite thermodynamique l'énergie libre par degré de liberté pour un ensemble d'apprentissage particulier $\mathcal{L}_\alpha$ est identique à la moyenne sur tous les ensembles d'apprentissage de même taille. Seule une fraction de mesure nulle des ensembles d'apprentissage peut ne pas vérifier cette relation. D'autres propriétés comme la qualité de l'apprentissage R peuvent elles aussi être automoyennantes dans la limite thermodynamique. L'automoyennage est une hypothèse qui consiste à supposer que la variance de la quantité considérée est nulle si $N = +\infty$. Cette hypothèse sera confirmée par des simulations numériques dans l'application du deuxième Chapitre.

L'intérêt de (1.11) est qu'en faisant la moyenne sur les ensembles d'apprentissage, on se débarrasse de la dépendance des propriétés par rapport à $\mathcal{L}_\alpha$. Le prix à payer est que l'on ne peut faire des prédictions que dans la limite thermodynamique.

La limite thermodynamique entraîne l'existence possible de transitions de phases lorsque α varie. Le problème présenté au troisième Chapitre illustre bien ce cas, puisqu'on y prédit des transitions du premier et du second ordre. Il est à noter que ces transitions sont essentiellement dues à la taille infinie du système. Pour de nombreux systèmes physiques, le nombre de particules est suffisamment grand (de l'ordre du nombre d'Avogadro $\mathcal{N} \sim 10^{23}$) pour que les transitions soient nettement visibles, notamment, des sauts dans les fonctions thermodynamiques pour les transitions du premier ordre. Dans les applications réalistes qui nous intéressent, la taille du système dépasse rarement des valeurs de l'ordre de 1000. Le comportement lors de ces transitions est alors généralement beaucoup plus *lisse* que celui prédit dans la limite thermodynamique. Pour ces systèmes, les effets de taille finie sont importants comme nous l'avons mis en évidence dans le problème étudié au deuxième Chapitre.

L'étude des propriétés de l'apprentissage est donc ramenée à celle de calculer la moyenne de l'énergie libre par rapport aux ensembles d'apprentissage de même taille dans la limite thermodynamique :

$$f \equiv \lim_{\substack{N\to+\infty\\ P\to+\infty\\ \alpha=P/N}} \frac{1}{N}\,\overline{F(\beta;\mathcal{L}_\alpha)}. \tag{1.12}$$

Il est bien évident que cette énergie libre f dépend de la perturbation V^* de la distribution des exemples, de la fonction de coût (ou plus simplement du potentiel V qui détermine cette fonction de coût) et de la taille réduite α de l'ensemble d'apprentissage. C'est l'étude de ces dépendances qui nous intéresse par la suite. Lors

du calcul de l'énergie libre, des paramètres d'ordre pertinents dans le problème vont naturellement apparaître, notamment, le paramètre R qui permet de caractériser la qualité de l'apprentissage.

1.4.1 Méthode des répliques

Le calcul de la moyenne de l'énergie libre sur tous les ensembles d'apprentissage de même taille est difficile, puisqu'il revient à moyenner le logarithme de la fonction de partition (1.8). Une approche fructueuse connue sous le nom de méthode des répliques permet dans certains cas de résoudre cette difficulté [81, 38, 39, 40, 41]. Cette méthode consiste à utiliser la relation suivante :

$$\overline{\ln Z} = \lim_{n\to 0} \frac{1}{n} \ln \overline{Z^n} \tag{1.13}$$

qui permet de relier la moyenne du logarithme de la fonction de partition à la moyenne de ses moments. La moyenne des moments de Z est d'un calcul plus aisé pour les valeurs entières de n. En effet, dans ce cas, cela revient à considérer un système composé de n répliques indépendantes du système initial, c'est-à-dire, correspondant à n directions $\mathbf{J}_a$ avec $a = 1, \cdots, n$ pour un même ensemble d'apprentissage $\mathcal{L}_\alpha$. La moyenne sur les ensembles d'apprentissage permet d'éliminer les variables gelées au détriment de l'introduction d'un couplage entre répliques. Ce couplage entre répliques est la signature du fait que bien qu'indépendantes, les répliques possédaient le même ensemble d'apprentissage $\mathcal{L}_\alpha$. Ce comportement est identique à celui du problème des verres de spins avec des constantes de couplages aléatoires et gelées [81, 111, 63]. Le calcul de $\ln \overline{Z^n}$ du système de répliques couplées est possible pour des valeurs entières de n. Toutefois, des difficultés surviennent lors du prolongement analytique à des valeurs réelles et, en particulier, lors de la limite $n \to 0$.

L'énergie libre par degré de liberté s'écrit d'après les équations (1.12) et (1.13) :

$$f = \lim_{\substack{N\to+\infty \\ P\to+\infty \\ \alpha=P/N}} \lim_{n\to 0} -\frac{1}{\beta N n} \ln \overline{Z^n\left(\beta;\mathcal{L}_\alpha\right)}. \tag{1.14}$$

Exprimons tout d'abord la fonction de partition Z (1.8) à la puissance n pour n entier :

$$Z^n\left(\beta;\mathcal{L}_\alpha\right) = \prod_{a=1}^{n} \int dP(\mathbf{J}_a) \exp\left(-\beta E\left(\mathbf{J}_a;\mathcal{L}_\alpha\right)\right). \tag{1.15}$$

En remplaçant l'énergie par son expression (1.6) et en introduisant les variables $\lambda_a^\mu = \mathbf{J}_a \cdot \boldsymbol{\xi}^\mu$ ainsi que leurs variables conjuguées $\widetilde{\lambda}_a^\mu$:

$$\begin{aligned} Z^n\left(\beta;\mathcal{L}_\alpha\right) &= \prod_{a=1}^{n} \int dP(\mathbf{J}_a) \prod_{\mu=1}^{P} \prod_{a=1}^{n} \frac{1}{2\pi} \iint d\lambda_a^\mu d\widetilde{\lambda}_a^\mu \\ &\quad \exp\left(i \sum_{a\mu} \widetilde{\lambda}_a^\mu \left(\lambda_a^\mu - \mathbf{J}_a \cdot \boldsymbol{\xi}^\mu\right) - \beta \sum_{a\mu} V\left(\lambda_a^\mu\right) \right). \end{aligned} \tag{1.16}$$

La moyenne sur tous les ensembles d'apprentissage de taille réduite α correspond à intégrer sur tous les exemples $\boldsymbol{\xi}^\mu$ avec leur distribution (1.2). En introduisant les variables $\gamma^\mu = \mathbf{B} \cdot \boldsymbol{\xi}^\mu$ et leurs variables conjuguées $\widetilde{\gamma}^\mu$, l'intégrale sur les exemples $\boldsymbol{\xi}^\mu$ se réduit à de simples intégrales gaussiennes :

$$\begin{aligned}\overline{Z^n(\beta;\mathcal{L}_\alpha)} &= \prod_{a=1}^{n}\int dP(\mathbf{J}_a)\prod_{a\mu}\frac{1}{2\pi}\iint d\lambda_a^\mu d\widetilde{\lambda}_a^\mu \prod_\mu \frac{1}{2\pi}\iint d\gamma^\mu d\widetilde{\gamma}^\mu \qquad (1.17)\\ &\exp\left(i\sum_{a\mu}\widetilde{\lambda}_a^\mu\lambda_a^\mu + i\sum_\mu \widetilde{\gamma}^\mu\gamma^\mu - \beta\sum_{a\mu}V(\lambda_a^\mu) - \sum_\mu V^*(\gamma^\mu)\right)\\ &\prod_{\mu=1}^{P}\int\frac{d\boldsymbol{\xi}^\mu}{(2\pi)^{N/2}}\exp\left(-\frac{\boldsymbol{\xi}^\mu\cdot\boldsymbol{\xi}^\mu}{2} - i\left(\sum_{a=1}^{n}\widetilde{\lambda}_a^\mu\mathbf{J}_a + \widetilde{\gamma}^\mu\mathbf{B}\right)\cdot\boldsymbol{\xi}^\mu\right).\end{aligned}$$

Les intégrales gaussiennes couplent les directions des différentes répliques $\mathbf{J}_a$ entre elles ainsi qu'avec la direction privilégiée $\mathbf{B}$. Après intégration sur les exemples, on obtient :

$$\begin{aligned}\overline{Z^n(\beta;\mathcal{L}_\alpha)} &= \prod_{a=1}^{n}\int dP(\mathbf{J}_a)\prod_{a\mu}\frac{1}{2\pi}\iint d\lambda_a^\mu d\widetilde{\lambda}_a^\mu \prod_\mu \frac{1}{2\pi}\iint d\gamma^\mu d\widetilde{\gamma}^\mu \qquad (1.18)\\ &\prod_\mu\exp\left(i\sum_a\widetilde{\lambda}_a^\mu\lambda_a^\mu + i\widetilde{\gamma}^\mu\gamma^\mu - \beta\sum_a V(\lambda_a^\mu) - V^*(\gamma^\mu)\right)\\ &\prod_\mu\exp\left(-\frac{1}{2}\sum_{a,b}\widetilde{\lambda}_a^\mu\widetilde{\lambda}_b^\mu\mathbf{J}_a\cdot\mathbf{J}_b - \sum_a\widetilde{\lambda}_a^\mu\widetilde{\gamma}^\mu\mathbf{J}_a\cdot\mathbf{B} - \frac{\widetilde{\gamma}^{\mu\,2}}{2}\right).\end{aligned}$$

Nous allons maintenant introduire les variables :

$$q_{ab} = \mathbf{J}_a\cdot\mathbf{J}_b \quad a<b, \qquad (1.19)$$

$$R_a = \mathbf{B}\cdot\mathbf{J}_a \quad a=1,\cdots,n. \qquad (1.20)$$

Les directions $\mathbf{J}_a$ étant normées : $\mathbf{J}_a\cdot\mathbf{J}_a = 1$, il n'est pas nécessaire d'introduire des variables q_{aa}. Il est intéressant de remarquer que la variable R_a correspond au produit scalaire entre la direction privilégiée et la direction de la réplique a. Dans la limite de température nulle qui va nous intéresser par la suite, la direction $\mathbf{J}_a$ correspond à la direction qui minimise l'énergie, ou encore, la fonction de coût. Les paramètres R_a seront alors directement reliés à la qualité de l'apprentissage R introduite précédemment. Les paramètres q_{ab} correspondent aux produits scalaires entre deux répliques. Ces paramètres sont analogues aux paramètres q_{ab} introduit lors de l'étude des verres de spins par la méthode des répliques [81, 111, 63]. Dans ce dernier cas, le paramètre q_{ab} caractérise la similitude des états de spins entre les deux répliques a et b. Par analogie, le paramètre R_a peut alors être considéré comme une aimantation moyenne suivant la direction privilégiée.

Introduisons les variables conjuguées $N\phi_{ab}$ et $N\epsilon_a$ des paramètres R_a et q_{ab} ainsi qu'une représentation intégrale de la fonction $\delta(\mathbf{J}_a\cdot\mathbf{J}_a - 1)$ qui apparaît dans $dP(\mathbf{J}_a)$ (Eq.(1.9)).

$$\begin{aligned}\overline{Z^n(\beta;\mathcal{L}_\alpha)} &= \prod_a \frac{N}{2\pi}\iint d\mathbf{J}_a d\psi_a \prod_a \frac{N}{2\pi}\iint dR_a d\epsilon_a \prod_{a<b}\frac{N}{2\pi}\iint dq_{ab}d\phi_{ab} \\ &\exp\left(iN\sum_{a<b}\phi_{ab}(\mathbf{J}_a\cdot\mathbf{J}_b - q_{ab}) + iN\sum_a \epsilon_a(\mathbf{B}\cdot\mathbf{J}_a - R_a)\right) \\ &\exp\left(iN\sum_a \psi_a(\mathbf{J}_a\cdot\mathbf{J}_a - 1)\right)\left\{\frac{1}{2\pi}\iint d\gamma d\widetilde{\gamma}\prod_a\frac{1}{2\pi}\iint d\lambda_a d\widetilde{\lambda}_a\right. \\ &\exp\left(i\sum_a \widetilde{\lambda}_a\lambda_a + i\widetilde{\gamma}\gamma - \beta\sum_a V(\lambda_a) - V^*(\gamma)\right) \\ &\left.\exp\left(-\sum_{a<b}q_{ab}\widetilde{\lambda}_a\widetilde{\lambda}_b - \frac{1}{2}\sum_a\widetilde{\lambda}_a^2 - \sum_a R_a\widetilde{\lambda}_a\widetilde{\gamma} - \frac{\widetilde{\gamma}^2}{2}\right)\right\}^P . \qquad (1.21)\end{aligned}$$

La première partie de l'expression (1.21) est une partie purement géométrique puisqu'elle dépend des contraintes entre les directions $\mathbf{J}_a$ et $\mathbf{B}$ mais pas de la distribution des exemples ni de la fonction de coût. Elle peut être assimilée à la partie entropique du système physique correspondant. Il est important de remarquer que chaque argument des diverses exponentielles qui composent cette première partie est proportionnel à la taille N parce que nous avons introduit des variables conjuguées proportionnelles à N. Ceci est nécessaire pour que ces termes, et les contraintes qu'ils représentent, ne disparaissent pas dans la limite thermodynamique. On peut remarquer qu'une autre façon, utilisée habituellement, d'obtenir le même résultat, est de normaliser les vecteurs $\mathbf{J}_a$ et $\mathbf{B}$ à $\sqrt{N}$. Alors, $dP(\mathbf{J}_a) = \delta(\mathbf{J}_a\cdot\mathbf{J}_a - N)$, et les variables conjuguées sont d'ordre 1.

Du fait de l'indépendance des exemples de l'ensemble d'apprentissage, la deuxième partie de l'expression (1.21) s'exprime comme une fonction dépendante d'un seul exemple à la puissance P où P est le nombre d'exemples de l'ensemble d'apprentissage. Cette deuxième partie peut être assimilée à la partie énergétique du système.

La limite thermodynamique devient alors plus naturelle. En effet, le terme entropique est proportionnel à N tandis que le terme énergétique est proportionnel à P. Afin de ne pas obtenir un résultat trivial, il est utile de considérer une limite où ces deux termes divergent en gardant un rapport fini ce qui revient à considérer la taille réduite $\alpha = P/N$ de l'ensemble d'apprentissage constante. Le rapport α remplacera, par la suite, le nombre d'exemples et les propriétés d'apprentissage seront données en fonction de ce paramètre.

Il est possible de réécrire la moyenne du moment d'ordre n de la fonction de partition Z en faisant apparaître explicitement le terme énergétique $\alpha\mathcal{E}$ et le terme entropique $\mathcal{S}$:

$$\overline{Z^n(\beta;\mathcal{L}_\alpha)} = \int \mathcal{D}\psi\,\mathcal{D}\phi\,\mathcal{D}\mathcal{Q}\,\mathcal{D}\epsilon\,\mathcal{D}\mathcal{R}\exp\left(-\beta N n\mathcal{F}(\psi_a,\phi_{ab},q_{ab},\epsilon_a,R_a)\right) \quad (1.22)$$

avec :

$$\mathcal{F}(\psi_a,\phi_{ab},q_{ab},\epsilon_a,R_a) = \alpha\mathcal{E}(q_{ab},R_a) - T\mathcal{S}(\psi_a,\phi_{ab},q_{ab},\epsilon_a,R_a) \quad (1.23)$$

et où $\mathcal{D}\psi\,\mathcal{D}\phi\,\mathcal{D}\mathcal{Q}\,\mathcal{D}\epsilon\,\mathcal{D}\mathcal{R}$ représentent les variables d'intégration ψ_a,ϕ_{ab},q_{ab},ϵ_a et R_a ainsi que des constantes d'intégration non pertinentes pour la suite.

Le terme énergétique s'écrit :

$$
\begin{aligned}
n\beta\mathcal{E} &= -\ln\left\{\frac{1}{2\pi}\iint d\gamma d\widetilde{\gamma}\prod_a \frac{1}{2\pi}\iint d\lambda_a d\widetilde{\lambda}_a \right. \\
&\quad \exp\left(i\sum_a \widetilde{\lambda}_a\lambda_a + i\widetilde{\gamma}\gamma - \beta\sum_a V(\lambda_a) - V^*(\gamma)\right) \\
&\quad \left.\exp\left(-\sum_{a<b} q_{ab}\widetilde{\lambda}_a\widetilde{\lambda}_b - \frac{1}{2}\sum_a \widetilde{\lambda}_a^2 - \sum_a R_a\widetilde{\lambda}_a\widetilde{\gamma} - \frac{\widetilde{\gamma}^2}{2}\right)\right\},
\end{aligned}
\tag{1.24}
$$

et le terme entropique est :

$$
\begin{aligned}
n\mathcal{S} &= -i\sum_{a<b}\phi_{ab}q_{ab} - i\sum_a \epsilon_a R_a - i\sum_a \psi_a \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\ln\left\{\prod_a\int dJ_{ia}\,\exp\left(iN\sum_{a<b}\phi_{ab}J_{ia}J_{ib}\right)\right. \\
&\quad \left.\exp\left(iN\sum_a \epsilon_a B_i J_{ia} + iN\sum_a \psi_a J_{ia}^2\right)\right\}.
\end{aligned}
\tag{1.25}
$$

La limite thermodynamique permet de simplifier le calcul des intégrales de l'équation (1.22) par l'utilisation de la méthode du col. Cette méthode permet d'estimer une intégrale de la façon suivante :

$$
\int dx \exp\left(-Ng(x)\right) \sim \exp\left(-N\min_x g(x) + O(\ln N)\right) \tag{1.26}
$$

dans la limite où N est grand. Cette méthode est suffisante puisqu'elle permet d'extraire la partie finie de l'énergie libre par degré de liberté dans la limite thermodynamique. Toutefois, l'application de la méthode du col au calcul de (1.22) correspond à faire une inversion de limites entre n et N (*cf.* (1.14)). Cette difficulté mathématique ainsi que les problèmes dus au prolongement analytique des valeurs entières de n aux valeurs réelles et à la limite $n \to 0$ ne seront pas discutés ici [81]. La modification surprenante qui en découle est le fait que pour les variables dépendantes de deux répliques différentes (a et b) le minimum doit être remplacé par un maximum pour que la solution reste stable. Ceci est dû au fait que le nombre de ces variables devient négatif pour des valeurs de n comprises entre 0 et 1. Pour simplifier, et éviter de préciser entre maximum et minimum pour toutes les variables, on parlera désormais d'extremum. En résumé, nous admettrons que l'inversion des limites n'entraîne pas de conséquences, ce qui permet de calculer l'énergie libre (1.14) par la méthode du col :

$$
f = \underset{\psi_a,\phi_{ab},q_{ab},\epsilon_a,R_a}{\text{extremum}} \mathcal{F}\left(\psi_a,\phi_{ab},q_{ab},\epsilon_a,R_a\right). \tag{1.27}
$$

Le calcul des différentes intégrales de l'équation (1.22) se ramène à chercher l'extremum de l'expression (1.23) par rapport aux variables ψ_a,ϕ_{ab},q_{ab},ϵ_a et R_a. Dans le cas où plusieurs extrema existent, celui correspondant à l'énergie libre f la plus faible est à considérer.

1.4.2 Hypothèse de symétrie des répliques

Le nombre de variables étant important, la recherche de l'extremum de $\mathcal{F}$ reste un problème difficile. Il est possible de restreindre la recherche de cet extremum à

des solutions qui conservent la symétrie de la fonction à extremiser. La symétrie la plus simple à considérer est l'invariance par rapport à la permutation des différentes répliques. Les répliques sont identiques ; il n'y a aucune raison *a priori* d'introduire des paramètres différents pour chaque réplique. Il est alors possible de réduire la recherche aux variables suivantes :

$$
\begin{aligned}
i\psi_a &= G, && && (1.28)\\
q_{ab} &= q, && a<b && (1.29)\\
i\phi_{ab} &= F, && a<b && (1.30)\\
R_a &= R, && && (1.31)\\
i\epsilon_a &= E. && && (1.32)
\end{aligned}
$$

Cette hypothèse est connue sous le nom de *symétrie des répliques*. Il est possible d'étudier la cohérence de la solution ainsi obtenue. Il suffit pour cela d'effectuer une analyse de stabilité locale de la solution. Cette étude de stabilité consiste simplement à regarder les petites perturbations autour de la solution et à vérifier que la solution correspond à un minimum ou un maximum selon la variable considérée. Deux stabilités différentes sont à considérer. La première dite longitudinale consiste à regarder les perturbations conservant la symétrie des répliques, par exemple, $R \rightarrow R + \delta R$. Cette stabilité est assez facilement vérifiable. La stabilité dite transverse est, par contre, plus difficile à vérifier. Elle consiste à regarder des perturbations ne conservant plus la symétrie des répliques. Pour étudier cette stabilité, il convient de regarder les valeurs propres de la matrice Hessienne, matrice des dérivées secondes, de $\mathcal{F}(\psi_a,\phi_{ab},q_{ab},\epsilon_a,R_a)$. Les valeurs propres doivent être positives pour un minimum et négatives pour un maximum. Un exemple du calcul de la stabilité transverse est développé par de Almeida et Thouless [2] dans le cas des verres de spins. Pour le problème de l'apprentissage, ce calcul a été effectué pour la première fois par Gardner [39, 40, 41]. La condition obtenue concerne le signe du déterminant de la matrice Hessienne. Un changement de signe de ce déterminant signale l'apparition d'une instabilité. Cette condition est nécessaire mais pas suffisante. En effet, deux valeurs propres peuvent changer de signe simultanément.

Nous démontrerons, par la suite, que la condition de stabilité locale est bien vérifiée dans le problème général d'apprentissage optimal traité dans cette partie de la thèse. Cela ne garantie pas que d'autres solutions localement stables mais ne vérifiant pas l'hypothèse de symétrie des répliques existent. Ces solutions sont appelées solutions avec brisure de symétrie des répliques [81]. L'existence possible de telles solutions sera discutée à propos du problème abordé au troisième Chapitre.

Avec l'hypothèse de symétrie des répliques, les expressions (1.24) de l'énergie et (1.25) de l'entropie se simplifient :

$$
\begin{aligned}
n\beta\mathcal{E} = &-\ln\left\{\frac{1}{2\pi}\iint d\gamma d\widetilde{\gamma}\exp\left(-\frac{\widetilde{\gamma}^2}{2}+i\widetilde{\gamma}\gamma-V^*(\gamma)\right)\right. \qquad (1.33)\\
&\prod_a \frac{1}{2\pi}\iint d\lambda_a d\widetilde{\lambda}_a \exp\left(i\sum_a \widetilde{\lambda}_a\lambda_a-\beta\sum_a V(\lambda_a)\right)\\
&\left.\exp\left(-q\sum_{a<b}\widetilde{\lambda}_a\widetilde{\lambda}_b-\frac{1}{2}\sum_a\widetilde{\lambda}_a^2-R\sum_a\widetilde{\lambda}_a\widetilde{\gamma}\right)\right\},
\end{aligned}
$$

$$
\begin{aligned}
n\mathcal{S} &= -\frac{1}{2}n(n-1)Fq - nER - nG \qquad (1.34)\\
&\quad +\frac{1}{N}\sum_i \ln\left\{\prod_a \int dJ_{ia}\ \exp\left(+NF\sum_{a<b} J_{ia}J_{ib}\right)\right.\\
&\quad \left.\exp\left(+NE\sum_a B_i J_{ia} + NG\sum_a J_{ia}^2\right)\right\}.
\end{aligned}
$$

Il reste à effectuer la limite $N \to +\infty$ pour l'entropie ainsi que le prolongement analytique de l'énergie et de l'entropie à des valeurs de n réelles. N'oubliant pas le fait que seule la limite $n \to 0$ nous intéresse, on se limitera à un développement autour de $n = 0$ et on ne gardera que le terme linéaire en n.

Traitons tout d'abord le cas de l'entropie $\mathcal{S}$. Les variables J_{ia} couplent les différentes répliques entre elles. Pour les découpler, on utilise les identités suivantes :

$$
\begin{aligned}
\sum_{a<b} J_{ia}J_{ib} &= \frac{1}{2}\left(\sum_a J_{ia}\right)^2 - \frac{1}{2}\sum_a J_{ia}^2, \qquad (1.35)\\
\exp\left(\frac{\lambda x^2}{2}\right) &= \sqrt{\frac{\lambda}{2\pi}}\int dz\ \exp\left(-\frac{\lambda z^2}{2} - \lambda xz\right). \qquad (1.36)
\end{aligned}
$$

Avec $x = \sum_a J_{ia}$ et $\lambda = NF$, ceci permet de réécrire (1.34) :

$$
\begin{aligned}
n\mathcal{S} &= -\frac{1}{2}n(n-1)Fq - nER - nG + \frac{1}{N}\sum_i \ln\left(\int dz \prod_a \int dJ_{ia}\right. \qquad (1.37)\\
&\quad \left.\exp\left\{-\frac{N}{2}Fz^2 - \frac{N}{2}(F-2G)\sum_a J_{ia}^2 + N(EB_i - Fz)\sum_a J_{ia}\right\}\right).
\end{aligned}
$$

L'intégration sur les variables J_{ia} et sur z ne pose alors plus de problèmes. En ne gardant que les termes dominants dans la limite $n \to 0$ et en utilisant le fait que $\mathbf{B}\cdot\mathbf{B} = 1$, à une constante additive près (non pertinente pour la suite), on obtient :

$$
\mathcal{S} = \frac{1}{2}\left(Fq - 2ER - 2G - \ln(F-2G) + \frac{F+E^2}{F-2G}\right). \qquad (1.38)
$$

On voit ici le résultat discuté plus haut, que la contribution de Fq change de signe lorsque $n < 1$, par rapport à (1.37). Les variables E, F et G n'intervenant que dans le terme entropique, il est possible de les déterminer grâce aux conditions d'extremum :

$$
\frac{\partial\mathcal{S}}{\partial E} = \frac{\partial\mathcal{S}}{\partial F} = \frac{\partial\mathcal{S}}{\partial G} = 0. \qquad (1.39)
$$

Ces conditions se réduisent aux équations suivantes :

$$
\begin{aligned}
R &= \frac{E}{F-2G}\ , \qquad (1.40)\\
q &= \frac{F+E^2}{2(F-2G)^2}\ , \qquad (1.41)\\
1-q &= \frac{1}{F-2G}\ . \qquad (1.42)
\end{aligned}
$$

En remplaçant ces variables E, F et G dans l'expression (1.38), on obtient :

$$\mathcal{S} = \frac{1}{2}\left(\frac{1-R^2}{1-q} + \ln(1-q)\right). \tag{1.43}$$

Le calcul de $\beta\mathcal{E}$ est assez similaire à celui de $\mathcal{S}$. Il faut en effet découpler les variables $\widetilde{\lambda}_a$. Pour cela, on utilise l'identité suivante :

$$\sum_{a<b} \widetilde{\lambda}_a \widetilde{\lambda}_b = \frac{1}{2}\left(\sum_a \widetilde{\lambda}_a\right)^2 - \frac{1}{2}\sum_a \widetilde{\lambda}_a^2 \tag{1.44}$$

et on introduit la variable $y = \sum_a \widetilde{\lambda}_a$ ainsi que sa variable conjuguée z :

$$\begin{aligned} n\beta\mathcal{E} &= -\ln\left\{\frac{1}{2\pi}\iint d\gamma d\widetilde{\gamma}\frac{1}{2\pi}\iint dydz\prod_a \frac{1}{2\pi}\iint d\lambda_a d\widetilde{\lambda}_a \right. \\ &\quad \exp\left(-\frac{\widetilde{\gamma}^2}{2} + i\widetilde{\gamma}\gamma - V^*(\gamma) + iyz - Rz\widetilde{\gamma} - q\frac{z^2}{2}\right) \\ &\quad \left.\exp\left(-\frac{1}{2}(1-q)\sum_a \widetilde{\lambda}_a^2 + i\sum_a \widetilde{\lambda}_a(\lambda_a - y) - \beta\sum_a V(\lambda_a)\right)\right\}. \end{aligned} \tag{1.45}$$

Les répliques étant découplées, le calcul se poursuit facilement :

$$\begin{aligned} n\beta\mathcal{E} &= -\ln\left\{\iint Dt\,Dy\,\exp\left(-V^*(\gamma)\right)\right. \\ &\quad \left.\left[\int \frac{d\lambda}{\sqrt{2\pi(1-q)}}\exp\left(-\frac{(\lambda - y\sqrt{q})^2}{2(1-q)} - \beta V(\lambda)\right)\right]^n\right\} \end{aligned} \tag{1.46}$$

$$\begin{aligned} &= -n\iint Dt\,Dy\,\exp\left(-V^*(\gamma)\right) \\ &\quad \ln\left[\int \frac{d\lambda}{\sqrt{2\pi(1-q)}}\exp\left(-\frac{(\lambda - y\sqrt{q})^2}{2(1-q)} - \beta V(\lambda)\right)\right] \end{aligned} \tag{1.47}$$

avec :

$$\gamma = \frac{t\sqrt{q-R^2} + yR}{\sqrt{q}}. \tag{1.48}$$

En introduisant dans (1.27) les expressions (1.43) et (1.47), qui correspondent à l'extremum de (1.23) par rapport aux variables ψ_a,ϕ_{ab} et ϵ_a dans l'hypothèse de symétrie des répliques, on obtient l'énergie libre f :

$$f = \underset{q,R}{\text{extr}}\,\mathcal{F}(q,R) \tag{1.49}$$

où $\mathcal{F}$, donné par (1.23), s'écrit :

$$\mathcal{F}(q,R) = \alpha\mathcal{E}(q,R) - T\mathcal{S}(q,R) \tag{1.50}$$

$$\begin{aligned} &= -\frac{1-R^2}{2\beta(1-q)} - \frac{1}{2\beta}\ln(1-q) - \frac{\alpha}{\beta}\iint Dt\,Dy\exp\left(-V^*(\gamma)\right) \\ &\quad \ln\left[\int \frac{d\lambda}{\sqrt{2\pi(1-q)}}\exp\left(-\frac{(\lambda - y\sqrt{q})^2}{2(1-q)} - \beta V(\lambda)\right)\right]. \end{aligned} \tag{1.51}$$

Ce résultat bien connu [38, 39, 110, 101, 120, 126] est le point de départ des travaux décrits dans cette thèse.

1.4.3 Interprétation des paramètres d'ordre

Les propriétés du minimum de la fonction de coût sont données par l'énergie libre (1.49) dans la limite de température nulle. Toutefois, il est intéressant de noter que certains algorithmes d'apprentissage ne cherchent pas à déterminer la direction qui minimise l'énergie, mais échantillonnent la direction $\mathbf{J}$ parmi toutes les directions possibles avec la distribution de Gibbs (1.7).

La constante de normalisation Z de la distribution correspond à la fonction de partition précédemment introduite par l'équation (1.8). En particulier, l'algorithme dit *de Gibbs* [110, 101, 120] consiste, à prendre comme potentiel V, la perturbation V^* de la distribution des exemples et à faire l'échantillonnage à la température $T = 1$. On reviendra à cet algorithme au troisième Chapitre.

Il convient tout d'abord d'expliquer la signification des deux paramètres d'ordre R (1.31) et q (1.29) qui sont apparus naturellement lors du calcul dans l'hypothèse de symétrie des répliques. Par définition (1.20), le paramètre R n'est rien d'autre que le produit scalaire entre la direction privilégiée $\mathbf{B}$ et la direction $\mathbf{J}$, distribuée selon (1.7). Dans la limite de température nulle, cette distribution se réduit à un pic $\delta(\mathbf{J}-\mathbf{J}^*)$ centré sur la direction $\mathbf{J}^*$ qui minimise la fonction de coût. Le paramètre R correspond alors au produit scalaire entre la direction privilégiée et la direction $\mathbf{J}^*$, c'est-à-dire, qu'il correspond bien au paramètre $R = \mathbf{B} \cdot \mathbf{J}^*$ que l'on avait introduit pour caractériser la qualité de l'apprentissage.

La signification du paramètre q est plus complexe puisque celui-ci fait intervenir deux répliques différentes (*cf.* (1.19)). Le paramètre q est le produit scalaire entre deux directions distribuées selon (1.7). Remarquons que si la température est nulle, cette distribution se réduit à un pic $\delta(\mathbf{J} - \mathbf{J}^*)$ centré sur la direction $\mathbf{J}^*$, ce qui entraîne $q = 1$. Or, dans l'expression (1.51) de $\mathcal{F}$, il apparaît le produit $c \equiv \beta(1-q)$, qui est singulier lorsque $T \to 0$, puisque $\beta \to +\infty$ et $q \to 1$. Donc il est clair que le paramètre pertinent lors de la limite de température nulle est c. Afin de comprendre le sens de ce nouveau paramètre, nous allons considérer le problème à température finie. Avant de faire l'hypothèse de symétrie des répliques, on peut définir c_{ab} qui, s'écrit :

$$c_{ab} = \frac{(\mathbf{J}_a - \mathbf{J}_b)^2}{2T} = \beta\,(1 - q_{ab})\,. \tag{1.52}$$

Les directions $\mathbf{J}_a$ et $\mathbf{J}_b$ sont distribuées selon (1.7). Pour les faibles températures, on peut considérer que seules les directions $\mathbf{J}$ dont la fonction de coût est proche du minimum possèdent une probabilité non nulle. La proximité de ce minimum est caractérisée par une fenêtre de largeur T autour du minimum. Plus la température décroît, plus l'ensemble des directions accessibles autour de $\mathbf{J}^*$ se réduit, en d'autres termes, les fluctuations autour de $\mathbf{J}^*$ se réduisent. D'après (1.52), le paramètre c correspond au rapport entre le carré de ces fluctuations et la température. Un paramètre c grand correspond à des fluctuations importantes. En particulier, si le minimum est continuement dégénéré, c diverge ce qui reflète qu'il y a un continuum de directions qui minimisent l'énergie. Par contre, lorsque c est très faible, le puit au voisinage du minimum est très étroit.

1.4.4 Limite de température nulle

Dans la suite de cette thèse, nous nous intéressons aux propriétés d'apprentissage par minimisation d'une fonction de coût. Elles sont données par l'énergie libre dans la limite $\beta \to +\infty$. En utilisant une nouvelle fois la méthode du col, par rapport à β cette fois, la fonction $\mathcal{F}$ (1.51) s'exprime en fonction de R et $c = \beta(1 - q)$ de la façon suivante :

$$\mathcal{F}(R,c) = -\frac{1-R^2}{2\,c} + \alpha \iint Dt\,Dy \exp\left(-V^*(\gamma)\right) W\left(\lambda(y,c),y,c\right) \tag{1.53}$$

avec :

$$\gamma = t\sqrt{1-R^2} + yR \tag{1.54}$$

et $\lambda(y,c)$ la valeur pour laquelle la fonction :

$$W(\lambda,y,c) = \frac{(\lambda-y)^2}{2c} + V(\lambda) \tag{1.55}$$

est minimale. Les paramètres R et c sont déterminés par les conditions d'extremum :

$$\frac{\partial \mathcal{F}}{\partial R} = \frac{\partial \mathcal{F}}{\partial c} = 0. \tag{1.56}$$

Après quelques manipulations simples, il peut être montré que R et c sont solutions des équations suivantes :

$$1-R^2 = \alpha \iint Dt\,Dy\, \exp\left(-V^*(\gamma)\right) \left(\lambda(y,c)-y\right)^2, \tag{1.57}$$

$$R\sqrt{1-R^2} = \alpha \iint t\,Dt\,Dy\, \exp\left(-V^*(\gamma)\right) \left(\lambda(y,c)-y\right). \tag{1.58}$$

Il est évident que R et c dépendent de la perturbation V^* de la distribution des exemples (1.2), du potentiel V qui détermine la fonction de coût (1.6) et de la taille réduite α de l'ensemble d'apprentissage. Nous avons montré que le paramètre R est le produit scalaire entre la direction privilégiée $\mathbf{B}$ et la direction $\mathbf{J}^*$ qui minimise la fonction de coût. C'est le paramètre qui détermine la qualité de l'apprentissage. La méthode des répliques permet donc de déterminer la qualité de l'apprentissage en fonction du problème posé (perturbation V^* et taille réduite α) et de la fonction de coût utilisée pour le résoudre (potentiel V). Le résultat intéressant de la limite thermodynamique est que la qualité R est indépendante de l'ensemble d'apprentissage particulier utilisé pour l'apprentissage.

1.5 Détermination du potentiel optimal

Les calculs précédents permettent de calculer les performances d'un algorithme (défini par la fonction de coût ou plus particulièrement par le potentiel V qui la définit) pour un problème d'apprentissage donné (défini par la perturbation V^* dans la direction privilégiée) et ceci en fonction de la taille réduite α de l'ensemble d'apprentissage dont on dispose pour effectuer l'apprentissage. Ces performances consistent essentiellement dans la détermination du produit scalaire R entre la direction privilégiée recherchée $\mathbf{B}$ et la direction obtenue après apprentissage $\mathbf{J}^*$. Cette dernière correspond au minimum de la fonction de coût. Comme nous l'avons vu précédemment, l'apprentissage est d'autant plus performant que ce produit scalaire est proche de 1 ou encore que la direction obtenue après apprentissage est proche de celle recherchée. Une question évidente se pose alors : étant donné un problème (la fonction V^*) et une quantité d'exemples (représentée par α) est-il possible de trouver un *potentiel optimal* V_{opt} ? C'est-à-dire un potentiel pour lequel R serait le plus proche possible de 1.

La réponse à cette question a été apportée premièrement par Kinouchi et Caticha [62, 61] ainsi que par nous-mêmes [21, 26] indépendamment, dans le cadre

de l'apprentissage d'une tâche de classification linéairement séparable par un perceptron simple (*cf.* Chap.2). Dans le cas général, le résultat a été obtenu par Van den Broeck et Reimann [120]. Notre point de vue a été de considérer la qualité R comme une fonctionnelle du potentiel V avec la perturbation V^* et la taille réduite α considérées comme des paramètres fixes. Grâce à une maximisation fonctionnelle de R, il est possible de déterminer le potentiel optimal $V_{\rm opt}$.

Les équations (1.57) et (1.58) déterminent R et c. Nous allons montrer qu'il est possible de s'affranchir du paramètre c. En effet, le minimum de la fonction de coût reste inchangé si le potentiel V est multiplié par un facteur $a > 0$. Ce facteur multiplicatif peut être vu comme un changement d'échelle de l'énergie. Pour obtenir une fonction de partition invariante par ce changement d'échelle, il faut que le produit βE reste invariant, c'est-à-dire, que la température soit aussi multipliée par le facteur a. Puisque le paramètre c est proportionnel à β, le changement d'échelle d'énergie n'est pas transparent pour c: c est transformé en c/a. Un changement d'échelle judicieux consistant à prendre $a = c$ permet de fixer l'échelle d'énergie pour laquelle $c = 1$. Nous allons donc chercher le potentiel $V_{\rm opt}$ qui maximise R, mesuré dans les unités définies par $c = 1$. L'équation (1.57) peut alors être considérée comme la définition de R tandis que l'équation (1.58) sera considérée comme une contrainte sur ce paramètre. Le moyen de tenir compte d'une contrainte lors d'une maximisation est d'introduire un paramètre de Lagrange μ. Il suffit alors de maximiser R (ou encore de minimiser $1 - R^2$) par rapport à la fonction V avec pour expression de $1 - R^2$:

$$\begin{aligned} 1 - R^2 &= \alpha \iint Dt\, Dy\, \exp\left(-V^*(\gamma)\right) \left(\lambda(y) - y\right)^2 \\ &+ \mu \left\{ R\sqrt{1 - R^2} - \alpha \iint t\, Dt\, Dy\, \exp\left(-V^*(\gamma)\right)\, \left(\lambda(y) - y\right) \right\}. \end{aligned} \tag{1.59}$$

avec $\gamma = \gamma(t,y)$ donné par (1.54) et $\lambda(y,c)$ la fonction qui minimise (1.55). $\lambda(y,c)$ est écrit $\lambda(y)$ par simplicité puisque $c = 1$. La dépendance par rapport au potentiel V est implicite dans la fonction $\lambda(y)$. Cette dépendance implicite permet de remplacer la maximisation de R par rapport à V par une maximisation par rapport à une autre fonction $g(y) \equiv \lambda(y) - y$. Cette maximisation est beaucoup plus aisée puisque $1 - R^2$ apparaît alors comme une simple intégrale quadratique de cette nouvelle fonction. Une fois obtenue la fonction $g(y)$ pour laquelle R est maximal, il faudra s'assurer que cette fonction permet de remonter au potentiel optimal $V_{\rm opt}$.

$1 - R^2$ est minimal lorsque la dérivée fonctionnelle de $1 - R^2$ par rapport à $g(y_0)$ s'annule. La dérivée fonctionnelle de $1 - R^2$ s'écrit :

$$\begin{aligned} \frac{\delta\left(1 - R^2\right)}{\delta\, g(y_0)} &= 2\,\alpha \iint Dt\, Dy\, \exp\left(-V^*(\gamma)\right) g(y)\, \delta(y - y_0) \\ &\quad - \mu\,\alpha \iint Dt\, Dy\, \exp\left(-V^*(\gamma)\right)\, \delta(y - y_0). \end{aligned} \tag{1.60}$$

La fonction g, pour laquelle (1.60) s'annule, se déduit alors aisément :

$$g(y) = \frac{\mu}{2}\, \frac{\displaystyle\int t\, Dt\, \exp\left(-V^*(\gamma)\right)}{\displaystyle\int Dt\, \exp\left(-V^*(\gamma)\right)}. \tag{1.61}$$

Les paramètres R et μ sont déterminés par les équations (1.57) et (1.58) après avoir introduit l'expression de g donnée par l'équation (1.61).

$$1 - R^2 = \frac{\alpha\mu^2}{4} \int Dy \, \frac{\left[\int t\, Dt \, \exp\left(-V^*(\gamma)\right)\right]^2}{\int Dt \exp\left(-V^*(\gamma)\right)}, \tag{1.62}$$

$$R\sqrt{1-R^2} = \frac{\alpha\mu}{2} \int Dy \, \frac{\left[\int t\, Dt \, \exp\left(-V^*(\gamma)\right)\right]^2}{\int Dt \exp\left(-V^*(\gamma)\right)}. \tag{1.63}$$

De ces équations, on peut déduire que :

$$\frac{\mu}{2} = \frac{\sqrt{1-R^2}}{R}. \tag{1.64}$$

En remplaçant le paramètre de Lagrange μ dans (1.62) ou (1.63), on obtient l'équation qui détermine la qualité optimale R_{opt} du problème général de l'apprentissage d'une direction privilégiée :

$$\alpha = R_{\text{opt}}^2 \left[\int Dy \, \left\{ \frac{\left[\int t\ Dt\ \exp\left(-V^*(\gamma)\right)\right]^2}{\int Dt\ \exp\left(-V^*(\gamma)\right)} \right\} \right]^{-1}, \tag{1.65}$$

avec :

$$\gamma = t\sqrt{1 - R_{\text{opt}}^2} + yR_{\text{opt}}. \tag{1.66}$$

À ce niveau, il est intéressant de remarquer que l'expression (1.65) permet de trouver α en fonction de R_{opt}. Afin d'obtenir, R_{opt} pour une valeur donnée de α, il faut inverser cette expression. Nous verrons au Chapitre 3 que cette inversion n'est pas triviale. Il existe en effet des problèmes pour lesquels l'inverse de (1.65) n'est pas unique. Dans ce cas, il faut choisir la solution correspondante à la valeur maximale de R_{opt} puisque le calcul a été mené pour obtenir cette valeur maximale. L'existence de plusieurs solutions lors de l'inversion entraîne aussi l'existence de transitions du premier ordre de la qualité de l'apprentissage R_{opt} en fonction de la taille réduite α de l'ensemble d'apprentissage.

Revenons maintenant à la détermination du potentiel optimal V_{opt}. En supposant cette fonction suffisamment dérivable, il est possible d'exprimer sa dérivée comme une fonction de g. En effet, d'après l'équation (1.55), la fonction $\lambda(y)$ est celle qui minimise $W(\lambda,y,c)$ avec $c = 1$. Cette condition peut s'écrire :

$$\frac{\partial W}{\partial \lambda}(\lambda(y),y,c) = 0 = \lambda(y) - y + V'_{\text{opt}}(\lambda(y)). \tag{1.67}$$

La fonction $g(y)$ vérifie donc :

$$g(y) \equiv \lambda(y) - y = -V'_{\text{opt}}(\lambda). \tag{1.68}$$

Ceci permet de remonter au potentiel V_{opt} connaissant $g(y)$. D'après l'équation (1.61), g ne dépend que de V^* et de $R_{\text{opt}}(\alpha)$ puisque R_{opt} est donné par (1.65) et μ est une fonction de R_{opt} d'après (1.64). Puisque $\alpha = \alpha(R_{\text{opt}})$, à chaque taille réduite α correspond un potentiel V_{opt} différent. Pour remonter à V_{opt}, il suffit d'inverser la fonction $\lambda(y)$. Cela impose toutefois comme condition que cette fonction soit bien inversible. Dans les deux applications présentées aux Chapitres 2 et 3, cette condition est vérifiée.

1.6 Cohérence de l'hypothèse de symétrie

L'hypothèse de symétrie des répliques utilisée pour calculer l'énergie libre permet de simplifier la recherche de l'extremum de l'énergie libre $\mathcal{F}$ en réduisant le nombre de paramètres à seulement deux : R et q (ou, dans la limite de température nulle, R et c). Il est toutefois indispensable de vérifier que la solution ainsi obtenue est bien l'extremum (maximum pour c et minimum pour R).

Dans cette section, nous allons vérifier uniquement la stabilité locale de la solution. C'est-à-dire que nous allons vérifier que la solution est bien un extremum de l'énergie libre par rapport à de petites perturbations mais on ne vérifiera pas que l'on a obtenu le minimum absolu de l'énergie libre. Une première condition est la stabilité longitudinale. Elle consiste à regarder les perturbations $R + \delta R$ et $c + \delta c$ pour vérifier que $\mathcal{F}(R,c)$ est bien un minimum vis-à-vis de R et un maximum vis-à-vis de c. Cette stabilité est appelée longitudinale car les perturbations considérées conservent toujours la symétrie des répliques. La stabilité transverse est plus difficile à vérifier. Elle consiste à regarder la stabilité par rapport à de faibles perturbations ne conservant pas la symétrie des répliques. Il faut pour cela revenir à l'expression de l'énergie libre (1.23) en fonction de toutes les variables précédemment introduites (ψ_a,ϕ_{ab},q_{ab},ϵ_a et R_a). Cette condition de stabilité est ramenée à une condition sur le déterminant de la matrice Hessienne de l'énergie libre (matrice des dérivées secondes) [2, 39, 40]. Je ne détaille pas ici ce calcul un peu complexe mais je rappelle la condition obtenue par Reimann et Van den Broeck [101] dans le cas général :

$$1 > \alpha \iint Dy\, Dt\, \exp\left(-V^*(\gamma)\right) \left(\lambda'(y) - 1\right)^2, \tag{1.69}$$

avec γ défini par (1.66) et $\lambda'(y)$ la dérivée de $\lambda(y)$, obtenue par minimisation de l'équation (1.55), par rapport à y.

Il est important de souligner que cette stabilité n'est qu'une stabilité locale. Elle ne permet, en effet, en aucun cas d'assurer qu'il n'existe pas une autre solution localement stable avec ou sans brisure de symétrie des répliques dont l'énergie libre serait inférieure. Nous verrons, dans le Chapitre 3, qu'il est possible d'obtenir pour une même valeur de α deux extrema localement stables $f(R_1,q_1)$ et $f(R_2,q_2)$ de l'énergie libre (1.49) correspondante à l'algorithme de Gibbs. Dans ce cas, la solution d'énergie libre la plus faible doit être conservée.

Il est possible de réécrire la condition de stabilité (1.69) en utilisant les équations (1.65) et (1.61) définissant R_{opt} et $g(y)$ pour aboutir à l'expression suivante [22] :

$$\frac{R^2_{\text{opt}}(1 - R^2_{\text{opt}})}{\alpha} \frac{d\ \alpha}{dR^2_{\text{opt}}} > 0. \tag{1.70}$$

Pour montrer l'équivalence des deux conditions (1.69) et (1.70), je vais introduire la notation suivante :

$$\langle t^n \rangle_y \equiv \int t^n\, Dt\, \exp\left(-V^*(\gamma)\right) \tag{1.71}$$

avec γ définit par (1.66). L'équation (1.65) déterminant R_{opt} et l'équation (1.61) déterminant $g(y)$ s'écrivent alors simplement :

$$R^2_{\text{opt}} = \alpha \int Dy\, \frac{\langle t \rangle_y^2}{\langle 1 \rangle_y}, \tag{1.72}$$

$$g(y) = \frac{\sqrt{1 - R^2_{\text{opt}}}}{R_{\text{opt}}} \frac{\langle t \rangle_y}{\langle 1 \rangle_y}. \tag{1.73}$$

La condition (1.69) s'écrit quant à elle :

$$1 > \alpha \frac{1 - R_{\rm opt}^2}{R_{\rm opt}^2} \int Dy \left(\frac{d}{dy} \frac{\langle t \rangle_y}{\langle 1 \rangle_y} \right)^2 \langle 1 \rangle_y . \tag{1.74}$$

Pour exprimer de manière plus simple cette condition, je vais exprimer la dérivée de $\langle t^n \rangle_y$ par rapport à y :

$$\frac{d}{dy} \langle t^n \rangle_y = \frac{R_{\rm opt}}{\sqrt{1 - R_{\rm opt}^2}} \left(\langle t^{n+1} \rangle_y - n \langle t^{n-1} \rangle_y \right) . \tag{1.75}$$

On en déduit l'expression suivante pour la condition (1.69) :

$$\int Dy \frac{\langle t \rangle_y^2}{\langle 1 \rangle_y} > R_{\rm opt}^2 \int Dy \left(\frac{\langle t^2 \rangle_y}{\langle 1 \rangle_y} - 1 - \frac{\langle t \rangle_y^2}{\langle 1 \rangle_y^2} \right)^2 \langle 1 \rangle_y . \tag{1.76}$$

Il reste maintenant à calculer :

$$\frac{R_{\rm opt}^2}{\alpha^2} \frac{d\ \alpha}{dR_{\rm opt}} = \int Dy \left(\frac{2}{R_{\rm opt}} \frac{\langle t \rangle_y^2}{\langle 1 \rangle_y} - 2 \frac{\langle t \rangle_y}{\langle 1 \rangle_y} \frac{d\langle t \rangle_y}{dR_{\rm opt}} + \frac{\langle t \rangle_y^2}{\langle 1 \rangle_y^2} \frac{d\langle 1 \rangle_y}{dR_{\rm opt}} \right) . \tag{1.77}$$

Pour cela, on utilise :

$$\frac{d\langle t \rangle_y}{dR_{\rm opt}} = -\frac{y}{\sqrt{1 - R_{\rm opt}^2}} \left(\langle 1 \rangle_y - \langle t^2 \rangle_y \right) + \frac{R_{\rm opt}}{1 - R_{\rm opt}^2} \left(2 \langle t \rangle_y - \langle t^3 \rangle_y \right) , \tag{1.78}$$

$$\frac{d\langle 1 \rangle_y}{dR_{\rm opt}} = \frac{y}{\sqrt{1 - R_{\rm opt}^2}} \langle t \rangle_y + \frac{R_{\rm opt}}{1 - R_{\rm opt}^2} \left(\langle 1 \rangle_y - \langle t^2 \rangle_y \right) . \tag{1.79}$$

En remplaçant ces deux expressions dans l'équation (1.77), en intégrant par parties sur y pour éliminer les termes proportionnels à y et en regroupant les termes correctement, on obtient :

$$\left(1 - R_{\rm opt}^2 \right) \frac{R_{\rm opt}^3}{2\alpha^2} \frac{d\ \alpha}{dR_{\rm opt}} = \int Dy \frac{\langle t \rangle_y^2}{\langle 1 \rangle_y} - R_{\rm opt}^2 \int Dy \left(\frac{\langle t \rangle_y^2}{\langle 1 \rangle_y^2} - 1 - \frac{\langle t^2 \rangle_y}{\langle 1 \rangle_y} \right)^2 \langle 1 \rangle_y . \tag{1.80}$$

Cette relation permet donc de remplacer la condition de stabilité locale (1.69) par la condition (1.70). Cette dernière condition est beaucoup plus parlante que l'expression (1.69). En effet, dès que $\alpha > 0$ alors la condition de stabilité locale de la solution symétrique par rapport à la permutation des répliques est équivalente à ce que $R_{\rm opt}$ soit une fonction croissante par rapport à la taille réduite α de l'ensemble d'apprentissage.

Pour résumer, on vient de démontrer que la condition de stabilité est équivalente à ce que la qualité optimale de l'apprentissage soit une fonction croissante de la quantité d'exemples dont on dispose pour cet apprentissage.

Notons que lorsque la fonction $\alpha(R_{\rm opt})$ donnée par (1.65) n'est pas monotone, plusieurs solutions différentes pour $R_{\rm opt}$ existent pour une même valeur de α. Dans ce cas, il y en a généralement au moins une qui est instable et qui ne doit pas être prise en compte. Cette solution n'est toutefois jamais la solution $R_{\rm opt}$ de qualité maximale. Dans les applications que nous avons considérées, la solution correspondante à la valeur maximale de $R_{\rm opt}$ vérifie toujours la condition de stabilité locale.

Chapitre 2

Apprentissage Supervisé

Dans ce chapitre, nous allons aborder un type particulier d'apprentissage. Dans le cas général précédemment présenté, l'apprentissage consistait à trouver des caractéristiques générales d'une distribution d'exemples. Dans ce chapitre, nous allons nous intéresser à ce que l'on appelle généralement l'*apprentissage supervisé*. L'apprentissage supervisé ne consiste plus à déterminer la distribution des exemples, qui sera supposée connue, mais consiste à classer les exemples. Ici encore, l'on dispose d'un ensemble d'apprentissage qui constitue l'information disponible. Cette fois-ci, à chaque exemple de l'ensemble d'apprentissage est associé une classe. On supposera que cette classe est donnée par ce que l'on appelle un *professeur* qui nous est inconnu. Le but de l'apprentissage supervisé est de trouver un *élève* qui classe correctement les exemples, c'est-à-dire, qui leur donne la même classe que le professeur.

Deux questions différentes peuvent avoir de l'intérêt. La première consiste à trouver un élève qui classe correctement les exemples de l'ensemble d'apprentissage sans se préoccuper de la réponse donnée par l'élève à d'autres exemples. C'est ce que l'on appelle l'apprentissage par cœur. La seconde consiste, à partir des exemples de l'ensemble d'apprentissage uniquement, à trouver un élève capable de classer de la même façon que le professeur des exemples qui ne font pas partie de l'ensemble d'apprentissage. Cette propriété est dite de généralisation. Nous allons nous intéresser dans ce chapitre à la deuxième question. Ces deux questions seront abordées à nouveau dans la deuxième partie de la thèse.

2.1 Présentation du problème

Dans ce chapitre, nous allons nous intéresser à un cas particulier d'apprentissage supervisé, le plus simple que l'on puisse imaginer. Pour cela, on va commencer par supposer que la dimension de l'espace des données est N et que la distribution des exemples est connue et se réduit à une simple gaussienne normale dans toutes les directions :

$$P(\boldsymbol{\xi}) = \frac{1}{(2\pi)^{N/2}} \exp\left(-\frac{\boldsymbol{\xi}\cdot\boldsymbol{\xi}}{2}\right). \tag{2.1}$$

La restriction dans le choix de la distribution des exemples simplifie l'étude du cas particulier qui nous intéresse dans ce chapitre. Dans les applications réalistes, la distribution des exemples est généralement inconnue. Les hypothèses effectuées sur cette densité peuvent s'avérer essentielles. Une discussion sur le choix de celle-ci sera effectuée dans la deuxième partie de la thèse où nous tenterons d'étudier des problèmes plus réalistes.

Plusieurs hypothèses sont effectuées sur le professeur, c'est-à-dire, sur le type de classification que l'élève doit apprendre. La première consiste à considérer que les exemples appartiennent à deux classes distinctes, que l'on appellera $+1$ et -1. Cette restriction sur le nombre de classes n'est pas importante. Premièrement, de nombreuses applications réalistes comportent uniquement deux classes. C'est le cas, par exemple, lorsque la classe correspond à la réponse par oui ou non à une question posée. Un problème, étudié dans notre laboratoire, est l'aide au diagnostic médical du cancer du sein [115]. Les deux classes sont alors : sujet sain ou sujet malade. Dans ce problème, un exemple est constitué de diverses caractéristiques médicales concernant des prélèvements cytologiques (épaisseur de l'échantillon, uniformité de la taille des cellules, de la forme, etc.). Chaque caractéristique correspond à une dimension de l'espace des données. Deuxièmement, s'il existe plus de deux classes différentes, il est toujours possible de séparer une classe par rapport aux autres. Le problème se ramène alors à la séparation en deux classes.

La deuxième hypothèse que nous ferons sur le professeur est beaucoup plus restrictive. Elle consiste à considérer que les exemples sont linéairement séparables. C'est à dire que le professeur consiste en un hyperplan de dimension $N-1$ qui sépare les exemples de classe $+1$ des exemples de classe -1. Cette restriction peut se formaliser simplement en disant qu'il existe une direction privilégiée $\mathbf{B}$ telle que tous les exemples ayant un produit scalaire positif avec cette direction seront classés $+1$ tandis que les autres seront classés -1. Cette direction privilégiée est la direction perpendiculaire à l'hyperplan séparateur. Nous avons supposé, de plus, que cet hyperplan passe par l'origine. Avec cette restriction, les deux classes ont une probabilité égale. En particulier, un exemple $-\boldsymbol{\xi}$ a une classe opposée à celle de l'exemple $\boldsymbol{\xi}$. Cette symétrie entre les exemples $\boldsymbol{\xi}$ et $-\boldsymbol{\xi}$ est nécessaire par la suite pour reformuler ce problème dans le cadre général du chapitre précédent. Si l'on note $\tau(\boldsymbol{\xi})$, la classe donnée par le professeur à un exemple $\boldsymbol{\xi}$, alors celle-ci s'écrit :

$$\tau(\boldsymbol{\xi}) = \mathrm{sign}(\mathbf{B}\cdot\boldsymbol{\xi}). \tag{2.2}$$

Dans la formulation du problème général, on supposait que la distribution des exemples possédait une symétrie axiale, dont la direction privilégiée $\mathbf{B}$, correspondante à l'axe de symétrie, était inconnue. Ici, nous supposons connu le fait que les exemples sont linéairement séparables. Ceci introduit naturellement une symétrie axiale qui permet de considérer l'apprentissage (supervisé) de la classification comme un cas particulier de la recherche d'une direction privilégiée, formulé au Chapitre 1. L'hyperplan séparateur, ou encore, la direction privilégiée $\mathbf{B}$ perpendiculaire à cet hyperplan séparateur et qui caractérise le professeur, est inconnue.

Pour ce problème, l'apprentissage supervisé se réduit à déterminer la direction privilégiée $\mathbf{B}$ à partir de l'ensemble d'apprentissage $\mathcal{L}_\alpha$, constitué de P exemples et des classes correspondantes, attribuées par le professeur :

$$\mathcal{L}_\alpha = \{\boldsymbol{\xi}^\mu, \tau(\boldsymbol{\xi}^\mu)\}_{\mu=1,\cdots,P}\,. \tag{2.3}$$

Il est évident que la procédure la plus simple pour l'apprentissage consiste à chercher un élève ayant les mêmes caractéristiques que le professeur, c'est-à-dire, classifiant les exemples par une séparation linéaire. Cela revient à chercher une direction $\mathbf{J}$ telle que la classe que l'élève donne à un exemple $\boldsymbol{\xi}$ soit donnée par :

$$\sigma(\boldsymbol{\xi}) = \mathrm{sign}\left(\mathbf{J}\cdot\boldsymbol{\xi}\right). \tag{2.4}$$

Comme nous l'avons vu dans l'introduction, ce type de classifieur est appelé *perceptron* [104, 105, 84]. Il constitue la brique élémentaire des réseaux de neurones. Ces réseaux de neurones sont très utilisés pour l'apprentissage de tâches de classification [3, 42, 52, 58, 93, 126]. Il est donc essentiel de comprendre en détail la brique

élémentaire qui les constituent. Il peut cependant apparaître surprenant que 40 ans après leur invention, de nouveaux résultats restent à découvrir mais cela peut aussi expliquer la richesse de ces réseaux de neurones pour l'apprentissage.

2.2 Notions d'erreur de classification

L'existence d'une classe pour chaque exemple permet d'introduire la notion d'erreur de classification des exemples. Un exemple $\boldsymbol{\xi}$ sera considéré mal classé si la classe attribuée par l'élève diffère de celle du professeur, c'est-à-dire, si $\sigma(\boldsymbol{\xi}) \neq \tau(\boldsymbol{\xi})$. L'erreur de classification d'un exemple dépend à la fois de l'élève et du professeur :

$$e(\boldsymbol{\xi},\mathbf{J},\mathbf{B}) = \Theta(-\sigma(\boldsymbol{\xi})\tau(\boldsymbol{\xi})). \tag{2.5}$$

$\Theta(x)$ est la fonction de Heaviside qui vaut 1 pour $x > 0$ et 0 sinon. Cette notion d'erreur de classification d'un exemple va permettre de caractériser l'apprentissage. Il est possible de définir deux notions d'erreur différentes pour caractériser la qualité de l'apprentissage [110, 126].

La première notion est appelée *erreur d'apprentissage*. Elle est définie comme la fraction des exemples de l'ensemble d'apprentissage qui sont mal classés par l'élève :

$$\varepsilon_t(\mathbf{J},\mathcal{L}_\alpha) = \frac{1}{P}\sum_{\mu=1}^{P} e(\boldsymbol{\xi}^\mu,\mathbf{J},\mathbf{B}). \tag{2.6}$$

Cette erreur dépend de l'élève et de l'ensemble d'apprentissage mais aussi du professeur. Cette dernière dépendance a été omise pour raison de simplicité.

La deuxième notion est appelée *erreur de généralisation*. Cette erreur est la moyenne de l'erreur de classification d'un exemple sur tout l'espace des exemples. La distribution utilisée pour calculer cette moyenne est la même que celle qui a servi à déterminer les exemples de l'ensemble d'apprentissage.

$$\varepsilon_g(\mathbf{J}) = \int e(\boldsymbol{\xi},\mathbf{J},\mathbf{B})\ P(\boldsymbol{\xi})\ d\boldsymbol{\xi}. \tag{2.7}$$

Cette erreur de généralisation dépend de l'élève et du professeur (là encore la dépendance vis-à-vis du professeur est omise) mais ne dépend pas de l'ensemble d'apprentissage. Il est possible d'obtenir une expression directe en fonction de $\mathbf{J}$ et $\mathbf{B}$ de cette erreur de généralisation [126] :

$$\varepsilon_g(\mathbf{J}) = \frac{1}{\pi}\arccos(\mathbf{J}\cdot\mathbf{B}). \tag{2.8}$$

En effet, un exemple $\boldsymbol{\xi}$ est mal classé par l'élève si le produit scalaire $\boldsymbol{\xi}\cdot\mathbf{J}$ entre l'exemple $\boldsymbol{\xi}$ et la direction de l'élève $\mathbf{J}$ est de signe opposé au produit scalaire $\boldsymbol{\xi}\cdot\mathbf{B}$ entre l'exemple et la direction du professeur $\mathbf{B}$. L'ensemble des exemples mal classés forment deux secteurs de l'espace des données d'angle $\theta = \arccos(\mathbf{J}\cdot\mathbf{B})$ comme la figure 2.1 le représente dans le plan des vecteurs $\mathbf{J}$ et $\mathbf{B}$. La distribution des exemples étant invariante par rotation autour de l'origine, la fraction de volume occupée par ces deux secteurs de l'espace des données correspond à l'erreur de généralisation. En effet, chaque secteur d'angle donné correspond à la même fraction d'exemples. Cette relation (2.8) pour l'erreur de généralisation ne serait plus valable si l'on supposait, comme pour le problème général du Chapitre 1, que la distribution des exemples est perturbée selon la direction privilégiée. L'équation (2.8) montre que, de la même façon que pour l'apprentissage non supervisé présenté au chapitre précédent, la qualité de l'apprentissage d'un élève $\mathbf{J}$ est caractérisée par le produit scalaire $R = \mathbf{J}\cdot\mathbf{B}$.

FIG. 2.1 – *Représentation dans le plan* (**J**,**B**) *des deux secteurs correspondants aux exemples mal classés par l'élève* **J** *pour un professeur* **B** *(parties grisées).*

L'intérêt dans ce problème d'apprentissage supervisé ne consiste pas à minimiser l'erreur d'apprentissage ε_t car de nombreuses directions permettent d'obtenir une erreur d'apprentissage nulle [124]. De plus, de nombreux algorithmes permettent de déterminer de telles directions [110, 126, 6, 94, 48, 100, 20, 77]. Par contre, la minimisation de l'erreur de généralisation ε_g est un problème beaucoup plus difficile. L'erreur de généralisation dépend de tous les exemples possibles alors que nous ne disposons que du nombre restreint des exemples de l'ensemble d'apprentissage pour déterminer la direction privilégiée. Ce nombre restreint entraîne, comme dans le chapitre précédent, que l'apprentissage parfait n'est généralement pas envisageable. Dans le cas d'un apprentissage supervisé, l'apprentissage parfait consisterait à obtenir une erreur de généralisation nulle. Ceci n'étant pas possible, nous allons essayer de minimiser cette erreur. On peut remarquer que minimiser l'erreur de généralisation revient à maximiser le produit scalaire entre la direction de l'élève **J** et celle du professeur **B** d'après l'équation (2.8). Le but de l'apprentissage supervisé consiste alors à déterminer une direction $\mathbf{J}^*$ pour l'élève qui soit la plus proche possible de la direction **B**. Le problème est similaire à celui présenté dans le premier chapitre. Il est d'ailleurs possible de reformuler ce problème d'apprentissage supervisé comme un problème non supervisé.

2.3 Reformulation du problème supervisé

La reformulation du problème d'apprentissage supervisé en un problème non supervisé a été proposée récemment par Reimann et Van den Broeck [101]. Cette reformulation consiste à considérer non pas le couple de variables $\{\boldsymbol{\xi},\tau(\boldsymbol{\xi})\}$ mais la variable $\boldsymbol{\zeta} = \tau\boldsymbol{\xi}$. La distribution de cette variable s'exprime de la façon suivante :

$$P(\boldsymbol{\zeta}|\mathbf{B}) = \int d\boldsymbol{\xi}\, P(\boldsymbol{\xi})\, \delta\left(\boldsymbol{\zeta} - \tau(\boldsymbol{\xi})\boldsymbol{\xi}\right) \tag{2.9}$$

avec $P(\boldsymbol{\xi})$ donnée par (2.1) et $\tau(\boldsymbol{\xi})$ par (2.2). La dépendance vis-à-vis de **B** provient de l'expression de $\tau(\boldsymbol{\xi})$. La distribution de $\boldsymbol{\zeta}$ peut se mettre sous la forme générale introduite au premier chapitre :

$$P(\boldsymbol{\zeta}|\mathbf{B}) = \frac{1}{(2\pi)^{N/2}} \exp\left(-\frac{\boldsymbol{\zeta}\cdot\boldsymbol{\zeta}}{2} - V^*(\mathbf{B}\cdot\boldsymbol{\zeta})\right), \tag{2.10}$$

$$= \frac{1}{(2\pi)^{N/2}} \exp\left(-\frac{\boldsymbol{\zeta}\cdot\boldsymbol{\zeta}}{2}\right) 2\,\Theta\left(\mathbf{B}\cdot\boldsymbol{\zeta}\right). \tag{2.11}$$

La détermination de la direction privilégié $\mathbf{B}$ à partir de l'ensemble d'apprentissage :

$$\mathcal{L}_\alpha = \{\boldsymbol{\zeta}^\mu\}_{\mu=1,\cdots,P} \tag{2.12}$$

avec les exemples $\boldsymbol{\zeta}^\mu$ distribués selon la distribution (2.11) est un cas particulier du problème de l'apprentissage non supervisé présenté au premier chapitre. De la distribution de $\boldsymbol{\zeta}$, on déduit la perturbation V^* :

$$V^*(\lambda) = \begin{cases} -\ln 2 & \text{si } \lambda > 0, \\ +\infty & \text{sinon.} \end{cases} \tag{2.13}$$

Sachant que l'on s'intéresse à l'apprentissage optimal, il faut s'assurer que les deux formulations, supervisée et non supervisée, du problème sont équivalentes. Cette propriété est garantie par la symétrie qu'il existe entre les exemples $\boldsymbol{\xi}$ et $-\boldsymbol{\xi}$: la classe de l'exemple $\boldsymbol{\xi}$ est l'opposée de la classe de $-\boldsymbol{\xi}$.

Il faut noter que l'étude des propriétés de l'apprentissage optimal pour le problème supervisé a commencé avant la reformulation de ce problème en un problème non supervisé.

Opper et Haussler [92] ont utilisé une méthode basée sur la définition d'apprentissage bayesien qui leur a permis d'obtenir les caractéristiques de l'apprentissage optimal dans la limite thermodynamique ($N \to +\infty$ et $P \to +\infty$ avec $\alpha = P/N$ constant). En effet, un classifieur bayesien est équivalent à un ensemble infini d'élèves qui votent, chacun ayant appris le même ensemble d'apprentissage. La classe attribuée à un exemple nouveau est celle de la majorité des élèves. Ces auteurs ont notamment déterminé l'erreur de généralisation minimale que l'on peut obtenir étant donné une taille réduite α, à partir des propriétés d'un ensemble infini d'élèves ayant appris à l'aide de l'algorithme de Boltzmann (appelé algorithme de Gibbs dans les publications plus récentes) [52, 53, 114]. Cet algorithme consiste à sélectionner de manière équiprobable une direction parmi celles pour lesquelles l'erreur d'apprentissage ε_t est nulle. Puisqu'on considère des élèves ayant la même structure que le professeur, cet ensemble de directions, appelé l'*espace des versions*, n'est pas vide puisqu'il contient la direction privilégiée $\mathbf{B}$. Leur approche ne permet toutefois pas d'en déduire un algorithme simple pour déterminer la direction $\mathbf{J}^*$ pour laquelle l'erreur de généralisation est optimale. En fait, cette approche ne démontre pas l'existence d'une telle direction optimale car elle ne démontre pas qu'il existe un élève dont la classification coïncide avec le résultat du vote, quelque soit l'exemple.

Watkin [124] a donné la preuve de l'existence d'une direction optimale en montrant que cette direction était la moyenne de toutes les directions correspondantes aux élèves ayant appris à l'aide de l'algorithme de Gibbs, c'est-à-dire, la moyenne des directions de l'espace des versions. Un argument de convexité de cet espace permet de prouver que la direction optimale vérifie, elle aussi, la propriété de correspondre à une erreur d'apprentissage nulle. Cette procédure nécessite toujours la moyenne sur une infinité de directions, et ne permet toujours pas d'obtenir un algorithme simple pour déterminer la direction optimale.

L'approche variationnelle que nous avons présentée au premier chapitre permet d'obtenir un algorithme simple capable de déterminer la direction optimale,

qui minimise l'erreur de généralisation. Cette approche variationnelle permet de déterminer le potentiel optimal $V_{\rm opt}$. La direction optimale est alors obtenue en minimisant la fonction de coût correspondante à ce potentiel. Cette minimisation s'effectue aisément par une simple descente de gradient. Cette approche a été proposée indépendamment par Kinouchi et Caticha [62] ainsi que par nous-mêmes [21, 26, 25] pour le cas particulier de l'apprentissage supervisé, avant d'être généralisée au problème de l'apprentissage non supervisé par Reimann et Van den Broeck [101, 120].

2.4 Erreur de généralisation minimale

Revenons maintenant à l'erreur de généralisation minimale. La reformulation du problème d'apprentissage supervisé a permis de déterminer la perturbation V^*, donnée par l'équation (2.13), correspondante au problème équivalent d'apprentissage non supervisé. Cette perturbation permet de calculer $R_{\rm opt}$ à partir de l'équation (1.65). On obtient :

$$\alpha = \pi R_{\rm opt}^2 \left\{ \int_{-\infty}^{+\infty} Dy \, \frac{\exp\left(-y^2/T^2\right)}{H(-y/T)} \right\}^{-1} \tag{2.14}$$

avec :

$$T^2 = \frac{1 - R_{\rm opt}^2}{R_{\rm opt}^2}, \tag{2.15}$$

$$H(x) = \int_x^{+\infty} Dy, \tag{2.16}$$

$$Dy = \exp\left(-\frac{y^2}{2}\right) \frac{dy}{\sqrt{2\pi}}. \tag{2.17}$$

De cette valeur optimale $R_{\rm opt}$, on déduit l'erreur de généralisation minimale :

$$\varepsilon_g(\alpha) = \frac{1}{\pi} \arccos\left(R_{\rm opt}(\alpha)\right). \tag{2.18}$$

Cette erreur de généralisation minimale est représentée sur la figure 2.2 en trait plein. Les carrés correspondent à l'erreur de généralisation obtenue par des simulations numériques.

Il est intéressant de faire quelques remarques sur cette erreur de généralisation minimale. Tout d'abord, il est à noter que $R_{\rm opt} > 0$ pour toute valeur de $\alpha > 0$. En conséquence, l'erreur de généralisation est strictement inférieure à $1/2$. En fait, pour les petites valeurs de α,

$$R_{\rm opt}(\alpha) \sim \sqrt{\frac{2\,\alpha}{\pi}}, \tag{2.19}$$

$$\varepsilon_g(\alpha) \sim \frac{1}{2} - \sqrt{\frac{2\,\alpha}{\pi^3}}. \tag{2.20}$$

Ceci signifie que l'on apprend dès que la taille réduite α de l'ensemble d'apprentissage est non nulle. Il faut toutefois noter que cette taille réduite est le rapport entre le nombre P d'exemples et la dimension N de l'espace des données. Dans la limite thermodynamique ($N \to +\infty$ et $P \to +\infty$), une taille réduite α même très faible correspond à une infinité d'exemples dans l'ensemble d'apprentissage.

FIG. 2.2 – *Erreur de généralisation optimale ε_g en fonction de la taille réduite α de l'ensemble d'apprentissage. Les carrés représentent les valeurs obtenues par des simulations numériques décrites dans le paragraphe suivant. Les barres d'erreur sont plus petites que la taille des carrés.*

Une autre remarque intéressante est que R_{opt} tend vers 1 lorsque α diverge. L'erreur de généralisation optimale s'annule lorsque $\alpha \to +\infty$. Le comportement de l'erreur de généralisation minimale est donné par :

$$\varepsilon_g(\alpha) \sim \frac{0.442}{\alpha}. \tag{2.21}$$

De nombreux algorithmes permettent d'obtenir des directions pour lesquelles l'erreur de généralisation s'annule lorsque α diverge. Dans certains cas [20, 48, 52, 124], l'erreur de généralisation décroît aussi proportionnellement à $1/\alpha$. Seul le coefficient de proportionnalité est différent et supérieur à 0.442. Deux algorithmes simples, dont on reparlera par la suite, possèdent cette propriété. Il s'agit premièrement de l'algorithme dit de Gibbs [6, 53, 92, 101, 114, 124]. Celui-ci consiste à choisir de manière aléatoire et équiprobable une direction parmi celles pour lesquelles l'erreur d'apprentissage ε_t est nulle. Cette solution correspond à une erreur de généralisation $\varepsilon_g \sim 0.625/\alpha$ lorsque α diverge [92, 124].

Le deuxième algorithme, dit de stabilité maximale, consiste à déterminer une direction particulière de l'espace des versions [65, 67, 48]. Pour définir cette direction particulière, il me faut tout d'abord introduire ce que l'on appelle la *stabilité* d'un exemple. Cette stabilité est la distance de l'exemple à l'hyperplan séparateur correspondant à l'élève. Cette distance γ peut s'écrire à l'aide de la direction $\mathbf{J}^*$ de l'élève :

$$\gamma = \tau \boldsymbol{\xi} \cdot \mathbf{J}^*. \tag{2.22}$$

La direction de stabilité maximale est celle pour laquelle la stabilité de l'exemple de l'ensemble d'apprentissage le plus proche de l'hyperplan séparateur est maximale. Cette solution possède une erreur de généralisation qui s'annule comme $\varepsilon_g \sim 0.5005/\alpha$ lorsque α diverge [48]. Nous reviendrons à cette notion de stabilité et,

FIG. 2.3 – *Potentiel optimal $V_{\rm opt}$ pour différentes valeurs de T. $T = 0.5$,1 et 2 de gauche à droite. Ces valeurs correspondent respectivement à $\alpha = 1.67$,1.04 et 0.61.*

notamment, à la distribution des stabilités des exemples de l'ensemble d'apprentissage. Cette quantité, accessible par la méthode des répliques, permet de comprendre les caractéristiques de la direction optimale.

2.5 Caractéristiques du potentiel optimal

Nous allons maintenant discuter des caractéristiques du potentiel optimal $V_{\rm opt}$. Celui-ci est obtenu par intégration de $V'_{\rm opt}(\lambda)$ défini par l'équation (1.68). On obtient :

$$V_{\rm opt}(\lambda) = \int_{y(\lambda)}^{+\infty} g(z)\,(1 + g'(z))\,dz \tag{2.23}$$

où $y(\lambda)$ est la fonction inverse de $\lambda(y)$ définie par $g(y) \equiv \lambda(y) - y$. L'introduction de V^*, donné par (2.13), dans (1.61) permet de calculer $g(y)$:

$$g(y) = T^2 \frac{d}{dy} \ln H\left(-\frac{y}{T}\right) \tag{2.24}$$

La figure 2.3 représente le potentiel optimal pour diverses valeurs de T ($T =$ 0.5,1 et 2). Ces valeurs de T correspondent à des valeurs de $\alpha = 1.67$,1.04 et 0.61 respectivement.

Les propriétés du minimum de la fonction de coût ne dépendent pas de l'origine des énergies choisie pour le potentiel optimal. En effet, un changement de cette origine correspond à l'ajout d'une constante indépendante des exemples. L'origine des énergies a été choisie arbitrairement de telle façon que $V_{\rm opt}$ s'annule lorsque $\lambda \to +\infty$.

Il est intéressant de noter que le potentiel $V_{\rm opt}$ est infini pour les valeurs négatives de λ. Il diverge logarithmiquement pour les petites valeurs de $\lambda > 0$:

$$V_{\text{opt}}(\lambda) \sim -T^2 \ln(\lambda). \tag{2.25}$$

Cette propriété entraîne une divergence de la fonction de coût pour toute direction $\mathbf{J}$ telle qu'il existe un exemple $\boldsymbol{\xi}^\mu$ de l'ensemble d'apprentissage pour lequel $\lambda^\mu = \boldsymbol{\zeta}^\mu \cdot \mathbf{J} = \tau(\boldsymbol{\xi}^\mu)\, \boldsymbol{\xi}^\mu \cdot \mathbf{J} \leq 0$. Sachant que $\tau(\boldsymbol{\xi}^\mu) = \text{sign}(\boldsymbol{\xi}^\mu \cdot \mathbf{B})$, $\lambda^\mu < 0$ signifie que l'élève $\mathbf{J}$ classe mal l'exemple $\boldsymbol{\xi}^\mu$. De tels élèves n'appartiennent pas à l'espace des versions. La fonction de coût optimale diverge donc pour toute direction n'ayant pas une erreur d'apprentissage nulle. Ceci confirme le fait que la direction optimale est une direction pour laquelle l'erreur d'apprentissage est nulle. Cette divergence entraîne des complications algorithmiques lors de la minimisation de la fonction de coût. Cette minimisation peut être effectuée par une simple descente de gradient, à condition que la condition initiale corresponde à une direction pour laquelle l'erreur d'apprentissage est nulle. Dans le cas contraire, la fonction de coût serait infinie. En pratique, il faudra, dans un premier temps, utiliser un algorithme permettant de trouver une direction sans erreur d'apprentissage. Pour cela, nous utiliserons, lors des simulations numériques, l'algorithme minimerror [100, 115, 116]. Dans un second temps, nous utiliserons une descente de gradient pour minimiser la fonction de coût optimale. Cette minimisation sera d'ailleurs facilitée par la convexité du potentiel optimal. En effet, il existe alors un unique minimum.

Une dernière remarque concerne le paramètre T défini par (2.15). Il dépend de la taille réduite α par l'intermédiaire de R_{opt}, et correspond à la portée du potentiel optimal pour les valeurs positives de λ. En effet, le potentiel s'annule de manière exponentielle pour les grandes valeurs de λ :

$$V_{\text{opt}} \sim \frac{T^3}{\lambda} \exp\left(-\frac{\lambda^2}{2T^2}\right). \tag{2.26}$$

La portée du potentiel optimal est décroissante par rapport à α ($T \sim 1.39/\alpha$ lorsque $\alpha \to +\infty$). Ce résultat n'est pas surprenant puisque plus le nombre d'exemples de l'ensemble d'apprentissage est élevé plus il existe d'exemples proches de l'hyperplan séparateur. Or, ce sont ces exemples qui caractérisent le mieux la direction privilégiée $\mathbf{B}$.

2.6 Distribution des stabilités

Une quantité intéressante, accessible par la méthode des répliques, est la distribution des stabilités des exemples de l'ensemble d'apprentissage. La stabilité γ^μ d'un exemple $\boldsymbol{\xi}^\mu$ s'exprime comme le produit scalaire entre $\boldsymbol{\zeta}^\mu$ et la direction de l'élève $\mathbf{J}^*$:

$$\gamma^\mu = \boldsymbol{\zeta}^\mu \cdot \mathbf{J}^* = \tau(\boldsymbol{\xi}^\mu)\ \boldsymbol{\xi}^\mu \cdot \mathbf{J}^*. \tag{2.27}$$

Ce paramètre est négatif pour les exemples mal classés par l'élève. En effet, dans ce cas $\sigma(\boldsymbol{\xi}^\mu) = \text{sign}(\boldsymbol{\xi}^\mu \cdot \mathbf{J}^*)$ et $\tau(\boldsymbol{\xi}^\mu)$ sont de signes opposés. Dans le cas où l'exemple est bien classé, la stabilité correspond alors à la distance de l'exemple à l'hyperplan séparateur de l'élève. Comme nous l'avons vu précédemment, l'erreur d'apprentissage de l'élève optimal est nulle. Tous les exemples de l'ensemble d'apprentissage étant bien classés, les stabilités de ces exemples sont positives.

La distribution des stabilités des exemples de l'ensemble d'apprentissage par rapport à l'élève optimal s'exprime de la façon suivante :

$$\rho_{\text{opt}}(\gamma) = \frac{1}{P} \sum_{\mu=1}^{P} \delta(\gamma - \gamma^\mu). \tag{2.28}$$

FIG. 2.4 – *Distribution des stabilités des exemples de l'ensemble d'apprentissage par rapport à l'élève optimal pour plusieurs valeurs de α. La stabilité d'un exemple est la distance de cet exemple à l'hyperplan séparateur de l'élève optimal. Pour comparer, il est représenté la distribution des distances des exemples à l'hyperplan séparateur du professeur.*

avec γ^μ défini par (2.27). La méthode des répliques permet de déterminer cette distribution [1, 60, 4, 66, 49, 48]. Celle-ci s'exprime de la façon suivante [21, 25, 26] :

$$\rho_{\rm opt}(\gamma) = 2\int_{-\infty}^{+\infty} Dy\, H\left(-\frac{y}{T}\right)\, \delta\left(\lambda(y)-\gamma\right), \tag{2.29}$$

$$= \sqrt{\frac{2}{\pi}}\, \exp\left(-\frac{y^2(\gamma)}{2}\right)\, H\left(-\frac{y(\gamma)}{T}\right) y'(\gamma) \tag{2.30}$$

avec $\lambda(y)$ définie par l'équation (2.24), $y(\lambda)$ la fonction inverse de $\lambda(y)$ et enfin $y'(\lambda)$ la dérivée de $y(\lambda)$ par rapport à λ.

Nous allons comparer la distribution $\rho_{\rm opt}(\gamma)$ à celle des distances des exemples de l'ensemble d'apprentissage par rapport à l'hyperplan séparateur correspondant au professeur. Cette distribution $\rho_{\rm prof}$ est celle de la variable $\gamma^\mu = \boldsymbol{\zeta}^\mu \cdot \mathbf{B}$. Elle se déduit directement de la distribution des exemples (2.11); elle est nulle pour les valeurs négatives de γ et le double de la gaussienne normale pour les valeurs positives :

$$\rho_{\rm prof}(\gamma) = \sqrt{\frac{2}{\pi}}\, \exp\left(-\frac{\gamma^2}{2}\right)\, \Theta(\gamma). \tag{2.31}$$

La figure 2.4 illustre l'évolution de la distribution $\rho_{\rm opt}(\gamma)$ en fonction de la taille réduite α de l'ensemble d'apprentissage. La distribution $\rho_{\rm opt}(\gamma)$ est représentée pour plusieurs valeurs différentes de α (α = 1,2,4,6 et 14). La distribution $\rho_{\rm prof}(\gamma)$ est aussi représentée afin de comparer.

Plusieurs caractéristiques intéressantes de la distribution $\rho_{\rm opt}(\gamma)$ sont à noter. La première est l'existence d'une région apparemment exempte d'exemples pour les faibles distances. Cette région n'est en réalité pas totalement vide d'exemples

comme pourrait le laisser penser la figure 2.4. La densité $\rho_{\rm opt}(\gamma)$ est non nulle pour toute valeur positive de γ. Toutefois, cette distribution s'annule exponentiellement pour les faibles valeurs de γ :

$$\rho_{\rm opt}(\gamma) \sim \frac{T}{\pi\gamma} \exp\left(-\frac{T^2}{2R_{\rm opt}^2\gamma^2}\right). \tag{2.32}$$

Cette région, presque vide d'exemples, est d'autant plus étendue que la taille réduite α est faible. Ce comportement est à rapprocher de celui de l'élève de stabilité maximale [48]. Comme il a été mentionné à la fin du paragraphe 2.4, l'orientation de l'hyperplan de stabilité maximale est celle qui maximise la stabilité des exemples les plus proches de l'hyperplan. Cette stabilité, notée $\kappa_{\rm max}$, est appelée stabilité ou marge maximale. Ainsi une région cette fois-ci complètement exempte d'exemples existe au voisinage de l'hyperplan, pour $\gamma < \kappa_{\rm max}$. La distribution des stabilités des exemples de l'ensemble d'apprentissage pour la solution de stabilité maximale, $\rho_{\rm SM}(\gamma)$, est simple. Elle se réduit à une distribution ρ_1 pour $\gamma > \kappa_{\rm max}$ et à un pic $\delta(\kappa_{\rm max} - \gamma)$ à la valeur $\kappa_{\rm max}$ qui contient le reste des exemples :

$$\rho_{\rm SM}(\gamma) = \rho_1(\gamma;\alpha)\,\Theta(\gamma - \kappa_{\rm max}) + \rho_0(\alpha)\,\delta(\gamma - \kappa_{\rm max}) \tag{2.33}$$

avec :

$$\rho_1(\gamma;\alpha) = \sqrt{\frac{2}{\pi}}\,H\left(-\frac{\gamma R_{\rm SM}}{\sqrt{1-R_{\rm SM}^2}}\right)\exp\left(-\frac{\gamma^2}{2}\right). \tag{2.34}$$

$\rho_0(\alpha)$ est déterminé par la normalisation de la distribution $\rho_{\rm SM}$. Le paramètre $R_{\rm SM}$ est l'analogue de $R_{\rm opt}$ pour la solution de stabilité maximale, c'est-à-dire, $R_{\rm SM} = \mathbf{J}_{\rm SM} \cdot \mathbf{B}$ où $\mathbf{J}_{\rm SM}$ est la direction de l'élève de stabilité maximale. $\kappa_{\rm max}$ et $R_{\rm SM}$ dépendent de α. Notons que cette solution de stabilité maximale n'est pas optimale au sens de l'optimalité que l'on a considéré (erreur de généralisation minimale). Par contre, cette solution possède de nombreux avantages, notamment, au niveau de la stabilité de la classification vis-à-vis de la corruption des exemples de l'ensemble d'apprentissage par du bruit. Cette notion sera discutée dans la deuxième partie de la thèse.

La deuxième remarque que l'on peut faire sur $\rho_{\rm opt}(\gamma)$ est l'accord avec $\rho_{\rm prof}(\gamma)$ pour les grandes valeurs de γ. Cet accord est de plus en plus important lorsque la taille réduite α augmente. Dans la limite d'ensembles d'apprentissage de taille réduite infinie, on retrouve d'ailleurs que les deux distributions sont identiques : l'apprentissage est parfait et l'erreur de généralisation nulle.

La troisième remarque concerne la région intermédiaire. Un maximum de la distribution $\rho_{\rm opt}(\gamma)$ est observé. Il reflète l'absence d'exemples pour de petites valeurs de γ ; ces exemples se trouvent repoussés à des valeurs de γ plus grandes. Le maximum de la distribution est à peu près constant tandis que sa position tend vers l'origine lorsque la taille réduite diverge. Si on note $\gamma_{\rm M}$ la position de ce maximum, on trouve $\gamma_{\rm M} \sim 1.77/\alpha$ pour les grandes valeurs de α. Cette valeur est à comparer avec la valeur de la stabilité maximale $\kappa_{\rm max} \sim 1.004/\alpha$ de l'élève de stabilité maximale. Le maximum de $\rho_{\rm opt}(\gamma)$ est plus éloigné de l'origine que le pic $\delta(\kappa_{\rm max} - \gamma)$ de $\rho_{\rm SM}(\gamma)$.

La conclusion que l'on peut tirer de ces caractéristiques ainsi que de la comparaison avec la distribution des stabilités de l'élève de stabilité maximale concerne la position de l'élève optimal dans l'espace des versions. Cet espace des versions est une partie connexe et convexe de l'espace des directions $\mathbf{J}$ (la sphère de rayon 1 de l'espace de dimension N). L'élève de stabilité maximale est situé à l'intérieur de l'espace des versions, le plus loin possible de la surface de l'espace des versions.

Par contre, du fait de l'existence d'exemples à une distance aussi faible que possible (même si ceux-ci ont une probabilité extrêmement faible), on peut conclure que l'élève optimal se trouve à la bordure de l'espace des versions. Le professeur est situé sur la surface même de l'espace des versions puisque la distribution $\rho_{\rm prof}(\gamma)$ est finie pour $\gamma = 0$. Comme nous l'avons déjà dit, Watkin [124] a démontré que l'élève optimal était l'isobarycentre de l'espace des versions. Il peut paraître contradictoire que l'élève optimal se trouve à la bordure de l'espace des versions et qu'il soit l'isobarycentre de cet espace. Il ne faut toutefois pas oublier que ces résultats sont obtenus dans la limite thermodynamique, c'est-à-dire, dans la limite d'un espace des données de dimension infinie. Dans ce cas, le poids de la surface est prépondérant par rapport à celui du volume. La distribution $\rho_{\rm opt}(\gamma)$ obtenue par des simulations numériques à des valeurs finies de N montrent l'existence d'une région de faibles stabilités totalement vide d'exemples. Cette région disparaît lorsque la dimension N de l'espace des données diverge.

2.7 Simulations numériques

Dans cette section nous allons présenter les résultats de simulations numériques qui ont été effectuées sur un simple ordinateur de travail du laboratoire ainsi que sur la machine parallèle T3D du centre de calcul du CEA Grenoble pour les plus grands systèmes.

Dans ces simulations, on choisit un professeur et on tire un ensemble d'apprentissage constitué de P exemples et de leurs classes, données par le professeur. On utilise alors la fonction de coût (1.6) avec le potentiel optimal $V_{\rm opt}$ donné par (2.23) pour déterminer l'élève optimal qui minimise cette fonction de coût. Cette minimisation est effectuée par une descente de gradient.

Le but de ces simulations est double. Tout d'abord, il est intéressant de vérifier la pertinence de la méthode des répliques et des hypothèses effectuées dans la limite thermodynamique comme la symétrie des répliques ou l'automoyennage de l'énergie libre et des propriétés d'apprentissage telles que $R_{\rm opt}(\alpha)$ ou $\varepsilon_g(\alpha)$. Le deuxième intérêt de ces simulations est une étude plus systématique des effets dus à la dimension finie de l'espace des données. Ces effets ne sont pas accessibles aisément par la méthode des répliques et peuvent s'avérer essentiels dans les applications pratiques où la taille de l'espace des données est rarement supérieure à 1000.

2.7.1 Description de l'algorithme utilisé

Plusieurs valeurs de la taille réduite α de l'ensemble d'apprentissage ont été considérées (α = 1,2,4,6,8,10 et 14). Pour chacune de ces valeurs, de 7 à 10 valeurs différentes de la dimension de l'espace des données ont été considérées. Les valeurs de α et N étant fixées, le nombre d'exemples de l'ensemble d'apprentissage est donné par $P = \alpha N$.

La distribution des exemples, considérée pour ces simulations, est différente de celle présentée au paragraphe 2.1 (Eq.(2.1)). Nous avons considéré les exemples comme des vecteurs dont les composantes sont binaires ($+1$ ou -1 avec la même probabilité $1/2$). Cette modification de la distribution des exemples n'affecte pas les résultats obtenus dans la limite thermodynamique. En effet, les résultats analytiques sont valables pour toutes les distributions des exemples telles que :

$$\int \xi_i \; P(\boldsymbol{\xi}) d\boldsymbol{\xi} \;\; = \;\; \langle \xi_i \rangle = 0, \tag{2.35}$$

$$\int \xi_i \xi_j \; P(\boldsymbol{\xi}) d\boldsymbol{\xi} \;\; = \;\; \langle \xi_i \xi_j \rangle = \delta_{ij}. \tag{2.36}$$

où ξ_i est la composante i de $\boldsymbol{\xi}$ et δ_{ij} est le symbole de Kroneker qui vaut 0 si $i \neq j$ et 1 si $i = j$.

Même si les résultats obtenus dans la limite thermodynamique sont les mêmes pour les distributions gaussienne et binaire, on peut s'attendre à ce que les effets de taille finie dépendent du choix de la distribution. Nous avons choisi la distribution qui nous paraissait la plus simple à mettre en œuvre.

La direction privilégiée $\mathbf{B}$ qui est un vecteur unitaire de l'espace des données a été choisie aléatoirement pour chaque ensemble d'apprentissage. Les valeurs de N et P étant données, la direction optimale a été déterminée pour un grand nombre d'ensembles d'apprentissage $\mathcal{L}_\alpha$ différents afin d'obtenir les propriétés moyennes de l'erreur de généralisation. Le nombre d'ensembles d'apprentissage $M(N,P)$ a été choisi de telle façon que les barres d'erreurs sur la moyenne de l'erreur de généralisation soient à peu près identiques pour chaque couple (N,P). Dans tous les cas analysés, les barres d'erreurs sont inférieures à 1%.

La moyenne de l'erreur de généralisation est calculée de la façon suivante :

$$\varepsilon_g(\alpha,N) = \frac{1}{M}\sum_{i=1}^{M} \varepsilon_g(\mathcal{L}_\alpha^i) \tag{2.37}$$

avec $\mathcal{L}_\alpha^i$ l'ensemble d'apprentissage numéro i et :

$$\varepsilon_g(\mathcal{L}_\alpha) = \frac{1}{\pi}\arccos\left(\mathbf{B}\cdot\mathbf{J}(\mathcal{L}_\alpha)\right) \tag{2.38}$$

où $\mathbf{J}(\mathcal{L}_\alpha)$ est la direction qui minimise la fonction de coût optimale pour l'ensemble d'apprentissage $\mathcal{L}_\alpha$. La variance de cette erreur de généralisation est :

$$\sigma^2(\alpha,N) = \frac{1}{M}\sum_{i=1}^{M}\left(\varepsilon_g\left(\mathcal{L}_\alpha\right) - \varepsilon_g\left(\alpha,N\right)\right)^2. \tag{2.39}$$

Les barres d'erreur de $\varepsilon_g(\alpha,N)$ sont estimées à $\sigma(\alpha,N)/\sqrt{M}$.

L'hypothèse d'automoyennage de l'erreur de généralisation $\varepsilon_g(\mathcal{L}_\alpha)$ ou du produit scalaire $R = \mathbf{B}\cdot\mathbf{J}(\mathcal{L}_\alpha)$ est équivalente à celle de l'indépendance de ces mêmes paramètres vis-à-vis de l'ensemble d'apprentissage $\mathcal{L}_\alpha$ dans la limite thermodynamique. Cette hypothèse est vérifiée si la variance $\sigma(\alpha,N)$ s'annule lorsque $N \to +\infty$. Les barres d'erreur étant dépendantes de cette variance, on s'attend à ce que le nombre d'ensembles d'apprentissage nécessaire pour que l'erreur soit inférieure à 1% diminue lorsque N augmente. Dans la pratique, $M(N,P)$ est compris entre 500 pour les grandes valeurs de N et 20 000 pour les plus faibles.

Comme il a déjà été vu précédemment, la fonction de coût optimale est infinie pour les directions $\mathbf{J}$ pour lesquelles l'erreur d'apprentissage est non nulle. Il faut donc initialiser correctement la direction $\mathbf{J}(0)$ avant de commencer la minimisation de la fonction de coût. Cette initialisation est effectuée par l'intermédiaire d'un algorithme appelé Minimerror, qui a été développé dans notre laboratoire [100, 115]. Dans les cas d'ensembles d'apprentissage linéairement séparables, cet algorithme permet d'obtenir une direction $\mathbf{J}(0)$ pour laquelle l'erreur d'apprentissage est nulle. Il est basé sur la minimisation de la fonction de coût définie par le potentiel :

$$V(\lambda) = \frac{1}{2}\left(1 - \tanh\left(\frac{\beta\lambda}{2}\right)\right). \tag{2.40}$$

Lorsque le paramètre β diverge, cette fonction de coût se réduit à l'erreur d'apprentissage. Au cours de la minimisation, le paramètre β est adapté.

Revenons maintenant à la descente de gradient de la fonction de coût optimale. La direction $\mathbf{J}(k)$ obtenue à l'itération k est modifiée de la façon suivante :

$$\mathbf{J} = \mathbf{J}(k) - \epsilon(k)\,\delta\mathbf{J}, \tag{2.41}$$

$$\delta\mathbf{J} = \sum_{\mu=1}^{P} \frac{dV_{\rm opt}}{d\lambda}\left(\tau\left(\boldsymbol{\xi}^{\mu}\right)\,\boldsymbol{\xi}^{\mu}\cdot\mathbf{J}(k)\right)\,\tau\left(\boldsymbol{\xi}^{\mu}\right)\,\boldsymbol{\xi}^{\mu}, \tag{2.42}$$

$$\mathbf{J}(k+1) = \frac{\mathbf{J}}{\sqrt{\mathbf{J}\cdot\mathbf{J}}}. \tag{2.43}$$

Comme pour toute descente de gradient, la direction de l'élève à l'itération $k+1$ est déduite de celle à l'itération k par le simple retrait d'un vecteur $\delta\mathbf{J}$ (Eq.(2.42)) proportionnel à la dérivée de la fonction de coût. Afin que la direction de l'élève reste sur la sphère de rayon 1, on effectue une normalisation de celle-ci après chaque itération (Eq.(2.43)). Lorsque la direction de l'élève s'approche de celle qui minimise la fonction de coût, la dérivée de la fonction de coût s'annule. De la même façon, $\delta\mathbf{J}_{\perp}$, la partie de $\delta\mathbf{J}$ perpendiculaire à la direction $\mathbf{J}(k)$:

$$\delta\mathbf{J}_{\perp} \equiv \delta\mathbf{J} - \mathbf{J}(k)\,\delta\mathbf{J}\cdot\mathbf{J}(k) \tag{2.44}$$

est un vecteur dont la norme s'annule. Le critère d'arrêt de la minimisation, qui a été choisi, est $\delta\mathbf{J}_{\perp}\cdot\delta\mathbf{J}_{\perp} \leq 10^{-14}$.

Le paramètre $\epsilon(k)$ contrôle la vitesse de convergence. Si ce paramètre est trop faible, le pas à chaque itération est faible et le temps de convergence est très long. Par contre, si ce paramètre est trop grand, on observe des oscillations autour du minimum et on peut ne jamais converger. Pour éviter ces deux problèmes et d'avoir à ajuster, pour chaque couple (N,P), le paramètre $\epsilon(k)$, nous avons adapté ce paramètre au cours de la simulation. Pour cela, à l'itération $k+1$, nous avons calculé la fonction de coût pour trois valeurs différentes de ϵ : $5\epsilon(k)$,$\epsilon(k)$ et $\epsilon(k)/2$. Le paramètre $\epsilon(k+1)$ est celui des trois qui produit la direction $\mathbf{J}(k+1)$ pour laquelle la fonction de coût est la plus faible. Cette adaptation à chaque itération a permis d'accélérer la convergence significativement et de ne pas avoir à se préoccuper de la valeur initiale $\epsilon(0)$. Cette valeur initiale est égale à 10^{-2} pour toutes les simulations.

2.7.2 Erreur de généralisation de l'élève optimal

Les simulations numériques nous ont permis de confirmer les résultats obtenus par la méthode des répliques, dans la limite thermodynamique, pour l'erreur de généralisation optimale. Les carrés de la figure 2.2 représentent les extrapolations linéaires en fonction de $1/N$ de l'erreur de généralisation moyenne $\varepsilon_g(\alpha,N)$ pour toutes les valeurs de α simulées. La figure 2.5 représente $\varepsilon_g(\alpha,N)$ pour trois valeurs de la taille réduite α (α = 8,10 et 14) et pour les valeurs de N considérées. Représentées en fonction de $1/N$, ces moyennes s'extrapolent linéairement vers la valeur théorique prédite par la méthode des répliques. La figure 2.5 ne représente les résultats des simulations que pour trois valeurs de α sur toutes celles étudiées mais les résultats sont similaires pour les valeurs non représentées. Ces extrapolations linéaires de l'erreur de généralisation moyenne en fonction de $1/N$ permettent de déterminer les effets de la dimension finie de l'espace des données. Les corrections par rapport à la valeur théorique sont négatives. Au premier ordre en $1/N$, on peut écrire :

$$\varepsilon_g(\alpha,N) = \varepsilon_g(\alpha,+\infty) - \frac{\phi(\alpha)}{\alpha N}. \tag{2.45}$$

Le fait que les corrections soient négatives n'est pas surprenant. En effet, les composantes des exemples étant à valeurs binaires, il existe 2^N exemples différents possibles. Le nombre P d'exemples de l'ensemble d'apprentissage est αN. La fraction

FIG. 2.5 – *Erreur de généralisation moyenne pour* $\alpha = 8{,}10$ *et* 14 *en fonction de* $1/N$. *Les barres d'erreur sont inférieures à la taille des symboles. Les symboles pleins correspondent aux valeurs théoriques. Les droites représentent les extrapolations linéaires des simulations numériques.*

$\alpha N/2^N$ des exemples utilisés pour l'apprentissage est donc une fonction décroissante lorsque N augmente. Ceci peut expliquer les corrections négatives. L'argument suivant peut expliquer le fait que ces corrections sont en $1/N$. Le professeur impose que la fonction booléenne que l'on cherche à apprendre soit linéairement séparable. Or, les fonctions booléennes linéairement séparables sont une petite fraction des 2^{2^N} fonctions booléennes possibles. Le nombre des fonctions linéairement séparables est proportionnel à 2^{N^2}, d'après un calcul de Peretto [99]. Le nombre de *bits* nécessaire pour spécifier une fonction booléenne, sachant qu'elle est linéairement séparable, est donc proportionnel à N^2 et non pas à 2^N comme pour une fonction booléenne quelconque. La fraction entre le nombre d'exemples et le nombre de bits nécessaire pour spécifier la fonction booléeenne correspondante au professeur est alors proportionnelle à $1/N$. Cet argument n'est plus valable pour des exemples qui seraient distribués selon une gaussienne normale.

La différence entre les effets de taille finie entre une distribution des exemples gaussienne et une distribution binaire est importante dans le cas de l'apprentissage d'une fonction booléenne quelconque. Les figures 2.6 représentent les histogrammes de la probabilité $P_{\rm LS}(\alpha,N)$ qu'un ensemble de taille réduite α dont les classes sont distribuées aléatoirement soit linéairement séparable pour deux types de distributions (distributions gaussienne et binaire). Ces histogrammes ont été déterminés par des simulations numériques. La probabilité $P_{\rm LS}$ est représentée pour plusieurs valeurs de la dimension de l'espace des données ($N = 4{,}10{,}20{,}40$). Pour une distribution gaussienne des exemples, les courbes se croisent toutes à $\alpha = 2$ pour une probabilité $P_{\rm LS} = 1/2$. Pour cette distribution, il a été ajouté aux simulations numériques, les prédictions théoriques faites par Cover [28]. La valeur $\alpha = 2$ correspond à la capacité d'un perceptron simple déterminée par Cover [28] puis par Gardner avec la méthode des répliques [38, 39] dans la limite thermodynamique. Pour la distribution binaire des exemples, les courbes ne se croisent pas toutes au même point. La valeur de la taille réduite $\alpha(N)$ pour laquelle $P_{\rm LS} = 1/2$ montre des corrections négatives par rapport à $\alpha(\infty) = 2$. On s'attend à ce que les effets

FIG. 2.6 – *Histogrammes de la probabilité P_{LS} qu'un ensemble d'apprentissage de taille réduite α dont les classes sont aléatoires soit linéairement séparable. La figure de gauche correspond à une distribution des exemples gaussienne et la figure de droite à une distribution binaire. Les histogrammes ont été obtenus par des simulations numériques pour diverses dimensions de l'espace des données. Dans le cas de la distribution gaussienne, il a été ajouté les prédictions théoriques obtenues par Cover.*

de taille finie diffèrent aussi suivant la distribution des exemples pour le problème supervisé considéré mais aucune simulation numérique avec une distribution gaussienne n'ayant été faite, nous ne pouvons pas confirmer ce résultat.

La variance $\sigma^2(\alpha,N)$, déterminée pour toutes les valeurs de α et N simulées, est représentée sur la figure 2.7. Cette variance s'annule proportionnellement à $1/N$ lorsque $N \to +\infty$ quelque soit la valeur de α. Dans cette limite, tous les ensembles d'apprentissage produisent des élèves ayant la même erreur de généralisation, car la variance de $\varepsilon_g(\alpha,N)$ s'annule. Ce comportement est celui attendu dans le cas d'un automoyennage de l'erreur de généralisation. L'hypothèse d'automoyennage utilisée pour déterminer l'erreur de généralisation optimale par la méthode des répliques dans la limite thermodynamique est donc vérifiée.

Afin de caractériser les corrections de taille finie de la variance $\sigma(\alpha,N)$, nous allons introduire la fonction $\psi(\alpha)$:

$$\sigma^2(\alpha,N) = \frac{\psi(\alpha)}{\alpha N}. \tag{2.46}$$

Les deux fonctions ϕ et ψ, définies par (2.45) et (2.46) respectivement, sont représentées en fonction de α sur la figure 2.8. Ces deux fonctions présentent un changement de comportement significatif pour une valeur proche de $\alpha = 2$. La fonction ϕ est croissante pour $\alpha < 2$ puis devient pratiquement constante pour $\alpha > 2$. Le changement de comportement pour la fonction ψ est encore plus significatif. Celle-ci est croissante pour $\alpha < 2$ puis décroît pour $\alpha > 2$. Cette observation sur le changement de comportement de ces deux fonctions reste inexpliquée. Toutefois, il est intéressant de remarquer que cette valeur $\alpha = 2$ n'est pas anodine. En effet, elle correspond à la taille réduite maximale de l'ensemble d'apprentissage qu'il est possible de séparer linéairement lorsque le professeur correspond à une fonction booléenne quelconque [38, 39, 40]. Cette taille réduite maximale est appelée capacité du classifieur linéaire. Il serait intéressant de savoir s'il existe un lien entre les effets de taille finie de l'apprentissage d'une fonction booléenne linéairement séparable et

FIG. 2.7 – *Variance de l'erreur de généralisation pour toutes les valeurs de α simulées (α = 1,2,4,6,8,10 et 14) en fonction de $1/N$. Les barres d'erreur sont inférieures à la taille des symboles. Les droites représentent les extrapolations linéaires des simulations numériques.*

la capacité du classifieur linéaire pouvant expliquer le changement de comportement des fonctions ϕ et ψ à cette valeur particulière $\alpha = 2$.

2.7.3 Distribution des stabilités

Nous allons maintenant discuter des effets de taille finie sur la distribution des stabilités des exemples de l'ensemble d'apprentissage.

La figure 2.9 représente les histogrammes ρ_{opt} des distances des exemples de l'ensemble d'apprentissage à l'hyperplan séparateur de l'élève et à celui du professeur pour $N = 100$ et $\alpha = 4$. Les distributions théoriques $\rho_{\text{opt}}(\gamma)$ et $\rho_{\text{prof}}(\gamma)$ sont représentées en traits pleins. On observe un excellent accord entre les résultats des simulations numériques et la prédiction théorique par la méthode des répliques dans la limite thermodynamique. Pour la dimension de l'espace des données représentée ($N = 100$), aucun effet de taille finie n'est visible à l'échelle de la figure.

La figure 2.10 représente les histogrammes des stabilités pour des valeurs plus faibles de N ($N = 20$ et 65) pour une taille réduite $\alpha = 6$ de l'ensemble d'apprentissage. Pour comparer, nous avons aussi représenté les prédictions théoriques des distributions des stabilités de l'élève $\rho_{\text{opt}}(\gamma)$ et des distances des exemples à l'hyperplan séparateur du professeur $\rho_{\text{prof}}(\gamma)$. Aucun effet de taille finie n'étant visible pour $\rho_{\text{prof}}(\gamma)$, les histogrammes des distributions pour N fini n'ont pas été représentées pour ne pas surcharger la figure. Différents effets de taille finie sont visibles pour $\rho_{\text{opt}}(\gamma)$ à ces faibles valeurs de N. Le premier concerne une diminution ainsi qu'un déplacement vers les valeurs plus importantes de γ du pic de la distribution des stabilités. Aucun effet de taille finie n'est visible pour les très grandes stabilités, pour lesquelles $\rho_{\text{opt}}(\gamma)$ coïncide avec $\rho_{\text{prof}}(\gamma)$. Pour les petites valeurs de stabilité, la région où la distribution est négligeable semble réduite. Toutefois, il faut noter que $\rho_{\text{opt}}(\gamma)$ ne s'annule pas pour $\gamma > 0$ dans la limite thermodynamique. D'après les simulations numériques, pour les faibles valeurs de N, il existe une région où la distribution des stabilités est strictement nulle. Cette région se réduit lorsque N

FIG. 2.8 – *Fonctions $\phi(\alpha)$ et $\psi(\alpha)$ correspondantes aux termes d'ordre $1/P$ des corrections de taille finie de l'erreur de généralisation (fonction ϕ figure de gauche) et de sa variance (fonction ψ figure de droite) pour les différentes valeurs de α étudiées ($\alpha = 1,2,4,6,8,10$ et 14).*

augmente pour disparaître à la limite thermodynamique.

FIG. 2.9 – *Histogrammes des distributions des stabilités des exemples de l'ensemble d'apprentissage (carrés) et des distances des exemples à l'hyperplan séparateur du professeur (croix) pour* $N = 100$ *et* $\alpha = 4$. *Les courbes en trait plein correspondent aux distributions théoriques (*$N = +\infty$ *et* $\alpha = 4$*).*

FIG. 2.10 – *Histogrammes des distributions des stabilités des exemples de l'ensemble d'apprentissage pour* $N = 20$ *(carrés) et* $N = 65$ *(croix) avec* $\alpha = 6$. *Les courbes correspondent aux distributions théoriques des stabilités (trait plein) et des distances à l'hyperplan du professeur (tirets) (*$N = +\infty$ *et* $\alpha = 6$*).*

Chapitre 3

Apprentissage non supervisé

Dans ce chapitre, nous allons revenir à un problème particulier d'apprentissage non supervisé. Dans de nombreuses applications réalistes, les exemples sont distribués en plusieurs amas. Une fois encore, nous allons parler de la distribution des exemples lorsqu'il faudrait parler, rigoureusement, de densité de probabilité. Déterminer les paramètres de la distribution des amas à partir d'un ensemble d'exemples est une tâche souvent très complexe. Ce problème a été précédemment étudié par de nombreux auteurs : Barkai, Seung et Sompolinsky [10, 9], Biehl et Mietzner [11, 12], Lootens et Van den Broeck [69], Marangi, Biehl et Solla [71], Meir [76], Rose, Gurewitz et Fox [103], Watkin et Nadal [125], etc. Nous nous sommes restreint, comme bon nombre de ces auteurs, au problème de deux amas [22, 23, 47]. Les résultats obtenus et, notamment, les effets des symétries de la distribution des exemples sur les performances de l'apprentissage peuvent sans aucun doute se généraliser aux cas où le nombre d'amas est plus important.

Pour le problème à deux amas, nous avons montré qu'il existe trois phases d'apprentissage, généralement bien distinctes, en fonction de la quantité d'exemples disponibles. La plupart de ces phases sont séparées par des transitions du premier et du second ordre de la performance de l'apprentissage optimal. Nous verrons que ces résultats soulèvent une controverse sur les performances optimales qu'il n'a pas encore été possible de résoudre.

3.1 Présentation du problème

Nous allons supposer que nous disposons d'un ensemble d'apprentissage $\mathcal{L}_\alpha$, constitué de P points d'un espace de données de dimension N :

$$\mathcal{L}_\alpha = \{\boldsymbol{\xi}^\mu\}_{\mu=1,\cdots,P} \,. \tag{3.1}$$

Ces points sont distribués indépendamment les uns des autres selon une distribution possédant deux maxima. Le nuage des points de l'ensemble d'apprentissage se répartit selon deux amas centrés aux maxima de la distribution. Nous allons considérer les deux amas placés de manière symétrique par rapport à l'origine. Il apparaît alors naturellement une direction privilégiée qui est la direction passant par les centres de ces amas. La distribution des exemples selon les directions perpendiculaires à la direction privilégiée sera supposée être une gaussienne normale. La distribution selon la direction privilégiée sera représentée, quant à elle, par la superposition de deux gaussiennes. Une telle distribution s'écrit sous la forme (1.2) introduite au premier chapitre, qu'on réécrit ici :

$$P(\boldsymbol{\xi}) = \frac{1}{(2\pi)^{N/2}} \exp\left(-\frac{\boldsymbol{\xi}\cdot\boldsymbol{\xi}}{2} - V^*(\lambda)\right) \tag{3.2}$$

avec $\lambda = \boldsymbol{\xi}\cdot\mathbf{B}$ où $\mathbf{B}$ correspond à la direction privilégiée. La perturbation V^* est définie par :

$$\begin{aligned} P(\lambda) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\lambda^2}{2} - V^*(\lambda)\right) && (3.3)\\ &= \frac{1}{2\sigma\sqrt{2\pi}} \left\{ \exp\left(-\frac{(\lambda-\rho)^2}{2\sigma^2}\right) + \exp\left(-\frac{(\lambda+\rho)^2}{2\sigma^2}\right)\right\}. && (3.4) \end{aligned}$$

Les paramètres ρ et σ caractérisent la distribution selon la direction privilégiée. Ces deux paramètres seront supposés connus. ρ reflète l'écartement des deux amas et σ la largeur de chaque gaussienne. La distribution des exemples (3.2) avec une perturbation définie par (3.4) est une généralisation de celle étudiée par Biehl et Mietzner [11, 12] et par Watkin et Nadal [125] à des valeurs quelconques de σ. Ces auteurs avaient limité leur étude à $\sigma = 1$. Le fait que la variance σ des gaussiennes dans la direction privilégiée puisse être différente de celle des gaussiennes dans les autres directions est essentiel pour obtenir des transitions de phases du premier ordre de la performance de l'apprentissage optimal. Ayant supposé identiques les poids des deux amas de la distribution des exemples, les directions $\mathbf{B}$ et $-\mathbf{B}$ sont équivalentes. Nous verrons que cette symétrie de la distribution a des conséquences importantes sur la performance optimale de l'apprentissage.

Les paramètres ρ et σ étant supposés connus, la détection des amas se réduit à la détermination de la direction privilégiée $\mathbf{B}$. Le problème d'apprentissage non supervisé que nous venons de présenter est donc un cas particulier de celui présenté au premier chapitre. L'apprentissage consiste à trouver une direction $\mathbf{J}^*$ aussi proche que possible de la direction privilégiée $\mathbf{B}$. La performance de cet apprentissage est caractérisée par $R = \mathbf{B}\cdot\mathbf{J}^*$. Les résultats du premier chapitre permettent de trouver la valeur optimale de ce paramètre dans la limite thermodynamique ($N \to +\infty$ et $P \to +\infty$ avec $\alpha = P/N$) en fonction de la taille réduite α de l'ensemble d'apprentissage. Suivant la formulation générale du Chapitre 1, il est possible de déterminer le potentiel optimal qui permet de déterminer la direction optimale par minimisation de la fonction de coût correspondante.

En particulier, l'introduction de $V^*(\lambda)$ déduit à partir de (3.4), dans (1.65), permet d'obtenir $R_{\rm opt}$, qui est donné par l'expression suivante :

$$\alpha = \frac{A^2 R_{\rm opt}^2}{(1-R_{\rm opt}^2)} \left\{ \int G^2(y; R_{\rm opt},\rho,\sigma,A) Dy \right\}^{-1} \tag{3.5}$$

avec :

$$\begin{aligned} A &= 1 - R_{\rm opt}^2(1-\sigma^2), && (3.6)\\ Dy &= \exp\left(-\frac{y^2}{2}\right)\frac{dy}{\sqrt{2\pi}}, && (3.7)\\ G(y; R,\rho,\sigma,A) &= \rho\tanh\left(\frac{\rho R x}{A}\right) - R\left(1-\sigma^2\right)x, && (3.8)\\ x &= y\sqrt{A} - \rho R. && (3.9) \end{aligned}$$

La fonction G ne coïncide pas avec la fonction $g(y)$ (1.61) introduite au premier chapitre. L'inversion de l'équation (3.5) permet d'obtenir la *courbe d'apprentissage*

optimal $R_{\rm opt}(\alpha)$ pour un couple de paramètres (ρ,σ) donné. Le potentiel optimal est déterminé à partir de l'équation suivante :

$$V'_{\rm opt}(\lambda(y)) = -\frac{1-R_{\rm opt}^2}{R_{\rm opt}}\, G(y; R_{\rm opt},\rho,\sigma,A) \tag{3.10}$$

avec $\lambda(y)$ définie par :

$$\lambda(y) = y + \frac{1-R_{\rm opt}^2}{R_{\rm opt}}\, G(y; R_{\rm opt},\rho,\sigma,A). \tag{3.11}$$

L'inversion de la fonction $\lambda(y)$ et l'équation (3.10) permettent de calculer le potentiel optimal $V_{\rm opt}(\lambda)$ qui dépend du couple (ρ,σ) mais aussi de la taille réduite α de l'ensemble d'apprentissage.

3.2 Courbes d'apprentissage optimal

Nous allons discuter, dans cette section, les différents types de courbes d'apprentissage optimal que l'on peut obtenir en fonction des paramètres ρ et σ.

Avant de commencer, nous allons rappeler le comportement de la courbe d'apprentissage optimal pour les faibles valeurs de $R_{\rm opt}$ dans le cas très général proposé par Reimann et Van den Broeck [101, 120]. Pour cela, il est possible d'effectuer un développement limité de l'équation (1.65) autour de $R_{\rm opt} = 0$. Ce développement fait intervenir les différents moments de la distribution des exemples (1.3) ou (3.3) selon la direction privilégiée $\mathbf{B}$:

$$\langle \lambda^n \rangle = \int D\lambda\, \lambda^n\, \exp\left(-V^*(\lambda)\right). \tag{3.12}$$

Dans le cas où $\langle\lambda\rangle \neq 0$, le développement au premier ordre donne :

$$R_{\rm opt} \simeq \langle\lambda\rangle\sqrt{\alpha}. \tag{3.13}$$

Il est intéressant de remarquer que $R_{\rm opt}$ croît alors dès $\alpha = 0$. De plus, une simple moyenne des exemples de l'ensemble d'apprentissage est une estimation quasi-optimale de la direction privilégiée. Cette règle d'apprentissage, appelée la règle de Hebb [54, 119, 1, 4, 127, 126, 101, 55], donne :

$$\mathbf{J}^* = \frac{C}{P}\sum_{\mu=1}^{P} \boldsymbol{\xi}^\mu \tag{3.14}$$

avec C la constante de normalisation. Celle-ci est déterminée par :

$$1 = \frac{C^2}{P^2}\sum_{\mu,\nu} \boldsymbol{\xi}^\mu\cdot\boldsymbol{\xi}^\nu \simeq \frac{C^2}{P}\left((P-1)\langle\lambda\rangle^2 + (N-1+\langle\lambda^2\rangle)\right) \tag{3.15}$$

en remplaçant les produits scalaires entre les exemples par leurs moyennes. Dans la limite thermodynamique et pour les ensembles d'apprentissage de faible taille réduite ($\alpha \ll 1$), la constante de normalisation se réduit à $C \simeq \sqrt{\alpha}$. Le produit scalaire entre la direction $\mathbf{J}^*$ et la direction privilégiée $\mathbf{B}$ est alors :

$$R = \mathbf{J}^*\cdot\mathbf{B} \simeq \langle\lambda\rangle\sqrt{\alpha}. \tag{3.16}$$

Dans le cas où $\langle\lambda\rangle = 0$, la règle de Hebb n'est alors plus efficace. Contrairement au cas précédent, le produit scalaire entre les exemples et la direction privilégiée est une variable aléatoire de moyenne nulle. L'apprentissage est plus difficile dans ce cas. On observe ce que l'on appelle un *retard à l'apprentissage* [125]. La valeur

optimale R_{opt} est nulle pour les tailles réduites $\alpha < \alpha_0$ et non nulle au dessus de α_0. Le développement limité de l'équation (1.65) autour de $R_{\text{opt}} = 0$ permet de prédire une croissance continue de R_{opt} uniquement à partir d'un seuil α_c :

$$R_{\text{opt}} \simeq \sqrt{\frac{\alpha - \alpha_c}{\alpha_c \langle \lambda^2 \rangle}} \tag{3.17}$$

pour $\alpha > \alpha_c$ avec $\alpha_c = (1 - \langle \lambda^2 \rangle)^{-2}$. Cette valeur seuil α_c ne correspond pas nécessairement à α_0, la taille réduite à partir de laquelle $R_{\text{opt}} > 0$. En effet, nous le verrons par la suite une transition du premier ordre de la courbe d'apprentissage optimal peut exister pour $\alpha_1 < \alpha_c$. Dans ce cas, $\alpha_0 = \alpha_1$ et non pas α_c.

Il est intéressant de remarquer que si les deux premiers moments de la distribution des exemples dans la direction privilégiée sont identiques à ceux dans les directions perpendiculaires, c'est-à-dire, $\langle \lambda \rangle = 0$ et $\langle \lambda^2 \rangle = 1$, non seulement il y a retard à l'apprentissage mais $\alpha_c = +\infty$. Toutefois, une distribution n'est pas déterminée uniquement par ses deux premiers moments. Une distribution, dont les deux premiers moments sont identiques à ceux d'une gaussienne normale, n'est pas nécessairement une gaussienne normale. Pour de telles distributions, la performance optimale $R_{\text{opt}}(\alpha)$ ne peut croître continuement puisque $\alpha_c = +\infty$. Cette remarque, qui nous a fait penser à l'existence possible de transitions du premier ordre pour certaines distributions des exemples, a motivé l'étude présentée dans ce chapitre.

Le cas particulier de la distribution des exemples (3.2), formée de deux gaussiennes dans la direction privilégiée, est un exemple de distribution où le phénomène décrit précédemment se produit. Il faut pour cela considérer des largeurs σ des gaussiennes dans la direction privilégiée différentes de celle dans les autres directions. Les deux premiers moments sont alors :

$$\langle \lambda \rangle = 0, \tag{3.18}$$

$$\langle \lambda^2 \rangle = \rho^2 + \sigma^2. \tag{3.19}$$

Dans les études précédentes, mentionnées au début du chapitre, les auteurs ont généralement considéré $\sigma = 1$, ce qui restreint la famille des distributions des exemples à celles pour lesquelles il n'y a pas de transition du premier ordre.

Dans le plan des paramètres (ρ,σ), la ligne $\rho^2 + \sigma^2 = 1$ correspond à un retard à l'apprentissage pour lequel $\alpha_c = +\infty$. Pour ces valeurs de paramètres, on s'attend à l'existence d'une transition du premier ordre entre $R_{\text{opt}} = 0$ pour $\alpha < \alpha_1$ et $R_{\text{opt}} > 0$ pour $\alpha \geq \alpha_1$ avec α_1 fini, donc une discontinuité de R_{opt} en α_1. En fait, il est possible de découper le plan (ρ,σ) en plusieurs régions correspondantes chacune à un type particulier de courbe d'apprentissage optimal $R_{\text{opt}}(\alpha)$.

Tout d'abord, il existe une région de ce plan (ρ,σ) où se produit le comportement habituel précédemment obtenu par d'autres auteurs [11, 12, 125]. La courbe d'apprentissage optimal R_{opt} montre une transition du second ordre à α_c entre une région de retard à l'apprentissage ($R_{\text{opt}} = 0$) et une région où R_{opt} croît continuement à partir de 0. Un exemple de ce comportement est représenté sur la figure 3.1. Cette figure correspond à la courbe d'apprentissage optimal pour les paramètres $\rho = 1.4$ et $\sigma = 0.5$.

Pour des valeurs des paramètres (ρ,σ) un peu plus proches de la ligne $\rho^2 + \sigma^2 = 1$, la courbe d'apprentissage optimal n'est plus une courbe univaluée pour toute valeur de α. Il apparaît, en effet, une plage de valeurs de la taille réduite α pour lesquelles il existe trois solutions différentes à l'équation (3.5). La figure 3.2 représente la courbe d'apprentissage optimal pour les paramètres $\rho = 1.2$ et $\sigma = 0.5$. Cette courbe présente une transition du second ordre à $\alpha_c = 2.10$ similaire à celle du cas précédent. R_{opt} croît continûement de α_c jusqu'à α_1. À partir de α_1, l'équation (3.5) possède deux nouvelles branches de solutions pour R_{opt}, supérieures à celle

FIG. 3.1 – *Courbe d'apprentissage optimal* $R_{\rm opt}(\alpha)$ *pour* $\rho = 1.4$ *et* $\sigma = 0.5$. *Cette courbe montre une transition du second ordre entre une phase de retard à l'apprentissage* ($R_{\rm opt} = 0$) *et une phase pour laquelle* $R_{\rm opt}$ *croît continuement à partir de* 0. *La taille réduite* α_c *de l'ensemble d'apprentissage à la transition est* $\alpha_c = 0.68$.

déjà existante. L'une de ces nouvelles branches est une branche de pente négative et donc instable par rapport à une brisure de symétrie des répliques, d'après l'étude de stabilité locale présentée au paragraphe 1.6 du premier chapitre. Les deux autres branches étant croissantes, il reste deux solutions (localement) stables possibles pour $R_{\rm opt}$. L'apprentissage optimal consistant à trouver la valeur maximale de R pour une taille réduite α donnée, il convient de sélectionner la branche supérieure. Il apparaît alors un saut dans la courbe d'apprentissage optimal à α_1, de $R_{\rm opt} \simeq 0.40$ à $R_{\rm opt} \simeq 0.76$. Ceci correspond à une transition du premier ordre. L'approche bayesienne de l'apprentissage, supposé optimal, prédit une transition du premier ordre à $\alpha_G > \alpha_1$. Cette contradiction entre l'approche variationnelle de l'apprentissage optimal présentée au premier chapitre et l'approche bayesienne sera discutée dans la section suivante.

Dans le plan (ρ,σ), les lignes qui séparent les deux régions que nous venons de décrire (dans l'une $R_{\rm opt}(\alpha)$ est continu, dans l'autre $R_{\rm opt}(\alpha)$ possède une transition du premier ordre entre deux valeurs de $R_{\rm opt}$ finies), correspondent aux paramètres (ρ,σ) pour lesquels la courbe d'apprentissage optimal possède une tangente verticale à α_1. Une expression analytique simple pour ces deux lignes, l'une correspondant à des distributions avec $\langle\lambda^2\rangle > 1$ et l'autre à $\langle\lambda^2\rangle < 1$, n'a pas pu être déterminée. De même, contrairement à la position de la transition du second ordre située à $\alpha_c = \left(1 - \langle\lambda^2\rangle\right)^{-2}$, il n'existe pas d'expression analytique simple pour la position α_1 de la transition du premier ordre.

Lorsque les paramètres (ρ,σ) se rapprochent encore de la ligne $\rho^2 + \sigma^2 = 1$, la taille réduite α_1 devient plus faible que α_c. La transition du second ordre disparaît, masquée par la transition du premier ordre. La figure 3.3 représente la courbe d'apprentissage optimal pour les paramètres $\rho = 1.1$ et $\sigma = 0.5$. La ligne $\rho^2+\sigma^2 = 1$ est contenue dans cette région puisque $\alpha_c = +\infty$ sur cette ligne. Dans le plan (ρ,σ), les deux lignes séparant cette région des régions précédentes sont déterminées par

FIG. 3.2 – *Courbe d'apprentissage optimal* $R_{\text{opt}}(\alpha)$ *pour* $\rho = 1.2$ *et* $\sigma = 0.5$ *(trait plein). Cette courbe montre une transition du second ordre puis une transition du premier ordre entre* $R_{\text{opt}} \simeq 0.40$ *et* $R_{\text{opt}} \simeq 0.76$*. La première transition se produit à* $\alpha_c = 2.10$ *et la deuxième à* $\alpha_1 = 2.49$*. À partir de* α_1*, il existe trois solutions possibles pour* R_{opt}*. Une des solutions correspond à une solution instable (pointillés) et les deux autres sont stables. La solution optimale correspond à la solution stable de plus grand* R_{opt} *(trait plein), l'autre solution stable est représentée avec des tirets. La valeur* $\alpha_G = 2.52$ *correspond à la position de la transition du premier ordre prédite par l'approche bayesienne de l'apprentissage.*

la condition $\alpha_1 = \alpha_c$. Une fois encore, aucune expression analytique simple n'a pu être obtenue pour ces lignes.

Suivant le couple de paramètres (ρ,σ) considéré, il existe donc trois types de courbes d'apprentissage optimal différentes. Les régions de l'espace des paramètres correspondantes à ces trois types de courbes sont représentées sur la figure 3.4. Les lignes séparatrices ont été déterminées numériquement.

Dans la région contenant la ligne $\rho^2 + \sigma^2 = 1$ (représentée en gris clair sur la figure), la courbe d'apprentissage optimal possède une transition du premier ordre uniquement où l'on passe de $R_{\text{opt}} = 0$ à une valeur de R_{opt} finie. Dès que la taille réduite de l'ensemble d'apprentissage dépasse α_1, la direction apprise est proche de la direction privilégiée $\mathbf{B}$ que l'on veut déterminer. La transition du second ordre qui apparaît lorsque la stratégie d'apprentissage n'utilise que des informations sur le second moment de la distribution des exemples, est masquée par la transition du premier ordre qui la précède ($\alpha_1 < \alpha_c$). La ligne $\rho^2 + \sigma^2 = 1$ est au centre de cette région. Ceci est cohérent avec le fait que sur cette ligne, la transition du second ordre ne peut pas avoir lieu puisque $\alpha_c = +\infty$.

De chaque coté de cette région se situe une région où la transition du second ordre à α_c est suivie par une transition du premier ordre à α_1 entre deux valeurs finies de R_{opt}. Cette région est représentée en gris foncé sur la figure.

Enfin, la plus grande région des paramètres (région blanche) correspond à une courbe d'apprentissage optimal possédant uniquement une transition du second ordre, c'est-à-dire, avec le retard à l'apprentissage habituel observé par Biehl et Mietzner [11, 12] et Watkin et Nadal [125] sur la ligne $\sigma = 1$.

FIG. 3.3 – *Courbe d'apprentissage optimal* $R_{\rm opt}(\alpha)$ *pour* $\rho = 1.1$ *et* $\sigma = 0.5$ *(trait plein). Cette courbe montre uniquement une transition du premier ordre, à* $\alpha_1 = 3.79$. *À partir de* α_1, *il existe trois solutions possibles pour* $R_{\rm opt}$. *Une des solutions correspond à une solution instable (pointillés) et les deux autres sont stables. La solution optimale correspond à la solution stable de plus grand* $R_{\rm opt}$ *(trait plein), l'autre solution stable est représentée avec des tirets. La transition du premier ordre, prédite par l'approche bayesienne, est à* $\alpha_G = 4.07$. *La transition du second ordre à* $\alpha_c = 4.73$ *est masquée par la transition du premier ordre.*

Les régions marquées par un S seront discutées par la suite lors de l'étude de l'évolution de la forme du potentiel optimal en fonction de α pour des paramètres (ρ,σ) donnés.

FIG. 3.4 – *Régions des diverses courbes d'apprentissage optimal rencontrées pour le problème de la double gaussienne en fonction des paramètres ρ et σ. La région gris clair correspond aux paramètres pour lesquels la courbe d'apprentissage optimal possède une transition du premier ordre, la région en gris foncé correspond à l'existence d'une transition du second ordre puis d'une transition du premier ordre. Enfin, la région blanche correspond à l'existence d'une transition du second ordre. La ligne pointillée représente la ligne $\rho^2 + \sigma^2 = 1$. Les deux lignes en tirets séparent les régions marquées d'un S pour lesquelles le potentiel optimal possède un unique extrema pour toutes les valeurs de α. Les trois carrés correspondent aux paramètres des courbes d'apprentissage optimal représentées sur les figures précédentes ($(\rho,\sigma) = (1.4,0.5),(1.2,0.5)$ et $(1.1,0.5)$ de haut en bas).*

3.3 Approche bayesienne

J'ai présenté jusqu'ici une approche variationnelle de l'apprentissage optimal pour obtenir le potentiel optimal ainsi que les performances optimales représentées par la fonction $R_{\rm opt}(\alpha)$. Je vais maintenant présenter une autre approche appelée bayesienne qui permet aussi un apprentissage optimal. Ces deux approches aboutissent à la même équation (1.65) définissant $R_{\rm opt}$. Elles paraissent donc satisfaisantes et cohérentes. C'est, en effet, le cas pour toutes les applications précédemment étudiées, comme par exemple, l'apprentissage supervisé du deuxième chapitre. Dans ces cas, l'équation (1.65) est inversible et les deux approches coïncident. Il n'existe qu'une seule solution possible $R_{\rm opt}(\alpha)$ pour une valeur donnée de α. Nous avons mis en évidence, dans le cas de la double gaussienne de la section précédente, qu'il y a une large région de paramètres (ρ,σ) pour lesquels plusieurs solutions différentes de $R_{\rm opt}$ existent, pour une même valeur de α. Dans ce cas, les deux approches diffèrent. La transition du premier ordre qui résulte de l'existence de ces solutions multiples est prédite à des valeurs différentes de la taille réduite α suivant l'approche considérée.

3.3.1 Apprentissage de Gibbs

Nous allons tout d'abord développer l'approche bayesienne. Cette approche est basée sur la règle de Bayes [32, 14]. Celle-ci relie la probabilité de $\mathcal{Y}$ étant donné $\mathcal{X}$ notée $P(\mathcal{Y}|\mathcal{X})$ connaissant la probabilité de $\mathcal{X}$ étant donné $\mathcal{Y}$:

$$P(\mathcal{Y}|\mathcal{X}) = \frac{P(\mathcal{X}|\mathcal{Y})P(\mathcal{Y})}{P(\mathcal{X})}. \tag{3.20}$$

La forme fonctionnelle de la probabilité d'obtenir un ensemble d'apprentissage $\mathcal{L}_\alpha$ à partir d'une direction $\mathbf{J}$ est connue. En effet, nous avons supposé les exemples $\boldsymbol{\xi}^\mu$ distribués indépendamment les uns des autres et la perturbation V^*, introduite au premier chapitre, connue. L'algorithme de Gibbs est basé sur la connaissance de cette forme fonctionnelle :

$$P(\mathcal{L}_\alpha|\mathbf{J}) = \prod_{\mu=1}^{P} P(\boldsymbol{\xi}^\mu|\mathbf{J}) \tag{3.21}$$

où $P(\boldsymbol{\xi}^\mu|\mathbf{J})$ est donnée par l'équation (1.2) introduite au premier chapitre avec $\mathbf{J}$ à la place de $\mathbf{B}$.

Dans le cas de l'apprentissage de la direction privilégiée $\mathbf{B}$ à partir de l'ensemble d'apprentissage $\mathcal{L}_\alpha(\mathbf{B})$, nous indiquerons explicitement, lorsqu'il est nécessaire pour la compréhension, la dépendance de l'ensemble d'apprentissage vis-à-vis de $\mathbf{B}$, la direction à partir de laquelle il a été obtenu. En statistique, $P(\mathcal{L}_\alpha(\mathbf{B})|\mathbf{J})$, où la forme fonctionnelle (3.21) est considérée pour l'ensemble d'apprentissage particulier $\mathcal{L}_\alpha(\mathbf{B})$ disponible pour l'apprentissage, s'appelle la vraisemblance de $\mathbf{J}$ [32]. L'apprentissage de Gibbs consiste à utiliser cette vraisemblance ainsi que la règle de Bayes pour en déduire une distribution sur les directions $\mathbf{J}$:

$$P_{\rm G}(\mathbf{J}|\mathcal{L}_\alpha(\mathbf{B})) = \frac{P(\mathcal{L}_\alpha(\mathbf{B})|\mathbf{J})P_{\rm G}(\mathbf{J})}{P_{\rm G}(\mathcal{L}_\alpha(\mathbf{B}))}. \tag{3.22}$$

Nous supposerons la distribution $P_{\rm G}(\mathbf{J})$, dite probabilité *a priori* de $\mathbf{J}$, uniforme sur la sphère de rayon 1 :

$$P_{\rm G}(\mathbf{J}) = \delta\left(\mathbf{J}\cdot\mathbf{J} - 1\right). \tag{3.23}$$

C'est la même hypothèse qui avait déjà été utilisée pour l'approche variationnelle (*cf.* Eq.(1.9)). Elle correspond à supposer que la direction privilégiée $\mathbf{B}$, à partir de laquelle l'ensemble d'apprentissage a été tiré, peut être n'importe quelle direction de manière équiprobable, et traduit notre ignorance sur cette direction. $P_{\rm G}(\mathcal{L}_\alpha(\mathbf{B}))$ permet la normalisation de la distribution $P_{\rm G}(\mathbf{J}|\mathcal{L}_\alpha(\mathbf{B}))$. Elle n'est donc rien d'autre que l'intégrale sur toutes les directions $\mathbf{J}$ possibles de $P(\mathcal{L}_\alpha(\mathbf{B})|\mathbf{J})P_{\rm G}(\mathbf{J})$:

$$P_{\rm G}(\mathcal{L}_\alpha(\mathbf{B})) = \int P(\mathcal{L}_\alpha(\mathbf{B})|\mathbf{J})P_{\rm G}(\mathbf{J})\, d\,\mathbf{J}. \tag{3.24}$$

Comme il a déjà été vu au premier chapitre, choisir une direction $\mathbf{J}^*$ de manière aléatoire selon la distribution (3.22) correspond à l'apprentissage de Gibbs. Les propriétés d'apprentissage correspondantes sont données par $P_{\rm G}(\mathcal{L}_\alpha(\mathbf{B}))$. Plus particulièrement, dans la limite thermodynamique, le logarithme de $P_{\rm G}(\mathcal{L}_\alpha(\mathbf{B}))$ est supposé indépendant de l'ensemble d'apprentissage, ou autrement dit, automoyennant. Par comparaison avec (1.8), ce logarithme peut être considéré comme (l'opposé de) l'énergie libre $F_{\rm G}$ d'un système à température $T = 1$ et dont l'énergie n'est autre que la fonction de coût (1.6) introduite au premier chapitre, avec pour potentiel $V = V^*$:

$$E_{\rm G}(\mathbf{J}; \mathcal{L}_\alpha(\mathbf{B})) = \sum_{\mu=1}^{P} V^*(\mathbf{J}\cdot\boldsymbol{\xi}^\mu). \tag{3.25}$$

La méthode des répliques est utilisée pour calculer la moyenne de l'énergie libre $F_{\rm G}$ sur tous les ensembles d'apprentissage possibles obtenus à partir de la direction privilégiée $\mathbf{B}$ et de même taille réduite $\alpha = P/N$:

$$\overline{F_{\rm G}} = -\int \ln\left(P_{\rm G}(\mathcal{L}_\alpha)\right) P(\mathcal{L}_\alpha|\mathbf{B}) d\mathcal{L}_\alpha, \tag{3.26}$$

$$= -\lim_{n\to 0}\frac{1}{n}\ln\left\{\int d\mathcal{L}_\alpha \prod_{a=1}^{n}\int d\mathbf{J}_a P_{\rm G}(\mathbf{J}_a) P(\mathcal{L}_\alpha|\mathbf{J}_a) P(\mathcal{L}_\alpha|\mathbf{B})\right\}, \tag{3.27}$$

avec la notation suivante :

$$d\mathcal{L}_\alpha = \prod_{\mu=1}^{P} d\boldsymbol{\xi}^\mu. \tag{3.28}$$

Il est intéressant de noter que l'intégrale sur les ensembles d'apprentissage s'effectue avec la distribution $P(\mathcal{L}_\alpha|\mathbf{B})$ pour tenir compte du fait que l'ensemble d'apprentissage est obtenu à partir de la direction privilégiée $\mathbf{B}$.

Lors du calcul de la moyenne de l'énergie libre $F_{\rm G}$, l'hypothèse de symétrie des répliques permet de réduire le nombre de paramètres d'ordre à deux : le produit scalaire $q = \mathbf{J}_a \cdot \mathbf{J}_b$ entre deux directions $\mathbf{J}_a$ et $\mathbf{J}_b$ et le produit scalaire $R = \mathbf{J}_a \cdot \mathbf{B}$ entre une direction $\mathbf{J}_a$ et la direction privilégiée $\mathbf{B}$. La symétrie, qui existe entre les directions $\mathbf{J}_a$ et la direction privilégiée $\mathbf{B}$ dans l'intégrant de (3.27), entraîne l'égalité des deux paramètres d'ordre : $q_{\rm G} = R_{\rm G}$. Le paramètre $R_{\rm G}(\alpha)$ ainsi obtenu caractérise les propriétés de l'apprentissage de Gibbs de la même façon que $R_{\rm opt}(\alpha)$ caractérise celles de l'apprentissage optimal. Il est possible de montrer que $R_{\rm G}$ satisfait la même équation (1.65) que $R^2_{\rm opt}$ [101].

3.3.2 Performance de l'estimateur bayesien

Il est possible d'augmenter les performances de l'apprentissage de Gibbs si on considère non plus une direction choisie avec la distribution (3.22) mais la moyenne de ces directions qui sera appelée l'*estimateur bayesien* :

$$\mathbf{J}_{\mathrm{B}}(\mathcal{L}_\alpha(\mathbf{B})) \equiv C \int P_{\mathrm{G}}(\mathbf{J}|\mathcal{L}_\alpha(\mathbf{B}))\, \mathbf{J}\, d\,\mathbf{J}. \tag{3.29}$$

C est une constante de normalisation qu'il faut introduire parce que la moyenne des directions $\mathbf{J}$ n'est pas normée à un. L'estimateur bayesien dépend de l'ensemble d'apprentissage $\mathcal{L}_\alpha(\mathbf{B})$ donné pour l'apprentissage. La performance de cet estimateur bayesien est donnée par :

$$R_{\mathrm{B}}(\mathcal{L}_\alpha) \equiv \mathbf{J}_{\mathrm{B}}(\mathcal{L}_\alpha) \cdot \mathbf{B} = C \int P_{\mathrm{G}}(\mathbf{J}|\mathcal{L}_\alpha(\mathbf{B}))\, \mathbf{J} \cdot \mathbf{B}\, d\,\mathbf{J}. \tag{3.30}$$

En remplaçant, dans l'intégrale (3.30), le produit scalaire $\mathbf{J} \cdot \mathbf{B}$ entre une direction $\mathbf{J}$ donnée par (3.22) et la direction privilégiée $\mathbf{B}$ par sa valeur typique R_{G}, dans la limite thermodynamique, on obtient $R_{\mathrm{B}} = CR_{\mathrm{G}}$. Dans cette limite, la performance $R_{\mathrm{B}}(\mathcal{L}_\alpha)$ est supposée indépendante de l'ensemble d'apprentissage, de même que R_{G} et q_{G}. C'est ce que l'on a déjà appelé précédemment la propriété d'automoyennage. La constante de normalisation C est déterminée par :

$$1 = C^2 \iint P_{\mathrm{G}}(\mathbf{J}_1|\mathcal{L}_\alpha) P_{\mathrm{G}}(\mathbf{J}_2|\mathcal{L}_\alpha)\, \mathbf{J}_1 \cdot \mathbf{J}_2\, d\,\mathbf{J}_1\, d\,\mathbf{J}_2 = C^2 q_{\mathrm{G}}, \tag{3.31}$$

où on a remplacé le produit scalaire $\mathbf{J}_1 \cdot \mathbf{J}_2$ par sa valeur typique q_{G} dans la limite thermodynamique. On en déduit que le produit scalaire entre l'estimateur bayesien $\mathbf{J}_{\mathrm{B}}$ et la direction privilégiée $\mathbf{B}$ satisfait $R_{\mathrm{B}} = R_{\mathrm{G}}/\sqrt{q_{\mathrm{G}}}$ qui, compte tenu que $R_{\mathrm{G}} = q_{\mathrm{G}}$ donne :

$$R_{\mathrm{B}} = \sqrt{R_{\mathrm{G}}}. \tag{3.32}$$

Puisque R_{G} satisfait la même équation (1.65) que R^2_{opt}, on est tenté de déduire que $R_{\mathrm{B}} = R_{\mathrm{opt}}$ et que R_{B} est optimal. Comme nous le verrons par la suite ceci peut poser un problème lorsque l'équation satisfaite par R_{opt} admet plusieurs solutions pour une même valeur de α, ce qui arrive au voisinage des transitions du premier ordre que nous avons décrites au paragraphe 3.2.

La question qui se pose est la suivante : pourquoi l'estimateur bayesien est optimal ?

La réponse à cette question a été donnée par Watkin [124, 125] et précisée par Reimann et Van den Broeck [101] dans le cas général. Il suffit pour cela d'introduire la notion suivante de qualité de l'apprentissage :

$$Q(\mathbf{J}^*(\mathcal{L}_\alpha)) \equiv \int P_{\mathrm{G}}(\mathbf{J}|\mathcal{L}_\alpha)\, h(\mathbf{J} \cdot \mathbf{J}^*(\mathcal{L}_\alpha))\, d\,\mathbf{J} \tag{3.33}$$

avec $\mathbf{J}^*(\mathcal{L}_\alpha)$ la direction obtenue par un algorithme donné à partir de l'ensemble d'apprentissage $\mathcal{L}_\alpha(\mathbf{B})$ et $h(x)$ une fonction strictement croissante sur $[-1,1]$. Cette qualité de l'apprentissage n'est en rien reliée à la performance $R^*(\mathcal{L}_\alpha) \equiv \mathbf{B} \cdot \mathbf{J}^*(\mathcal{L}_\alpha)$ de l'algorithme considéré.

Dans le cas qui nous intéresse, un choix naturel pour la fonction $h(x)$ est $h(x) = |x|$. La valeur absolue est rendue indispensable à cause de la symétrie qu'il existe entre les directions $\mathbf{B}$ et $-\mathbf{B}$. La fonction $h(x)$ est donc strictement croissante uniquement sur l'intervalle $[0,1]$. Cette restriction supplémentaire n'est pas essentielle. Pour s'en convaincre, il suffit de penser au problème similaire du système d'Ising sous champ magnétique nul en dimension supérieure ou égale à 2. Il existe pour ce problème une symétrie entre les spins $+1$ et les spins -1. Si la moyenne thermodynamique est effectuée sous champ magnétique nul, l'aimantation est nulle quelque soit la température du fait de cette symétrie. Par contre, si l'on effectue la moyenne thermodynamique en supposant qu'il existe un champ magnétique non nul qui privilégie une direction par rapport à l'autre et que l'on prend la limite de

champ nul après la limite thermodynamique, alors l'aimantation est non nulle pour les températures inférieures à une température critique T_c. Dans notre cas, il est possible de briser faiblement la symétrie entre $\mathbf{B}$ et $-\mathbf{B}$ en supposant par exemple que le poids de chaque amas n'est pas $1/2$ mais $1/2+\epsilon$ et $1/2-\epsilon$ respectivement. ϵ joue alors le rôle du faible champ magnétique. Il suffit alors d'effectuer la limite thermodynamique avant la limite $\epsilon \to 0$ pour retrouver des résultats cohérents. De même que pour le système d'Ising, cette symétrie étant connue, il est inutile de considérer la brisure de symétrie représentée par ϵ. Il suffit de limiter la recherche des solutions R_{G} dans l'intervalle [0,1] et non [−1,1] comme on se limite à la recherche d'une aimantation positive uniquement pour le système d'Ising. Par la suite, on supposera, de même, que tous les produits scalaires entre directions $\mathbf{J}$ sont positifs et l'on oubliera la valeur absolue pour $h(x)$.

Il est assez facile de montrer que la qualité (3.33) est maximale pour l'estimateur bayesien $\mathbf{J}_{\mathrm{B}}(\mathcal{L}_\alpha)$. Considérons la direction $\mathbf{J}^*(\mathcal{L}_\alpha)$, obtenue par un algorithme d'apprentissage donné, à partir de l'ensemble d'apprentissage $\mathcal{L}_\alpha(\mathbf{B})$. D'après la définition de $\mathbf{J}_{\mathrm{B}}$ et en utilisant l'équation (3.29) :

$$
\begin{aligned}
Q(\mathbf{J}_{\mathrm{B}}(\mathcal{L}_\alpha)) - Q(\mathbf{J}^*(\mathcal{L}_\alpha)) &= \int P_{\mathrm{G}}(\mathbf{J}|\mathcal{L}_\alpha)\,(\mathbf{J}_{\mathrm{B}}\cdot\mathbf{J} - \mathbf{J}^*\cdot\mathbf{J})\,d\,\mathbf{J}, \qquad (3.34)\\
&= C^{-1}\,(\mathbf{J}_{\mathrm{B}} - \mathbf{J}^*)\cdot\mathbf{J}_{\mathrm{B}}. \qquad (3.35)
\end{aligned}
$$

Comme les directions $\mathbf{J}_{\mathrm{B}}$ et $\mathbf{J}^*$ sont normées à un : $\mathbf{J}_{\mathrm{B}}\cdot\mathbf{J}_{\mathrm{B}} = 1$ et $R = \mathbf{J}_{\mathrm{B}}\cdot\mathbf{J}^* \leq 1$. De cela, on déduit l'inégalité suivante :

$$
Q(\mathbf{J}_{\mathrm{B}}(\mathcal{L}_\alpha)) - Q(\mathbf{J}^*(\mathcal{L}_\alpha)) = C^{-1}(1-R) \geq 0. \qquad (3.36)
$$

La qualité $Q(\mathbf{J}_{\mathrm{B}}(\mathcal{L}_\alpha))$ est donc supérieure à celle correspondante à la direction obtenue par n'importe quel algorithme à partir de $\mathcal{L}_\alpha$. Toutefois, comme nous l'avons déjà dit, la qualité ne correspond pas à la performance de l'algorithme.

Les performances $R_{\mathrm{B}}(\mathcal{L}_\alpha)$, de l'estimateur bayesien, et $R^*(\mathcal{L}_\alpha)$, de l'algorithme ayant donné $\mathbf{J}^*(\mathcal{L}_\alpha)$, sont généralement moyennées sur tous les ensembles d'apprentissage $\mathcal{L}_\alpha$ possibles tirés à partir de $\mathbf{B}$:

$$
\begin{aligned}
\overline{R_{\mathrm{B}}}(\mathbf{B}) &\equiv \int P(\mathcal{L}_\alpha|\mathbf{B})\,\mathbf{B}\cdot\mathbf{J}_{\mathrm{B}}\,d\mathcal{L}_\alpha, \qquad (3.37)\\
\overline{R^*}(\mathbf{B}) &\equiv \int P(\mathcal{L}_\alpha|\mathbf{B})\,\mathbf{B}\cdot\mathbf{J}^*\,d\mathcal{L}_\alpha. \qquad (3.38)
\end{aligned}
$$

Ces *performances moyennes* dépendent *a priori* de la direction privilégiée $\mathbf{B}$.

Nous allons maintenant montrer à partir de l'inégalité (3.36) que $\overline{R_{\mathrm{B}}}(\mathbf{B}) \geq \overline{R^*}(\mathbf{B})$. Nous allons pour cela intégrer l'inégalité (3.36) sur tous les ensembles d'apprentissage possibles $\mathcal{L}_\alpha$ avec la distribution $P_{\mathrm{G}}(\mathcal{L}_\alpha)$ (3.24) :

$$
\int \left(Q(\mathbf{J}_{\mathrm{B}}(\mathcal{L}_\alpha)) - Q(\mathbf{J}^*(\mathcal{L}_\alpha))\right) P_{\mathrm{G}}(\mathcal{L}_\alpha)\,d\mathcal{L}_\alpha \geq 0. \qquad (3.39)
$$

En effectuant cette intégration sur tous les ensembles d'apprentissage possibles avec la distribution (3.24), on oublie le fait que l'ensemble d'apprentissage $\mathcal{L}_\alpha$ disponible pour l'apprentissage a été tiré à partir de la direction privilégiée $\mathbf{B}$. En conséquence, on ne peut pas directement en déduire des résultats sur les performances moyennes $\overline{R_{\mathrm{B}}}(\mathbf{B})$ et $\overline{R^*}(\mathbf{B})$ qui sont des moyennes sur tous les ensembles d'apprentissage avec la distribution $P(\mathcal{L}_\alpha|\mathbf{B})$ et non pas $P_{\mathrm{G}}(\mathcal{L}_\alpha)$.

En remplaçant la qualité Q par son expression (3.33) avec $h(x) = x$ et en utilisant la règle de Bayes (3.24) pour remplacer $P_{\mathrm{G}}(\mathbf{J}|\mathcal{L}_\alpha)P_{\mathrm{G}}(\mathcal{L}_\alpha)$ par $P(\mathcal{L}_\alpha|\mathbf{J})P_{\mathrm{G}}(\mathbf{J})$, on obtient alors l'inégalité suivante :

$$\int d\mathcal{L}_\alpha \int d\,\mathbf{J}\ (\mathbf{J}\cdot\mathbf{J}_\mathrm{B} - \mathbf{J}\cdot\mathbf{J}^*)\, P(\mathcal{L}_\alpha|\mathbf{J}) P_\mathrm{G}(\mathbf{J}) \geq 0. \tag{3.40}$$

Il suffit maintenant d'inverser les intégrales sur les ensembles d'apprentissage $\mathcal{L}_\alpha$ et sur les directions $\mathbf{J}$ pour reconnaître les performances moyennes :

$$\int d\mathcal{L}_\alpha P(\mathcal{L}_\alpha|\mathbf{J})\, \mathbf{J}\cdot\mathbf{J}_\mathrm{B} \quad = \quad \overline{R_\mathrm{B}}(\mathbf{J}), \tag{3.41}$$

$$\int d\mathcal{L}_\alpha P(\mathcal{L}_\alpha|\mathbf{J})\, \mathbf{J}\cdot\mathbf{J}^* \quad = \quad \overline{R^*}(\mathbf{J}). \tag{3.42}$$

L'inégalité (3.40) n'est alors rien d'autre que l'inégalité suivante :

$$\int d\,\mathbf{J} P_\mathrm{G}(\mathbf{J}) \left(\overline{R_\mathrm{B}}(\mathbf{J}) - \overline{R^*}(\mathbf{J})\right) \geq 0, \tag{3.43}$$

avec $\overline{R^*}(\mathbf{J})$, la performance moyenne de l'algorithme considéré et $\overline{R_\mathrm{B}}(\mathbf{J})$, la performance moyenne de l'estimateur bayesien pour un apprentissage dont la direction privilégiée n'est pas $\mathbf{B}$ mais $\mathbf{J}$.

Dans la limite thermodynamique, en supposant que les performances moyennes $\overline{R_\mathrm{B}}(\mathbf{J})$ et $\overline{R^*}(\mathbf{J})$ sont indépendantes de la direction $\mathbf{J}$ alors on obtient l'inégalité $\overline{R^*} \leq \overline{R_\mathrm{B}}$. Cette hypothèse d'indépendance des performances moyennes vis-à-vis de la direction privilégiée est due au fait que le probabilité *a priori* $P_\mathrm{G}(\mathbf{J})$ est cohérente avec la direction privilégiée.

On vient de démontrer l'optimalité de la performance moyenne de l'estimateur bayesien. Toutefois, dans la pratique, nous avons accès uniquement à la performance pour un ensemble d'apprentissage $\mathcal{L}_\alpha$ donné. En supposant la performance $R_\mathrm{B}(\mathcal{L}_\alpha)$ de l'estimateur bayesien et celle $R^*(\mathcal{L}_\alpha)$ de l'algorithme considéré automoyennantes dans la limite thermodynamique, alors, de l'inégalité sur les performances moyennes, on en déduit :

$$R^*(\mathcal{L}_\alpha) \leq R_\mathrm{B}(\mathcal{L}_\alpha). \tag{3.44}$$

En résumé, l'optimalité de la performance de l'approche bayesienne a été obtenue avec les hypothèses suivantes. Premièrement, la performance bayesienne $R_\mathrm{B}(\mathcal{L}_\alpha)$ et la performance $R^*(\mathcal{L}_\alpha)$ obtenue à partir de l'algorithme considéré sont automoyennantes dans la limite thermodynamique. Deuxièmement, la distribution *a priori* $P_\mathrm{G}(\mathbf{J})$ des directions possibles est cohérente avec la direction privilégiée $\mathbf{B}$, c'est-à-dire, que cette direction $\mathbf{B}$ est une direction quelconque. Il faut aussi noter l'inversion des intégrales sur $\mathbf{J}$ et $\mathcal{L}_\alpha$ qui, dans la limite thermodynamique, peut ne pas s'avérer correcte.

3.4 Controverse sur la position de la transition du premier ordre

Nous allons maintenant discuter de la position de la transition du premier ordre pour les deux approches et montrer que, selon l'approche considérée, la transition se produit à des valeurs de α différentes. Ceci a suscité une controverse sur l'optimalité de l'approche bayesienne. Nous essaierons de donner des pistes pour expliquer ce désaccord entre les deux approches.

3.4.1 Résultats analytiques

Pour déterminer la position de la transition du premier ordre, dans le cas de l'approche bayesienne, il faut revenir à l'énergie libre $\overline{F_{\rm G}}$ (3.26). Cette énergie libre est donnée, après calcul par la méthode des répliques, par :

$$\lim_{\substack{N\to+\infty \\ P\to+\infty \\ \alpha=P/N}} \frac{1}{N}\,\overline{F_{\rm G}} = \operatorname*{extr}_{R,q} f_{\rm G}(R,q) = f_{\rm G}(R_{\rm G},q_{\rm G}) \tag{3.45}$$

avec :

$$\begin{aligned} f_{\rm G}(R,q) &= -\frac{1}{2}\left\{\frac{1-R^2}{1-q}+\ln(1-q)\right\} \\ &\quad -\alpha\iint Dx\,Dy\,\exp\left(-V^*(x)\right)\ln\left\{\int Dz\,\exp\left(-V^*(t)\right)\right\}, \end{aligned} \tag{3.46}$$

$$t = xR + y\sqrt{q-R^2} + z\sqrt{1-q}. \tag{3.47}$$

Cette fonction $f_{\rm G}(R,q)$ possède plusieurs extrema $\{R_{\rm G},q_{\rm G}\}$ différents avec $R_{\rm G} = q_{\rm G}$ pour certaines valeurs de α. La solution, pour laquelle $f_{\rm G}(R_{\rm G},R_{\rm G})$ est la plus faible, doit être choisie. Comme pour toute transition du premier ordre, lorsque deux nouveaux extrema apparaissent en $\alpha = \alpha_1$ dans l'énergie libre $f_{\rm G}$, aucun de ces deux extrema ne correspond au minimum absolu de l'énergie libre. L'un des extrema correspond à une solution instable et l'autre correspond seulement à un minimum local. Ce n'est que pour $\alpha = \alpha_G > \alpha_1$ que la nouvelle branche stable devient un minimum absolu de l'énergie libre et que la branche initiale devient un minimum relatif. Le changement de branche entraîne une discontinuité de $R_{\rm G}(\alpha)$ et l'apparition d'une transition du premier ordre à α_G pour la performance de l'apprentissage de Gibbs.

Nous allons faire, ici, une digression pour montrer la stabilité locale de la solution obtenue avec l'hypothèse de symétrie des répliques. L'étude de cette stabilité montre que la même condition s'applique pour la solution de l'apprentissage de Gibbs $R_{\rm G}(\alpha)$ que pour la solution optimale de l'approche variationnelle $R_{\rm opt}(\alpha)$, c'est-à-dire, que la condition de stabilité locale peut s'écrire :

$$\frac{dR_{\rm G}}{d\alpha} > 0. \tag{3.48}$$

Dans le cadre général de l'apprentissage d'une direction privilégiée présenté au premier chapitre, Reimann et Van den Broeck ont établi que la condition de stabilité locale s'exprime [101] :

$$1 > \alpha\iint Dx\,Dy\,\exp\left(-V^*(x)\right)\left[\rho(x,y)\right]^2 \tag{3.49}$$

avec :

$$\rho(x,y) = \frac{\displaystyle\iint Dz\,Dz'\left(1-\frac{(z-z')^2}{2}\right)\exp\left(-V^*(t)-V^*(t')\right)}{\displaystyle\iint Dz\,Dz'\exp\left(-V^*(t)-V^*(t')\right)}, \tag{3.50}$$

avec t et t' définis par (3.47) où pour t', z est remplacé par z'. R et q sont remplacés par $R_{\rm G}$. En effectuant le changement de variable suivant :

$$x' = x\sqrt{1-R_{\rm G}} - y\sqrt{R_{\rm G}}, \tag{3.51}$$
$$y' = x\sqrt{R_{\rm G}} + y\sqrt{1-R_{\rm G}}, \tag{3.52}$$

alors $\rho(x,y)$ et la condition de stabilité locale s'expriment de la façon suivante :

$$\rho(x,y) = \rho(y') = 1 - \frac{\langle t^2\rangle_{y'}}{\langle 1\rangle_{y'}} + \frac{\langle t\rangle_{y'}^2}{\langle 1\rangle_{y'}^2}, \tag{3.53}$$

$$\int Dy' \, \frac{\langle t\rangle_{y'}^2}{\langle 1\rangle_{y'}} > R_{\rm G} \int Dy' \left[\rho(y')\right]^2 \langle 1\rangle_{y'}, \tag{3.54}$$

où l'on a utilisé les notations (1.71) introduites au premier chapitre et l'équation déterminant $R_{\rm G}$ pour remplacer α. Cette dernière équation est l'équation (1.72) où l'on a remplacé $R_{\rm opt}$ par $\sqrt{R_{\rm G}}$. La condition de stabilité locale (3.54) est identique à l'équation (1.76) si l'on remplace $\sqrt{R_{\rm G}}$ par $R_{\rm opt}$. Il en résulte que la démonstration effectuée pour la condition de stabilité locale de l'apprentissage optimal avec l'approche variationnelle est valable aussi pour l'apprentissage de Gibbs. La condition de stabilité locale se réduit alors à :

$$\frac{dR_{\rm G}}{d\alpha} > 0. \tag{3.55}$$

Il est intéressant de remarquer que cette démonstration ne fait en aucun cas appel au fait que l'on a considéré une densité de probabilité des exemples sous la forme d'une double gaussienne. Cette démonstration, valable quelque soit la perturbation V^* dans (1.2), est un résultat original qui n'avait pas été noté par Reimann et Van den Broek [101].

Revenons à la position de la transition du premier ordre suivant les deux approches possibles : l'approche bayesienne et l'approche variationnelle. Une transition du premier ordre est prédite pour l'apprentissage de Gibbs à α_G. La performance de l'apprentissage bayesien est obtenue par la relation $R_{\rm B} = \sqrt{R_{\rm G}}$. On en déduit que la transition du premier ordre pour l'apprentissage bayesien s'effectue pour la même taille réduite α_G que pour l'apprentissage de Gibbs. Dans le cas de l'approche variationnelle, il existe un potentiel optimal $V_{\rm opt}$ permettant d'obtenir la performance $R_{\rm opt}(\alpha_1)$ correspondant à la solution la plus élevée de l'équation (1.65) pour $\alpha = \alpha_1$, où α_1 est la taille réduite pour laquelle il apparaît trois solutions différentes à l'équation (1.65). L'étude de la stabilité locale par rapport à la brisure de symétrie des répliques a permis de mettre en évidence la stabilité de cette solution. La performance $R_{\rm opt}(\alpha)$ pour l'approche variationnelle possède alors une discontinuité à α_1 et la transition du premier ordre s'effectue à $\alpha_1 < \alpha_G$. Une question se pose alors naturellement : les potentiels optimaux obtenus par l'approche variationnelle permettent-ils l'obtention des états métastables de l'approche bayesienne pour $\alpha_1 < \alpha < \alpha_G$?

Dans les paragraphes qui suivent, nous allons essayer de comprendre les raisons de la différence de position de la transition du premier ordre pour les deux approches étudiées.

3.4.2 Résultats numériques

Comme on vient de le voir dans le cas de l'apprentissage de Gibbs, il se peut qu'il existe plusieurs solutions métastables pour une même valeur de α. Le choix entre les deux solutions stables s'effectue par l'intermédiaire de leur valeurs de l'énergie libre. La solution d'énergie libre minimale est retenue. Nous avons essayé de regarder si

un tel phénomène se produit pour les potentiels optimaux obtenus par l'approche variationnelle.

Avec l'approche variationnelle, pour chaque valeur de α, nous avons obtenu un potentiel optimal qui dépend de la valeur de α considérée. La direction $\mathbf{J}_{\text{opt}}$, qui minimise la fonction de coût (1.6) correspondante à ce potentiel, a une performance $R_{\text{opt}} = \mathbf{J}_{\text{opt}} \cdot \mathbf{B}$ maximale. Pour certaines valeurs de α, nous avons obtenu plusieurs potentiels, permettant d'obtenir des directions avec des performances différentes. La comparaison de l'énergie libre de ces différentes solutions n'a aucun sens dans ce cas, car elles correspondent à des fonctions de coût différentes. Cela reviendrait à comparer l'énergie libre de deux systèmes physiques différents. Par contre, il est intéressant et même indispensable de vérifier que l'énergie libre correspondante à un potentiel optimal ne possède pas plusieurs solutions stables pour la valeur de α qui a permis de déterminer ce potentiel optimal. Pour vérifier cela, il faut revenir à l'étude générale de ces potentiels optimaux. Cette étude consiste à considérer l'énergie libre donnée par l'équation (1.53) correspondant au potentiel optimal. Les paramètres R et c sont déterminés par les équations (1.57) et (1.58). Une seule solution stable nous est connue : $R = R_{\text{opt}}$ et $c = 1$. La question est de savoir si il existe d'autres solutions stables pour la valeur de α pour laquelle le potentiel optimal a été déterminé. Une étude analytique des potentiels optimaux s'est avérée inextricable du fait de l'absence d'une expression analytique simple pour ces potentiels optimaux. Nous avons donc effectué une étude numérique de différents potentiels optimaux. Nous représentons ici seulement ceux correspondant au couple de paramètres $(\rho,\sigma) = (1.1,0.5)$.

Les performances de potentiels optimaux sont représentées sur la figure 3.5 pour les valeurs de $R_{\text{opt}}(\alpha) = 0.1,0.2,\cdots,0.8,0.84$ et 0.87. Dans tous les cas où la solution $R_{\text{opt}}(\alpha)$ est localement stable par rapport à la brisure de symétrie des répliques, ces performances sont tangentes à la courbe d'apprentissage optimal aux points $(\alpha,R_{\text{opt}}(\alpha))$ pour lesquels ils ont été optimisés. Les performances des potentiels optimaux correspondant à des valeurs instables ($R = 0.5,0.6$ et 0.7) sont inaccessibles avec ces potentiels, comme il se doit. En effet, ils présentent une instabilité par rapport à la brisure de symétrie des répliques au voisinage de la partie instable de la courbe R_{opt}. Il est aussi intéressant de noter que, pour les faibles valeurs de la taille réduite α, il n'existe aucune solution $R > 0$ stable ce qui est consistant avec le retard à l'apprentissage observé pour la courbe d'apprentissage optimal $R_{\text{opt}}(\alpha)$. De plus, la solution $R = 0$ est instable aussi pour les faibles valeurs de α. Il apparaît donc un région pour laquelle aucune solution stable ne peut être obtenue avec l'hypothèse de symétrie des répliques. Ceci implique bien évidemment l'existence d'une solution avec brisure de symétrie des répliques. Le potentiel optimal n'ayant pas une expression analytique simple, il est compliqué de chercher des solutions brisant la symétrie des répliques. L'existence d'une solution avec brisure de symétrie, d'énergie plus faible que la solution symétrique pour $\alpha_1 < \alpha < \alpha_G$ est une possibilité envisageable pour réconcilier les deux approches.

3.4.3 Simulations numériques

Afin de tester la position de la transition du premier ordre et l'existence possible d'une solution avec brisure de symétrie des répliques, nous avons effectué des simulations numériques pour différents potentiels optimaux correspondant au couple de paramètres $(\rho,\sigma) = (1.1,0.5)$.

L'idée étant de tester la position de la transition du premier ordre, nous avons simulé deux valeurs différentes de α. La première $\alpha = 3.86$ est encadrée par $\alpha_1 = 3.79$ et $\alpha_G = 4.07$. La deuxième $\alpha = 4.71$ est supérieure à α_G et proche de $\alpha_c = 4.73$. Le potentiel optimal correspondant à chacune de ces deux valeurs de la taille réduite possède deux minima. En conséquence, la fonction de coût pour un ensemble d'ap-

FIG. 3.5 – *Performances des potentiels optimaux correspondant aux différentes valeurs de $R_{\rm opt}$ indiquées à droite de la figure. Ces performances sont tangentes à la courbe d'apprentissage optimal pour $R_{\rm opt}$ justifiant l'optimalité des potentiels pour ces valeurs. Les paramètres ρ et σ sont* 1.1 *et* 0.5 *respectivement.*

prentissage donné présente de nombreux minima locaux. Cette remarque est un argument en faveur d'une possible brisure de symétrie des répliques. Toutefois, la présence de nombreux minima n'est pas suffisante pour prouver l'existence de cette brisure. En effet, le minimum correspondant au bassin d'attraction du professeur est privilégié par rapport aux autres. Les simulations montrent que la direction correspondante au minimum local de l'énergie libre située à l'intérieure du bassin d'attraction de la direction privilégiée $\mathbf{B}$ possède un produit scalaire avec $\mathbf{B}$ qui correspond à la valeur de $R_{\rm opt}(\alpha)$ prédite par l'approche variationnelle à de faibles effets de taille finie près. Par contre, les directions correspondantes aux autres minima ont un produit scalaire avec $\mathbf{B}$ proche de 0. De plus, il s'avère que le minimum correspondant au bassin d'attraction de $\mathbf{B}$ est le minimum absolu de la fonction de coût pour un grand nombre d'ensembles d'apprentissage.

Nous avons déterminé le pourcentage p d'ensembles d'apprentissage pour lesquels le minimum absolu de la fonction de coût correspond au minimum du bassin d'attraction de la direction privilégiée. Si ce pourcentage tend vers 100%, dans la limite thermodynamique, alors on s'attend à ce qu'il n'y ait pas de brisure de symétrie des répliques et que la performance soit celle prédite par l'approche variationnelle.

Afin d'obtenir le pourcentage p, nous avons tout d'abord minimisé la fonction de

FIG. 3.6 – *Pourcentage p d'ensembles d'apprentissage pour lesquels le minimum absolu de la fonction de coût optimale correspond au minimum du bassin d'attraction de la direction privilégiée. Les triangles vers le haut correspondent à des simulations numériques correspondant à $\alpha = 4.71$ et les triangles vers le bas à $\alpha = 3.86$. Pour le couple de paramètre $(\rho,\sigma) = (1.1,0.5)$ considéré, $\alpha_1 = 3.79$, $\alpha_G = 4.07$ et $\alpha_c = 4.73$.*

coût optimale avec, comme condition initiale, la direction privilégiée **B** et nous avons utilisé de faibles déplacements au cours de la descente de gradient afin d'obtenir le minimum correspondant au bassin d'attraction de **B**. Ensuite, afin de déterminer si ce minimum est le minimum absolu de la fonction de coût, nous avons effectué une minimisation pour différentes conditions initiales aléatoires. Pour se convaincre que l'on a obtenu le minimum absolu de la fonction de coût, nous avons testé un nombre de conditions initiales suffisant pour obtenir plusieurs fois le minimum correspondant au bassin d'attraction de **B**. Cette condition n'est pas suffisante mais elle est nécessaire. En effet, si l'on teste un nombre de conditions initiales aléatoires insuffisant pour retrouver un des minima connus, il y a peu de chance pour que l'on ait détecté le minimum absolu.

La figure 3.6 représente l'évolution du pourcentage p pour les deux valeurs de α simulées en fonction de $1/N$. Premièrement, il faut noter les faibles valeurs de N simulées. Celles-ci sont comprises entre 4 et 40. La raison de ces faibles valeurs est simple. Le nombre de minima locaux augmente rapidement avec N, ainsi que le temps de convergence de chaque minimisation. Le temps nécessaire pour les simulations devient vite prohibitif. Ayant noté cette limitation, l'évolution du pourcentage p en fonction de $1/N$ tend à faire penser que, pour $\alpha = 4.71 > \alpha_G$, ce pourcentage tend vers 100% dans la limite thermodynamique et 0% pour $\alpha = 3.86 < \alpha_G$. Ce comportement est un argument en faveur de la position de la transition du premier ordre prédite par l'approche bayesienne puisque, pour $\alpha < \alpha_G$, les simulations semblent montrer un désaccord avec les prédictions de l'approche variationnelle. Toutefois, plusieurs limitations à ces simulations sont à noter. Les faibles valeurs de N simulées est l'une d'elles. La deuxième est le comportement de p pour $\alpha = 4.71$. Ce pourcentage décroît en fonction de $1/N$ pour toutes les valeurs de N simulées excepté pour la dernière valeur $N = 40$. Le comportement de ce pourcentage n'est donc pas monotone. Cette remarque peut aussi bien s'appliquer pour le comporte-

ment du pourcentage pour $\alpha = 3.86$ à des valeurs de N supérieures à celles simulées. De tels comportements non monotones ont déjà été observés par Derrida, Griffiths et Prügel-Bennett [30] ainsi que par Schröder et Urbanczik [109] (dans un commentaire au papier de Nadler et Fink [89]) pour d'autres problèmes d'apprentissage.

L'apprentissage de Gibbs consiste à tirer aléatoirement une direction avec la densité de probabilité (3.22). Des simulations numériques ont permis de mettre en évidence un nombre important de maxima locaux de cette densité de probabilité. L'existence de ces maxima locaux ne permet pas la convergence rapide d'un algorithme simple de Monte-Carlo pour échantillonner les directions $\mathbf{J}$. Les maxima sont suffisamment importants pour empêcher l'algorithme de les visiter tous. Ceci pourrait faire penser à l'existence d'une solution avec brisure de symétrie des répliques pour l'apprentissage de Gibbs.

En résumé de ces simulations numériques, aucune preuve évidente ne peut être obtenue de l'existence ou de l'absence d'une solution avec brisure de symétrie des répliques pour l'approche variationnelle comme pour l'approche bayesienne de l'apprentissage optimal.

3.4.4 Conclusions actuelles sur la controverse

Récemment, Herschkowitz et Nadal [56, 57] ont mis en évidence la possibilité d'obtenir la transition du premier ordre pour l'apprentissage de Gibbs pour des tailles réduites $\alpha < \alpha_G$. L'idée est que la présence d'une solution métastable pour les tailles réduites comprises entre α_1 et α_G permet d'envisager l'obtention de cette solution métastable. Physiquement, il est bien connu qu'il est possible d'obtenir des états métastables. C'est le cas par exemple pour un liquide surfondu à des températures inférieures à la température de solidification. Un résultat surprenant obtenu par ces auteurs est le fait que l'on ne puisse pas obtenir l'état métastable pour toutes les valeurs de α comprises entre α_1 et α_G mais seulement pour $\alpha > \alpha_2$. Pour les paramètres $\rho = 1.2$ et $\sigma = 0.5$ étudiés plus particulièrement par ces auteurs, ils ont montré que $\alpha_2 = 2.515$ alors que $\alpha_1 = 2.49$ et $\alpha_G = 2.527$ [56, 57].

En conclusion, il apparaît que les deux approches de l'apprentissage optimal considérées ne sont pas cohérentes l'une par rapport à l'autre dans le cas où l'hypothèse de symétrie des répliques est considérée. Ce résultat est d'autant plus surprenant que prises séparément les deux approches apparaissent cohérentes vis-à-vis de cette hypothèse. Il semble nécessaire de rechercher des solutions brisant la symétrie des répliques pour réconcilier les deux approches. Il se pose alors la question de savoir si les transitions du premier ordre obtenues quelque soit l'approche considérée seront alors conservées.

3.5 Interprétation des 3 phases d'apprentissage

Dans cette section, nous allons revenir à l'étude des courbes d'apprentissage optimal en fonction de la taille réduite α de l'ensemble d'apprentissage. Le but est de comprendre les différentes phases de l'apprentissage optimal, qui sont généralement séparées par des transitions du premier ou du second ordre. La position de la transition du premier ordre, discutée dans la section précédente, n'est pas déterminante pour cette discussion, l'essentiel étant que, quelque soit l'approche utilisée, cette transition semble exister.

Pour expliquer les caractéristiques des différentes phases, je vais introduire, premièrement, une notion d'apprentissage de la composante principale [102, 90, 91] puis une formulation supervisée du problème de la double gaussienne pour laquelle on déterminera les propriétés de l'apprentissage optimal.

3.5.1 Apprentissage de la composante principale

L'apprentissage de la composante principale d'un ensemble de points, qui dans le contexte de ce travail est appelé ensemble d'apprentissage, consiste à déterminer l'axe correspondant à la plus grande variance de l'ensemble d'apprentissage. Si la distribution des exemples a une variance plus faible suivant l'axe d'anisotropie, le problème est celui de trouver la plus petite composante. Par abus de langage, nous appelons composante principale l'axe d'anisotropie dans les deux cas. Pour cela, il est assez simple de se rendre compte [102] que, si la distribution des points a une moyenne nulle, comme nous l'avons supposé dans ce chapitre, la recherche du minimum de la fonction de coût suivante :

$$E(\mathbf{J};\mathcal{L}_\alpha) = \epsilon \sum_{\mu=1}^{P} \left(\mathbf{J}\cdot\boldsymbol{\xi}^\mu\right)^2 \tag{3.56}$$

permet de résoudre ce problème avec $\epsilon = -1$ pour un axe d'anisotropie de plus grande variance et $\epsilon = 1$ sinon. En effet, si $\epsilon = -1$, la direction $\mathbf{J}_{\rm cp}$ qui minimise la fonction de coût (3.56) est la direction selon laquelle la projection des exemples est aussi grande que possible. Si les exemples sont symétriques par rapport à l'origine, cette direction correspond bien à celle de plus grande variance de l'ensemble d'apprentissage. Il est intéressant de noter que la fonction de coût correspond à un potentiel $V(\lambda) = \epsilon\lambda^2$ quadratique.

La distribution des exemples considérée dans ce chapitre possède une symétrie par rapport à l'origine, une variance unité suivant les $N-1$ directions perpendiculaires à $\mathbf{B}$ et une variance $\langle\lambda^2\rangle = \rho^2+\sigma^2$ suivant la direction $\mathbf{B}$. Dans ces conditions, la détermination de la direction $\mathbf{J}_{\rm cp}$ (de plus grande variance si $\langle\lambda^2\rangle > 1$ et de plus faible variance si $\langle\lambda^2\rangle < 1$) est une estimation intéressante de la direction $\mathbf{B}$.

La méthode des répliques permet le calcul de la performance $R_{\rm cp} = \mathbf{J}_{\rm cp}\cdot\mathbf{B}$ de l'apprentissage de la composante principale dans la limite thermodynamique [102, 125] :

$$R_{\rm cp}(\alpha) = \sqrt{\frac{\alpha-\alpha_c}{\alpha+1/(\langle\lambda^2\rangle-1)}} \tag{3.57}$$

pour $\alpha > \alpha_c$ avec $\alpha_c = (\langle\lambda^2\rangle-1)^{-2}$. La première remarque que l'on peut faire est qu'il existe un retard à l'apprentissage jusqu'à $\alpha = \alpha_c$. La taille réduite α_c est la même que celle à la transition du second ordre observée dans les courbes d'apprentissage optimal (*cf.* (3.17)). De plus, le développement de $R_{\rm opt}(\alpha)$ autour de α_c est identique à celui obtenu dans le cas de l'apprentissage optimal (*cf.* (3.17)). Enfin, la ligne $\rho^2+\sigma^2 = 1$ correspond à des paramètres (ρ,σ) pour lesquels l'apprentissage de la composante principale n'est pas possible, car elle n'est pas définie : toutes les directions ont la même variance unité. On peut aussi remarquer qu'en dehors de cette ligne, l'apprentissage de la composante principale est asymptotiquement parfait : $R_{\rm cp} \to 1$ lorsque $\alpha \to +\infty$.

3.5.2 Scénario supervisé

Nous allons maintenant supposer que l'ensemble d'apprentissage est constitué non seulement des exemples mais aussi d'une variable binaire $\tau^\mu = \pm 1$ qui caractérise l'amas d'où provient l'exemple $\boldsymbol{\xi}^\mu$. Ainsi, le problème devient un cas particulier d'apprentissage supervisé, où les exemples ne sont pas linéairement séparables. Ce problème est de nature différente de ceux abordés au deuxième chapitre, car ici on ne cherche pas à classer les exemples, mais à déterminer la direction privilégiée $\mathbf{B}$.

On supposera $\tau^\mu = +1$ si $\boldsymbol{\xi}^\mu$ provient de l'amas centré en $+\rho\mathbf{B}$ et -1 si $\boldsymbol{\xi}^\mu$ provient de celui centré en $-\rho\mathbf{B}$. Cette variable binaire apporte une information supplémentaire par rapport au problème non supervisé. La performance de l'apprentissage optimal du problème supervisé sera une borne supérieure de la performance du problème non supervisé.

Le nouvel ensemble d'apprentissage $\widetilde{\mathcal{L}}_\alpha$ est constitué des couples $(\boldsymbol{\xi}^\mu,\tau^\mu)$ avec $\mu = 1,\cdots,P$. La densité de probabilité d'un exemple constitué du couple $(\boldsymbol{\xi},\tau)$ est donnée par :

$$P((\boldsymbol{\xi},\tau)|\mathbf{B}) = \frac{1}{(2\pi)^{N/2}} \exp\left(-\frac{\boldsymbol{\xi}\cdot\boldsymbol{\xi}}{2} - \widetilde{V}^*\left(\tau\boldsymbol{\xi}\cdot\mathbf{B}\right)\right) \tag{3.58}$$

avec $\widetilde{V}^*$ définie par :

$$\begin{aligned} P(\gamma) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\gamma^2}{2} - \widetilde{V}^*(\gamma)\right) && (3.59)\\ &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\gamma-\rho)^2}{2\sigma^2}\right). && (3.60) \end{aligned}$$

La symétrie des deux amas (position symétrique par rapport à l'origine et poids identique de chaque amas) permet de reformuler ce problème supervisé comme un problème non supervisé pour la variable $\boldsymbol{\zeta} = \tau\boldsymbol{\xi}$. Cette reformulation, proposée par Reimann et Van den Broeck [101], a été décrite au deuxième chapitre. On peut remarquer que le nouveau problème non supervisé correspond à une densité de probabilité des exemples $\boldsymbol{\zeta}^\mu$ gaussienne simple. La densité de probabilité dans la direction privilégiée $\mathbf{B}$ correspond à une gaussienne de moyenne ρ et de variance σ alors que, dans les autres directions, la moyenne est nulle et la variance unité.

Le problème non supervisé d'une distribution simple gaussienne a été étudié de manière très détaillée par Reimann, Van den Broeck et Bex [102]. La performance optimale $R_{\text{sup}} = \mathbf{J}_{\text{sup}}\cdot\mathbf{B}$ de ce problème est donnée par :

$$\begin{aligned} R_{\text{sup}} &= \left\{\frac{1-\alpha\left(1-2(\rho^2+\sigma^2+\sigma^4+\rho^2\sigma^2)\right)-\sqrt{Q}}{2(1-\sigma^2)\left(1-\alpha(1-\rho^2\sigma^2)\right)}\right\}^{1/2}, && (3.61)\\ Q &= \left(1-\alpha\left(1-2\sigma^2+\sigma^4+\sigma^2\rho^2\right)\right)^2 + 4\alpha\rho^2\sigma^2. && (3.62) \end{aligned}$$

L'existence d'une moyenne non nulle de la densité de probabilité de $\boldsymbol{\zeta}$ selon la direction privilégiée $\mathbf{B}$ entraîne que R_{sup} croît en fonction de α dès $\alpha = 0$. Cette croissance est continue pour toutes valeurs de α et $R_{\text{sup}} \to 1$ lorsque $\alpha \to +\infty$, garantissant un apprentissage asymptotiquement parfait.

3.5.3 Discussion des phases de l'apprentissage optimal

Revenons, maintenant, à l'apprentissage optimal du problème non supervisé. La distribution des exemples est donnée par (3.2) avec la perturbation V^* définie par (3.4), et est constituée de deux gaussiennes dans la direction $\mathbf{B}$ que l'on veut déterminer. Les figures 3.7 illustrent les différentes phases d'apprentissage observées Chaque figure correspond à un couple de paramètres (ρ,σ) : $(\rho,\sigma) = (1.4,0.5)$ (figure *a*), (1.2,0.5) (figure *b*) et (1.1,0.5) (figure *c*). En plus de la courbe d'apprentissage optimal du problème non supervisé $R_{\text{opt}}(\alpha)$ (courbe en trait plein), il a été représenté la courbe d'apprentissage de la composante principale $R_{\text{cp}}(\alpha)$ (courbe en tirets) ainsi que la courbe d'apprentissage optimal du problème supervisé $R_{\text{sup}}(\alpha)$ (courbe

FIG. 3.7 – *Courbes d'apprentissage pour trois couples de paramètres* (ρ,σ). $(\rho,\sigma) = (1.4,0.5)$ *figure a,* $(\rho,\sigma) = (1.2,0.5)$ *figure b et* $(\rho,\sigma) = (1.1,0.5)$ *figure c. Les courbes en trait plein correspondent à l'apprentissage optimal du problème non supervisé* $R_{\text{opt}}(\alpha)$*, les courbes en tirets à l'apprentissage de la composante principale* $R_{\text{cp}}(\alpha)$ *et les courbes en pointillés à l'apprentissage optimal du problème supervisé* $R_{\text{sup}}(\alpha)$.

en pointillés). Comme on l'a déjà mentionné, ces courbes constituent des bornes inférieure et supérieure aux courbes de l'apprentissage optimal non supervisé.

La première des phases observées dans le cas de l'apprentissage optimal, c'est-à-dire, la phase correspondante aux plus petites valeurs de α, est une absence totale d'apprentissage ($R_{\rm opt} = 0$). Cette phase de retard à l'apprentissage, est présente pour toutes les valeurs des paramètres (ρ,σ) possibles. Comme on l'a vu précédemment, ce retard à l'apprentissage est dû au fait que la distribution des exemples a une moyenne nulle ($\langle\lambda\rangle = 0$), et il n'est pas limité au problème de la double gaussienne. En général, une symétrie entre les directions $\mathbf{B}$ et $-\mathbf{B}$ entraîne la nullité de tous les moments d'ordre impair de la distribution des exemples et induit par conséquent un retard à l'apprentissage. Toutefois, ce retard se produit aussi en l'absence de cette symétrie, comme par exemple, pour une distribution des exemples pour laquelle $\langle\lambda\rangle = 0$ mais $\langle\lambda^3\rangle \neq 0$.

La deuxième phase d'apprentissage est similaire à l'apprentissage de la composante principale. $R_{\rm opt}(\alpha)$ croît à partir de $R_{\rm opt} = 0$ de manière continue en fonction de α pour $\alpha > \alpha_c$. Cette deuxième phase n'est pas observée pour tous les paramètres (ρ,σ). Elle apparaît uniquement dans les cas où il existe une transition du second ordre dans la courbe d'apprentissage optimal $R_{\rm opt}(\alpha)$, c'est-à-dire, pour toutes les valeurs des paramètres (ρ,σ) qui ne se situent pas dans la région gris clair de la figure 3.4. Cela correspond, par exemple, aux cas a et b de Fig 3.7. Dans ces deux cas, la courbe d'apprentissage de la composante principale est très proche de celle de l'apprentissage optimal. Dans le cas a, les deux courbes s'écartent progressivement et continûement à des valeurs finies de α puisqu'il n'existe pas de transition du premier ordre. Par contre, pour le cas b, la similitude entre les deux courbes s'arrête dès que la taille réduite α atteint la valeur α_1 correspondant à la transition du premier ordre entre une valeur de $R_{\rm opt}$ proche de $R_{\rm pc}$ et une valeur proche de $R_{\rm sup}$. La différence entre les deux cas a et b est que, dans le cas a on a un passage continu (*cross-over*) d'une solution proche de l'apprentissage de la composante principale à une solution proche de l'apprentissage du problème supervisé alors que, dans le cas b, ce passage s'effectue par un saut à $\alpha = \alpha_1$. Nous verrons dans la section suivante que cette phase intermédiaire de l'apprentissage correspond bien à l'apprentissage de la composante principale. En effet, la forme du potentiel optimal pour cette phase d'apprentissage est proche de celle d'un potentiel quadratique.

Dans le cas c, la phase d'apprentissage similaire à l'apprentissage de la composante principale n'est pas présente, puisqu'elle est masquée par la transition du premier ordre entre une phase d'absence totale d'apprentissage ($R_{\rm opt} = 0$) et une phase où la valeur de $R_{\rm opt}$ est proche de $R_{\rm sup}$, la performance de l'apprentissage optimal du problème supervisé. La raison de l'absence de phase d'apprentissage proche de l'apprentissage de la composante principale est évidente : la variance de la distribution des exemples dans la direction privilégiée est trop proche de celle dans les autres directions pour pouvoir être détectée.

La troisième phase d'apprentissage correspond aux grandes valeurs de la taille réduite α. Les performances de cette phase d'apprentissage sont proches de celles de l'apprentissage optimal du problème supervisé. Cette fois-ci, l'apprentissage non supervisé optimal ne se limite pas à déterminer la direction de plus grande (ou de plus faible) variance de l'ensemble d'apprentissage mais détecte la structure proprement dite de la distribution des exemples. Ceci se reflète d'ailleurs dans l'allure du potentiel optimal, qui comporte deux minima proches de $\lambda = \pm\rho$, comme nous le verrons dans le paragraphe suivant.

3.5.4 Comportement du potentiel optimal

Nous allons discuter dans cette section l'allure du potentiel optimal en fonction de ρ et σ ainsi que son évolution avec la taille réduite α.

Je vais commencer par une remarque évidente mais valable quelques soient les paramètres (ρ,σ) : pour la première phase d'apprentissage, celle correspondant à une absence totale d'apprentissage, il n'est pas possible de définir un potentiel optimal.

Il est intéressant de remarquer que ce n'est pas la position des paramètres (ρ,σ) dans une région correspondante à l'une des différentes courbes d'apprentissage optimal possibles représentées sur la figure 3.4 qui déterminent l'allure du potentiel optimal. Par contre, la forme du potentiel optimal est fortement corrélée avec les différentes phases de l'apprentissage introduites dans la section précédente. Ceci est naturel : nous avons déjà mentionné que, étant donné un problème (c'est-à-dire un couple de paramètres (ρ,σ)), il y a un potentiel optimal différent pour chaque valeur de α.

Il est possible de distinguer quatre régions distinctes dans l'espace des paramètres (ρ,σ) qui correspondent à une évolution différente du potentiel optimal en fonction de α.

Un premier critère pour l'évolution du potentiel optimal consiste à regarder le comportement de $V_{\rm opt}(\lambda)$ pour les grandes valeurs de λ. Ce comportement dépend uniquement de σ :

$$V_{\rm opt}(\lambda) \sim \begin{cases} -\lambda^2 & \text{si} \quad \sigma > 1, \\ +\lambda^2 & \text{si} \quad \sigma < 1. \end{cases} \tag{3.63}$$

Ce comportement quadratique fait naturellement penser à l'apprentissage de la composante principale, dont le potentiel est $\pm\lambda^2$. La question qui se pose alors est pourquoi le changement de comportement du potentiel optimal ne s'effectue pas sur la ligne $\rho^2 + \sigma^2 = 1$. En effet, cette ligne sépare les paramètres (ρ,σ) pour lesquels la variance dans la direction privilégiée est supérieure à l'unité de ceux pour lesquels elle est inférieure à l'unité. Elle sépare ainsi les paramètres pour lesquels le potentiel utilisé pour l'apprentissage de la composante principale est λ^2 de ceux correspondants au potentiel $-\lambda^2$. En fait, le changement de comportement aux grandes valeurs de λ peut se comprendre si l'on regarde le rapport entre la distribution des exemples dans la direction privilégiée de celle dans une direction perpendiculaire. Ce rapport, dans la limite des grandes valeurs de λ, est nul si $\sigma < 1$ et infini si $\sigma > 1$.

Le premier critère permet de mettre en évidence les régions marquées d'un S sur la figure 3.4 pour signifier que le potentiel optimal reste avec un unique extremum en $\lambda = 0$ quelque soit la valeur de α. La région correspondante à $\sigma > 1$ a un potentiel optimal concave ($\epsilon = -1$) quelque soit ρ, et la région correspondante à $\sigma < 1$ et $\rho < 1$ a un potentiel convexe ($\epsilon = 1$). Le potentiel optimal évolue faiblement en fonction de la taille réduite α. Celui-ci reste assez proche du potentiel $\epsilon\lambda^2$. Pour les paramètres (ρ,σ) correspondant à ces deux régions, la largeur des gaussiennes dans la direction privilégiée est suffisamment grande (région pour laquelle $\sigma > 1$) ou suffisamment faible (région pour laquelle $\sigma < 1$) pour permettre que l'apprentissage de la composante principale soit pratiquement optimal. La courbe d'apprentissage optimal dans ces régions des paramètres (ρ,σ) ne montre pas de transition du premier ordre entre un type d'apprentissage proche de celui de la compostante principale et un autre proche de celui de l'apprentissage supervisé optimal.

La région restante peut se diviser en deux régions distinctes séparées par la ligne $\rho^2 + \sigma^2 = 1$. Dans ces deux dernières régions, la dépendance du potentiel optimal en fonction de la taille réduite α est importante.

Nous allons d'abord discuter de la région où la direction recherchée est celle de plus faible variance. Elle correspond à des paramètres (ρ,σ) tels que $\rho^2 + \sigma^2 < 1$. La figure 3.8 qui correspond aux paramètres $\rho = 0.65$ et $\sigma = 0.5$ permet d'inférer l'évolution du potentiel optimal en fonction de α dans cette région. Ce couple de paramètres (ρ,σ) se situe dans la région blanche immédiatement au dessus de la région

FIG. 3.8 – *Potentiel optimal pour trois valeurs de la taille réduite :* $\alpha = 11.2,14.9$ *et* 18.3. *Les paramètres* $\rho = 0.65$ *et* $\sigma = 0.5$ *pour lesquels sont calculés ces potentiels optimaux correspondent à un point situé dans la région blanche qui n'est pas marquée d'un S, au dessous de la ligne* $\rho^2 + \sigma^2 = 1$ *de l'espace des paramètres* (ρ,σ). *Deux minima apparaissent à l'origine et s'écartent jusqu'à des valeurs proches de* $\pm\rho$ *lorsque* α *augmente.*

($\sigma < 1$) marquée d'un S. Pour la plus petite valeur de α pour laquelle le potentiel optimal est représenté, celui-ci est très similaire au potentiel $+\lambda^2$, et ne comporte qu'un seul minimum. Cela explique pourquoi la phase d'apprentissage correspondante est semblable à celle de l'apprentissage de la composante principale. Lorsque la valeur de α augmente, le potentiel optimal évolue jusqu'à posséder deux minima, qui apparaissent autour de $\lambda = 0$ et s'éloignent progressivement pour atteindre une valeur proche de $\lambda = \pm\rho$. Dans ce cas, la courbe d'apprentissage optimal ne possède pas de transition du premier ordre. L'évolution du potentiel optimal est continue. Dans le cas où la transition du premier ordre existe, l'évolution du potentiel optimal à la transition est abrupte et l'on passe généralement d'un potentiel qui ne possédait qu'un seul minimum à un potentiel possédant deux minima déjà assez proches de $\lambda = \pm\rho$. Il n'a toutefois pas pu être montré rigoureusement que les deux minima ne sont pas apparus avant la transition du premier ordre de $R_{\text{opt}}(\alpha)$.

La dernière région est située entre les lignes $\rho^2 + \sigma^2 = 1$ et $\sigma = 1$. La figure 3.9, correspondant aux paramètres $\rho = 1.2$ et $\sigma = 0.5$, montre le comportement du potentiel optimal dans cette région. Ces valeurs de ρ et σ correspondent au carré intermédiaire de la figure 3.4. La courbe d'apprentissage correspondant à ces paramètres est représentée sur la figure 3.2. Les deux premières valeurs de α ($\alpha =$ 2.23 et 2.30) correspondent à des valeurs plus faibles que $\alpha_1 = 2.49$, position de la transition du premier ordre, tandis que la troisième $\alpha = 3.00$ correspond à une valeur supérieure. Il est possible de remarquer que même dans la phase correspondant à l'apprentissage de la composante principale, c'est-à-dire $\alpha_c < \alpha < \alpha_1$, le potentiel optimal possède deux minima. Toutefois, ces minima sont situés très loin de l'origine (il faut souligner, à ce sujet, le changement d'échelle pour λ par rapport à la figure précédente). Le potentiel optimal peut être considéré comme similaire au potentiel $-\lambda^2$ dans la région où va se situer la grande majorité des exemples. En effet, très peu d'exemples seront concernés par la région du potentiel au delà des deux minima.

FIG. 3.9 – *Potentiel optimal pour trois valeurs de la taille réduite :* $\alpha = 2.23$,2.35 *et* 3.00. *Les paramètres* $\rho = 1.2$ *et* $\sigma = 0.5$ *pour lesquels sont calculés ces potentiels optimaux correspondent à un point situé dans la région gris foncé de l'espace des paramètres* (ρ,σ). *Les deux premières valeurs de* α *correspondent à la région intermédiaire entre la transition du second ordre et la transition du premier ordre. La dernière valeur de* α *correspond à la région après transition du premier ordre.*

Cela n'est plus vrai après la transition du premier ordre. En effet, pour $\alpha = 3.00$, les minima se sont considérablement rapprochés de l'origine et se situent à des valeurs assez proches de $\pm\rho$. Dans le cas où la transition du premier ordre n'existe pas (région blanche), le rapprochement de ces minima est continu.

Je voudrais faire quelques remarques générales à partir de cette étude du potentiel optimal. Tout d'abord, la transition du premier ordre ne semble pas être reliée directement à l'apparition des deux minima dans le potentiel optimal. En effet, les deux dernières régions montrent la possibilité d'une apparition continue de deux minima à partir de l'origine, ou le rapprochement continu de deux minima à partir de l'infini, pour des paramètres pour lesquels la courbe d'apprentissage optimal ne possède pas de transition du premier ordre. Cette remarque soulève encore une fois la question de la possibilité de brisure de symétrie des répliques. Si l'on ne considère que les régions où la transition du premier ordre n'existe pas, les deux approches de l'apprentissage optimal, présentées dans ce mémoire, que sont l'approche variationnelle et l'approche bayesienne, sont cohérentes et l'hypothèse de symétrie des répliques ne semble alors pas poser de difficulté. Toutefois, l'existence de deux minima pour le potentiel optimal entraîne l'existence d'un grand nombre de minima locaux pour la fonction de coût correspondante.

La conclusion est que la brisure de symétrie des répliques n'est pas directement reliée à l'existence de minima locaux et que l'apparition de minima locaux ne peut donc pas prouver l'existence d'une solution avec brisure de symétrie des répliques. La controverse entre les deux approches reste entière, aucun argument simple pouvant nous permettre de conclure sur l'existence ou non d'une brisure de symétrie des répliques pour l'une ou l'autre des approches, voire même pour les deux.

Deuxième partie

Au-delà du cas uniaxial : étude de deux approches

Introduction

Dans la première partie de la thèse, nous nous sommes concentrés sur l'apprentissage d'une direction de l'espace des données à partir d'un ensemble d'apprentissage. Nous avons supposé que cette direction est la seule inconnue du problème. Deux types de problèmes ont été abordés : la détermination de la direction privilégiée de la densité de probabilité des exemples dans le cas d'un apprentissage non supervisé et la détection de la direction perpendiculaire à l'hyperplan séparateur dans le cas d'un apprentissage supervisé.

Dans la deuxième partie de la thèse, nous allons aborder des problèmes d'apprentissage plus complexes, c'est-à-dire, pour lesquels une simple direction de l'espace des données n'est pas suffisante pour résoudre le problème. Nous allons nous consacrer à l'apprentissage supervisé, c'est-à-dire, que nous allons considérer qu'à chaque exemple est associé une classe. On suppose que le professeur qui attribue la classe à chaque exemple a une structure plus complexe qu'un hyperplan séparateur de l'espace des données. Deux approches différentes, proposées pour résoudre ce type de problèmes complexes, vont être abordées, celle des machines à exemples supports et celle de la machine de parité. Dans ces deux approches, on fait une transformation de l'espace des données vers un nouvel espace, souvent appelé de représentations internes. La transformation doit être telle que les exemples dans le nouvel espace soient linéairement séparables. Ainsi, la difficulté de trouver une surface séparatrice complexe dans l'espace des données est ramené à celle de trouver l'application adéquate entre l'espace des données et celui des représentations.

L'approche des machines à exemples supports utilise un pré-traitement des exemples qui consiste à étendre l'espace des données à un espace de plus grande dimension. Il conviendra de choisir judicieusement la transformation permettant de passer de l'espace des données à l'espace des représentations afin que l'ensemble d'apprentissage soit linéairement séparable dans ce dernier. Notre contribution est une des premières études de cette approche avec les outils de la Physique Statistique.

Dans la deuxième approche considérée, on construit des représentations internes en augmentant progressivement le nombre de dimensions, ce qui correspond à un apprentissage par étapes successives des paramètres à adapter. La procédure est plus simple à expliquer dans le langage des réseaux de neurones en couches. Alors, les composantes des vecteurs dans l'espace des données sont les N entrées du réseau. Les états binaires des neurones cachées peuvent être interprétés comme les représentations internes des données. Le neurone de sortie attribue des classes aux représentations internes. Chaque étape de l'apprentissage consiste à déterminer un nouveau perceptron simple. Les classes successives que l'on se propose d'apprendre avec celui-ci à chaque étape doivent corriger les erreurs des étapes précédentes. Ainsi, on génère des représentations internes binaires pour chaque région de l'espace des données, définies par l'ensemble des hyperplans séparateurs correspondant aux perceptrons simples. Chaque perceptron attribue une classe aux points en fonction du côté où ils se trouvent par rapport à l'hyperplan séparateur correspondant. La classe finale d'un point est le produit des différentes classes données par les perceptrons simples. Cette procédure d'apprentissage correspond à un algorithme incrémental

pour la machine de parité. Nous nous intéresserons plus particulièrement à la quantité d'exemples de classes aléatoires qu'il est possible d'apprendre avec un nombre donné de perceptrons.

Chapitre 1

Machine à exemples supports

Dans ce chapitre, je vais présenter une des premières études faites avec des outils de la Physique Statistique de ce que je nomme en français les *machines à exemples supports* (encore appelées en anglais Support Vector Machines). Ces machines constituent une approche possible pour l'apprentissage de tâches complexes. Nous nous limiterons par la suite à des tâches de classification. J'entends alors par complexes, des tâches pour lesquelles les exemples ne sont pas linéairement séparables. L'apprentissage ne peut alors pas se restreindre à la recherche d'un simple hyperplan séparateur. Il faut rechercher des surfaces séparatrices plus complexes que de simples hyperplans.

Une des possibilités pour déterminer de telles surfaces séparatrices est d'utiliser les machines à exemples supports introduites par Vapnik [122, 123, 16, 50, 74, 17, 27]. L'idée à la base de ces machines est d'utiliser une transformation non linéaire de l'espace des données vers un espace de dimension beaucoup plus importante que l'on appellera *espace des représentations* (en anglais, on l'appelle *features space*). Si ce nouvel espace est suffisamment étendu et la transformation bien choisie, les images par cette transformation des exemples de l'ensemble d'apprentissage sont linéairement séparables dans l'espace des représentations. De nombreux algorithmes permettent alors de déterminer un hyperplan séparateur classifiant correctement tous les exemples de l'ensemble d'apprentissage. Parmi ces hyperplans séparateurs, celui de stabilité maximale, introduit dans la première partie de la thèse (Chap.2), définit la machine à exemples supports.

De nombreuses applications réalistes utilisent les machines à exemples supports pour l'apprentissage de tâches complexes de classification. La reconnaissance de caractères manuscrits en est un bon exemple [122, 123, 16, 50]. De manière surprenante, l'erreur de généralisation, c'est-à-dire, la probabilité que la machine à exemples supports classe de manière incorrecte un exemple ne faisant pas partie de l'ensemble d'apprentissage, est très faible dans ces applications. Aucune étude théorique satisfaisante ne permet d'expliquer les raisons des faibles erreurs de généralisation obtenues. Les résultats que nous avons obtenus pour une famille de transformations particulières permettent de donner des pistes sur les raisons de ces faibles erreurs de généralisation des machines à exemples supports [24].

1.1 Présentation du problème

Nous allons supposer que nous disposons d'un ensemble d'apprentissage $\mathcal{L}_\alpha$ constitué de P exemples. Un exemple est un point $\boldsymbol{\xi}$ de l'espace des données de dimension N. La densité de probabilité des exemples sera supposée gaussienne normale dans toutes les directions :

$$P(\boldsymbol{\xi}) = \frac{1}{(2\pi)^{N/2}} \exp\left(-\frac{\boldsymbol{\xi}^2}{2}\right). \tag{1.1}$$

À chaque exemple $\boldsymbol{\xi}^\mu$ est associé une classe $\tau^\mu = \pm 1$. L'ensemble d'apprentissage est donné par :

$$\mathcal{L}_\alpha = \{\boldsymbol{\xi}^\mu, \tau^\mu\}_{\mu=1,\cdots,P}. \tag{1.2}$$

Le professeur qui classifie les exemples est supposé inconnu. Généralement, ce professeur ainsi que le nombre d'exemples sont tels que l'ensemble d'apprentissage ne peut pas être séparé linéairement, c'est-à-dire, qu'il n'existe pas d'hyperplan de l'espace des données pour lequel les exemples classés $+1$ soient d'un côté et les exemples classés -1 de l'autre côté de l'hyperplan. Une possibilité pour remédier à ce problème consiste à projeter l'espace des données dans un espace des représentations de plus grande dimension par l'intermédiaire d'une transformation qui permet de rendre les images des exemples de l'ensemble d'apprentissage linéairement séparables. Parmi les hyperplans de l'espace des représentations qui permettent de séparer les images des exemples de l'ensemble d'apprentissage, la machine à exemples supports correspond à l'hyperplan séparateur de stabilité maximale [122, 123, 65, 48] présenté au chapitre 2.

Dans la suite, nous allons considérer une famille de transformation définie de la façon suivante :

$$\boldsymbol{\xi} \to \Phi(\boldsymbol{\xi}) \equiv \{\boldsymbol{\xi}, \phi(\lambda_1)\boldsymbol{\xi}, \cdots, \phi(\lambda_k)\boldsymbol{\xi}\}. \tag{1.3}$$

Cette fonction Φ transforme un point de l'espace des données en un point de l'espace des représentations de dimension $(1+k)N$. Les variables λ_i correspondent à $\boldsymbol{\xi} \cdot \mathbf{B}_i$ pour $i = 1, \cdots, k \leq N$ où les vecteurs $\mathbf{B}_i$ forment un ensemble de vecteurs orthonormés ($\mathbf{B}_i \cdot \mathbf{B}_j = \delta_{ij}$). La fonction ϕ est une fonction à valeurs réelles.

Le choix des vecteurs $\mathbf{B}_i$ est arbitraire. Ils peuvent, par exemple, correspondre aux k premières coordonnées de l'espace des données : $\mathbf{B}_i = \mathbf{e}_i$ où $\mathbf{e}_1 = (1,0,\cdots,0)$, $\mathbf{e}_2 = (0,1,0,\cdots,0)$, etc. Dans la limite thermodynamique, $N \to +\infty$ et $P \to \infty$ avec la taille réduite $\alpha = P/N$ de l'ensemble d'apprentissage constante, tout choix aléatoire de k vecteurs normés vérifie les conditions d'orthogonalité avec une probabilité un. Ils constituent un choix possible de vecteurs $\mathbf{B}_i$.

Le choix de la fonction ϕ est lui aussi arbitraire. Nous nous limiterons cependant à des fonctions impaires. Avec cette restriction, les corrélations entre les différents vecteurs $\phi_i\boldsymbol{\xi}$ avec $\phi_0 \equiv 1$ et $\phi_i \equiv \phi(\lambda_i)$ pour $i \geq 1$ sont nulles. Deux fonctions seront particulièrement étudiées : $\phi(\lambda) = \lambda$ et $\phi(\lambda) = \text{sign}(\lambda)$.

Pour $k = 0$, l'espace des représentations coïncide avec l'espace des données. Les hyperplans séparateurs sont des hyperplans de l'espace des données et la machine à exemples supports correspond au Perceptron de Stabilité Maximale [65, 48]. À l'opposé, le cas $k = N$ avec le choix particulier $\phi(\lambda) = \lambda$ correspond à rechercher des surfaces séparatrices quadratiques dans l'espace des données. Plus généralement, la non linéarité de la transformation Φ permet la recherche d'hyperplans séparateurs dans l'espace des représentations qui correspondent à des surfaces séparatrices non linéaires dans l'espace des données. Cependant, il faut bien souligner que la transformation (1.3) introduit des redondances dans les représentations. Par exemple, si $\phi(\lambda) = \lambda$ et $\mathbf{B}_i = \mathbf{e}_i$, il est évident que les nouvelles représentations $\lambda_i\boldsymbol{\xi}$ et $\lambda_j\boldsymbol{\xi}$ ne sont pas toutes indépendantes. En effet, les produits $\xi_i\xi_j$ apparaissent dans les deux représentations. Dans les résultats présentés dans ce chapitre, nous avons négligé ces corrélations. Elles sont négligeables dans la limite thermodynamique pourvu que $k \ll N$. Notre approche est donc analogue à celle utilisée pour l'étude du modèle de Hopfield dilué [29], valable pour $k \sim \ln N$.

Un hyperplan de l'espace des représentations est déterminé par le vecteur $\mathbf{W} = \{\mathbf{w}_0,\mathbf{w}_1,\cdots,\mathbf{w}_k\}$ perpendiculaire à cet hyperplan, où les vecteurs $\mathbf{w}_i$, de dimension N, correspondent aux différentes *représentations*. Nous nous limiterons à des hyperplans passant par l'origine. La classe d'un exemple $\boldsymbol{\xi}$ est donnée par :

$$\sigma(\boldsymbol{\xi}) \equiv \text{sign}\left(\Phi(\boldsymbol{\xi}) \cdot \mathbf{W}\right). \tag{1.4}$$

La classe de cet exemple étant indépendante de la norme du vecteur $\mathbf{W}$, nous considérerons uniquement les vecteurs $\mathbf{W}$ normés à $(1+k)N$. Par contre, aucune restriction supplémentaire ne sera imposée sur la norme des vecteurs $\mathbf{w}_i$.

Un exemple de l'ensemble d'apprentissage est bien classé si $\sigma(\boldsymbol{\xi}^\mu) = \tau^\mu$. Cette condition peut aussi s'écrire :

$$\gamma^\mu \equiv \tau^\mu \frac{\Phi(\boldsymbol{\xi}^\mu) \cdot \mathbf{W}}{\sqrt{(1+k)N}} \geq 0. \tag{1.5}$$

La variable γ^μ est appelée stabilité. La valeur absolue $|\gamma^\mu|$ correspond à la distance de l'image $\Phi(\boldsymbol{\xi}^\mu)$ de l'exemple $\boldsymbol{\xi}^\mu$ à l'hyperplan séparateur (hyperplan perpendiculaire à $\mathbf{W}$ et passant par l'origine).

Les vecteurs $\mathbf{W}$ qui vérifient la condition (1.5) pour tous les exemples de l'ensemble d'apprentissage séparent linéairement les images $\Phi(\boldsymbol{\xi}^\mu)$ des exemples $\boldsymbol{\xi}^\mu$ possédant une sortie $\tau^\mu = +1$ de ceux possédant une sortie $\tau^\mu = -1$. Parmi tous les vecteurs $\mathbf{W}$ satisfaisant ces conditions, nous allons privilégier celui de stabilité maximale. L'hyperplan correspondant est la machine à exemples supports. La marge ou stabilité d'un vecteur $\mathbf{W}$ est définie par :

$$\kappa(\mathbf{W}) = \inf_\mu \gamma^\mu. \tag{1.6}$$

Le vecteur $\mathbf{W}^*$ de stabilité maximale est donnée par la relation suivante :

$$\kappa_{\max} = \kappa(\mathbf{W}^*) = \max_{\mathbf{W}} \kappa(\mathbf{W}). \tag{1.7}$$

L'un des principaux intérêts de ce vecteur particulier est qu'il peut être défini avec une partie seulement des exemples de l'ensemble d'apprentissage. Ces exemples sont appelés *exemples supports* (ES) et ce sont les exemples dont la stabilité γ^μ est égale à $\kappa_{\max}$. Le vecteur $\mathbf{W}^*$ s'écrit comme une combinaison linéaire des exemples supports :

$$\mathbf{W}^* = \sum_{\mu \in ES} a^\mu \tau^\mu \Phi(\boldsymbol{\xi}^\mu) \tag{1.8}$$

avec a^μ des paramètres positifs. L'apprentissage consiste à déterminer ces paramètres a^μ et les exemples qui sont supports [122, 123, 43, 44].

1.2 Calcul des propriétés d'apprentissage

Nous allons nous intéresser aux propriétés d'apprentissage de ces machines à exemples supports en fonction de la transformation Φ considérée, et pour divers professeurs différents. Nous allons considérer en particulier l'apprentissage de deux tâches très différentes.

La première est l'apprentissage d'une tâche linéairement séparable dans l'espace des données. Ce problème est celui que l'on a résolu de manière optimale dans le deuxième chapitre de la première partie de la thèse. Cette fois-ci, le but de l'étude est de comprendre comment se détériorent les propriétés de généralisation lorsqu'on utilise une machine de moins en moins adaptée à la tâche.

La seconde tâche considérée correspond à un problème où les classes des exemples sont des variables aléatoires. Dans ce cas, nous ne nous intéresserons plus à l'erreur de généralisation, qui n'a pas de sens pour cette tâche, mais à la capacité d'apprendre correctement les exemples de l'ensemble d'apprentissage. On s'attend à ce que cette capacité augmente avec la complexité de la machine à exemples supports.

1.2.1 Méthode des répliques

Les propriétés d'apprentissage des machines à exemples supports peuvent être déterminées par la méthode des répliques dans la limite thermodynamique. Cette limite correspond à $N \to +\infty$ et $P \to +\infty$ avec la taille réduite $\alpha = P/N$ de l'ensemble d'apprentissage fixée. Dans cette limite, les propriétés d'apprentissage deviennent indépendantes de l'ensemble d'apprentissage $\mathcal{L}_\alpha$ considéré.

La fonction de coût appropriée à l'étude de l'hyperplan de stabilité maximale est la suivante [65, 48] :

$$E(\mathbf{W},\mathcal{L}_\alpha,\kappa) = \sum_{\mu=1}^{P} \Theta(\kappa - \gamma^\mu) \tag{1.9}$$

où Θ est la fonction de Heaviside. Cette fonction de coût (1.9) compte le nombre d'exemples de l'ensemble d'apprentissage qui possèdent une stabilité γ^μ plus faible que κ. La plus grande valeur de κ pour laquelle la fonction de coût est nulle correspond à la stabilité maximale $\kappa_{\max}$. Le vecteur $\mathbf{W}^*$ correspondant est donné par :

$$E(\mathbf{W}^*,\mathcal{L}_\alpha,\kappa_{\max}) = 0. \tag{1.10}$$

Cette relation caractérise de manière unique le vecteur $\mathbf{W}^*$. Par définition de $\kappa_{\max}$, pour $\kappa > \kappa_{\max}$, il n'existe pas de vecteur $\mathbf{W}$ satisfaisant cette condition. Pour $\kappa < \kappa_{\max}$, il existe une infinité de solutions possibles, en fait, un volume non vide de l'espace des directions $\mathbf{W}$, qui sera appelé par la suite volume des solutions. Pour $\kappa = \kappa_{\max}$, ce volume se réduit à un point correspondant à $\mathbf{W}^*$.

De même que pour la première partie de la thèse, la fonction de coût est considérée comme une énergie et on introduit une température fictive $T = 1/\beta$. La fonction de partition correspondante s'écrit alors :

$$Z(\mathcal{L}_\alpha,\kappa) = \int dP(\mathbf{W}) \exp\left(-\beta E(\mathbf{W},\mathcal{L}_\alpha,\kappa)\right). \tag{1.11}$$

La densité de probabilité des vecteurs $\mathbf{W}$ est considérée uniforme sur la surface de la sphère de rayon $\sqrt{(1+k)N}$:

$$dP(\mathbf{W}) = \delta(\mathbf{W}\cdot\mathbf{W} - (1+k)N)\; d\mathbf{W}. \tag{1.12}$$

L'énergie libre correspondant à la fonction de partition (1.11) est supposée auto-moyennante dans la limite thermodynamique. Sa valeur pour un ensemble d'apprentissage particulier $\mathcal{L}_\alpha$ est alors égale à la moyenne sur tous les ensembles d'apprentissage de même taille réduite α avec une probabilité un. Puisque l'on recherche les propriétés du vecteur $\mathbf{W}^*$ pour lequel la fonction de coût est minimale (valeur de la fonction de coût nulle), on considère la limite de température nulle de l'énergie libre.

Dans la limite thermodynamique et la limite de température nulle, l'énergie libre s'écrit :

$$f(\kappa,\alpha,k) = -\lim_{\beta\to+\infty} \lim_{\substack{N\to+\infty \\ P\to+\infty \\ \alpha=P/N}} \frac{1}{\beta N} \overline{\ln Z}. \tag{1.13}$$

La barre au dessus du logarithme de la fonction de partition symbolise la moyenne sur tous les ensembles d'apprentissage de taille réduite α, comme il a été discuté dans la première partie de cette thèse.

Dans la limite de température nulle, la fonction de partition n'est rien d'autre que la fraction de volume occupé par les vecteurs $\mathbf{W}$ de fonction de coût nulle. En conséquence, si l'énergie libre $f(\kappa,\alpha,k)$ est nulle, alors il existe un vecteur $\mathbf{W}$ pour lequel $E(\mathbf{W},\mathcal{L}_\alpha,\kappa) = 0$. Par contre, si $f(\kappa,\alpha,k) > 0$, aucun vecteur $\mathbf{W}$ ne satisfait cette condition. $\kappa_{\max}(\alpha,k)$ est la plus grande valeur de κ pour laquelle $f(\kappa,\alpha,k) = 0$.

Le calcul de l'énergie libre f s'effectue grâce à la méthode des répliques, en introduisant la relation suivante :

$$\overline{\ln Z} = \lim_{n\to 0} \frac{1}{n} \ln \overline{Z^n}. \tag{1.14}$$

Je ne vais pas détailler le calcul de l'énergie libre, celui-ci est assez similaire à celui présenté au premier chapitre de la première partie. Par contre, je vais présenter les paramètres d'ordre introduits, ainsi que l'énergie libre $f(\kappa,\alpha,k)$ obtenue dans les deux cas : celui de l'apprentissage d'une tâche linéairement séparable par une machine à exemples supports, et celui de l'apprentissage d'un ensemble d'exemples dont les classes sont aléatoires. Je présenterai d'abord les deux calculs, que je discuterai par la suite.

1.2.2 Tâche linéairement séparable

Nous allons supposer que la tâche est linéairement séparable dans l'espace des données. Le professeur est alors un hyperplan séparateur, défini par le vecteur normal $\mathbf{B}$. La classe d'un exemple $\boldsymbol{\xi}$ est donnée par :

$$\tau(\boldsymbol{\xi}) = \operatorname{sign}\left(\mathbf{B}\cdot\boldsymbol{\xi}\right). \tag{1.15}$$

Pour cette tâche, quelque soit la transformation Φ (1.3) et quelque soit la taille réduite α de l'ensemble d'apprentissage, il existe des vecteurs $\mathbf{W}$ pour lesquels $E(\mathbf{W},\mathcal{L}_\alpha,0) = 0$. Le vecteur proportionnel à $(\mathbf{B},\mathbf{0},\cdots,\mathbf{0})$ en est un exemple. La stabilité maximale $\kappa_{\max}(\alpha,k)$ est donc strictement positive pour toutes les valeurs de α.

Le paramètre $\kappa_{\max}$ n'est pas la seule quantité intéressante pour caractériser les propriétés de l'apprentissage, il est aussi possible de déterminer l'erreur de généralisation $\varepsilon_g(\alpha,k)$ et la distribution $\rho(\gamma;\alpha,k)$ des stabilités des exemples de l'ensemble d'apprentissage. Une fois encore, je précise que l'on appelle distribution ce qu'il faudrait nommer rigoureusement densité de probabilité.

L'erreur de généralisation est la probabilité que la solution $\mathbf{W}^*$ classifie incorrectement, c'est-à-dire, différemment du professeur $\mathbf{B}$, un exemple ne faisant pas partie de l'ensemble d'apprentissage. Cette erreur de généralisation se calcule de la manière suivante :

$$\varepsilon_g(\alpha,k) = \int \Theta\left(-\sigma(\boldsymbol{\xi})\tau(\boldsymbol{\xi})\right) P(\boldsymbol{\xi}) d\boldsymbol{\xi}. \tag{1.16}$$

La machine à exemples supports est de moins en moins adaptée à la tâche à apprendre lorsque k augmente, on s'attend donc à ce que l'erreur de généralisation soit une fonction croissante de k.

La distribution des stabilités n'est autre que la densité de probabilité de la variable γ^μ (1.5) calculée pour le vecteur $\mathbf{W}^*$ et les exemples de l'ensemble d'apprentissage. Cette distribution va aussi permettre de calculer le nombre d'exemples supports P_{ES}.

Nous allons maintenant introduire les différents paramètres d'ordre qui apparaissent lors du calcul de l'énergie libre par la méthode des répliques :

$$R_a = \frac{\mathbf{w}_{0,a}\cdot\mathbf{B}}{\|\mathbf{w}_{0,a}\|\|\mathbf{B}\|}, \tag{1.17}$$

$$v_{i,a} = \frac{\mathbf{w}_{i,a}\cdot\mathbf{w}_{i,a}}{N}, \tag{1.18}$$

$$c_{i,ab} = \lim_{\beta\to+\infty}\beta\,\frac{(\mathbf{w}_{i,a}-\mathbf{w}_{i,b})^2}{2N} \qquad a<b. \tag{1.19}$$

avec $i=1,\cdots,k$ et $\mathbf{W}_a$ et $\mathbf{W}_b$ les vecteurs correspondant aux répliques a et b respectivement. Les produits scalaires $\mathbf{w}_{i,a}\cdot\mathbf{w}_{j,b}/N$ pour $i\neq j$ sont supposés négligeables ainsi que $\mathbf{B}\cdot\mathbf{w}_{i,a}/N$ pour $i\geq 1$. Ces hypothèses sont réalistes lorsque $k\ll N$, car les corrélations entre $\phi(\lambda_i)\boldsymbol{\xi}$ et $\phi(\lambda_j)\boldsymbol{\xi}$ pour $i\neq j$ sont nulles dans la limite thermodynamique. On remarquera l'importance du choix d'une fonction $\phi(\lambda)$ impaire pour ce résultat.

Par la suite, on utilise l'hypothèse de symétrie des répliques :

$$R_a = R, \tag{1.20}$$

$$v_{i,a} = v_i, \tag{1.21}$$

$$c_{i,ab} = c_i. \tag{1.22}$$

Le paramètre R est l'analogue de celui introduit dans la première partie. Il reflète la qualité de l'apprentissage car la direction $\mathbf{B}$ est celle du professeur. Le paramètre c_i est une généralisation du paramètre c de la première partie. En effet, dans le cas présent, la solution $\mathbf{W}^*$ comporte $1+k$ vecteurs de dimension N, $\{\mathbf{w}_0,\cdots,\mathbf{w}_k\}$. Chaque paramètre c_i reflète l'importance des fluctuations de l'un de ces vecteurs dans la limite de température nulle. Dans le cas où la fonction de coût a un continuum de minima dégénérés (plus précisément lorsque la fonction de coût est minimale à l'intérieur d'un volume fini), les fluctuations décroissent très lentement avec la température et les paramètres c_i sont infinis. C'est le cas notamment pour $\kappa<\kappa_{\max}$, où il existe un volume fini de solutions $\mathbf{W}$ avec une fonction de coût nulle. $\kappa_{\max}$ peut alors être vue comme la plus grande stabilité κ pour laquelle c_i est infini. Les paramètres v_i ne possèdent pas d'analogues dans les problèmes traités dans la première partie. La signification de ces paramètres est toutefois assez naturelle. Seule la norme du vecteur $\mathbf{W}^*$ est fixée, la norme des différents vecteurs $\mathbf{w}_i$ n'est pas contrainte. Le paramètre v_i correspond à la norme de $\mathbf{w}_i$. Lorsque la taille réduite de l'ensemble d'apprentissage diverge, on s'attend à ce que la direction $\mathbf{W}^*$ correspondant à la solution de stabilité maximale soit proportionnelle au professeur, c'est-à-dire, $\mathbf{W}^*\sim(\mathbf{B},\mathbf{0},\cdots,\mathbf{0})$. Pour cette direction $v_0=1+k$ et $v_i=0$ pour $i\geq 1$.

Nous allons utiliser une symétrie du problème pour réduire encore le nombre de paramètres d'ordre. Les performances de l'apprentissage sont indépendantes du choix des vecteurs $\mathbf{B}_i$ et notamment de leur étiquetage. En effet, ces performances sont invariantes par rapport à la permutation des vecteurs $\mathbf{B}_i$. En supposant la même symétrie pour les paramètres c_i et v_i alors $c_i=c_1$ et $v_i=v_1$ pour $i>1$. Il ne reste plus alors que 4 paramètres indépendants : R, $\widetilde{c}_0=c_0/(1+k)$, $\widetilde{c}_1=c_1/c_0$ et $\widetilde{v}_1=v_1/v_0$. Le paramètre v_0 est déterminé par la condition de normalisation du vecteur $\mathbf{W}^*$:

$$\frac{\mathbf{W}^*\cdot\mathbf{W}^*}{N}=1+k=v_0+k\widetilde{v}_1v_0. \tag{1.23}$$

L'énergie libre $f(\kappa,\alpha,k)$ s'exprime alors simplement comme l'extremum sur ces 4 paramètres d'ordre d'une fonction $g(\kappa,\alpha,k;R,\widetilde{c}_0,\widetilde{c}_1,\widetilde{v}_1)$, avec :

$$
\begin{aligned}
g(\kappa,\alpha,k;R,\widetilde{c}_0,\widetilde{c}_1,\widetilde{v}_1) &= -\frac{\widetilde{c}_1(1-R^2)+k\widetilde{v}_1}{2\widetilde{c}_0\widetilde{c}_1(1+k\widetilde{v}_1)} \\
&+ \frac{\alpha}{\widetilde{c}_0}\prod_{i=1}^{k}\int D\lambda_i \int_{\kappa a-b}^{\kappa a} H\left(-\frac{yR}{\sqrt{e}}\right)\frac{(\kappa-y/a)^2}{e}Dy \\
&+ 2\alpha\prod_{i=1}^{k}\int D\lambda_i \int_{-\infty}^{\kappa a-b} H\left(-\frac{yR}{\sqrt{e}}\right)Dy.
\end{aligned} \tag{1.24}
$$

Dy et $H(x)$ sont définies par les équations (2.17) et (2.16) de la première partie de la thèse :

$$
Dy = \exp\left(-\frac{y^2}{2}\right)\frac{dy}{\sqrt{2\pi}}, \tag{1.25}
$$

$$
H(x) = \int_x^{+\infty} Dy. \tag{1.26}
$$

a,b et e s'expriment de la façon suivante :

$$
a = \sqrt{\frac{1+k\widetilde{v}_1}{e+R^2}}, \tag{1.27}
$$

$$
b = a\left[\widetilde{c}_0\left(1+\widetilde{c}_1\sum_{i=1}^{k}\phi_i^2\right)\right]^{1/2}, \tag{1.28}
$$

$$
e = 1-R^2+\widetilde{v}_1\sum_{i=1}^{k}\phi_i^2, \tag{1.29}
$$

avec $\phi_0 = 1$ et $\phi_i = \phi(\lambda_i)$ pour $i > 0$.

Une fois l'énergie libre f (1.13) calculée pour une valeur quelconque de κ, la stabilité maximale $\kappa_{\max}$ est déterminée comme étant la plus grande valeur de κ pour laquelle f s'annule comme nous l'avons discuté précédemment. Cela correspond aussi à la plus grande valeur de κ pour laquelle $\widetilde{c}_0 = +\infty$ puisque $f = 0$ si $\widetilde{c}_0 = +\infty$.

L'erreur de généralisation s'exprime simplement par la formule suivante :

$$
\varepsilon_g(\alpha,k) = \frac{1}{\pi}\prod_{i=1}^{k}\int D\lambda_i \arccos\left(\frac{R}{\sqrt{e+R^2}}\right). \tag{1.30}
$$

Lorsque $k = 0$, on retrouve bien l'expression $\varepsilon_g = (1/\pi)\arccos(R)$ du chapitre 2 de la première partie. La distribution des stabilités des exemples de l'ensemble d'apprentissage s'écrit :

$$
\rho(\gamma;\alpha,k) = \rho_1(\gamma;\alpha,k)\,\Theta(\gamma-\kappa_{\max}) + \rho_0(\alpha,k)\,\delta(\gamma-\kappa_{\max}) \tag{1.31}
$$

avec :

$$
\rho_1(\gamma;\alpha,k) = \prod_{i=1}^{k}\int D\lambda_i \sqrt{\frac{2}{a\pi}}\,H\left(-\frac{\gamma R}{\sqrt{e}}\right)\exp\left(-\frac{\gamma^2}{2a}\right) \tag{1.32}
$$

et $\rho_0(\alpha,k)$ est telle que ρ soit normée à un. On peut noter aussi que ρ_0 n'est rien d'autre que la fraction d'exemples qui sont supports, c'est-à-dire, $P_{\mathrm{ES}} = P\rho_0$.

1.2.3 Tâche aléatoire

Nous allons maintenant considérer que la classe de chaque exemple est choisie de manière aléatoire avec une probabilité 1/2 pour la classe +1 et une probabilité 1/2 pour la classe −1. L'étude de ce problème permet d'évaluer la capacité de la machine.

L'énergie libre calculée par la méthode des répliques pour cette tâche aléatoire se déduit aisément de celle du problème précédent. Dans le cas présent, il n'existe pas de direction privilégiée **B**. Cette absence est prise en compte en imposant $R = 0$. L'énergie libre devient alors l'extremum de la fonction :

$$\begin{aligned} g(\kappa,\alpha,k;\widetilde{c}_0,\widetilde{c}_1,\widetilde{v}_1) &= -\frac{\widetilde{c}_1 + k\widetilde{v}_1}{2\,\widetilde{c}_0\,\widetilde{c}_1(1+k\widetilde{v}_1)} + \alpha \prod_{i=1}^{k} \int D\lambda_i \int_{-\infty}^{\kappa a - b} Dy \\ &\quad + \frac{\alpha}{2\widetilde{c}_0} \prod_{i=1}^{k} \int D\lambda_i \int_{\kappa a - b}^{\kappa a} Dy\ \frac{(\kappa - y/a)^2}{e}. \end{aligned} \tag{1.33}$$

a,b et e s'expriment de la façon suivante :

$$a = \sqrt{\frac{1+k\widetilde{v}_1}{e}}, \tag{1.34}$$

$$b = a\left[\widetilde{c}_0\left(1+\widetilde{c}_1\sum_{i=1}^{k}\phi_i^2\right)\right]^{1/2}, \tag{1.35}$$

$$e = 1+\widetilde{v}_1\sum_{i=1}^{k}\phi_i^2. \tag{1.36}$$

Pour la tâche aléatoire considérée, l'erreur de généralisation n'a pas de signification pertinente, la classe de chaque exemple étant choisie aléatoirement. Cette erreur est 1/2 quelque soit la taille réduite α et quelque soit la complexité, représentée par k, de la machine à exemples supports. La distribution des stabilités des exemples de l'ensemble d'apprentissage est donnée aussi par l'équation (1.31) avec :

$$\rho_1(\gamma;\alpha,k) = \prod_{i=1}^{k} \int D\lambda_i\ \frac{1}{\sqrt{2a\pi}} \exp\left(-\frac{\gamma^2}{2a}\right). \tag{1.37}$$

Une fois encore, ρ_0 représente la fraction d'exemples qui sont supports et $P_{\mathrm{ES}} = P\rho_0$.

1.3 Discussion des propriétés d'apprentissage

Après avoir présenté le calcul des propriétés d'apprentissage d'une machine à exemples supports pour deux tâches différentes, je vais maintenant discuter des résultats obtenus en fonction de la taille réduite α de l'ensemble d'apprentissage et de la transformation Φ utilisée. Cette transformation est principalement caractérisée par la nombre k de représentations introduites, et qui détermine la complexité de la machine, et la fonction $\phi(\lambda)$. Nous essaierons de tirer de ces résultats quelques conclusions générales sur les machines à exemples supports.

1.3.1 Tâche linéairement séparable

Considérons tout d'abord la fonction $\phi(\lambda) = \mathrm{sign}(\lambda)$. Dans ce cas, l'extremum de la fonction g (1.24) correspond à $\widetilde{c}_1 = 1$ et $\widetilde{v}_1 = 1 - R^2$. Après introduction de ces valeurs dans la fonction g, il est possible de remarquer l'identité suivante :

FIG. 1.1 – *Erreur de généralisation d'une machine à exemples supports avec* $\phi(\lambda) = \text{sign}(\lambda)$ *en fonction de* α *pour plusieurs valeurs de* k *(*$k = 0,1,2,5$ *et* 10*). Pour* $k = 0$*, la machine à exemples supports correspond à un perceptron de stabilité maximale.*

$$g(\kappa,\alpha,k;R,\widetilde{c}_0) = g\left(\kappa,\frac{\alpha}{1+k},0;\widetilde{R},\widetilde{c}_0\right) \tag{1.38}$$

où la partie droite de l'équation correspond au cas linéaire $k = 0$, qui n'est autre que le perceptron de stabilité maximale, avec le paramètre :

$$\widetilde{R} = \frac{R}{\sqrt{1+k(1-R^2)}} \tag{1.39}$$

à la place de R. Alors, toutes les propriétés peuvent être déduites de celles du perceptron de stabilité maximale [48].

La conséquence pour l'erreur de généralisation est la suivante :

$$\varepsilon_g(\alpha,k) = \varepsilon_g\left(\frac{\alpha}{1+k},0\right). \tag{1.40}$$

Lorsque la tâche à apprendre est linéairement séparable dans l'espace des données, l'erreur de généralisation d'une machine à exemple support dans un espace des représentations augmenté de k nouvelles représentations est donc la même que celle du perceptron de stabilité maximale avec un ensemble d'apprentissage de taille réduite $1 + k$ fois plus faible. La figure 1.1 représente l'erreur de généralisation en fonction de α pour plusieurs valeurs différentes de k ($k = 0,1,2,5$ et 10).

L'effet entropique dû à l'introduction de nouvelles représentations entraîne donc une augmentation significative de l'erreur de généralisation. La conclusion que l'on peut tirer de ce résultat est qu'une machine trop complexe ne peut pas apprendre

correctement une tâche linéairement séparable (tout au moins dans la limite thermodynamique lorsque α est fini). Ce résultat doit être nuancé car dans le cadre de nos calculs, $k \ll N$. Pour remédier à ce problème, on doit introduire dans la fonction de coût, une pénalité qui favorise les solutions moins complexes. Une approche proposée récemment équivaut à attribuer aux diverses représentations $\mathbf{w}_i$ pour $i > 0$ des coûts qui augmentent avec leurs normes [31, 132].

Il est intéressant de remarquer que lorsque $\alpha \to +\infty$ alors $R \to 1$ et $\widetilde{v}_1 \to 0$ ce qui montre bien que malgré l'extension de l'espace des données à un espace de plus grande dimension, la tâche linéairement séparable est apprise parfaitement dans la limite $\alpha \to +\infty$. Par contre, on peut aussi remarquer que les fluctuations des nouvelles représentations sont identiques à celles de la représentation 0 correspondant à l'espace des données puisque $\widetilde{c}_1 = 1$. Ce sont ces fluctuations importantes qui sont responsables du fait que l'erreur de généralisation est croissante avec k. En limitant les fluctuations des nouvelles représentations, en imposant par exemple un rapport constant et fixé à l'avance entre la norme $\|\mathbf{w}_0\|$ de la partie linéaire et celles $\|\mathbf{w}_i\|$ $(i > 0)$ des k nouvelles représentations, il est possible de limiter cet effet néfaste sur l'erreur de généralisation [31, 132].

Le comportement de l'erreur de généralisation dans le cas où les représentations internes dans (1.3) sont définies avec la fonction $\phi(\lambda) = \lambda$ est identique au cas précédent dans la limite des grands ensembles d'apprentissage ($\alpha \gg 1 + k$), c'est-à-dire,

$$\varepsilon_g(\alpha,k) \sim C\frac{1+k}{\alpha}. \tag{1.41}$$

Il est intéressant de noter que cette limite ($\alpha \ll 1 + k$) garde un sens malgré la condition $k \ll N$ car nous regardons les propriétés d'apprentissage dans la limite thermodynamique ($N \to +\infty$). La constante C est 0.5005 pour la fonction $\phi(\lambda) = \text{sign}(\lambda)$ [48] et en diffère légèrement pour la fonction $\phi(\lambda) = \lambda$. On peut supposer que pour toutes les fonctions impaires ϕ, le comportement sera analogue.

Il est intéressant de noter que pour les faibles tailles de l'ensemble d'apprentissage ($\alpha \ll 1 + k$), il est possible de montrer que la fonction $\phi(\lambda) = \text{sign}(\lambda)$ est la fonction qui possède la plus faible erreur de généralisation pour une valeur de k donnée.

L'étude de la distribution des stabilités des exemples de l'ensemble d'apprentissage permet d'analyser le comportement de la stabilité maximale $\kappa_{\max}$ ainsi que celui de la fraction d'exemples supports ρ_0. Cette dernière quantité est essentielle dans la théorie proposée par Vapnik [122, 123] pour démontrer l'intérêt de ces machines. En effet, celui-ci a montré que la fraction d'exemples supports était une borne supérieure de l'erreur de généralisation :

$$\varepsilon_g(\alpha,k) \leq \rho_0(\alpha,k) = \frac{P_{\text{ES}}}{P}. \tag{1.42}$$

L'intérêt des machines à exemples supports est principalement dû à deux observations qui semblent se retrouver dans bon nombre d'applications réalistes. Premièrement, le nombre d'exemples supports est faible par rapport au nombre total d'exemples dans l'ensemble d'apprentissage. Deuxièmement, ce nombre d'exemples supports semble varier faiblement avec la complexité de la machine à exemples supports utilisée, c'est-à-dire, la dimension de l'espace des représentations. La conclusion de ces deux observations est que l'erreur de généralisation obtenue dans ces applications réalistes est faible. Malheureusement, aucune approche théorique n'a pu à ce jour expliquer les raisons du faible nombre d'exemples supports.

Considérons d'abord la transformation (1.3) définie par $\phi(\lambda) = \text{sign}(\lambda)$ pour laquelle on peut exprimer la stabilité maximale à k donné en fonction de sa valeur à $k = 0$, correspondant à la stabilité maximale du perceptron simple [48] :

FIG. 1.2 – *Stabilité maximale d'une machine à exemples supports avec* $\phi(\lambda) = \mathrm{sign}(\lambda)$ *en fonction de* α *pour plusieurs valeurs de* k *(*$k = 0,1,2,5$ *et* 10*). Pour* $k = 0$*, la machine à exemples supports correspond à un perceptron de stabilité maximale.*

$$\kappa_{\max}(\alpha,k) = \kappa_{\max}\left(\frac{\alpha}{1+k},0\right). \tag{1.43}$$

La figure 1.2 représente la stabilité maximale en fonction de α pour plusieurs valeurs de k ($k = 0,1,2,5$ et 10). Deux comportements limites différents sont intéressants à considérer. Le premier correspond à $\alpha \ll 1+k$,

$$\kappa_{\max} \sim \sqrt{\frac{1+k}{\alpha}}, \tag{1.44}$$

$$\rho_0(\alpha,k) \sim 1 - \sqrt{\frac{\alpha}{2\pi(1+k)}}\exp\left(-\frac{1+k}{\alpha}\right). \tag{1.45}$$

Dans ce cas, la borne (1.42), donnée par Vapnik, n'est pas pertinente puisque la fraction d'exemples supports borne l'erreur de généralisation par une quantité supérieure à 1/2. Cela se comprend aisément : si la fraction d'exemples supports est très proche de l'unité, presque tous les exemples sont utilisés pour déterminer la séparation. Cela signifie que l'apprentissage est plus proche d'un apprentissage par cœur des exemples que d'un réel apprentissage de la règle donnée par le professeur. Le deuxième comportement, plus intéressant, correspond à $\alpha \gg 1+k$,

$$\kappa_{\max} \sim 0.226\sqrt{2\pi}\,\frac{1+k}{\alpha}, \tag{1.46}$$

$$\rho_0(\alpha,k) \sim 0.952\,\frac{1+k}{\alpha}. \tag{1.47}$$

Dans ce cas, la borne (1.42), donnée par la fraction d'exemples supports, est pertinente car proche de l'erreur de généralisation $\varepsilon_g \sim 0.5005(1+k)/\alpha$. Le comportement inversement proportionnel à $\alpha/(1+k)$ est correctement prédit. Cependant, le coefficient de proportionnalité donné par la borne de Vapnik est près de deux fois plus important que le coefficient de l'erreur de généralisation. Il est toutefois important de noter que le nombre d'exemples supports n'est pas faible pour autant. En effet, $P_{\rm ES} = P\rho_0 \sim 0.952\,(1+k)N$, c'est-à-dire, que le nombre de supports est très proche de la dimension de l'espace des représentations. Ce résultat est très différent de celui obtenu dans diverses applications [16, 50, 74, 122, 123] où le nombre d'exemples supports semble saturer et non croître proportionnellement à la taille de l'espace des représentations.

Pour la fonction $\phi(\lambda) = \lambda$ étudiée, les deux comportements à $\alpha \gg 1+k$ et $\alpha \ll 1+k$ considérés, sont similaire à ceux obtenus pour $\phi(\lambda) = \text{sign}(\lambda)$. On peut donc s'attendre à ce que les machines à exemples supports dont les représentations sont définies par des fonctions $\phi(\lambda)$ impaires présentent ces mêmes comportements lors de l'apprentissage de tâches linéairement séparables.

1.3.2 Tâche aléatoire

Considérons maintenant le cas où les classes des points de l'espace des données sont attribués aléatoirement. L'étude de ce problème permet de déterminer la fraction maximale d'exemples α_c que la machine peut apprendre sans erreurs. Cette quantité α_c correspond à la taille réduite α pour laquelle $\kappa_{\max}(\alpha,k)$ s'annule. Cette valeur sera appelée capacité $\alpha_c(k)$ de la machine à exemples supports. L'erreur de généralisation n'est pas une quantité pertinente car elle reste constante et égale à $1/2$ pour toutes les valeurs de la taille réduite α et quelque soit la transformation Φ (1.3), c'est-à-dire, quelques soient k et la fonction $\phi(\lambda)$.

L'extremum de la fonction $g(\kappa,\alpha,k;\widetilde{c}_0,\widetilde{c}_1,\widetilde{v}_1)$ (1.33) pour $\kappa_{\max} = 0$ correspond aux valeurs $\widetilde{c}_0 = +\infty$ et $\widetilde{c}_1 = \widetilde{v}_1$ quelque soit la fonction $\phi(\lambda)$ considérée. Il est à noter toutefois que $\widetilde{v}_1$ n'est pas nécessairement égal à 1. La conclusion surprenante est que :

$$\alpha_c(k) = 2\,(1+k), \tag{1.48}$$

indépendante de la fonction $\phi(\lambda)$. Il est important de souligner que nous nous sommes limités aux fonctions $\phi(\lambda)$ impaires. Ceci entraîne notamment que, si les exemples sont en position générale dans l'espace des données, ce qui est vrai pour la distribution (1.1) des exemples considérée, alors les images des exemples sont aussi en position générale dans l'espace des représentations. Pour des exemples en position générale, Cover [28] a démontré par des méthodes géométriques que la capacité d'un perceptron simple ($k = 0$) est $\alpha_c(0) = 2$ et celle d'une machine quadratique ($k = N$ et $\phi(\lambda) = \lambda$) est $\alpha_c = N+1$. Le préfacteur 2 qui apparaît dans notre résultat si l'on remplace k par N provient du fait que nous avons négligé les corrélations entre les nouvelles représentations. Si $k = N$ et $\phi(\lambda) = \lambda$, tous les termes $\xi_i\xi_j$ pour $i \neq j$ apparaissent deux fois dans (1.3). Puisque les autres termes sont décorrélés, il est facile de trouver la correction à notre résultat. En effet, parmi les $N(1+k)$ composantes, nous en avons compté $k(k-1)/2$ deux fois. Si on les soustrait, on trouve les corrections en $1/N$ à la valeur que nous avons calculée. On obtient :

$$\alpha_c(k,N) = 2\left[(1+k) - \frac{k(k-1)}{2N}\right], \tag{1.49}$$

qui a le comportement prédit par Cover pour $k = N$ [28, 85]. Le calcul de la capacité de la machine à exemples supports pour la famille de transformations considérées

permet de trouver par la méthode des répliques des résultats en accord avec les approches géométriques.

Il est intéressant de comparer la capacité des machines à exemples supports que nous avons considérées à celles des perceptrons avec une couche cachée. Pour faire la classification, ces derniers introduisent k perceptrons simples, ou encore, k hyperplans séparateurs, dont les classes constituent une représentation interne de k variables binaires pour chaque point de l'espace des données. Parmi les perceptrons avec une couche cachée, deux machines particulières ont été très étudiées : la machine de parité et les comités. La classe d'un point de l'espace des données est le produit des k variables binaires de la représentation interne pour les machines de parité et la majorité des k variables binaires pour les comités. Dans ce dernier cas, la valeur de k doit être impaire pour qu'il existe toujours une majorité entre la classe $+1$ et la classe -1. Ces deux types de machines possèdent le même nombre de degrés de liberté kN. Dans la limite $1 \ll k \ll N$, il est donc comparable à celui des machines à exemples supports qui est $(1+k)N$, ou plus précisément $(1+k)N - k(k-1)/2$. Il est alors pertinent de comparer leurs capacités respectives. Dans les deux cas, les perceptrons avec une couche cachée possèdent une plus grande capacité que les machines à exemples supports. En effet, la capacité de la machine de parité est $\alpha_c(k) \sim k \ln k$ et celle des comités est $\alpha_c(k) \sim k\sqrt{\ln k}$ [86, 87, 68, 130, 131, 34, 118].

Nous allons présenter dans le prochain chapitre un algorithme particulier permettant d'obtenir une machine de parité. Cet algorithme incrémental permet d'augmenter progressivement le nombre de perceptrons simples de la couche cachée jusqu'à ce que l'erreur d'apprentissage s'annule. Nous étudierons le comportement de la capacité de la machine de parité apprenant avec cet algorithme particulier.

Je vais considérer maintenant les comportements de la stabilité $\kappa_{\max}$ et de la fraction d'exemples supports ρ_0 lorsque la taille réduite α de l'ensemble d'apprentissage est inférieure à la capacité $\alpha_c(k)$.

Tout d'abord, pour la fonction $\phi(\lambda) = \mathrm{sign}(\lambda)$, les paramètres d'ordre qui extremisent g (1.33) sont $\widetilde{v}_1 = \widetilde{c}_1 = 1$. Il en résulte :

$$g(\kappa,\alpha,k;\widetilde{c}_0,1,1) = (1+k)g\left(\kappa,\frac{\alpha}{1+k},0;\widetilde{c}_0,1,1\right). \tag{1.50}$$

De cette expression, on déduit que la stabilité maximale à k donné s'exprime en fonction de celle du perceptron simple ($k=0$) :

$$\kappa_{\max}(\alpha,k) = \kappa_{\max}\left(\frac{\alpha}{1+k},0\right). \tag{1.51}$$

La figure 1.3 représente la stabilité maximale en fonction de α pour plusieurs valeurs de k ($k = 0,1,2,5$ et 10). Dans ce cas, le comportement de $\kappa_{\max}$ et ρ_0 pour $\alpha \ll 1+k$ est [48] :

$$\kappa_{\max}(\alpha,k) \sim \sqrt{\frac{1+k}{\alpha}}, \tag{1.52}$$

$$\rho_0(\alpha,k) \sim 1 - \sqrt{\frac{\alpha}{2\pi(1+k)}}\,\exp\left(-\frac{1+k}{\alpha}\right), \tag{1.53}$$

le même que pour une tâche linéairement séparable. Le comportement pour α proche de la capacité $\alpha_c(k) = 2\,(1+k)$ est [48] :

$$\kappa_{\max}(\alpha,k) \sim \sqrt{\frac{\pi}{8}}\,\frac{\alpha_c - \alpha}{\alpha_c}, \tag{1.54}$$

$$\rho_0(\alpha,k) \sim \frac{1}{2} + \frac{\alpha_c - \alpha}{4\,\alpha_c}. \tag{1.55}$$

FIG. 1.3 – *Stabilité maximale d'une machine à exemples supports avec* $\phi(\lambda) = \text{sign}(\lambda)$ *en fonction de* α *pour plusieurs valeurs de* k *(*$k = 0,1,2,5$ *et* 10*). Pour* $k = 0$*, la machine à exemples supports correspond à un perceptron de stabilité maximale.*

La stabilité maximale s'annule linéairement à l'approche de la capacité $\alpha_c(k)$. La fraction d'exemples supports décroît pour atteindre $1/2$ à la capacité. Ceci est cohérent avec le fait que cette fraction est une borne supérieure de l'erreur de généralisation et que cette erreur est $1/2$ pour une tâche aléatoire. Lorsque la taille réduite de l'ensemble d'apprentissage est égale à la capacité, le nombre d'exemples apprenables est deux fois plus grand que la dimension de l'espace des représentations mais seulement la moitié sont supports. En effet, le nombre d'exemples supports $P_{\text{ES}} = \rho_0 P$ correspond à la dimension de l'espace des représentations. Le nombre d'exemples supports est alors maximal car il n'est pas possible de trouver un hyperplan contenant plus de points en position générale que la dimension de l'espace correspondant.

1.3.3 Remarques générales

Nous venons d'étudier deux tâches d'apprentissage diamétralement opposées dans le sens où l'une d'entre elles correspond à une tâche aisée, puisque linéairement séparable, tandis que l'autre correspond à une tâche inapprenable, puisque aléatoire. Pour la tâche linéairement séparable, la machine à exemples supports est de moins en moins adaptée lorsque la dimension de l'espace des représentations augmente. Les performances de l'apprentissage reflètent ce phénomène puisque l'erreur de généralisation, à α fixé, croît lorsque k augmente. Pour la tâche aléatoire, au contraire, la machine à exemples supports permet l'apprentissage par cœur d'un nombre de plus en plus important d'exemples lorsque k augmente. Au-delà de ces deux remarques évidentes, il est possible de déduire de cette étude quelques conclusions générales

sur les machines à exemples supports indépendantes de la tâche considérée.

La première des conclusions est que la stabilité maximale $\kappa_{\max}$ est une fonction croissante de la dimension de l'espace des représentations. Cette croissance permet une plus grande robustesse vis-à-vis du bruit dans les données après l'apprentissage. Je discuterai plus en détail cette notion de robustesse dans la prochaine section.

La deuxième conclusion, beaucoup moins évidente et qui va à l'encontre du résultat attendu, est le fait que le nombre d'exemples supports est une fonction croissante de la dimension de l'espace des représentations. Or, dans les applications réalistes rapportées [16, 50, 74, 122, 123], le nombre d'exemples supports est une fonction presque constante de l'espace des représentations. Il faut toutefois noter que dans ces applications réalistes, il est laissé la possibilité à quelques exemples d'être mal classés. La solution obtenue après apprentissage ne correspond alors pas à la solution de stabilité maximale. Cette modification est peut-être responsable de la limitation observée du nombre d'exemples supports. Une étude plus proche des applications réalistes semble nécessaire.

Il est intéressant aussi de noter le comportement de la stabilité maximale et du nombre d'exemples supports pour les faibles valeurs de la taille réduite α de l'ensemble d'apprentissage. Il n'existe aucune différence de comportement au premier ordre en α entre la tâche aléatoire et la tâche linéairement séparable. La théorie de Vapnik, qui est principalement basée sur le nombre d'exemples supports, ne permet pas de mettre en évidence que l'erreur de généralisation décroît, dans le cas de la tâche linéairement séparable, dès $\alpha = 0$. Ceci est dû au fait que la théorie de Vapnik considère le cas le pire alors que l'approche thermodynamique étudie le cas typique.

Dans la limite des grandes tailles réduites de l'ensemble d'apprentissage, il semble que l'augmentation de la dimension de l'espace des représentations ait un effet défavorable sur la généralisation. Le nombre de fonctions booléennes accessibles augmente avec la dimension de l'espace des représentations ce qui est indispensable pour espérer apprendre des tâches non linéairement séparables. Malheureusement, l'augmentation du nombre de fonctions accessibles fait que la probabilité d'obtenir l'une d'entre elles diminue. C'est ce phénomène qui est responsable de la croissance de l'erreur de généralisation lorsque k augmente, à α fixé, dans le cas de la tâche linéairement séparable. Les approches de Yoon et Oh [132] et de Dietrich, Opper et Sompolinsky [31] semblent corriger ce problème en multipliant par des poids les représentations introduites. Ces poids sont tels qu'ils permettent de décroître l'importance relative des représentations, proportionnellement à leurs dimensions. Dans le cas de la transformation (1.3) étudiée dans cette thèse, ceci revient à diviser par $\sqrt{k}$ les représentations avec $i \geq 1$. Ceci permet de limiter la décroissance de l'erreur de généralisation lors de l'augmentation de la dimension de l'espace des représentations. Il est intéressant de remarquer que dans les applications réalistes, il semblerait que ce phénomène n'ait pas été analysé.

1.4 Erreur de reconnaissance

Dans la plupart des applications réalistes, une des hypothèses les plus couramment utilisées, souvent de manière implicite, est que deux exemples proches dans l'espace des données appartiennent à la même classe avec une forte probabilité. Cette hypothèse peut être vue comme une hypothèse de continuité de la classification.

L'hypothèse de continuité est une des raisons pour lesquelles la direction de stabilité maximale est choisie parmi toutes les directions de l'espace des représentations ayant une erreur d'apprentissage nulle. Nous allons essayer de formaliser cette notion en introduisant une erreur de reconnaissance que l'on pourra aussi voir comme une mesure de la robustesse de la machine à exemples supports par rapport à des perturbations, qui peuvent être considérées comme un bruit, des exemples.

Supposons que la dimension de l'espace des représentations est suffisamment grande pour qu'il existe une direction de stabilité maximale positive, ou encore, que l'erreur d'apprentissage correspondant à cette direction soit nulle. Nous allons supposer maintenant que l'apprentissage a été effectué et que la direction de stabilité maximale est déterminée. L'erreur de reconnaissance est alors définie comme l'erreur commise sur un ensemble de points, désormais appelés ensemble de reconnaissance, constitué des exemples de l'ensemble d'apprentissage auxquels on a ajouté un bruit gaussien, mais dont les classes restent celles des exemples de l'ensemble d'apprentissage initial.

Les exemples de l'ensemble de reconnaissance sont perturbés de la façon suivante :

$$\boldsymbol{\xi}^\mu \to \boldsymbol{\zeta}^\mu \equiv \boldsymbol{\xi}^\mu + \boldsymbol{\eta}^\mu \tag{1.56}$$

avec $\boldsymbol{\eta}^\mu$ un vecteur distribué aléatoirement selon la densité de probabilité :

$$P(\boldsymbol{\eta}) = \frac{1}{\left(2\pi\Delta^2\right)^{N/2}} \exp\left(-\frac{\boldsymbol{\eta}^2}{2\Delta^2}\right). \tag{1.57}$$

Le paramètre Δ reflète le taux de bruit. Les classes de $\boldsymbol{\xi}^\mu$ et de $\boldsymbol{\zeta}^\mu$ sont supposées identiques. Cette hypothèse n'a un intérêt que dans la limite où la perturbation des exemples est faible ($\Delta \ll 1$) pour que l'hypothèse de continuité faite sur la classe réelle des exemples soit satisfaite avec une forte probabilité. La classe σ^μ donnée par la machine à exemples supports est quant à elle modifiée par l'ajout de bruit sur les exemples. Nous appellerons $\sigma^\mu(\Delta)$ cette nouvelle classe qui est définie par :

$$\sigma^\mu(\Delta) = \operatorname{sign}\left(\Phi(\boldsymbol{\xi}^\mu + \boldsymbol{\eta}^\mu) \cdot \mathbf{W}^*\right). \tag{1.58}$$

Ces notations étant introduites, l'erreur de reconnaissance est définie par :

$$\varepsilon_r(\Delta,\kappa_{\max},k) = \frac{1}{P}\sum_{\mu=1}^{P} \Theta\left(-\sigma^\mu(\Delta)\tau^\mu\right). \tag{1.59}$$

L'erreur de reconnaissance dépend implicitement de la taille réduite α de l'ensemble d'apprentissage par l'intermédiaire de $\kappa_{\max}(\alpha,k)$.

Regardons tout d'abord le cas du perceptron de stabilité maximale ($k = 0$), l'erreur de reconnaissance est :

$$\varepsilon_r(\Delta,\kappa,0) = H(-\kappa)H\left(\frac{\kappa}{\Delta}\right) + \int_\kappa^{+\infty} Dz\; H\left(\frac{z}{\Delta}\right). \tag{1.60}$$

La stabilité maximale $\kappa_{\max}$ a été notée κ pour simplifier la notation, et cela sera le cas par la suite.

Imaginons que $\kappa = 0$, alors dans ce cas, l'erreur de reconnaissance est proche de $1/4$ lorsque $\Delta \ll 1$. Ce résultat n'est pas surprenant si l'on considère que lorsque $\kappa = 0$, la distribution des stabilités des exemples de l'ensemble d'apprentissage initial est telle que la moitié d'entre eux ont une stabilité nulle. La perturbation de ces exemples entraîne une perturbation gaussienne de leur stabilité centrée sur la valeur avant perturbation. Celle-ci étant nulle, la moitié de ces exemples perturbés vont posséder une stabilité négative équivalente à une erreur de classification. Il existe donc bien près d'un quart des exemples perturbés mal classés.

Imaginons maintenant que κ est non nul. L'erreur de reconnaissance, dans la limite des faibles perturbations ($\Delta \ll 1$), est proportionnelle à :

$$\varepsilon_r(\Delta,\kappa,0) \sim \exp\left(-\frac{\kappa^2}{2\Delta^2}\right). \tag{1.61}$$

L'erreur de reconnaissance est d'autant plus faible que la stabilité est grande ce qui permet de justifier notamment le choix de la stabilité maximale. Il est aussi intéressant de noter la croissance très lente de l'erreur de reconnaissance lorsque la perturbation est augmentée. La question est de savoir sous quelles conditions ce résultat reste valable pour les machines à exemples supports, c'est-à-dire, pour $k > 0$.

Supposons tout d'abord que l'on a utilisé la fonction $\phi(\lambda) = \text{sign}(\lambda)$ et que $\kappa > 0$, alors pour $\Delta \ll 1$:

$$\varepsilon_r(\Delta,\kappa,k) \sim \Delta. \tag{1.62}$$

Ce résultat est indépendant de la valeur de la stabilité κ et surtout l'erreur de reconnaissance croît plus rapidement avec la perturbation Δ que dans le cas $k = 0$. Cette croissance rapide est due à la discontinuité de la fonction ϕ. En effet, cette discontinuité peut entraîner une importante modification de la stabilité même pour une faible perturbation. Ce n'est plus le cas pour une fonction ϕ continue comme $\phi(\lambda) = \lambda$ pour laquelle :

$$\varepsilon_r(\Delta,\kappa,k) \sim \exp\left(-h(k)\frac{\kappa}{\Delta}\right). \tag{1.63}$$

La fonction $h(k)$ est une fonction croissante de k. Le comportement (1.60) du perceptron de stabilité maximale n'est pas entièrement retrouvé. Toutefois, l'erreur de reconnaissance croît très faiblement par rapport à la perturbation et décroît lorsque la stabilité augmente.

En conclusion, la transformation non linéaire définissant la machine à exemples supports doit être continue si l'on veut obtenir une faible erreur de reconnaissance lors de la perturbation des exemples de l'ensemble d'apprentissage. Cette erreur de reconnaissance est principalement dépendante de la machine utilisée et de la distribution des stabilités des exemples. Elle ne dépend pas de la fonction à apprendre puisque on a supposé que les exemples avant et après perturbation étaient classés de façon identique par cette fonction.

1.5 Tentative de borne pour l'erreur de généralisation

1.5.1 Présentation de la borne

Je vais maintenant discuter la raison qui nous a amenés à introduire l'erreur de reconnaissance. L'idée est d'essayer de borner de manière intéressante l'erreur de généralisation. Borner de manière triviale l'erreur de généralisation consiste par exemple à la borner par 1/2. La tentative que nous allons présenter maintenant aboutit dans tous les cas étudiés à une borne triviale. Les résultats obtenus permettent toutefois de comprendre quelles sont les raisons qui sont susceptibles d'être responsables d'une erreur de généralisation faible.

L'erreur de généralisation est la probabilité de mal classer une donnée ne faisant pas partie des exemples de l'ensemble d'apprentissage. Séparons les données en deux sous-ensembles. On place dans le premier qu'on appellera *ensemble de proximité*, les données pour lesquelles il existe un exemple de l'ensemble d'apprentissage à une distance d inférieure à Δ. La distance d entre deux exemples $\boldsymbol{\xi}$ et $\boldsymbol{\xi}'$ est définie par :

$$d^2 = \frac{(\boldsymbol{\xi} - \boldsymbol{\xi}')^2}{N}. \tag{1.64}$$

Il est possible de borner l'erreur de classification d'une donnée du premier ensemble à l'aide de l'erreur de reconnaissance. Pour cela, il faut introduire une erreur

que l'on appellera erreur de proximité ε_p qui dépend de la fonction à apprendre mais qui ne dépend pas de la machine utilisée pour l'apprentissage. Cette erreur est la probabilité que deux points distant de moins de Δ n'appartiennent pas à la même classe. Elle peut être vue comme l'analogue de l'erreur de reconnaissance pour la fonction à apprendre. La probabilité de faire une erreur sur une donnée de l'ensemble de proximité est alors bornée simplement par :

$$\widetilde{\varepsilon}(\Delta) = \varepsilon_r(\Delta,\kappa,k)(1-\varepsilon_p(\Delta)) + (1-\varepsilon_r(\Delta,\kappa,k))\,\varepsilon_p(\Delta). \tag{1.65}$$

La première partie de cette somme correspond à des données qui ne sont pas classées par la machine comme l'exemple de l'ensemble d'apprentissage le plus proche alors que la fonction à apprendre les classe de manière identique. La deuxième partie de l'expression correspond à l'autre possibilité, inverse de la précédente, de faire une erreur de classification. Les deux erreurs ε_r et ε_p étant des fonctions croissantes de Δ de 0 à 1/2, il en est de même pour $\widetilde{\varepsilon}$.

La probabilité de faire une erreur sur les données ne faisant pas partie de l'ensemble de proximité est majorée trivialement par 1/2. En appelant $\varepsilon_d(\Delta,\alpha)$ la probabilité qu'un exemple soit dans l'ensemble de proximité, on obtient la borne suivante pour l'erreur de généralisation :

$$\varepsilon_g(\alpha) \leq \varepsilon_b(\Delta,\alpha) = \varepsilon_d(\Delta,\alpha)\,\widetilde{\varepsilon}(\Delta) + \frac{1}{2}(1-\varepsilon_d(\Delta,\alpha)). \tag{1.66}$$

La probabilité ε_d dépend à la fois de la distance Δ mais aussi de la taille réduite α de l'ensemble d'apprentissage. La probabilité ε_d est croissante par rapport à Δ. La borne ε_b est donc la somme de deux termes, l'un croissant et l'autre décroissant en fonction de Δ. La valeur de cette borne est 1/2 dans les deux limites $\Delta \to 0$ et $\Delta \to +\infty$ mais on peut s'attendre à l'existence d'une distance Δ_0 pour laquelle $\varepsilon_b(\Delta_0,\alpha)$ est minimale. Lorsque le nombre d'exemples augmente, on s'attend à ce que l'ensemble de proximité contienne de plus en plus d'exemples, c'est-à-dire que ε_d soit une fonction croissante de α. Il est alors possible que la borne $\varepsilon_b(\Delta_0,\alpha)$ soit une fonction décroissante par rapport à α et ne soit pas une borne triviale de l'erreur de généralisation.

1.5.2 Calcul de la borne pour les tâches étudiées

Calculons cette borne pour les deux tâches étudiées précédemment et pour une machine à exemples supports à représentations continues ($\phi(\lambda)$ continue). Nous verrons alors que tout ne se passe pas comme prévu dans la limite thermodynamique.

Pour la tâche aléatoire, il est évident que $\varepsilon_p(\Delta) = 1/2$ pour toute valeur de Δ puisque l'hypothèse de proximité n'est pas vérifiée dans ce cas. On aboutit logiquement à une borne triviale $\varepsilon_b(\Delta) = 1/2$ pour cette tâche, cohérente avec le fait que $\varepsilon_g = 1/2$.

En ce qui concerne la tâche linéairement séparable, on peut s'attendre à une borne non triviale puisque, cette fois-ci, l'erreur de généralisation n'est plus égale à 1/2 mais décroît lorsque α augmente pour s'annuler lorsque α diverge. L'ensemble de proximité est non vide et

$$\varepsilon_p(\Delta) = 2\int_0^{+\infty} Dx\; H\left(\frac{x}{\Delta}\right) = \frac{1}{\pi}\arccos\left(\frac{1}{\sqrt{1+\Delta^2}}\right). \tag{1.67}$$

Cette erreur s'annule lorsque $\Delta \to 0$ et atteint 1/2 lorsque $\Delta \to +\infty$. Dans la section précédente, on a montré que pour une machine à exemples supports continue, l'erreur de reconnaissance s'annule pour $\Delta \to 0$. Il ne reste donc plus qu'à déterminer la probabilité $\varepsilon_d(\Delta,\alpha)$. Malheureusement, cette probabilité n'est pas une fonction continue de Δ mais une fonction seuil. En fait, la distribution de distances d entre

deux exemples, dans la limite thermodynamique, se réduit à $\delta(d-\sqrt{2})$. On en déduit que

$$\varepsilon_d(\Delta,\alpha) = \Theta\left(\Delta - \sqrt{2}\right). \tag{1.68}$$

Pour que la borne soit non triviale, il faut donc $\Delta > \Delta_0 = \sqrt{2}$ et cela quelque soit la valeur de α. Pour cette valeur Δ_0, l'erreur de reconnaissance que l'on a définie précédemment n'a plus de sens, puisque tous les exemples de l'ensemble d'apprentissage sont à la même distance d'une donnée prise au hasard. On retrouve, comme pour la tâche aléatoire, que la borne ainsi déterminée est triviale ($\varepsilon_b = 1/2$).

1.5.3 Ouvertures possibles

La question qui se pose alors est de savoir si cette approche est totalement inadaptée à des applications réalistes ou si ce sont les deux tâches considérées qui sont très éloignées des applications réalistes.

Je vais essayer de donner quelques raisons en faveur de la deuxième option qui permettent de justifier l'intérêt de l'erreur de reconnaissance que nous avons introduite.

La première difficulté dans les deux tâches considérées réside en la distribution des exemples. Elle entraîne que, dans la limite thermodynamique, la distribution de distances entre paires d'exemples possède un unique pic δ. Par conséquent, ε_d est une fonction seuil à partir de laquelle on ne peut pas envisager une borne non triviale à l'erreur de généralisation. Toutefois, ce comportement est exclusivement dû à la distribution des exemples considérée. Or, celle-ci est la plus simple que l'on puisse imaginer, mais elle est très éloignée de ce que l'on peut trouver dans des problèmes réalistes. En effet, dans le cas de la tâche linéairement séparable étudiée, la probabilité de trouver des exemples est maximale sur la surface discriminante. Dans les problèmes réalistes, on s'attend au contraire à une densité de probabilité minimale sur cette surface.

Une des possibilités pour résoudre le problème de la densité de probabilité des exemples est de considérer non plus une simple gaussienne centrée à l'origine mais plusieurs gaussiennes centrées autour de points appelés des prototypes et qui sont situés non plus sur la surface séparatrice mais au contraire à l'intérieure de l'espace correspondant à une classe bien déterminée. Cette hypothèse semble beaucoup plus réaliste, elle ne suffit pourtant pas à résoudre le problème de la distribution des distances entre exemples dans la limite thermodynamique.

En présence de l prototypes $(\mathbf{S}_1, \cdots, \mathbf{S}_l)$, la densité de probabilité des exemples s'écrit de la façon suivante :

$$P(\boldsymbol{\xi}) = \frac{1}{l\,(2\pi\sigma^2)^{N/2}} \left\{ \sum_{i=1}^{l} \exp\left(-\frac{(\boldsymbol{\xi} - \mathbf{S}_i)^2}{2\sigma^2} \right) \right\}. \tag{1.69}$$

où σ est la largeur des gaussiennes des différents prototypes. La largeur de la gaussienne ainsi que le poids de chacun des prototypes sont supposés identiques pour chacun d'entre eux.

Deux cas différents sont à considérer. Le premier correspond à des prototypes dont la norme est finie. La distribution en double gaussienne introduite au troisième chapitre de la première partie peut être considérée comme une distribution de deux prototypes symétriques par rapport à l'origine. La distance entre ces prototypes est nulle dans la limite thermodynamique :

$$\lim_{N\to+\infty} \frac{(\mathbf{S}_i - \mathbf{S}_j)^2}{N} = 0. \tag{1.70}$$

Dans ce cas, la distribution $P(d)$ des distances entre les exemples reste triviale :

$$P(d) = \delta(d - \sqrt{2}\,\sigma). \tag{1.71}$$

La borne de l'erreur de généralisation est alors triviale pour les mêmes raisons que précédemment.

Dans le second cas, où la distance entre les prototypes est finie :

$$\lim_{N \to +\infty} \frac{(\mathbf{S}_i - \mathbf{S}_j)^2}{N} = d_0, \tag{1.72}$$

la distribution des distances $P(d)$ entre les exemples possède non plus un pic δ mais deux :

$$P(d) = \frac{1}{l}\,\delta(d - \sqrt{2}\,\sigma) + \frac{l-1}{l}\,\delta(d - d_0 - \sqrt{2}\,\sigma). \tag{1.73}$$

Cette distribution ne permet toutefois pas d'envisager une dépendance de $\varepsilon_d(\Delta,\alpha)$ vis-à-vis de la taille réduite α. En effet, le nombre d'exemples de l'ensemble d'apprentissage diverge dans la limite thermodynamique, il existe donc toujours au moins un exemple proche de chaque prototype d'où :

$$\varepsilon_d(\Delta,\alpha) = \Theta(\Delta - \sqrt{2}\,\sigma). \tag{1.74}$$

Toutefois, cette fois-ci, il est possible d'obtenir une borne non triviale à l'erreur de généralisation. En effet, pour $\Delta = \sqrt{2}\,\sigma$, l'erreur de reconnaissance a un sens si $\sigma \ll d_0$. Un exemple n'est proche que d'un seul prototype et l'erreur de reconnaissance est alors simplement la probabilité que la machine à exemples supports ne donne pas la même classe à l'exemple et au prototype.

Il semble plus intéressant d'envisager non pas un nombre fini de prototypes mais un nombre proportionnel à la dimension N de l'espace des données. Cette hypothèse n'est pas incohérente si l'on pense que le nombre d'exemples de l'ensemble d'apprentissage est lui aussi proportionnel à N. Le nombre de prototypes est alors infini dans la limite thermodynamique, on pourra le noter $P_0 = \alpha_0 N$. Dans ce cas, on s'attend à ce que la probabilité $\varepsilon_d(\Delta,\alpha)$ qu'une donnée soit dans l'ensemble de proximité dépende de α et qu'elle soit une fonction croissante de α avec un changement de comportement plus prononcé autour de α_0. Les calculs concernant cette hypothèse n'ont pas été menés à bien jusqu'à présent mais ils constituent une piste intéressante pour mieux comprendre les bonnes performances de généralisation des machines à exemples supports dans les applications réalistes.

Chapitre 2

Machine de parité : Algorithme incrémental

Le perceptron simple, présenté dans l'introduction et étudié dans le deuxième chapitre de la première partie, permet de classer uniquement dans le cas de tâches linéairement séparables. En conséquence, sa capacité à apprendre correctement un ensemble de points, dont les classes sont aléatoires, est limité. Cover [28] a montré que cette capacité est $\alpha_c \equiv P_c/N \to 2$ dans la limite thermodynamique avec P_c le nombre maximal d'exemples de classe aléatoire qu'il est possible d'apprendre correctement et N la dimension de l'espace des données. Une possibilité pour augmenter la capacité est de considérer un réseau de neurones avec une couche cachée constituée de perceptrons simples. L'intérêt de tels réseaux est de créer une représentation interne binaire de chaque point de l'espace des données. Cette représentation interne est constituée de la classe associée par chacun des perceptron simples. La sortie finale est une fonction booléenne quelconque. Cette fonction est généralement fixée et deux choix particuliers ont été étudiés de manière intensive : le plus simple est la majorité des classes (le classifieur correspondant est la machine de comité) et le deuxième est la parité (le classifieur correspondant est appelé machine de parité) [126].

Il est bien connu que toute fonction booléenne spécifiée sur un ensemble d'apprentissage de taille quelconque peut être réalisée avec un réseau de neurones possédant un nombre suffisant de perceptrons sur la couche cachée [107, 126]. Le problème d'apprentissage reste cependant entier. En effet, il reste à trouver une procédure permettant de déterminer le nombre de perceptrons simples et leurs poids réalisant la fonction booléenne particulière spécifiée par l'ensemble d'apprentissage. L'idée de base des algorithmes incrémentaux est de construire progressivement un réseau de neurones en ajoutant des perceptrons simples tant qu'il est nécessaire pour corriger les erreurs de classifications restantes. Cette idée a été mise en avant premièrement par Gallant [37]. Elle a ensuite été reprise par d'autres auteurs. Mézard et Nadal [80] ont présenté, par exemple, l'algorithme dit de pavage (*tiling algorithm*). Des modifications à cet algorithme ont été apportées, notamment, par Nadal et Sirat [88, 113] avec les arbres neuronaux (*neural trees*), Frean [36] avec l'algorithme *Upstart*, Ruján et Marchand [106, 72] avec l'algorithme séquentiel (*sequential algorithm*), Martinez et Estève [73] avec l'algorithme *Offset*, Biehl et Opper [13] avec l'algorithme de pavage de la machine de parité (*tilinglike algorithm*) et Torres-Moreno, Peretto et Gordon [117, 115] avec les algorithmes *NetLines* et *NetSpheres*.

Dans ce chapitre, nous allons nous intéresser à l'étude de l'algorithme d'apprentissage particulier étudié en [13, 117]. Cet algorithme incrémental permet d'ap-

prendre correctement tous les exemples de l'ensemble d'apprentissage grâce à une machine de parité. Une borne supérieure de la capacité de la machine de parité, indépendante de l'algorithme utilisé, a été obtenue pour la première fois par Mitchinson et Durbin [85]. Des calculs basés sur les méthodes de Physique Statistique ont permis de montrer que la capacité de la machine de parité sature cette borne : $\alpha_c(k) = k \ln k / \ln 2$ lorsque $k \to +\infty$ [7, 8]. Nous nous proposons de comparer la capacité de l'algorithme incrémental, dont l'intérêt réside dans l'existence de preuves de convergence [13, 46], par rapport à la capacité de la machine de parité. Nous allons déterminer le nombre k de perceptrons simples nécessaires pour que l'algorithme incrémental converge, c'est-à-dire que la machine de parité obtenue classe correctement tous les exemples de l'ensemble d'apprentissage.

2.1 Présentation de l'algorithme incrémental

Nous allons supposer que l'on dispose de l'ensemble d'apprentissage suivant :

$$\mathcal{L}_\alpha = \{\boldsymbol{\xi}^\mu, \tau^\mu\}_{\mu=1,\cdots,P} \tag{2.1}$$

constitué de P exemples $\boldsymbol{\xi}^\mu$ distribués selon la densité de probabilité :

$$P(\boldsymbol{\xi}) = \frac{1}{(2\pi)^{N/2}} \exp\left(-\frac{\boldsymbol{\xi}^2}{2}\right). \tag{2.2}$$

Les classes τ^μ sont choisies aléatoirement avec une probabilité $1/2$ pour la classe $+1$ et $1/2$ pour la classe -1.

La machine de parité est constituée des briques élémentaires que sont les perceptrons simples avec seuil. Un perceptron simple avec seuil est une généralisation du classifieur introduit au Chapitre 2 de la première partie de la thèse. Ses paramètres sont composés d'une direction de l'espace des données $\mathbf{J}$, normée à un, et d'un seuil θ. Un exemple $\boldsymbol{\xi}$ est classé par le perceptron de la façon suivante :

$$\sigma \equiv \text{sign}\left(\mathbf{J} \cdot \boldsymbol{\xi} - \theta\right). \tag{2.3}$$

La différence par rapport au perceptron décrit dans la première partie de la thèse est l'introduction d'un seuil. Ce seuil correspond à la distance à l'origine de l'hyperplan séparateur, suivant la direction $\mathbf{J}$. Le seuil permet au perceptron simple de classer les exemples selon les classes ± 1 avec une probabilité différente de $1/2$. Nous verrons par la suite que cette propriété est indispensable pour la convergence de l'algorithme considéré.

La première étape de l'apprentissage consiste à considérer un perceptron simple dont les paramètres sont constitués du vecteur $\mathbf{J}_1$ et du seuil θ_1. Ces paramètres sont adaptés afin de minimiser une fonction de coût de la forme :

$$E(\mathbf{J}_1, \theta_1; \mathcal{L}_\alpha) = \sum_{\mu=1}^{P} V\left(\lambda_1^\mu\right) \tag{2.4}$$

avec λ^μ la stabilité de l'exemple μ :

$$\lambda_1^\mu \equiv \tau^\mu \left(\mathbf{J}_1 \cdot \boldsymbol{\xi}^\mu - \theta_1\right). \tag{2.5}$$

La stabilité est la distance de l'exemple à l'hyperplan séparateur pour un exemple bien classé et le négatif de cette distance si l'exemple est mal classé.

Si l'on prend pour potentiel $V(\lambda)$ la fonction de Heaviside $\Theta(-\lambda)$, la fonction de coût n'est rien d'autre que l'erreur d'apprentissage du perceptron simple avec seuil. Nous formulons l'algorithme en terme d'une fonction V générale, car l'étude

analytique lorsque la fonction de coût est l'erreur d'apprentissage est souvent complexe. Il est plus aisé de considérer des fonctions de coût plus simples, par exemple, convexes qui ont alors un unique minimum.

Une fois la minimisation de la fonction de coût effectuée et les paramètres $\mathbf{J}_1^*$ et θ_1^* déterminés, l'erreur d'apprentissage ε_t^1 correspondante est calculée :

$$\varepsilon_t^1(\mathbf{J}_1^*,\theta_1^*;\mathcal{L}_\alpha) = \frac{1}{P}\sum_{\mu=1}^{P}\Theta\left(-\tau^\mu\sigma_1^\mu\right) \tag{2.6}$$

où σ_1^μ est la classe que le premier perceptron associe à l'exemple $\boldsymbol{\xi}^\mu$. Cette classe σ_1^μ dépend de l'exemple $\boldsymbol{\xi}^\mu$ mais aussi de $\mathbf{J}_1^*$ et θ_1^* d'après l'équation (2.3).

Si l'erreur d'apprentissage ε_t^1 est nulle, l'apprentissage s'arrête. Alors la classe associée par la machine de parité à une donnée quelconque est celle du perceptron simple.

Si au contraire l'erreur d'apprentissage ε_t^1 est non nulle, l'apprentissage continue par l'ajout et l'apprentissage d'un nouveau perceptron simple avec seuil. Toutefois, pour ce nouvel apprentissage, la classe des exemples de l'ensemble d'apprentissage $\mathcal{L}_\alpha$ que l'on souhaite apprendre n'est plus τ^μ mais $\tau_2^\mu = \sigma_1^\mu\tau^\mu$. La classe τ_2^μ que l'on se propose d'apprendre avec le deuxième perceptron est $+1$ si l'exemple a été bien appris par le perceptron précédent et -1 sinon. L'idée est de séparer les exemples bien appris des autres. Cela définit un nouvel ensemble d'apprentissage :

$$\mathcal{L}_\alpha^2 = \{\boldsymbol{\xi}^\mu,\tau_2^\mu\}_{\mu=1,\cdots,P}\,. \tag{2.7}$$

La classe τ_2^μ ne possède pas la même probabilité d'être égale à $+1$ ou à -1. En effet, la probabilité d'être égale à -1 est ε_t^1 et celle d'être égale à $+1$ est $1-\varepsilon_t^1$. L'intérêt de considérer des perceptrons simples avec seuil est alors évident. Pour l'apprentissage du premier perceptron, le seuil n'était pas essentiel puisque les classes $+1$ et -1 était équiprobables, mais ce n'est plus le cas pour l'apprentissage du deuxième perceptron.

L'apprentissage du deuxième perceptron $(\mathbf{J}_2^*,\theta_2^*)$ est effectué avec la même fonction de coût que celui du premier perceptron. On calcule ensuite l'erreur d'apprentissage ε_t^2 correspondante :

$$\varepsilon_t^2(\mathbf{J}_2^*,\theta_2^*,\mathcal{L}_\alpha^2) = \frac{1}{P}\sum_{\mu=1}^{P}\Theta\left(-\tau_2^\mu\sigma_2^\mu\right) \tag{2.8}$$

où, comme précédemment, σ_2^μ est la classe que le deuxième perceptron associe à l'exemple $\boldsymbol{\xi}^\mu$. Il est possible de montrer [13, 46] que si les exemples sont en position générale et si l'apprentissage consiste à minimiser l'erreur d'apprentissage, alors ε_t^2 est strictement plus faible que ε_t^1. Il suffit alors de recommencer cette procédure jusqu'à l'itération k pour laquelle $\varepsilon_t^k = 0$. Le nombre d'itérations k est fini, car $P\varepsilon_t^i$ est une suite d'entiers strictement décroissante. Il existe donc bien une itération k pour laquelle $\varepsilon_t^k = 0$, c'est-à-dire que l'algorithme converge.

L'erreur d'apprentissage du perceptron k est nulle et on peut démontrer [13, 73] que la machine de parité classe correctement tous les exemples de l'ensemble d'apprentissage. En effet, la classe τ_i^μ de l'exemple $\boldsymbol{\xi}^\mu$ que le perceptron i apprend a été définie de la façon suivante : $\tau_i^\mu = \tau_{i-1}^\mu\sigma_{i-1}^\mu$. La sortie σ^μ de la machine de parité est le produit de toutes les sorties des perceptrons :

$$\sigma^\mu \equiv \sigma_1^\mu\cdots\sigma_k^\mu. \tag{2.9}$$

La sortie du dernier perceptron est $\sigma_k^\mu = \tau_k^\mu$ puisque l'erreur d'apprentissage ε_t^k pour celui-ci est nulle. En remplaçant τ_k^μ par son expression, on aboutit à :

$$\begin{aligned}
\sigma^\mu &= \sigma_1^\mu \cdots \sigma_{k-1}^\mu \sigma_{k-1}^\mu \tau_{k-1}^\mu, \\
\sigma^\mu &= \sigma_1^\mu \cdots \sigma_{k-2}^\mu \tau_{k-1}^\mu, \\
&\vdots \\
\sigma^\mu &= \tau^\mu. \qquad (2.10)
\end{aligned}$$

La classe σ^μ donnée par la machine de parité à l'exemple $\boldsymbol{\xi}^\mu$ est bien celle de l'ensemble d'apprentissage initial. L'erreur d'apprentissage de la machine de parité est donc nulle.

2.2 Propriétés de l'algorithme incrémental

Nous allons maintenant nous intéresser à la détermination du nombre de perceptrons simples nécessaires pour que l'algorithme incrémental converge. Dans la limite thermodynamique, $N \to +\infty$ et $P \to +\infty$ avec la taille réduite $\alpha = P/N$ de l'ensemble d'apprentissage constante, ce nombre de perceptrons devient indépendant du choix particulier des exemples de l'ensemble d'apprentissage et ne dépend plus alors que de la taille réduite α.

Pour déterminer les propriétés d'apprentissage et particulièrement l'erreur d'apprentissage de chaque perceptron, nous allons utiliser l'approche de la Physique Statistique.

2.2.1 Approche par la Physique Statistique

Les perceptrons simples sont déterminés successivement par la minimisation d'une fonction de coût (2.4). Pour déterminer les propriétés du premier perceptron, notamment l'erreur d'apprentissage ε_t^1, nous allons utiliser la même approche que dans la première partie de la thèse. Nous introduisons l'inverse d'une température fictive β_1 et nous considérons la fonction de partition suivante :

$$Z_1(\beta_1,\mathcal{L}_\alpha) = \int P(\mathbf{J}_1) d\mathbf{J}_1 \int d\theta_1 \exp\left(-\beta_1 E(\mathbf{J}_1,\theta_1,\mathcal{L}_\alpha)\right). \qquad (2.11)$$

La variable θ_1 est une nouvelle variable dynamique qui s'ajoute à la direction $\mathbf{J}_1$. La densité de probabilité $P(\mathbf{J}_1)$ est uniforme sur la sphère de rayon 1 :

$$P(\mathbf{J}_1) = \delta\left(\mathbf{J}_1 \cdot \mathbf{J}_1 - 1\right) \qquad (2.12)$$

et la densité de probabilité de θ_1 est uniforme entre $-\sqrt{N}$ et $\sqrt{N}$. Dans la limite thermodynamique, cela revient à une densité uniforme sur l'espace tout entier. La constante de normalisation étant nulle dans cette limite, nous l'avons omise dans l'expression de la fonction de partition puisqu'elle ne modifie pas les propriétés thermodynamiques.

Les propriétés du premier perceptron sont obtenus en calculant l'énergie libre f, correspondant à la fonction de partition (2.11), dans la limite de température nulle ($\beta_1 \to +\infty$). Ce calcul est effectué par la méthode des répliques dans la limite thermodynamique :

$$f = -\lim_{\beta\to+\infty} \lim_{\substack{N\to+\infty \\ P\to+\infty \\ \alpha=P/N}} \lim_{n\to 0} \frac{1}{\beta n N} \ln \overline{Z_1^n(\beta,\mathcal{L}_\alpha)}. \qquad (2.13)$$

Le calcul des propriétés du deuxième perceptron et des suivants est un peu plus complexe. Nous introduisons toujours l'inverse d'une température fictive β_i

pour tenir compte du minimum de la fonction de coût dans la limite $\beta_i \to +\infty$. Toutefois, pour les propriétés du deuxième perceptron, il faut aussi prendre en compte la dépendance de la classe τ_2^μ vis-à-vis du premier perceptron. Pour cela, nous introduisons la fonction de partition Z_{12} à deux perceptrons :

$$Z_{12}(\beta_1,\beta_2,\mathcal{L}_\alpha) = \int P(\mathbf{J}_1)d\mathbf{J}_1 \int d\theta_1 \int P(\mathbf{J}_2)d\mathbf{J}_2 \int d\theta_2 \exp\left(-\beta_1 E(\mathbf{J}_1,\theta_1,\mathcal{L}_\alpha) - \beta_2 E(\mathbf{J}_2,\theta_2,\mathcal{L}_\alpha^2)\right). \tag{2.14}$$

Pour fixer les paramètres $(\mathbf{J}_1,\theta_1)$ aux valeurs correspondant au premier perceptron, il convient de prendre la limite $\beta_1 \to +\infty$. La fonction de partition Z_2 pour le deuxième perceptron s'écrit alors :

$$Z_2(\beta_2,\mathcal{L}_\alpha) = \lim_{\beta_1\to+\infty} \frac{Z_{12}(\beta_1,\beta_2,\mathcal{L}_\alpha)}{Z_1(\beta_1,\mathcal{L}_\alpha)}. \tag{2.15}$$

où l'on a divisé Z_{12} par Z_1 pour des raisons de normalisation. Cette méthode se généralise aisément à plus de deux perceptrons. Toutefois, le calcul de la fonction de partition à k perceptrons devient vite inextricable. À partir de ces fonctions de partition, il est théoriquement possible de calculer la distribution des stabilités des exemples pour chaque perceptron et d'en déduire l'erreur d'apprentissage ε_t^i qui n'est autre que l'intégrale sur les stabilités négatives de cette distribution.

Pour simplifier le calcul de Z_2, nous allons faire une hypothèse drastique. Celle-ci consiste à considérer les classes τ_2^μ de l'ensemble d'apprentissage $\mathcal{L}_\alpha^2$ comme étant distribuées aléatoirement avec une probabilité ε_t^1 pour la classe -1 et $1-\varepsilon_t^1$ pour la classe $+1$. Il est possible de montrer [128] que cette hypothèse revient simplement à négliger la corrélation entre les directions $\mathbf{J}_1^*$ et $\mathbf{J}_2^*$ dans le calcul de Z_{12}. West et Saad ont montré récemment que cette hypothèse n'est pas vérifiée [128]. Elle néglige les contraintes sur les classes τ_2^μ dues au premier perceptron. On peut s'attendre à ce que l'effet de ces contraintes soit défavorable, ce qui signifie que l'erreur d'apprentissage ε_t^2 calculée avec cette hypothèse est une borne inférieure de l'erreur d'apprentissage réelle. Cette remarque justifie la poursuite du calcul avec cette hypothèse même si elle n'est pas exacte. L'erreur d'apprentissage étant sous estimée à chaque étape, le nombre de perceptrons nécessaires à la convergence de l'algorithme obtenu avec cette hypothèse est une borne inférieure du nombre exact de perceptrons nécessaires.

L'hypothèse considérée, appelée par la suite *hypothèse de non corrélation des perceptrons*, permet de remplacer la fonction de partition Z_2 de l'équation (2.15) par une fonction de partition du même type que Z_1 (2.11) pour laquelle l'ensemble d'apprentissage est composé des exemples de $\mathcal{L}_\alpha$ mais dont les classes sont choisies aléatoirement avec une probabilité ε_t^1 pour la classe -1 et $1-\varepsilon_t^1$ pour la classe $+1$. La distribution des classes d'un tel ensemble d'apprentissage est appelée biaisée dans le sens où la probabilité de la classe $+1$ est différente de celle de la classe -1. Le biais est $1-2\varepsilon_t^1$.

2.2.2 Conditions de convergence

Le calcul du nombre de perceptrons nécessaires pour que l'erreur d'apprentissage de la machine de parité apprenant avec l'algorithme incrémental soit nulle, se réduit à l'étude de l'apprentissage d'un perceptron simple avec seuil, avec un ensemble d'apprentissage dont la distribution des classes est biaisée. Plus particulièrement, nous allons nous intéresser à l'erreur d'apprentissage $\mathcal{E}_t(\alpha,\varepsilon)$ d'un perceptron simple apprenant un ensemble d'apprentissage de taille réduite α et dont le biais de la distribution des classes est $1-2\varepsilon$. En effet, l'erreur d'apprentissage ε_t^i du perceptron i s'écrit alors simplement :

$$\varepsilon_t^i = \mathcal{E}_t(\alpha,\varepsilon_t^{i-1}). \tag{2.16}$$

où le biais $1-2\varepsilon_t^{i-1}$ du perceptron i est déterminé par l'erreur d'apprentissage du perceptron $i-1$.

Revenons, maintenant, à la convergence de l'algorithme incrémental considéré. La relation (2.16) qui existe entre l'erreur d'apprentissage du perceptron $i-1$ et i impose que la fonction de coût utilisée pour l'apprentissage du perceptron simple vérifie la condition suivante :

$$\mathcal{E}_t(\alpha,\varepsilon) < \varepsilon. \tag{2.17}$$

Cette condition permet de garantir que l'erreur d'apprentissage ε_t^i décroît à chaque itération. La condition d'arrêt de l'algorithme incrémental avec un nombre de perceptrons k fini est la suivante :

$$\varepsilon_t^k = \mathcal{E}_t(\alpha,\varepsilon_t^{k-1}) = 0. \tag{2.18}$$

Cette deuxième condition nécessite que pour toute taille réduite α, il existe une valeur $\varepsilon_0(\alpha)$ non nulle pour laquelle $\mathcal{E}_t(\alpha,\varepsilon_0) = 0$. En effet, dans la limite thermodynamique où P est infini, il n'est plus possible d'invoquer que la suite des entiers $P\varepsilon_t^i$ est strictement décroissante pour assurer que l'algorithme s'arrête avec un nombre de perceptrons k fini. Ceci signifie, entre autre, que la solution qui consiste à apprendre correctement un exemple à chaque itération comme le propose la démonstration de convergence de Biehl et Opper [13] utilisant un perceptron appelé *grand-mère*, ne nous satisfait pas dans la limite thermodynamique.

L'inverse $\alpha_0(\varepsilon)$ de $\varepsilon_0(\alpha)$ est la capacité du perceptron simple avec seuil dans le cas d'une distribution des classes avec un biais $1-2\varepsilon$. Cette capacité est la taille réduite maximale d'exemples de l'ensemble d'apprentissage que le perceptron est capable d'apprendre sans erreur d'apprentissage. La condition qu'il existe une valeur non nulle ε_0 pour toute valeur de α impose que $\alpha_0(\varepsilon)$ diverge lorsque ε s'annule. Cette condition n'est pas irréalisable. En effet, un des premiers calculs utilisant la méthode des répliques effectué par Gardner [38, 39, 40] a été celui de la capacité $\alpha_0(\varepsilon)$ d'un perceptron dont la distribution des classes est biaisée. Lorsque le biais est nul ($\varepsilon = 1/2$), correspondant à une distribution des classes symétrique, la capacité est $\alpha_0(1/2) = 2$. Lorsque le biais est maximal ($\varepsilon = 0$), les exemples sont tous classés de manière identique et, lorsque $\varepsilon \to 0$, la capacité diverge comme :

$$\alpha_0(\varepsilon) \sim -\frac{1}{\varepsilon \ln \varepsilon}. \tag{2.19}$$

2.3 Détermination du nombre de perceptrons

Connaissant l'erreur d'apprentissage $\mathcal{E}_t(\alpha,\varepsilon)$ d'un perceptron simple avec seuil en fonction du biais $1-2\varepsilon$ de la distribution des classes et de la taille réduite α de l'ensemble d'apprentissage, connaissant de plus la relation liant l'erreur d'apprentissage des perceptrons successifs de la machine de parité :

$$\varepsilon_t^i(\alpha) = \mathcal{E}_t(\alpha,\varepsilon_t^{i-1}), \tag{2.20}$$

il est possible de déterminer le nombre de perceptrons nécessaires à la convergence de l'algorithme incrémental, c'est-à-dire, le nombre nécessaire pour obtenir une erreur d'apprentissage nulle de la machine de parité. Cette condition s'écrit pour une distribution des classes initiales τ^μ symétrique ($\varepsilon = 1/2$) :

FIG. 2.1 – *Evolution de l'erreur d'apprentissage des perceptrons successifs. La courbe en trait plein représente l'erreur d'apprentissage $\mathcal{E}_t(\alpha,\varepsilon)$ d'un perceptron simple avec seuil pour une taille réduite α de l'ensemble d'apprentissage en fonction de ε (reflétant le biais $1-2\varepsilon$ de la distribution des classes). L'erreur d'apprentissage du premier perceptron est donnée par $\varepsilon_t^1 = \mathcal{E}_t(\alpha,1/2)$ et celle des suivants par $\varepsilon_t^{i+1} = \mathcal{E}_t(\alpha,\varepsilon_t^i)$. L'algorithme incrémental converge, dans ce cas, pour le sixième perceptron.*

$$\underset{k}{\circ} f_\alpha(1/2) = \underbrace{f_\alpha \circ \cdots \circ f_\alpha}_{k \text{ fois}} (1/2) = 0. \tag{2.21}$$

La fonction $f_\alpha(\varepsilon)$ correspond à $\mathcal{E}_t(\alpha,\varepsilon)$ pour laquelle la dépendance en α est indiquée en indice, et le symbole $\circ$ désigne la composition de fonctions. L'évolution de l'erreur d'apprentissage des différents perceptrons est représentée sur la figure 2.1 pour une erreur d'apprentissage $\mathcal{E}_t(\alpha,\varepsilon)$ donnée. Pour ce cas, l'algorithme incrémental converge après l'introduction du sixième perceptron.

Nous allons nous s'intéresser, par la suite, à la limite des grands ensembles d'apprentissage, c'est-à-dire, $\alpha \to +\infty$. Dans ce cas, la fonction $f_\alpha(\varepsilon)$ peut s'écrire, pour des fonctions de coût (2.4) efficaces, de la façon suivante :

$$f_\alpha(\varepsilon) \simeq \varepsilon - h(\alpha,\varepsilon) \tag{2.22}$$

et la fonction $h(\alpha,\varepsilon)$ est une fonction qui s'annule lorsque $\alpha \to +\infty$. Introduisons maintenant les variables :

$$x_i = \underset{i}{\circ} f_\alpha(1/2). \tag{2.23}$$

Avec cette définition :

$$
\begin{aligned}
x_0 &= \frac{1}{2}, && (2.24)\\
x_1 &= \frac{1}{2} - h(\alpha,x_0), && (2.25)\\
x_2 &= x_1 - h(\alpha,x_1) = \frac{1}{2} - \sum_{j=0}^{1} h(\alpha,x_j), && (2.26)\\
&\vdots &&\\
x_k &= 0 = \frac{1}{2} - \sum_{j=0}^{k-1} h(\alpha,x_j). && (2.27)
\end{aligned}
$$

On arrive donc à l'équation suivante :

$$\frac{1}{2} = \sum_{j=0}^{k-1} h(\alpha,x_j). \tag{2.28}$$

Nous allons exprimer la somme sur j comme une intégrale en passant au continuum. Introduisons la fonction $x(y)$ définie pour j/k avec $j = 0, \cdots ,k$ par :

$$x(j/k) = x_j. \tag{2.29}$$

La fonction $x(y)$, prolongée par continuité sur [0,1], est décroissante avec $x(0) = 1/2$ et $x(1) = 0$. Cette fonction est une fonction monotone bien définie puisque l'on se place dans la limite $\alpha \to +\infty$ pour laquelle le nombre k de perceptrons simples nécessaires à la convergence de l'algorithme incrémental diverge. Avec l'aide de la fonction $x(y)$, on peut réécrire l'équation (2.28) de la façon suivante :

$$\frac{1}{2} = \sum_{j=1}^{k-1} h(\alpha,x(j/k)) \simeq k \int_0^1 dy\, h(\alpha,x(y)). \tag{2.30}$$

On en déduit que le comportement de $k(\alpha)$ lorsque $\alpha \to +\infty$ est donné par :

$$k(\alpha) \simeq \frac{1}{2\displaystyle\int_0^1 dy\, h(\alpha,x(y))}. \tag{2.31}$$

Il reste maintenant à déterminer l'intégrale de $h(\alpha,x(y))$. Pour cela, on revient à la définition de x_j et on exprime x_{j+1} en fonction de x_j :

$$x_{j+1} - x_j = -h(\alpha,x_j). \tag{2.32}$$

Dans la limite du continuum, on obtient :

$$\frac{dx}{dy} = -k\, h(\alpha,x) = -\frac{h(\alpha,x)}{2\displaystyle\int_0^1 dy\, h(\alpha,x(y))}. \tag{2.33}$$

Cette équation différentielle permet de déterminer $x(y)$ connaissant $h(\alpha,x)$. Elle est simple à résoudre puisque les deux variables x et y peuvent se séparer :

$$dy = -\frac{2\,dx}{h(\alpha,x)} \int_0^1 dy\, h(\alpha,x(y)). \tag{2.34}$$

Il suffit alors d'intégrer les deux cotés de l'équation pour obtenir l'égalité suivante :

$$1 = 2 \int_0^{1/2} \frac{dx}{h(\alpha,x)} \int_0^1 dy\, h(\alpha,x(y)). \tag{2.35}$$

Le comportement de $k(\alpha)$ se réduit alors simplement à :

$$k(\alpha) \simeq \int_0^{1/2} \frac{dx}{h(\alpha,x)}. \tag{2.36}$$

Nous avons ainsi déterminé le comportement asymptotique du nombre $k(\alpha)$ de perceptrons simples nécessaires à la convergence de l'algorithme incrémental dans la limite des très grands ensembles d'apprentissage. $k(\alpha)$ dépend de la fonction $h(\alpha,x)$ qui elle même dépend de la fonction de coût (2.4) utilisée pour l'apprentissage de chaque perceptron simple. L'inversion de $k(\alpha)$ permet de déterminer la capacité $\alpha_c(k)$ de la machine de parité avec k perceptrons pour l'algorithme incrémental considéré. Cette capacité est une borne inférieure à la capacité générale, celle indépendante de l'algorithme utilisé pour l'apprentissage de la machine de parité. Nous allons, dans la suite, étudier la capacité que l'on peut atteindre pour diverses règles d'apprentissage, c'est-à-dire, différents potentiels V définissant la fonction de coût (2.4).

2.4 Calcul de l'erreur d'apprentissage d'un perceptron simple avec seuil

Les approximations introduites ramènent le calcul à celui de l'erreur d'apprentissage d'un perceptron simple avec seuil. Bien que le perceptron ait été étudié de manière très approfondie, il y a peu de résultats systématiques pour le perceptron avec seuil en fonction du biais dans la distribution des classes. Nous allons développer ce calcul général de l'erreur d'apprentissage $\mathcal{E}_t(\alpha,\varepsilon)$ d'un perceptron simple avec seuil, dans le cas où l'ensemble d'apprentissage est de taille réduite α et a un biais $1-2\varepsilon$ dans la distribution des classes, dans la limite thermodynamique et pour une fonction de coût (2.4) définie par un potentiel V quelconque. Les résultats de ce paragraphe seront repris ensuite pour évaluer les performances de deux algorithmes différents.

Le calcul de l'énergie libre (2.13) par la méthode des répliques est similaire à celui de la première partie. Nous utilisons l'hypothèse de symétrie des répliques, ce qui permet d'aboutir à l'énergie libre suivante :

$$f(\alpha,\varepsilon) = \lim_{\beta\to+\infty} \lim_{\substack{N\to+\infty \\ P\to+\infty \\ \alpha=P/N}} \lim_{n\to 0} -\frac{1}{\beta n N} \ln \overline{Z_1^n(\beta,\mathcal{L}_\alpha(\varepsilon))}, \tag{2.37}$$

$$= \operatorname*{extr}_{c,\theta} g(\alpha,\varepsilon,c,\theta). \tag{2.38}$$

$\mathcal{L}_\alpha(\varepsilon)$ correspond à un ensemble d'apprentissage contenant P exemples $\boldsymbol{\xi}^\mu$ distribués selon une densité de probabilité gaussienne (2.2). Les classes correspondantes $\tau^\mu = \pm 1$ ont une probabilité $1-\varepsilon$ pour la classe $\tau^\mu = +1$ et ε pour la classe $\tau^\mu = -1$. Le paramètre d'ordre c est identique à celui introduit dans la première partie de la thèse, c'est-à-dire :

$$c = \lim_{\beta\to+\infty} \beta(1 - \mathbf{J}_a \cdot \mathbf{J}_b) \tag{2.39}$$

avec $\mathbf{J}_a$ et $\mathbf{J}_b$ les directions correspondantes à deux répliques différentes. Le paramètre R n'est pas présent dans ce calcul car il n'existe pas de direction privilégiée.

Le paramètre θ correspond, quant à lui, au seuil. Les deux paramètres c et θ sont solutions des équations d'extremum :

$$\frac{\partial g}{\partial c} = \frac{\partial g}{\partial \theta} = 0 \tag{2.40}$$

avec :

$$\begin{aligned} g(\alpha,\varepsilon,c,\theta) &= -\frac{1}{2c} + \alpha\,(1-\varepsilon)\int \frac{dy}{\sqrt{2\pi}} \exp\left(-\frac{(y+\theta)^2}{2}\right) W[\lambda(y,c),y,c] \\ &\quad + \alpha\,\varepsilon \int \frac{dy}{\sqrt{2\pi}} \exp\left(-\frac{(y-\theta)^2}{2}\right) W[\lambda(y,c),y,c]. \end{aligned} \tag{2.41}$$

La fonction $\lambda(y,c)$ est celle qui minimise :

$$W[\lambda,y,c] = V(\lambda) + \frac{(\lambda - y)^2}{2\,c}. \tag{2.42}$$

La distribution $\rho(\gamma)$ des stabilités des exemples de l'ensemble d'apprentissage est donnée par :

$$\begin{aligned} \rho(\gamma) &= (1-\varepsilon)\int \frac{dy}{\sqrt{2\pi}} \exp\left(-\frac{(y+\theta)^2}{2}\right) \delta(\lambda(y,c)-\gamma) \\ &\quad + \varepsilon \int \frac{dy}{\sqrt{2\pi}} \exp\left(-\frac{(y-\theta)^2}{2}\right) \delta(\lambda(y,c)-\gamma). \end{aligned} \tag{2.43}$$

L'erreur d'apprentissage $\mathcal{E}_t(\alpha,\varepsilon)$ se calcule alors simplement en intégrant la fraction d'exemples qui ont des stabilités négatives et est donnée par l'expression suivante :

$$\begin{aligned} \mathcal{E}_t(\alpha,\varepsilon) &= (1-\varepsilon)\int \frac{dy}{\sqrt{2\pi}} \exp\left(-\frac{(y+\theta)^2}{2}\right) \Theta\left(-\lambda(y,c)\right) \\ &\quad + \varepsilon \int \frac{dy}{\sqrt{2\pi}} \exp\left(-\frac{(y-\theta)^2}{2}\right) \Theta\left(-\lambda(y,c)\right). \end{aligned} \tag{2.44}$$

2.5 Règles d'apprentissage particulières

Le but de l'apprentissage des perceptrons simples par la minimisation d'une fonction de coût est de donner l'erreur d'apprentissage la plus faible possible, afin que l'algorithme incrémental converge avec le plus petit nombre de perceptrons possible. La première des fonctions de coût que nous allons considérer est la plus simple d'entre elles. Il s'agit de l'erreur d'apprentissage. Cette fonction de coût est aussi celle qui possède les meilleurs performances par définition. Toutefois, l'hypothèse de symétrie des répliques utilisée pour calculer l'erreur d'apprentissage n'est pas vérifiée dès que celle-ci est non nulle [39, 40, 18, 49, 51]. Le résultat obtenu avec cette hypothèse est alors une borne inférieure de l'erreur réelle [35, 70, 129]. Là encore, comme pour l'hypothèse de non corrélation des perceptrons, la conséquence est que l'on détermine uniquement une borne inférieure sur le nombre $k(\alpha)$ de perceptrons nécessaires à une erreur d'apprentissage nulle de la machine de parité. Cette borne devient une borne supérieure pour la capacité $\alpha_c(k)$ d'une machine de parité à k perceptrons apprenant avec l'algorithme incrémental étudié.

La deuxième fonction de coût que nous allons considérer ne possède pas le problème de brisure de symétrie des répliques. Il s'agit d'une fonction de coût quadratique par morceaux [6, 49] :

$$V(\lambda) = (\kappa - \lambda)^2 \, \Theta(\kappa - \lambda) \, . \tag{2.45}$$

Cette fonction V convexe entraîne la convexité de la fonction de coût et l'existence d'un minimum unique de celle-ci. Cette fonction de coût possède l'avantage de satisfaire à l'hypothèse de symétrie des répliques pour toutes les valeurs des paramètres κ,α et ε. Le paramètre κ devra être choisi au mieux au cours de chaque minimisation de la fonction de coût pour obtenir une erreur d'apprentissage faible. Par exemple, il faut noter que le choix d'un paramètre κ identique pour chaque perceptron ne permet pas la convergence de l'algorithme incrémental. On reviendra sur les raisons de cette absence de convergence.

2.5.1 Minimisation de l'erreur d'apprentissage

L'erreur d'apprentissage correspond à un potentiel : $V(\lambda) = \Theta(-\lambda)$. La fonction $\lambda(y,c)$ qui minimise (2.42) est :

$$\lambda(y,c) = \begin{cases} y & y < -\sqrt{2c}, \\ 0 & -\sqrt{2c} < y < 0, \\ y & 0 < y. \end{cases} \tag{2.46}$$

Introduisant cette fonction $\lambda(y,c)$ dans (2.41), on en déduit :

$$\begin{aligned} g(\alpha,\varepsilon,c,\theta) &= -\frac{1}{2c} + \alpha(1-\varepsilon) \int_{-\infty}^{-\sqrt{2c}+\theta} Dy + \alpha\,\varepsilon \int_{-\infty}^{-\sqrt{2c}-\theta} Dy \\ &\quad + \frac{\alpha}{2\,c} \left\{ (1-\varepsilon) \int_{-\sqrt{2c}+\theta}^{\theta} (y-\theta)^2 Dy + \varepsilon \int_{-\sqrt{2c}-\theta}^{-\theta} (y+\theta)^2 Dy \right\}. \end{aligned} \tag{2.47}$$

Les conditions d'extremum (2.40) permettant de déterminer c et θ ainsi que l'erreur d'apprentissage $\mathcal{E}_t(\alpha,\varepsilon)$ (2.44) sont données par :

$$\frac{1}{\alpha} = (1-\varepsilon) \int_{-\sqrt{2c}+\theta}^{\theta} (y-\theta)^2 Dy + \varepsilon \int_{-\sqrt{2c}-\theta}^{-\theta} (y+\theta)^2 Dy, \tag{2.48}$$

$$0 = (1-\varepsilon) \int_{-\sqrt{2c}+\theta}^{\theta} (y-\theta) Dy - \varepsilon \int_{-\sqrt{2c}-\theta}^{-\theta} (y+\theta) Dy, \tag{2.49}$$

$$\mathcal{E}_t(\alpha,\varepsilon) = (1-\varepsilon) \int_{\sqrt{2c}-\theta}^{+\infty} Dy + \varepsilon \left(1 - \int_{-\sqrt{2c}-\theta}^{+\infty} Dy \right), \tag{2.50}$$

avec $Dy = \exp(-y^2/2)dy/\sqrt{2\pi}$.

Plusieurs remarques s'imposent sur la solution de ces équations. La première est qu'il existe une solution d'erreur d'apprentissage nulle pour $\alpha < \alpha_0(\varepsilon)$. Cette solution correspond à une valeur infinie de c pour laquelle l'énergie libre f est nulle. $\alpha_0(\varepsilon)$ est déterminée par les équations (2.48) et (2.49). Cette valeur n'est rien d'autre que la capacité d'un perceptron simple apprenant par minimisation du nombre d'erreurs un ensemble d'apprentissage avec des classes biaisées. Cette capacité a été calculée pour la première fois par Gardner [39, 40] avec la méthode des répliques.

Pour $\alpha > \alpha_0(\varepsilon)$, la solution correspond à une valeur de c finie. Il est intéressant de noter que la distribution $\rho(\gamma)$ (2.43) des stabilités des exemples de l'ensemble d'apprentissage, c'est-à-dire, de la variable $\lambda^\mu = \tau^\mu(\mathbf{J}^* \cdot \boldsymbol{\xi}^\mu - \theta^*)$ possède une région comprise entre deux régions de distribution non nulle où la distribution est nulle. Il a été montré par Bouten [18] que, dans ce cas, l'hypothèse de symétrie des répliques

n'était pas vérifiée. La brisure de symétrie des répliques est indispensable. Cependant, les corrections à l'erreur d'apprentissage, dues à cette brisure, sont assez faibles et montrent que l'hypothèse de symétrie des répliques sous estime l'erreur d'apprentissage réelle [35, 70, 129, 45]. Ceci permet de conclure que la capacité $\alpha_c(k)$ que nous allons déterminer n'est qu'une borne supérieure de la capacité réelle.

Nous allons commencer par déterminer la capacité $\alpha_c(2)$, correspondant à la plus grande taille réduite de l'ensemble d'apprentissage pour laquelle l'algorithme incrémental converge après l'introduction de deux perceptrons. Pour déterminer $\alpha_c(2)$, il faut tout d'abord calculer l'erreur d'apprentissage $\varepsilon_t^1 = \mathcal{E}_t(\alpha_c(2),1/2)$ du premier perceptron. Cette erreur d'apprentissage correspond à la valeur la plus grande pour laquelle $\mathcal{E}_t(\alpha_c(2),\varepsilon_t^1) = 0$ par définition de la capacité $\alpha_c(2)$. D'après l'équation (2.49), le seuil $\theta_1 = 0$ pour un biais nul de la distribution des classes ($\varepsilon = 1/2$). L'erreur d'apprentissage ε_t^1 se réduit à :

$$\varepsilon_t^1 = \int_{\sqrt{2c_1}}^{+\infty} Dy \tag{2.51}$$

avec c_1 solution de l'équation :

$$\frac{1}{\alpha_c(2)} = \int_0^{\sqrt{2c_1}} y^2\, Dy. \tag{2.52}$$

L'erreur d'apprentissage $\mathcal{E}_t$ est nulle lorsque c est infini. On en déduit que le paramètre $c_2 = +\infty$ ce qui simplifie les équations (2.48) et (2.49) pour le second perceptron :

$$\frac{1}{\alpha_c(2)} = (1-\varepsilon_t^1)\int_{-\infty}^{\theta_2}(y-\theta_2)^2 Dy + \varepsilon_t^1 \int_{-\infty}^{-\theta_2}(y+\theta_2)^2 Dy, \tag{2.53}$$

$$0 = (1-\varepsilon_t^1)\int_{-\infty}^{\theta_2}(y-\theta_2) Dy - \varepsilon_t^1 \int_{-\infty}^{-\theta_2}(y+\theta_2) Dy. \tag{2.54}$$

La résolution de ce système de 4 équations à 4 inconnues (c_1, ε_t^1, θ_2 et $\alpha_c(2)$) permet de déduire $\alpha_c(2) = 4.60$. Nous retrouvons ainsi le résultat de Biehl et Opper [13]. Ce résultat est très différent de la capacité générale de la machine de parité avec deux perceptrons (qui est indépendante de l'algorithme utilisé). En effet, Mézard et Patarnello [82] ont obtenu avec l'hypothèse de symétrie des répliques $\alpha_c(2) = 11.0$ et Barkai, Hansel et Kanter ont obtenu $\alpha_c(2) = 8.1$ avec un pas de brisure. Cette différence tend à montrer que la capacité de l'algorithme incrémental est très faible par rapport à la capacité de la machine de parité. Nous allons voir que ce résultat n'est plus vrai dans la limite des grands ensembles d'apprentissage.

En effet, nous allons nous intéresser au comportement du nombre $k(\alpha)$ de perceptrons nécessaires à la convergence de l'algorithme incrémental dans la limite des plus grandes tailles réduites α ($\alpha \to +\infty$). Pour cela, il convient de déterminer le comportement de $\mathcal{E}_t(\alpha,\varepsilon)$ dans la limite $\alpha \to +\infty$. On effectue l'hypothèse que $c \ll |\theta|$. Il est à noter que le seuil θ est négatif. Nous allons supposer dans les développements suivants que θ diverge tout en gardant le produit $a = \theta\sqrt{2c}$ fini. Avec ces hypothèses :

$$\int_{-\sqrt{2c}+\theta}^{\theta}(y-\theta)Dy = \exp\left(-\theta^2/2\right)\int_{-\sqrt{2c}}^{0} y\exp\left(-y\theta\right) Dy, \tag{2.55}$$

$$\simeq \exp\left(-\theta^2/2\right)\frac{2\,c}{\sqrt{2\pi}}\int_{-1}^{0} y\exp\left(-ya\right) dy, \tag{2.56}$$

$$\simeq \frac{\exp\left(-\theta^2/2\right)}{\theta^2\sqrt{2\pi}}\left[e^a\,(1-a)-1\right]. \tag{2.57}$$

De la même façon, on obtient :

$$\int_{-\sqrt{2c}-\theta}^{-\theta}(y+\theta)Dy \simeq \frac{\exp\left(-\theta^2/2\right)}{\theta^2\sqrt{2\pi}}\left[e^{-a}\left(1+a\right)-1\right], \tag{2.58}$$

$$\int_{-\sqrt{2c}+\theta}^{\theta}(y-\theta)^2Dy \simeq \frac{\exp\left(-\theta^2/2\right)}{\theta^3\sqrt{2\pi}}\left[e^{a}\left(2-2a+a^2\right)-2\right], \tag{2.59}$$

$$\int_{-\sqrt{2c}-\theta}^{-\theta}(y+\theta)^2Dy \simeq \frac{\exp\left(-\theta^2/2\right)}{\theta^3\sqrt{2\pi}}\left[2-e^{-a}\left(2+2a+a^2\right)\right]. \tag{2.60}$$

De cela, on en déduit une expression pour ε en fonction de a et une relation entre α,θ,ε et a :

$$\varepsilon = \frac{e^a\left(1-a\right)-1}{e^a\left(1-a\right)+e^{-a}\left(1+a\right)-2}, \tag{2.61}$$

$$\frac{1}{\alpha} = \frac{\exp\left(-\theta^2/2\right)}{\theta^3\sqrt{2\pi}}\left[e^{a}\left(2-2a+a^2\right)-2\right. \tag{2.62}$$

$$\left.-\,\varepsilon\left(e^{a}\left(2-2a+a^2\right)+e^{-a}\left(2+2a+a^2\right)-4\right)\right]. \tag{2.63}$$

Le comportement de ε en fonction de a est cohérent. En effet, ε est une fonction croissante de a. Lorsque $a \to -\infty$ alors $\varepsilon \to 0$, cela correspond à un seuil θ qui diverge lorsque les classes sont toutes identiques ($\varepsilon = 0$). Au contraire, lorsque $a = 0$ alors $\varepsilon = 1/2$ ce qui correspond bien à une absence de seuil $\theta = 0$ pour une distribution symétrique des classes.

Il reste maintenant à déterminer le développement de $\mathcal{E}_t(\alpha,\varepsilon)$. Pour cela, on utilise :

$$\int_{-\sqrt{2c}-\theta}^{+\infty}Dy \simeq -\frac{\exp\left(-\theta^2/2\right)}{\theta\sqrt{2\pi}}\exp(-a), \tag{2.64}$$

$$\int_{\sqrt{2c}-\theta}^{+\infty}Dy \simeq -\frac{\exp\left(-\theta^2/2\right)}{\theta\sqrt{2\pi}}\exp(a). \tag{2.65}$$

L'erreur d'apprentissage $\mathcal{E}_t(\alpha,\varepsilon)$ s'écrit alors :

$$\mathcal{E}_t(\alpha,\varepsilon) \simeq \varepsilon - \frac{\exp\left(-\theta^2/2\right)}{\theta\sqrt{2\pi}}\left((1-\varepsilon)e^{a}-\varepsilon\, e^{-a}\right). \tag{2.66}$$

Nous avons maintenant les ingrédients suffisants pour écrire cette erreur de la façon suivante :

$$\mathcal{E}_t(\alpha,\varepsilon) \simeq \varepsilon - \delta\, h(\varepsilon) \tag{2.67}$$

avec δ une fonction de α uniquement qui s'annule lorsque α diverge et $h(\varepsilon)$ une fonction de ε uniquement.

Il est possible de résumer les résultats obtenus de la façon suivante :

$$\varepsilon = F_1(a), \tag{2.68}$$

$$\frac{1}{\alpha} = \frac{\exp\left(-\theta^2/2\right)}{\theta^3\sqrt{2\pi}}F_2(a), \tag{2.69}$$

$$\mathcal{E}_t(\alpha,\varepsilon) = \varepsilon - \frac{\exp\left(-\theta^2/2\right)}{\theta\sqrt{2\pi}}F_3(a). \tag{2.70}$$

avec :

$$F_1(a) = \frac{e^a(1-a)-1}{2(\mathrm{ch}\, a - a\mathrm{sh}\, a - 1)}, \tag{2.71}$$

$$F_2(a) = \frac{a^2(a-\mathrm{sh}\, a)}{\mathrm{ch}\, a - a\,\mathrm{sh}\, a - 1}, \tag{2.72}$$

$$F_3(a) = \frac{a-\mathrm{sh}\, a}{\mathrm{ch}\, a - a\,\mathrm{sh}\, a - 1}. \tag{2.73}$$

Pour déterminer δ et $h(\varepsilon)$, on exprime θ en fonction de α d'après l'équation (2.69) :

$$\theta^2 \simeq 2\,\ln\alpha + O(\ln(\ln\alpha)). \tag{2.74}$$

Il ne reste plus qu'à utiliser l'équation (2.69) à nouveau pour obtenir :

$$\mathcal{E}_t(\alpha,\varepsilon) \simeq \varepsilon - \frac{\ln\alpha}{\alpha}\,\frac{2F_3(a)}{F_2(a)}. \tag{2.75}$$

Dans cette expression, on a négligé le terme $\ln(\ln\alpha)$ par rapport à $\ln\alpha$. δ et $h(\varepsilon)$ s'expriment de manière simple en fonction de α et a respectivement :

$$\delta = \frac{\ln\alpha}{\alpha}, \tag{2.76}$$

$$h(\varepsilon) = \frac{2}{a^2}. \tag{2.77}$$

Pour obtenir h comme une fonction de ε, il suffit d'inverser $F_1(a)$ pour obtenir $a(\varepsilon)$. Il est possible de calculer l'intégrale de $1/h(\varepsilon)$ pour obtenir un équivalent de k. Pour cela, il suffit de dériver F_1 :

$$F_1'(a) = \frac{a(\mathrm{sh}\, a - a)}{2(\mathrm{ch}\, a - a\,\mathrm{sh}\, a - 1)^2}. \tag{2.78}$$

L'équivalent de k s'écrit alors simplement :

$$k = \delta^{-1}\int_0^{1/2}\frac{d\varepsilon}{h(\varepsilon)}, \tag{2.79}$$

$$= \delta^{-1}\int_{-\infty}^{0}\frac{a^2}{2}F_1'[a]da, \tag{2.80}$$

$$= \delta^{-1}\int_{-\infty}^{0}\frac{a^3(\mathrm{sh}\, a - a)\,da}{4\,(\mathrm{ch}\, a - a\,\mathrm{sh}\, a - 1)^2}, \tag{2.81}$$

$$\simeq 0.475\ \delta^{-1}. \tag{2.82}$$

De ce résultat, on tire le comportement du nombre $k(\alpha)$ de perceptrons nécessaires à l'obtention d'une erreur d'apprentissage nulle ainsi que celui de la capacité $\alpha_c(k)$ pour une machine de parité à k perceptrons apprenant avec l'algorithme incrémental lorsque $\alpha \to +\infty$ et $k \to +\infty$:

$$k(\alpha) \simeq 0.475\,\frac{\alpha}{\ln\alpha}, \tag{2.83}$$

$$\alpha_c(k) \simeq 2.11\,k\ln k. \tag{2.84}$$

Le résultat de la capacité est à comparer à la borne supérieure donnée par Mitchinson et Durbin [85] pour la capacité de la machine de parité indépendamment de l'algorithme utilisé. Cette borne est :

$$\alpha_c(k) \leq k \frac{\ln k}{\ln 2} \simeq 1.44\, k \ln k. \tag{2.85}$$

La borne que nous venons de calculer est supérieure à celle de Mitchinson et Durbin. Toutefois, il ne faut pas oublier que les hypothèses utilisées pour obtenir ce résultat, l'hypothèse de non corrélation des perceptrons et celle de la symétrie des répliques, entraînent une surestimation de la capacité.

Il est intéressant de noter qu'avec les deux hypothèses utilisées nous arrivons à un équivalent dont seul le coefficient semble erroné. Les premiers calculs directs de la capacité d'une machine de parité utilisant la symétrie des répliques ne permettent pas de retrouver cet équivalent en $k \ln k$ mais un équivalent proportionnel à k^3 [7]. Seul un pas de brisure de symétrie des répliques permet de retrouver un comportement qui sature la borne donnée par Mitchinson et Durbin [7, 86, 87, 130].

2.5.2 Fonction de coût quadratique

Afin de nous affranchir du problème de la brisure de symétrie des répliques, nous allons maintenant étudier une autre fonction de coût, définie par le potentiel [6, 49] :

$$V(\lambda) = (\kappa - \lambda)^2\, \Theta(\kappa - \lambda). \tag{2.86}$$

Si l'ensemble d'apprentissage est linéairement séparable, le paramètre κ est la marge, introduit au chapitre précédent. Dans ce chapitre, nous nous intéressons à l'erreur d'apprentissage lorsque l'ensemble n'est pas linéairement séparable. Dans ce cas, comme nous le verrons par la suite, κ joue le rôle d'un paramètre ajustable qui permet de minimiser l'erreur d'apprentissage.

La fonction $\lambda(y,c)$ qui minimise W, donné par (2.42), est :

$$\lambda(y,c) = \begin{cases} \dfrac{y + 2c\kappa}{1 + 2c} & y < \kappa, \\ \\ y & \kappa < y. \end{cases} \tag{2.87}$$

De cette fonction, on déduit la fonction g (2.41), les conditions d'extremum (2.40) permettant de déterminer c et θ, ainsi que l'erreur d'apprentissage $\mathcal{E}_t(\alpha,\varepsilon)$ (2.44) :

$$\begin{aligned} g(\alpha,\varepsilon,c,\theta) &= -\frac{1}{2c} + \frac{\alpha(1-\varepsilon)}{1+2c} \int_{-\infty}^{\kappa+\theta} (\kappa - y + \theta)^2\, Dy \\ &\quad + \frac{\alpha\,\varepsilon}{1+2c} \int_{-\infty}^{\kappa-\theta} (\kappa - y - \theta)^2\, Dy, \end{aligned} \tag{2.88}$$

$$\frac{1}{\alpha}\left[\frac{1+2c}{2c}\right]^2 = (1-\varepsilon)\int_{-\infty}^{\kappa+\theta} (\kappa - y + \theta)^2 Dy + \varepsilon \int_{-\infty}^{\kappa-\theta} (\kappa - y - \theta)^2 Dy, \tag{2.89}$$

$$0 = (1-\varepsilon)\int_{-\infty}^{\kappa+\theta} (\kappa - y + \theta) Dy - \varepsilon \int_{-\infty}^{\kappa-\theta} (\kappa - y - \theta) Dy, \tag{2.90}$$

$$\mathcal{E}_t(\alpha,\varepsilon) = (1-\varepsilon)\int_{2c\kappa-\theta}^{+\infty} Dy + \varepsilon\left(1 - \int_{-2c\kappa-\theta}^{+\infty} Dy\right). \tag{2.91}$$

Regardons tout d'abord la solution obtenue pour $\kappa = 0$. D'après (2.90), on peut remarquer que θ devient une fonction de ε et l'erreur d'apprentissage une fonction de θ d'après (2.91). Dans le cas où c est infini, l'erreur d'apprentissage est nulle. D'après l'équation (2.89), c est infini pour $\alpha < \alpha_0(\varepsilon)$. On retrouve la même capacité $\alpha_0(\varepsilon)$ que dans le cas où la fonction de coût est l'erreur d'apprentissage. Ce résultat

est valable pour toute fonction de coût dont le potentiel V est nul pour $\lambda > 0$ et strictement positif pour $\lambda < 0$.

Au dessus de la capacité $\alpha_0(\varepsilon)$, l'erreur d'apprentissage est indépendante de α et vérifie la propriété suivante :

$$\mathcal{E}_t(\alpha,\varepsilon) = \varepsilon + (1-2\varepsilon)\int_{-\theta}^{+\infty} Dy > \varepsilon. \tag{2.92}$$

L'erreur d'apprentissage pour ε donné est supérieure à cette même valeur. On déduit aisément de (2.92), que $\varepsilon_t = 1/2$ est le point fixe attractif de la suite définie par :

$$\varepsilon_t^i = \mathcal{E}_t(\alpha,\varepsilon_t^{i-1}). \tag{2.93}$$

Ce résultat est contraire à celui de l'étude de la fonction de coût précédente pour laquelle le point fixe attractif est $\varepsilon = 0$, condition nécessaire pour que l'algorithme incrémental puisse converger. Il est nécessaire d'introduire le paramètre κ, qui va permettre de remédier à ce problème.

Regardons maintenant la solution obtenue pour une valeur fixée de κ. D'après l'équation (2.89), c s'annule lorsque α diverge. θ est une fonction de ε et κ par l'intermédiaire de l'équation (2.90). L'erreur d'apprentissage s'écrit alors :

$$\mathcal{E}_t(\alpha,\varepsilon) = \varepsilon + (1-2\varepsilon)\int_{-\theta}^{+\infty} Dy + O(2\,c\,\kappa) > \varepsilon. \tag{2.94}$$

Une fois encore, l'algorithme incrémental ne peut pas converger. Il faut supposer que κ est un paramètre ajustable et notamment qu'il diverge lorsque α diverge.

Nous allons faire les hypothèses suivantes en ce qui concerne le comportement des diverses variables lorsque α diverge :

$$\begin{aligned}
\kappa &\rightarrow +\infty, && (2.95)\\
\theta &\rightarrow -\infty, && (2.96)\\
c &\rightarrow 0, && (2.97)\\
\kappa+\theta &\rightarrow +\infty, && (2.98)\\
2c\kappa &\rightarrow 0. && (2.99)
\end{aligned}$$

Ces hypothèses permettent de réécrire les équations (2.89) et (2.90) de la manière suivante :

$$\begin{aligned}
1 &\simeq \alpha\left(\frac{2c}{1+2c}\right)^2\left((1-\varepsilon)(\kappa+\theta)^2 + \varepsilon(\kappa-\theta)^2 + (1-2\varepsilon)\right), && (2.100)\\
0 &\simeq (1-\varepsilon)(\kappa+\theta) - \varepsilon(\kappa-\theta). && (2.101)
\end{aligned}$$

De la deuxième équation, on obtient $\theta \simeq (1-2\varepsilon)\kappa$ ce qui permet de déterminer c :

$$\begin{aligned}
1 &\simeq \alpha\left(\frac{2c}{1+2c}\right)^2\left(4\,\kappa^2\varepsilon(1-\varepsilon) + (1-2\varepsilon)\right), && (2.102)\\
2c &\simeq \left(\left(4\alpha\kappa^2\varepsilon(1-\varepsilon) + \alpha(1-2\varepsilon)\right)^{1/2} - 1\right)^{-1}, && (2.103)\\
2c &\simeq \frac{1}{\sqrt{4\alpha\kappa^2\varepsilon(1-\varepsilon)}}\left(1 - \frac{1-2\varepsilon}{8\kappa^2\varepsilon(1-\varepsilon)} + O\left(\frac{1}{\kappa^4}\right)\right.\\
&\qquad \left. + \frac{1}{\sqrt{4\alpha\kappa^2\varepsilon(1-\varepsilon)}} + O\left(\frac{1}{\alpha\kappa^2}\right)\right). && (2.104)
\end{aligned}$$

L'erreur d'apprentissage s'écrit :

$$\mathcal{E}_t(\alpha,\varepsilon,\kappa) \simeq \varepsilon - \varepsilon\, H\left(\kappa(1-2\varepsilon) - 2c\kappa\right) + (1-\varepsilon) H\left(\kappa(1-2\varepsilon) + 2c\kappa\right). \tag{2.105}$$

avec :

$$H(x) = \int_x^{+\infty} Dy. \tag{2.106}$$

Puisque l'on recherche l'erreur d'apprentissage la plus faible possible, κ est déterminé par la condition de minimum de l'erreur d'apprentissage :

$$\frac{\partial \mathcal{E}_t}{\partial \kappa} = 0 \tag{2.107}$$

qui s'écrit :

$$\exp\left(4c\kappa^2(1-2\varepsilon)\right) \simeq \frac{1-\varepsilon}{\varepsilon}\left(\frac{1-2\varepsilon+\dfrac{\partial}{\partial\kappa}(2c\kappa)}{1-2\varepsilon-\dfrac{\partial}{\partial\kappa}(2c\kappa)}\right), \tag{2.108}$$

$$4c\kappa^2(1-2\varepsilon) \simeq \ln\left(\frac{1-\varepsilon}{\varepsilon}\right) + O\left(\frac{1}{\kappa^4}\right). \tag{2.109}$$

En remplaçant le paramètre c par son expression (2.104) :

$$\begin{aligned}\kappa \simeq\ & \frac{\sqrt{\alpha\varepsilon(1-\varepsilon)}}{1-2\varepsilon}\ln\left(\frac{1-\varepsilon}{\varepsilon}\right)\left(1 - \frac{1-2\varepsilon}{2\alpha\varepsilon(1-\varepsilon)}\left[\ln\left(\frac{1-\varepsilon}{\varepsilon}\right)\right]^{-1}\right.\\ & \left. + \frac{(1-2\varepsilon)^3}{8\alpha\varepsilon^2(1-\varepsilon)^2}\left[\ln\left(\frac{1-\varepsilon}{\varepsilon}\right)\right]^{-2} + O\left(\frac{1}{\alpha^2}\right)\right).\end{aligned} \tag{2.110}$$

Ayant déterminé κ, il est assez aisé de vérifier que les hypothèses (2.95) à (2.99) sont bien vérifiées. La détermination de κ permet d'exprimer l'erreur d'apprentissage en fonction de α et ε. Le développement de κ au premier ordre en $1/\alpha$ est indispensable puisque $\mathcal{E}_t$ dépend de l'exponentielle de κ^2 :

$$\begin{aligned}\mathcal{E}_t(\alpha,\varepsilon) \simeq\ & \varepsilon - \sqrt{\frac{\varepsilon(1-\varepsilon)}{2\pi}}\exp\left(-\frac{1}{2}\kappa^2(1-2\varepsilon)^2 - 2c^2\kappa^2\right) \\ & \left[(\kappa(1-2\varepsilon)-2c\kappa)^{-1} - (\kappa(1-2\varepsilon)+2c\kappa)^{-1} + O(\kappa^{-4})\right],\end{aligned} \tag{2.111}$$

$$\simeq \varepsilon - \sqrt{\frac{\varepsilon(1-\varepsilon)}{2\pi}}\frac{4c}{\kappa(1-2\varepsilon)^2}\exp\left(-\frac{1}{2}\kappa^2(1-2\varepsilon)^2\right). \tag{2.112}$$

La fonction $h(\alpha,\varepsilon)$ (2.22) s'écrit alors :

$$\begin{aligned}h(\alpha,\varepsilon) =\ & \frac{(1-2\varepsilon)^{-2}}{\alpha\varepsilon\,(1-\varepsilon)\sqrt{2\pi\alpha}}\exp\left(\frac{1-2\varepsilon}{2}\ln\left(\frac{1-\varepsilon}{\varepsilon}\right) - \frac{(1-2\varepsilon)^3}{8\,\varepsilon\,(1-\varepsilon)}\right) \\ & \exp\left(-\frac{\alpha}{2}\,\varepsilon\,(1-\varepsilon)\ln^2\left(\frac{1-\varepsilon}{\varepsilon}\right)\right),\end{aligned} \tag{2.113}$$

$$= \alpha^{-3/2} h_1(\varepsilon)\exp\left(-\frac{\alpha}{2}\,h_2(\varepsilon)\right). \tag{2.114}$$

De cette expression et compte tenu de (2.36), on en déduit le comportement de $k(\alpha)$ pour $\alpha \to +\infty$:

$$k(\alpha) \simeq \int_0^{1/2} \frac{d\varepsilon}{g(\alpha,\varepsilon)}, \tag{2.115}$$

$$\simeq \frac{2\,\alpha}{h_1(\varepsilon_0)} \sqrt{\frac{\alpha\pi}{-h_2''(\varepsilon_0)}} \exp\left(\frac{\alpha}{2}\, h_2(\varepsilon_0)\right). \tag{2.116}$$

L'intégrale est calculée par la méthode du col déjà utilisée précédemment. Le paramètre ε_0 vérifie la condition suivante :

$$h_2'(\varepsilon_0) = 0, \tag{2.117}$$

$$\ln\left(\frac{1-\varepsilon_0}{\varepsilon_0}\right) = \frac{2}{1-2\varepsilon_0}, \tag{2.118}$$

$$\varepsilon_0 \simeq 0.0832. \tag{2.119}$$

La capacité $\alpha_c(k)$ est donnée par :

$$\alpha_c(k) \simeq \frac{2}{h_2(\varepsilon_0)} \ln k, \tag{2.120}$$

$$\simeq 4.55 \ln k. \tag{2.121}$$

Les corrections à cet équivalent sont proportionnelles à $\ln(\ln k)$.

L'équivalent obtenu avec la fonction de coût quadratique est beaucoup plus faible que celui obtenu précédemment en prenant l'erreur d'apprentissage comme fonction de coût. Cette différence importante n'est pas due à l'absence de brisure de symétrie des répliques mais au fait que la fonction de coût n'est pas optimale pour le problème considéré qui consiste à minimiser le plus possible l'erreur d'apprentissage à chaque nouveau perceptron que l'on rajoute.

Des simulations numériques préliminaires de l'algorithme incrémental où l'apprentissage de chaque perceptron simple se fait par minimisation de la fonction de coût quadratique montrent une absence de convergence pour de nombreux ensembles d'apprentissage. Ceci peut s'expliquer par les divergences du seuil θ et du paramètre κ lorsque α diverge. Dans le cas d'un espace des données de dimension N finie, il existe une borne naturelle pour ces deux paramètres qui est de l'ordre de la norme des exemples, c'est-à-dire, $\sqrt{N}$. Au-dessus de cette borne, les exemples se trouvent tous du même côté de l'hyperplan. La solution devient alors triviale et ne permet plus de faire décroître *strictement* l'erreur d'apprentissage à chaque itération. Celle-ci reste constante et strictement positive.

Les simulations ont mis aussi en évidence qu'il existe des corrélations entre les directions des différents perceptrons. Ces corrélations ont été négligées pour le calcul de la capacité de l'algorithme incrémental. Aucun argument simple ne permet de prédire si les corrections dues à la prise en compte de ces corrélations modifient l'équivalent obtenu ou si elles affectent uniquement le préfacteur. De la même façon, les corrections apportées par un calcul de l'erreur d'apprentissage avec brisure de symétrie des répliques n'ont pas été déterminées et n'ont donc pas pu être estimées.

2.6 Conclusion

Nous avons déduit la capacité de la machine de parité pour cet algorithme particulier pour deux fonctions de coût différentes. Pour la fonction de coût correspondant à l'erreur d'apprentissage, le résultat obtenu permet de retrouver l'équivalent

$\alpha_c \sim k \ln k$ dans la limite $k \to +\infty$. Cet équivalent a été obtenu avec l'hypothèse de symétrie des répliques, contrairement au calcul de la capacité générale de la machine de parité, c'est-à-dire, indépendante de l'algorithme utilisé [7] qui nécessite la brisure de symétrie pour être obtenu. L'hypothèse de symétrie des répliques n'est toutefois pas vérifiée pour le calcul que nous avons effectué. De même, l'hypothèse de non corrélation des perceptrons simples entre eux n'est pas vérifiée mais on peut s'attendre, tout de même, à ce que les corrections apportées par la prise en compte de ces deux effets modifient uniquement le préfateur du comportement de la capacité dans la limite d'un grand nombre de perceptrons. Si tel était le cas, l'algorithme incrémental pourrait être considéré comme quasiment optimal dans le sens où il permettrait d'obtenir une capacité proche de la capacité générale de la machine de parité.

Conclusion générale

Dans cette thèse, nous avons présenté plusieurs études des performances de divers problèmes d'apprentissage supervisé ou non supervisé. Ces performances ont été déterminées à l'aide des méthodes de la Physique Statistique, et plus particulièrement, de la méthode des répliques.

Dans la première partie de la thèse, nous avons présenté l'apprentissage non supervisé de la détermination d'une direction privilégiée dans un espace de grande dimension. Nous nous sommes intéressés à la détermination de la direction optimale, c'est-à-dire, la direction la plus proche possible de la direction privilégiée ainsi qu'aux propriétés de cette direction optimale. Une méthode variationnelle nous a permis de déterminer une fonction de coût qui est minimale pour la direction optimale. Ces performances peuvent être déterminées par une approche bayesienne de l'apprentissage, qui permet de les prédire sans toutefois fournir un algorithme simple pour les atteindre. Au cours de cette thèse, nous avons démontré un résultat très général concernant la méthode des répliques : l'hypothèse de symétrie des répliques est cohérente, tant dans l'apprentissage de Gibbs que dans l'approche variationnelle, pour tous les problèmes d'apprentissage se ramenant à celui de la détermination d'une direction privilégiée dans l'espace de données. En effet, nous avons démontré que dans les deux approches, la solutions symétrique par rapport à des permutations des répliques est localement stable. La méthode variationnelle, utilisée dans cette première partie de la thèse, paraît suffisamment générale pour pouvoir être adaptée à d'autres problèmes d'apprentissage plus complexes.

Nous nous sommes intéressés au problème de l'apprentissage supervisé d'une tâche linéairement séparable. Les prédictions théoriques, obtenues avec l'approche variationnelle, ont été confirmées par des simulations numériques. Ces simulations ont permis de mettre en évidence la validité de l'hypothèse de symétrie des répliques mais aussi de déterminer les corrections dues à la taille finie de l'espace des données pour ce problème particulier d'apprentissage. Ces simulations présentent un changement du comportement des effets de taille finie pour une taille réduite de l'ensemble d'apprentissage correspondant à la capacité du perceptron simple. Ce résultat reste inexpliqué et il serait intéressant d'étudier analytiquement ces effets de taille finie afin d'expliquer ce résultat surprenant.

Nous nous sommes aussi intéressés à la détection de la direction reliant les centres de deux amas d'une densité de probabilité. L'approche variationnelle a permis de calculer les performances d'apprentissage optimal et de mettre en évidence l'existence de nombreuses transitions de phases pour ces performances en fonction de la taille réduite de l'ensemble d'apprentissage. L'existence de transitions de phases du premier ordre est un résultat original qui n'avait jusqu'à présent jamais été observé dans des problèmes d'apprentissage non supervisé comme celui-ci. Nous avons présenté le désaccord sur la position de la transition du premier ordre qui apparaît entre l'approche variationnelle et l'approche bayesienne de l'apprentissage optimal. Aucune réponse satisfaisante n'a pu être apportée à cette controverse jus-

qu'à présent.

Dans la deuxième partie de la thèse, nous avons étudié deux approches différentes pour l'apprentissage de tâches plus complexes. Nous nous sommes consacrés à des problèmes supervisés pour lesquels une séparation linéaire ne permet pas la classification de tous les exemples de l'ensemble d'apprentissage.

La première approche développée est l'étude avec les outils de la Physique Statistique des machines à exemples supports. L'étude de la stabilité maximale et du nombre d'exemples supports a permis de mettre en avant la nécessité de contraindre la norme des nouvelles représentations pour conserver des propriétés intéressantes en généralisation. L'introduction d'une notion d'erreur de reconnaissance permet d'envisager l'obtention d'une borne intéressante de l'erreur de généralisation. Toutefois, il faut pour cela utiliser des hypothèses plus réalistes sur la densité de probabilité des exemples. L'hypothèse d'une densité gaussienne semble, en effet, inadaptée. L'étude de densités de probabilité autour de prototypes paraît être une possibilité qu'il conviendrait de développer.

La deuxième approche étudiée concerne l'apprentissage incrémental d'une machine de parité. Nous avons déterminé le nombre de perceptrons simples nécessaires à la convergence de l'algorithme. Nous en avons déduit un résultat intéressant : l'algorithme incrémental permettrait d'atteindre la capacité de la machine de parité. Nos calculs ont été faits avec l'hypothèse de symétrie des répliques et de non corrélation des perceptrons, qui ne sont pas vérifiées. Un calcul avec la prise en compte des corrélations, et avec au moins un pas de brisure reste à faire, mais nous nous attendons à ce que le comportement asymptotique trouvé ne soit pas qualitativement modifié.

Du même auteur

1. E. Buffenoir, A. Coste, J. Lascoux, P. Degiovanni and A. Buhot
   Precise study of some number fields and Galois actions occurring in conformal field theory.
   Ann. Inst. Henri Poincaré 63, 41-79 (1995).
2. A. Buhot and M. B. Gordon
   Cost function and patterns distribution of the bayesian perceptron.
   Physics Letters A 228, 73-78 (1997).
3. A. Buhot, J.-M. Torres Moreno and M. B. Gordon
   Numerical simulations of an optimal algorithm for supervised learning.
   Proceedings ESANN'97, 151-156 (1997).
4. A. Buhot, J.-M. Torres Moreno and M. B. Gordon
   Finite size scaling of the Bayesian perceptron.
   Phys. Rev. E 55, 7434-40 (1997).
5. A. Buhot and M. B. Gordon
   Phase transitions in optimal unsupervised learning.
   Phys. Rev. E 57, 3326-33 (1998).
6. A. Buhot and W. Krauth
   Numerical Solution of Hard-Core Mixtures.
   Phys. Rev. Lett. 80, 3787-90 (1998).
   Phys. Rev. Focus 1, story 11, (1998), http://focus.aps.org/v1/st11.html
7. M. B. Gordon and A. Buhot
   Bayesian learning versus optimal learning.
   Physica A 257, 85-98 (1998).
8. A. Buhot and W. Krauth
   Phase Separation in Two-Dimensional Additive Mixtures.
   Phys. Rev. E 59, 2939-2941 (1999).
9. A. Buhot
   Packing fraction at the phase separation transition in hard-core mixtures.
   Phys. Rev. Lett. 82, 960-963 (1999).
10. A. Buhot and M. B. Gordon
    Statistical mechanics of support vector machines.
    Proceedings ESANN'99, 201-206 (1999).
11. A. Buhot and M. B. Gordon
    Detection of two Gaussion clusters.
    Proceedings ESANN'99, 327-332 (1999).

Bibliographie

[1] L. F. Abbott and T. B. Kepler. Universality in the space of interactions for network models. *J. Phys. A: Math. Gen.*, **22**:2031–2038, (1989).

[2] J. R. L. de Almeida and D. J. Thouless. Stability of the Sherrington-Kirkpatrick solution of a spin glass model. *J. Phys. A: Math. Gen.*, **11**:983–990, (1978).

[3] D. J. Amit. *Modeling Brain Function.* Cambridge University Press, Cambridge, (1989).

[4] D. J. Amit, M. R. Evans, H. Horner, and K. Y. M. Wong. Retrival phase diagrams for attractor neural networks with optimal interactions. *J. Phys. A: Math. Gen.*, **23**:3361–3381, (1990).

[5] D. J. Amit, H. Gutfreund, and H. Sompolinsky. Storage infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.*, **55**:1530–1533, (1985).

[6] J. K. Anlauf and M. Biehl. The AdaTron: An adaptive perceptron algorithm. *Europhys. Lett.*, **10**:687–692, (1989).

[7] E. Barkai, D. Hansel, and I. Kanter. Statistical mechanics of a multilayer neural network. *Phys. Rev. Lett.*, **65**:2312–2315, (1990).

[8] E. Barkai, D. Hansel, and H. Sompolinsky. Broken symetries in multilayered perceptrons. *Phys. Rev. A*, **45**:4146–4161, (1992).

[9] N. Barkai, H. S. Seung, and H. Sompolinsky. Scaling laws in learning of classification tasks. *Phys. Rev. Lett.*, **70**:3167–3170, (1993).

[10] N. Barkai and H. Sompolinsky. Statistical mechanics of the maximum-likelihood density estimation. *Phys. Rev. E*, **50**:1766–1769, (1994).

[11] M. Biehl and A. Mietzner. Statistical Mechanics of unsupervised learning. *Europhys. Lett.*, **24**:421–426, (1993).

[12] M. Biehl and A. Mietzner. Statistical Mechanics of unsupervised structure recognition. *J. Phys. A: Math. Gen.*, **27**:1885–1897, (1994).

[13] M. Biehl and M. Opper. Tilinglike learning in the parity machine. *Phys. Rev. A*, **44**:6888–6894, (1991).

[14] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, (1995).

[15] A. Blandin. Theories versus experiments in the spin glass systems. *J. de Physique : Colloques*, **C6**:1499–1516, (1978).

[16] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In ACM Press, editor, *Fifth Annual Wokshop on Computationnal Learning Theory*, (1992).

[17] L. Bottou and V. N. Vapnik. Local learning algorithm. *Neural Comp.*, **4**:888–900, (1992).

[18] M. Bouten. Replica symmetry instability in perceptron models. *J. Phys. A: Math. Gen.*, **27**:6021–6023, (1994).

[19] M. Bouten, L. Reimers, and B. Van Rompaey. Learning in the hypercube: A stepping stone to the binary perceptron. *Phys. Rev. E*, **58**:2378–2385, (1998).

[20] M. Bouten, J. Schietse, and C. Van den Broeck. Gradient descent learning in perceptrons : A review of its possibilities. *Phys. Rev. E*, **52**:1958–1967, (1995).

[21] A. Buhot and M. B. Gordon. Cost function and pattern distribution of the Bayesian perceptron. *Phys. Lett. A*, **228**:73–78, (1997).

[22] A. Buhot and M. B. Gordon. Phase transitions in optimal unsupervised learning. *Phys. Rev. E*, **57**:3326–3333, (1998).

[23] A. Buhot and M. B. Gordon. Detection of two gaussian clusters. In M. Verleysen, editor, *European Symposium on Artificial Neural Networks*, pages 327–332, (1999).

[24] A. Buhot and M. B. Gordon. Statistical Mechanics of Support Vector Machines. In M. Verleysen, editor, *European Symposium on Artificial Neural Networks*, pages 201–206, (1999).

[25] A. Buhot, J.-M. Torres Moreno, and M. B. Gordon. Finite size scaling of the Bayesian perceptron. *Phys. Rev. E*, **55**:7434–7440, (1997).

[26] A. Buhot, J.-M. Torres Moreno, and M. B. Gordon. Numerical simulations of an optimal algorithm for supervised learning. In M. Verleysen, editor, *European Symposium on Artificial Neural Networks*, pages 151–156, (1997).

[27] C. Cortes and V. N. Vapnik. Support-vector networks. *Machine Learning*, **20**:273–297, (1995).

[28] T. M. Cover. Geometrical and statistical properties of system of linear inequalities with applications in pattern recognition. *IEEE Trans. on Electronic Computers*, **14**:326–334, (1965).

[29] B. Derrida, E. Gardner, and A. Zippelius. An exactly solvable asymetric neural network model. *Europyhs. Lett.*, **4**:167–173, (1987).

[30] B. Derrida, R. B. Griffiths, and A. Prügel-Bennett. Finite-size effects and bounds for perceptron models. *J. Phys. A: Math. Gen.*, **24**:4907–4940, (1991).

[31] R. Dietrich, M. Opper, and H. Sompolinsky. Statistical mechanics of support vectors networks. *Phys. Rev. Lett.*, **82**:2975–2978, (1999).

[32] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, (1973).

[33] S. F. Edwards and P. W. Anderson. Theory of spin glasses. *J. Phys. F: Metal Phys.*, **5**:965–974, (1975).

[34] A. Engel, H. M. Köhler, F. Tschepke, H. Vollmayr, and A. Zippelius. Storage capacity and learning algorithms for two-layer neural networks. *Phys. Rev. A*, **45**:7590–7609, (1992).

[35] R. Erichsen and W. K. Theumann. Optimal storage of a neural network model: A replica symmetry-breaking solution. *J. Phys. A: Math. Gen.*, **26**:L61–L68, (1993).

[36] M. Frean. The Upstart algorithm: A method for constructing and training feedforward neural networks. *Neural Comp.*, **2**:198–209, (1990).

[37] S. I. Gallant. Three constructive algorithms for network learning. In *Proc. 8th Ann. Conf. Science Soc. (Amherst, MA, 15-17 August 1986)*, pages 652–660, (1986).

[38] E. Gardner. Maximum storage capacity in neural networks. *Europhys. Lett.*, **4**:481–485, (1987).

[39] E. Gardner. The phase space of interactions in neural network models. *J. Phys. A: Math. Gen.*, **21**:257–270, (1988).

[40] E. Gardner and B. Derrida. Optimal storage properties of neural network models. *J. Phys. A: Math. Gen.*, **21**:271–284, (1988).

[41] E. Gardner and B. Derrida. Three unfinished works on the optimal storage capacity of networks. *J. Phys. A: Math. Gen.*, **22**:1983–1994, (1989).

[42] J. Phys. A: Math. Gen. En l'honneur d'Elizabeth Gardner. *J. Phys. A: Math. Gen.*, **22**(12), (1989).

[43] F. Gerl and U. Krey. Storage capacity and optimal learning of Potts-model perceptrons by a cavity method. *J. Phys. A: Math. Gen.*, **27**:7353–7372, (1994).

[44] F. Gerl and U. Krey. A Kuhn-Tucker cavity method for generalization with applications to perceptrons with Ising and Potts neurons. *J. Phys. A: Math. Gen.*, **28**:6501–6516, (1995).

[45] F. Gerl and U. Krey. Replica symmetry breaking and the Kuhn-Tucker cavity method in simple and multilayer perceptrons. *J. Phys. I France*, **7**:303–327, (1997).

[46] M. B. Gordon. A convergence theorem for incremental learning with real-valued inputs. In *ICNN'96*, pages 381–386, (1996).

[47] M. B. Gordon and A. Buhot. Bayesian learning versus optimal learning. *Physica A*, **257**:85–98, (1998).

[48] M. B. Gordon and D. R. Grempel. Learning with a temperature dependant algorithm. *Europhys. Lett.*, **29**:257–262, (1995).

[49] M. Griniasty and H. Gutfreund. Learning and retrieval in attractor neural networks above saturation. *J. Phys. A: Math. Gen.*, **24**:715–734, (1991).

[50] I. M. Guyon, B. E. Boser, and V. N. Vapnik. Automatic capacity tuning of very large VC-dimension classifiers. In *Advances in Neural Information Processing Systems*, pages 147–155, (1993).

[51] G. Györgyi and P. Reimann. Parisi phase in a neuron. *Phys. Rev. Lett.*, **79**:2746–2749, (1997).

[52] G. Györgyi and N. Tishby. *Statistical theory of learning a rule*, pages 3–36. Theumann, W. K. and Köberle R., Neural networks and spin glasses, World Scientific edition, (1990).

[53] D. Hansel and H. Sompolinsky. Learning from examples in single-layer neural networks. *Europhys. Lett.*, **11**:687–692, (1990).

[54] D. O. Hebb. *The organization of behavior.* Wiley, New York, (1949).

[55] M. Heerema and W. A. van Leeuwen. Derivation of Hebb's rules. *J. Phys. A: Math. Gen.*, **32**:263–287, (1999).

[56] D. Herschkowitz and J.-P. Nadal. Unsupervised and supervised clustering: The mutual information between parameters and observations. In D. A. Cohn M. S. Kearns, S. A. Solla, editor, *Advances in Neural Information Processing Systems 11*, (1999).

[57] D. Herschkowitz and J.-P. Nadal. Unsupervised and supervised learning: Mutual information between parameters and observations. *Phys. Rev. E*, **59**:3344–3360, (1999).

[58] J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the theory of neural computation.* Addison-Wesley Publishing Company, Redwood City, CA, USA, (1991).

[59] J. J. Hopfield. Neural networks and physical systems with emergent collective computational properties. *Proc. Natl. Acad. Sci. USA*, **79**:2554–2588, (1982).

[60] T. B. Kepler and L. F. Abbott. Domains of attraction in neural networks. *J. Phys. France*, **49**:1657–1662, (1988).

[61] O. Kinouchi. *Aprendizagem ótima em percéptrons a partir de exemplos com ruido.* PhD thesis, Universidad de São Paulo, Instituto de Física, (1996).

[62] O. Kinouchi and N. Caticha. Learning algorithm that gives the Bayes generalization limit for perceptrons. *Phys. Rev. E*, **54**:R54–R57, (1996).

[63] S. Kirkpatrick and D. Sherrington. Infinite-ranged models of spin-glasses. *Phys. Rev. B*, **17**:4384–4403, (1978).

[64] H. Köhler, S. Diederich, W. Kinzel, and M. Opper. Learning algorithm for a neural network with binary synapses. *Z. Phys. B*, **78**:331–342, (1990).

[65] W. Krauth and M. Mézard. Storage capacity of memory networks with binary couplings. *J. Phys. France*, **50**:3057–3066, (1989).

[66] W. Krauth, M. Mézard, and J.-P. Nadal. Bassins of attraction in a Perceptron-like Neural Network. *Complex Systems*, **2**:387–408, (1988).

[67] W. Krauth and M. Opper. Critical storage capacity of the $J = \pm 1$ neural network. *J. Phys. A: Math. Gen.*, **22**:L519–523, (1989).

[68] C. Kwon and J.-H. Oh. Exact asymptotic estimates of the storage capacities of the committee machines with overlapping and non-overlapping receptive fields. In M. Verleysen, editor, *European Symposium on Artificial Neural Networks*, pages 157–162, (1997).

[69] E. Lootens and C. Van den Broeck. Analysing cluster formation by replica method. *Europhys. Lett.*, **30**:381–386, (1995).

[70] P. Majer, A. Engel, and A. Zippelius. Perceptrons above saturation. *J. Phys. A: Math. Gen.*, **26**:7405–7416, (1993).

[71] C. Marangi, M. Biehl, and S. A. Solla. Supervised learning from clustered input examples. *Europhys. Lett.*, **30**:117–122, (1995).

[72] M. Marchand, M. Golea, and P. Ruján. A convergence theorem for sequential learning in two-layer perceptrons. *Europhys. Lett.*, **11**:487–492, (1989).

[73] D. Martinez and D. Estève. The offset algorithm: Building and learning method for multilayer neural networks. *Europhys. Lett.*, **18**:95–100, (1992).

[74] N. Matić, I. M. Guyon, J. Denker, and V. N. Vapnik. Writer-adaptation for on-line handwritten character recognition. In IEEE Computer, editor, *In Second International Conference on Pattern Recognition and Document Analysis*, pages 187–191, (1993).

[75] W. S. McCulloch and W. Pitts. A logical calculus of ideas immanent in nervous activity. *Bull. Math. Biophys.*, **5**:115, (1943).

[76] R. Meir. Empirical risk minimization versus maximum-likelihood estimation: A case study. *Neural Computation*, **7**:144–157, (1995).

[77] R. Meir and J. F. Fontanari. Calculation of learning curves for inconsistent algorithms. *Phys. Rev. A*, **45**:8874–8884, (1992).

[78] R. Meir and J. F. Fontanari. Learning from examples in weight-constrained neural networks. *J. Phys. A: Math. Gen.*, **25**:1149–1168, (1992).

[79] S. Mertens and A. Engel. Vapnik-Chervonenkis dimension of neural networks with binary weights. *Phys. Rev. E*, **55**:4478–4488, (1997).

[80] M. Mézard and J.-P. Nadal. Learning in feedforward layered networks: The tiling algorithm. *J. Phys. A: Math. Gen.*, **22**:2191–2203, (1989).

[81] M. Mézard, G. Parisi, and M. A. Virasoro. *Spin glass theory and beyond*. World Scientific, Singapore, (1987).

[82] M. Mézard and S. Patarnello. On the capacity of feedforward and layered networks. *Unpublished*, , (1989).

[83] G. Milde and S. Kobe. An exact learning algorithm for autoassociative neural networks with binary couplings. *J. Phys. A: Math. Gen.*, **30**:2349–2352, (1997).

[84] M. L. Minsky and S. A. Papert. *Perceptrons*. MIT Press, Cambridge, MA, USA, (1969).

[85] G. J. Mitchison and R. M. Durbin. Bounds on the learning capacity of some multi-layer networks. *Biol. Cyber.*, **60**:345, (1989).

[86] R. Monasson and R. Zecchina. Weigth space structure and internal representations: A direct approach to learning and generalization in multilayer neural networks. *Phys. Rev. Lett.*, **75**:2432–2435, (1995).

[87] R. Monasson and R. Zecchina. Weigth space structure and internal representations: A direct approach to learning and generalization in multilayer neural networks. *Erratum Phys. Rev. Lett.*, **76**:2205, (1996).

[88] J.-P. Nadal. Study of a growth algorithm for a feedforward neural network. *Int. Journ. of Neur. Syst.*, **1**:55–59, (1989).

[89] W. Nadler and W. Fink. Finite size scaling in neural networks. *Phys. Rev. Lett.*, **78**:555–558, (1997).

[90] E. Oja. A simplified neuron model as a principal component analyser. *J. Math. Biology*, **15**:267–273, (1982).

[91] E. Oja. Neural networks, principal components, and subspaces. *Int. Journ. of Neur. Syst.*, **1**:61–68, (1989).

[92] M. Opper and D. Haussler. Generalization performance of Bayes optimal classification algorithm for learning a perceptron. *Phys. Rev. Lett.*, **66**:2677–2680, (1991).

[93] M. Opper and W. Kinzel. Statistical mechanics of generalisation. In J. L. van Hemmen E. Domany and K. Schulten, editors, *Models of Neural Networks III*, pages 151–209, (1996).

[94] M. Opper, W. Kinzel, J. Kleinz, and R. Nehl. On the ability of the optimal perceptron to generalise. *J. Phys. A: Math. Gen.*, **23**:L581–L586, (1990).

[95] G. Parisi. Toward a mean field theory for spin-glasses. *Phys. Rev. Lett.*, **73A**:203–205, (1979).

[96] G. Parisi. Magnetic properties of spin glasses in a new mean field theory. *J. Phys. A: Math. Gen.*, **13**:1887–1895, (1980).

[97] G. Parisi. The order parameter for spin glasses: A function on the interval $0-1$. *J. Phys. A: Math. Gen.*, **13**:1101–1112, (1980).

[98] G. Parisi. A sequence of approximated solutions to the S-K model for spin glasses. *J. Phys. A: Math. Gen.*, **13**:L115–L121, (1980).

[99] P. Peretto. *An Introduction to the Modeling of Neural Networks*. Cambridge University Press, Cambridge (UK), collection aléa-saclay edition, (1992).

[100] B. Raffin and M. B. Gordon. Learning and generalization with minimerror, a temperature dependent learning algorithm. *Neural Comp.*, **7**:1182, (1995).

[101] P. Reimann and C. Van den Broeck. Learning by examples from a nonuniform distribution. *Phys. Rev. E*, **53**:3989–3998, (1996).

[102] P. Reimann, C. Van den Broeck, and G. J. Bex. A Gaussian scenario for unsupervised learning. *J. Phys. A: Math. Gen.*, **29**:3521–3535, (1996).

[103] K. Rose, E. Gurewitz, and G. C. Fox. Statistical mechanics and phase transitions in clustering. *Phys. Rev. Lett.*, **65**:945–948, (1990).

[104] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Phys. Rev.*, **65**:386, (1958).

[105] F. Rosenblatt. *Principles of Neurodynamics - Perceptrons and the Theory of Brain*. Spartan Books, Washington D. C., (1961).

[106] P. Ruján and M. Marchand. Learning by activating neurons: A new approach to learning in neural networks. *Complex Systems*, **3**:229, (1989).

[107] D. E. Rumelhart and J. L. McClelland. *Parallel Distributed Processing*. Cambridge, MA: Bradford, (1986).

[108] J. Schietse, M. Bouten, and C. Van den Broeck. Training binary perceptrons by clipping. *Europhys. Lett.*, **32**:279–284, (1995).

[109] M. Schröder and R. Urbanczik. Comment on "Finite size scaling in neural networks". *Phys. Rev. Lett.*, **80**:4109, (1998).

[110] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Phys. Rev. A*, **45**:6056–6091, (1992).

[111] D. Sherrington and S. Kirkpatrick. Solvable model of spin-glass. *Phys. Rev. Lett.*, **35**:1792–1796, (1975).

[112] J.-C. Simon. La reconnaissance des formes à l'épreuve des faits. *La Recherche*, **312**:58–62, (1998).

[113] J. A. Sirat and J.-P. Nadal. Neural trees: A new tool for classification. *Network*, **1**:423–438, (1990).

[114] H. Sompolinsky, N. Tishby, and H. S. Seung. Learning from examples in large neural networks. *Phys. Rev. Lett.*, **65**:1683–1686, (1990).

[115] J.-M. Torres Moreno. *Apprentissage et généralisation par des réseaux de neurones : étude de nouveaux algorithmes constructifs*. PhD thesis, Institut National Polytechnique de Grenoble, (1997).

[116] J. M. Torres Moreno and M. B. Gordon. Efficient adaptive learning for classification tasks with binary units. *Neural Computation*, **10**:1007–1030, (1998).

[117] J. M. Torres Moreno, P. Peretto, and M. B. Gordon. An evolutive architecture coupled with optimal perceptron learning for classification. In M. Verleysen, editor, *European Symposium on Artificial Neural Networks*, pages 365–370, (1995).

[118] R. Urbanczik. Storage capacity of the fully-connected committee machine. *J. Phys. A: Math. Gen.*, **30**:L387–L392, (1997).

[119] F. Vallet. The Hebb rule for learning linearly separable boolean functions: Learning and generalization. *Europhys. Lett.*, **8**:747–751, (1989).

[120] C. Van den Broeck and P. Reimann. Unsupervised learning by examples: On-line versus off-line. *Phys. Rev. Lett.*, **76**:2188–2191, (1996).

[121] B. Van Rompaey. Precursor networks for training the binary perceptron. In M. Verleysen, editor, *European Symposium on Artificial Neural Networks*, pages 175–180, (1997).

[122] V. N. Vapnik. *The nature of statistical learning theory*. Springer Verlag, New York, (1995).

[123] V. N. Vapnik. *Statistical learning theory*. John Wiley & Sons, New York, (1998).

[124] T. L. H. Watkin. Optimal learning with a neural network. *Europhys. Lett.*, **21**:871–876, (1993).

[125] T. L. H. Watkin and J.-P. Nadal. Optimal unsupervised learning. *J. Phys. A: Math. Gen.*, **27**:1899–1915, (1994).

[126] T. L. H. Watkin, A. Rau, and M. Biehl. The statistical mechanics of learning a rule. *Rev. Mod. Phys.*, **65**:499–543, (1993).

[127] T. L. H. Watkin, A. Rau, D. Bollé, and J. van Mourik. Learning multi-class classification problems. *J. Phys. I France*, **2**:167–180, (1992).

[128] A. H. L. West and D. Saad. The statistical mechanics of constructive algorithms. *J. Phys. A: Math. Gen.*, **31**:8977–9021, (1998).

[129] W. Whyte and D. Sherrington. Replica-symmetry breaking in perceptrons. *J. Phys. A: Math. Gen.*, **29**:3063–3073, (1996).

[130] J.-H. Xiong, C. Kwon, and J.-H. Oh. Storage capacity of a fully-connected parity machine with continuous weights. *J. Phys. A: Math. Gen.*, **31**:7043–7049, (1998).

[131] J.-H. Xiong, J.-H. Oh, and C. Kwon. Weight space structure and the storage capacity of a fully connected committee machine. *Phys. Rev. E*, **56**:4540–4544, (1997).

[132] H. Yoon and J.-H. Oh. Learning of higher-order perceptrons with tunable complexities. *J. Phys. A: Math. Gen.*, **31**:7771–7784, (1998).

ABSTRACT

In this thesis, we study different properties of learning from examples with Statistical Mechanics tools and, particularly, with the replica trick. Supervised tasks, corresponding to a binary classification of data, and unsupervised tasks like the parametric estimation of a probability density function, are considered.

In the first part, a variational approach allows us to determine the optimal learning performance in the problem of learning an anisotropy direction, and to deduce a cost function which allows to obtain such optimal performance. In the case of the supervised learning of a linearly separable task, numerical simulations, that confirm our theoretical results, allow us to determine finite size effects. In the case of a probability density function composed of a mixture of two Gaussians, the optimal learning performance presents several phase transitions as a function of the size of the data set. These results raise a controversy between the variational theory and the Bayesian approach of the optimal learning.

In the second part, we study two different approaches used to learn complex classification tasks. We first consider the one of support vector machines. We study a family of such machines for which linear and quadratic separations are particular cases. The capacity, the typical value of the margin and the number of support vectors, are determined. The second approach is the one of a parity machine trained with an incremental learning algorithm. This algorithm constructs progressively a neural network with one hidden layer. The capacity of this algorithm is found to be close to the capacity of the parity machine.

Key words

Neural Networks

Perceptron

Learning

Bayesian Inference

Statistical Mechanics

Replica Trick

Phase Transitions

Disordered Systems

RÉSUMÉ

L'objet de cette thèse est l'étude de diverses propriétés d'apprentissage à partir d'exemples par des méthodes de Physique Statistique, notamment, par la méthode des répliques. Des tâches supervisées, correspondant à la classification binaire de données, ainsi que des tâches non supervisées, comme l'estimation paramétrique d'une densité de probabilité, sont considérées.

Dans la première partie, une approche variationnelle permet de déterminer la performance de l'apprentissage optimal d'une direction d'anisotropie, et de déduire une fonction de coût permettant d'obtenir ces performances optimales. Dans le cas de l'apprentissage supervisé d'une tâche linéairement séparable, des simulations numériques confirmant nos résultats théoriques ont permis de déterminer les effets de taille finie. Dans le cas d'une densité de probabilité constituée de deux gaussiennes, la performance de l'apprentissage optimal présente de nombreuses transitions de phases en fonction du nombre de données. Ces résultats soulèvent une controverse entre la théorie variationnelle et l'approche bayesienne de l'apprentissage optimal.

Dans la deuxième partie, nous étudions deux approches différentes de l'apprentissage de tâches de classification complexes. La première approche considérée est celle des machines à exemples supports. Nous avons étudié une famille de ces machines pour laquelle les séparateurs linéaire et quadratique sont deux cas particuliers. La capacité, les valeurs typiques de la marge et du nombre d'exemples supports, sont déterminées. La deuxième approche considérée est celle d'une machine de parité apprenant avec un algorithme incrémental. Cet algorithme construit progressivement un réseau de neurones à une couche cachée. La capacité théorique obtenue pour l'algorithme considéré est proche de celle de la machine de parité.

Mots clefs

Réseaux de Neurones	Physique Statistique
Perceptron	Méthode des Répliques
Apprentissage	Transitions de Phases
Inférence Bayesienne	Systèmes Désordonnés