MINES ParisTech

Travail d'option MAREVA
Mathématiques appliquées
(Robotique, Vision, Automatique)

ENS Paris-Saclay

Stage M2 recherche MVA
Mathématiques, Vision,
Apprentissage

Performance d'algorithmes d'apprentissage en grande dimension : entre matrices aléatoires et information

Hugues Souchard de Lavoreille

Travail supervisé par Romain Couillet & Steeve Zozor

15 septembre 2020

gipsa-lab

MINES ParisTech

PSL

école normale supérieure paris-saclay

Résumé

Lors de ce stage de recherche, nous nous sommes intéressés au problème de l'apprentissage dans des espaces de grande dimension. Ces espaces ont plusieurs particularités dues à leur géométrie peu intuitive, notamment le phénomène de concentration de la mesure, qui rend les algorithmes d'apprentissage statistique classiques moins efficaces dès que les données sont de taille non-négligeable, ce qui arrive déjà en pratique. Des travaux récents issus de deux domaines différents - la théorie des matrices aléatoires et la théorie de l'information - suggèrent que la première est capable de fournir des algorithmes quasi-optimaux dans ce cadre. Est-il possible d'atteindre l'optimalité et si ce n'est pas le cas, pourquoi ? Pour apporter un début de réponse à cette question, nous avons calculé le classifieur bayésien sur un mélange de gaussiennes de grande dimension, avant d'analyser la performance associée, d'abord dans le cas supervisé avant d'aborder le cas semi-supervisé qui est par essence plus difficile et qui nécessite le recours à des outils complexes issus de la physique statistique que nous avons commencé à passer en revue.

Summary

During this research internship, we tackled the problem of high-dimensional machine learning. High-dimensional spaces have several particularities due to their counter-intuitive geometrical properties, such as the concentration of measure phenomenon. This phenomenon makes classical statistical learning algorithmes less efficient as soon as the data has a non-negligible size, which already happens in practical applications. Recent works in two distinct research areas - random matrix theory and information theory - proved that the first one may bring quasi-optimal algorithms in this setting. May one reach optimality and, if the answer is no, why ? In order to shed some light on this question, we computed the bayesian classifier on a high-dimensional gaussian mixture, before analysing the related classification performance, firstly in a supervised setting before switching to the semi-supervised case. The latter is more complex by nature and it requires using advanced tools from the domain of statistical physics, which we finally started to review.

Remerciements

En cette difficile période d'épidémie de Covid-19, je tiens à remercier tous ceux qui ont permis de rendre la tenue de ce stage possible. Qui aurait pu croire qu'une semaine seulement après la signature de la convention de stage, le 17 mars 2020, le confinement généralisé serait décrété sur l'ensemble du territoire français, avec fermeture des frontières, annulation ou report de stages en cascade et crise économique ? Les impacts psychologiques de cette période seront probablement profonds même s'ils restent encore difficilement connus ou quantifiables. Deux mois de confinement ont conduit à une adoption massive du télétravail et au recentrage de nos vies sur la famille et les proches en même temps qu'à la rupture d'une bonne partie de nos relations sociales professionnelles. Je l'ai ressenti personnellement en devant accomplir l'intégralité des cinq mois de ce stage depuis la chambre de mon enfance, chez mes parents, à 600 kilomètres de l'endroit où il aurait dû se dérouler normalement. J'aimerais donc d'abord remercier ma famille de m'avoir accueilli à nouveau pendant ces quelques mois.

Si je connaissais déjà Romain grâce à l'excellent cours qu'il donne dans le cadre du master MVA à CentraleSupelec, je n'ai découvert mon deuxième encadrant Steeve que début juillet au cours d'un bref passage à Grenoble pour rencontrer enfin l'équipe. Nous n'avions échangé auparavant que par courriel et par audio à cause de la mauvaise qualité de nos connexions internet respectives. Ces conditions ont donné à ce stage une coloration étrange. J'ai pu à la fois bénéficier d'une très grande liberté tout en apprenant à connaître la routine et la monotonie du travail en solitaire dans une seule pièce. Si très vite, l'optimisme des débuts s'est envolé, Romain et Steeve ont toujours été là, malgré des conditions de travail parfois précaires, pour me redonner goût à mon stage en répondant aux questions parfois presque métaphysiques d'un élève en quête de sens. Leur soutien m'a été d'une grande aide et je leur en suis extrêmement reconnaissant.

Je tiens enfin à remercier infiniment Cosme qui m'a accueilli à bras ouvert lors de mes deux courts séjours à Grenoble alors que nous ne nous connaissions pas auparavant. Merci de m'avoir montré le quotidien d'un doctorant épanoui dans son travail, de m'avoir patiemment introduit le sujet de la concentration et de m'avoir partagé ton optimisme sans faille. J'en ai été et reste impressionné, ce d'autant plus depuis que je sais que la période de confinement a été destructrice pour nombre de jeunes chercheurs.

Je remercie enfin une dernière fois Steeve qui a relu en détail ce document.

Ce travail a été soutenu en partie par MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

Introduction

Le travail présenté dans ce manuscrit clôt trois années de scolarité à l'École des Mines de Paris dont la dernière a été effectuée en parallèle avec le master MVA (Mathématiques, Vision, Apprentissage) de l'École Normale Supérieure Paris-Saclay. Ce double cursus donne une coloration résolument mathématique à mon parcours d'ingénieur, car deux expériences de stage effectuées en année de césure m'ont fait ressentir le besoin de pousser plus loin ma formation généraliste, de façon à pouvoir m'adapter aux évolutions techniques qui se produiront nécessairement dans le futur. Je suis en effet persuadé que les réponses aux enjeux de demain, qu'ils soient environnementaux, techniques ou sociaux, passeront par une compréhension fine des phénomènes, par une modélisation adaptée et surtout par une analyse précise et approfondie des données associées. Alors même que les données ont pris, en quelques années à peine, une importance jusque-là inimaginable dans tous les secteurs d'activité, leur interprétation et leur traitement n'en sont pas pour autant toujours devenus aisés. Leur caractère massif nécessite le recours à des algorithmes optimisés et leur grande dimension rend leur visualisation difficile. Les mathématiques permettent d'analyser rigoureusement ces algorithmes et fournissent un point de vue renouvelé sur les données, afin de dépasser les intuitions initiales qui conduisent parfois à des erreurs d'interprétation. Elles permettent par exemple de réduire efficacement la dimension des données ou encore de construire des espaces dont la géométrie est adaptée aux données et phénomènes étudiés.

Un cours m'a particulièrement marqué cette année, celui donné par Jamal Najim et Romain Couillet dans le cadre du master MVA. Tous deux sont spécialistes de la théorie des grandes matrices aléatoires, qui est née grâce à des travaux initialement situés à l'intersection de la physique théorique et des mathématiques, lorsque le physicien hongrois Eugene Wigner puis les mathématiciens ukrainiens Marchenko et Pastur se sont intéressés au spectre de grandes matrices aléatoires hermitiennes. De telles matrices apparaissent naturellement en physique, pour étudier le spectre de noyaux lourds ou d'opérateurs hamiltoniens. Les observations ayant mené à cette théorie sont au premier abord très déroutantes : si l'on considère une suite de matrices aléatoires carrées dont la taille grandit et tend vers l'infini, la distribution des valeurs propres converge [20]. Autrement dit, plus la matrice est grande et complexe, plus le comportement de son spectre se régularise ! Évidemment, la distribution limite dépend de la structure de la suite de matrices aléatoires mais l'idée est là : certains grands systèmes complexes se simplifient asymptotiquement, de même qu'en vertu du théorème central limite, une somme de variables aléatoires indépendantes ressemble asymptotiquement en distribution à une gaussienne, ou encore de même qu'un gaz constitué de très nombreuses particules et ayant un nombre exponentiel de micro-états possède, à l'échelle macroscopique, un comportement régulier.

Les communautés du traitement du signal et de l'apprentissage automatique se sont progressivement approprié les avancées de la théorie des matrices aléatoires pour les appliquer aux systèmes complexes que forment les données massives, en prenant pour piste l'idée que des données en grand nombre et en grande dimension pourront conséquemment se simplifier. Si la grande dimension semble ouvrir un océan de possibilités, elle n'en demeure pas moins un domaine très difficile, notamment à cause du fait que l'intuition s'accoutume très peu à ces espaces. Les algorithmes qui ont rendu populaire l'intelligence artificielle ces dernières années se basent sur un raisonnement statistique qui prend pour postulat que la dimension des données est finie et donc petite. L'équipe de Romain Couillet, tuteur de mon stage, a cependant récemment montré que ces algorithmes ne fonctionnent pas comme prévu dès que la dimension des données est d'ordre supérieur à 10 ou 100 selon les contextes, ce qui est déjà souvent le cas en pratique.

Mon travail s'est donc situé exactement à mi-chemin entre le côté sombre et le côté clair de la

grande dimension, qui est à la fois une malédiction et une bénédiction. C'est avec le double-objectif de pouvoir développer mon esprit critique sur les algorithmes opérant sur des données de grande dimension et de contribuer à un domaine à l'intersection entre mathématiques, physique et informatique que je me suis décidé à effectuer ce travail d'option et de stage M2 au laboratoire Gipsa-lab de Grenoble, sous la direction des professeurs Romain Couillet, professeur à CentraleSupelec et porteur d'une chaire à l'Université Grenoble-Alpes, et Steeve Zozor, Directeur de recherche au CNRS et spécialiste de la théorie de l'information.

J'ai donc pris part aux recherches de l'équipe supervisée par Romain Couillet, hébergée par le pôle GAIA (Géométrie, Apprentissage, Information, Algorithmes) de Gipsa-lab. Le laboratoire Gipsa-lab est une unité mixte de recherche du CNRS et de Grenoble INP, qui compte en tout environ 350 membres dont 150 chercheurs permanents et 150 doctorants. Il est spécialisé et reconnu en automatique, traitement du signal et des images ainsi qu'en parole et en cognition. Les trois autres pôles (aux côté de GAIA) du laboratoire à compter de 2020 sont ainsi intitulés Automatique et Diagnostic, Sciences des Données, ainsi que Parole et Cognition. L'équipe à laquelle j'ai été associé se structure autour d'une chaire 3IA créée en 2019 et intitulée "LargeDATA". Sous l'impulsion du rapport de C. Villani, quatre instituts d'intelligence artificielle (3IA) ont été créés en 2019 dont l'un, appelé MIAI (Multidisciplinary Institute in Artificial Intelligence) est hébergé par l'Université Grenoble-Alpes et regroupe une trentaine de chaires. Celle de mon tuteur compte une trentaine de membres et collabore avec des acteurs variés, notamment le CEA et les entreprises ST-Microelectronics et Huawei. J'ai d'ailleurs eu l'occasion de participer à une réunion de travail avec Huawei début juillet, pendant laquelle j'ai pu présenter mon travail et assister à une présentation scientifique de l'entreprise, qui s'intéresse à la théorie des matrices aléatoires du fait du lien historique qu'a le domaine avec le secteur des télécommunications.

Table des matières

Notations

On notera les scalaires sans attribut particulier, les vecteurs en gras souligné et les matrices en gras souligné deux fois. Ceci vaut en particulier pour la matrice identité en dimension D :

$$a \in \mathbb{R} \qquad \underline{\boldsymbol{a}} \in \mathbb{R}^2 \qquad \underline{\underline{\boldsymbol{a}}} \in \mathbb{R}^{2\times 2} \qquad \underline{\underline{\boldsymbol{id}}}_D \in \mathbb{R}^{D\times D}$$

Dans la suite, $\mathbb{P}$ désignera la plupart du temps une mesure de probabilité et p une densité de probabilité. Par facilité, on adoptera parfois abusivement la notation continue à la place des mesures de probabilité appropriées dans le cas discret. On notera en général les variables aléatoires en majuscule et les réalisations associées (ou valeurs déterministes) en minuscule.

$$\mathbb{P}\left(A = a \,\middle|\, \underline{\underline{\boldsymbol{B}}} = \underline{\boldsymbol{b}}\right)$$

Lorsqu'il n'y a pas d'ambiguïté, on note les probabilités sous forme raccourcie

$$\mathbb{P}\left(\underline{\boldsymbol{a}} \mid \underline{\boldsymbol{b}}\right) \triangleq \mathbb{P}\left(\underline{\boldsymbol{A}} = \underline{\boldsymbol{a}} \mid \underline{\boldsymbol{B}} = \underline{\boldsymbol{b}}\right)$$

On utilise dans ce document les notations suivantes pour (respectivement) une densité de loi normale ainsi qu'une densité de loi normale multivariée en dimension D :

$$\mathcal{N}(x;\mu,\sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\mathcal{N}_D(\underline{x};\underline{\mu},\underline{\underline{\sigma}}) \triangleq \frac{1}{\sqrt{(2\pi)^D \left|\underline{\underline{\sigma}}\right|}} \exp\left(-\frac{1}{2}(\underline{x}-\underline{\mu})^T \underline{\underline{\sigma}}^{-1}(\underline{x}-\underline{\mu})\right)$$

On notera aussi $\mathcal{N}(\mu,\sigma^2)$ et $\mathcal{N}_D(\underline{\mu},\underline{\underline{\sigma}})$ les lois correspondantes.
On appelle $\Phi : \mathbb{R} \to]0,1[$ la fonction queue de gaussienne définie comme étant le complémentaire de la fonction de la répartition d'une loi normale centrée réduite, ou encore comme suit :

$$\Phi(x) = \int_x^{+\infty} \mathcal{N}(t;0,1)\,\mathrm{d}t$$

Pour $\alpha > 0$ fini, on notera $N \doteq_\alpha D$ le régime asymptotique

$$N \doteq_\alpha D \Leftrightarrow \begin{cases} N \longrightarrow +\infty \\ D \longrightarrow +\infty \\ \frac{N}{D} \longrightarrow \alpha \end{cases}$$

Le stage a conduit à l'écriture de morceaux de code en Python. Ceux-ci sont disponibles sur un dépôt gitlab.com. Dans la suite, toute référence à un fichier Python (généralement dans une police spéciale, de la forme `S00_script.py`) sera implicitement à relier à ce dépôt, disponible à l'adresse suivante. La liste complète de ces scripts est disponible en annexe.

`https://gitlab.com/h_sdl/rmt-bounds`

Chapitre 1

De l'apprentissage en grande dimension à la théorie de l'information des systèmes complexes

Résumé du chapitre

Les espaces de grande dimension possèdent une géométrie complexe peu intuitive du fait de l'abondance d'espace. Cette caractéristique affecte profondément les algorithmes d'apprentissage classiques lorsque ceux-ci opèrent sur des données de grande taille. Les travaux de Xiaoyi Mai et Romain Couillet ont récemment mis en évidence un algorithme efficace pour résoudre le problème de l'apprentissage semi-supervisé en grande dimension, et une publication de 2019 a montré que celui-ci était en fait quasi-optimal grâce à des arguments issus de la théorie de l'information et de la physique statistique. Le but de ce stage est d'investiguer l'optimalité de cet algorithme sous un angle mathématique et de comprendre le décalage entre le classifieur bayésien et celui issu de l'algorithme proposé.

Dans ce chapitre, nous allons replacer le stage dans son contexte pour comprendre ses motivations. Nous partirons des observations de l'équipe de Romain Couillet, qui a supervisé mon stage, sur le comportement des algorithmes d'apprentissage dans les espaces de grande dimension, et les relierons à des découvertes récentes par des équipes travaillant à l'interface entre physique statistique et théorie de l'information.

1.1 Concentration de la mesure dans les espaces de grande dimension

Les espaces de grande dimension sont résolument étranges pour le mathématicien novice. Si l'on se les imagine souvent en raisonnant sur des espaces visualisables (2 ou 3 dimensions) avant de généraliser, il se trouve que les intuitions ainsi construites sont souvent en réalité fausses ou tout du moins imprécises. Prenons une suite de vecteurs aléatoires $(X_D)_D$ où $X_D \sim \mathcal{N}_D(0, \underline{\underline{\boldsymbol{id}}}_D)$. Autrement dit, X_D suit une loi normale D-variée avec covariance identité. Quand D augmente, l'espace ambiant dans lequel est tiré X_D augmente en dimension. On est en général habitué à imaginer une distribution gaussienne comme sur l'image de gauche de la figure 1.1.

Observons maintenant le carré de la norme de X_D et rappelons que toutes les coordonnées sont indépendantes, ce qui permet d'approcher la distribution de notre observation par une loi normale (même si on connaît sa loi de façon exacte : il s'agit d'une loi χ^2 à D degrés de liberté) :

$$\|X_D\|^2 = X_{D,1}^2 + ... + X_{D,D}^2 \approx \mathcal{N}(D, 2D)$$

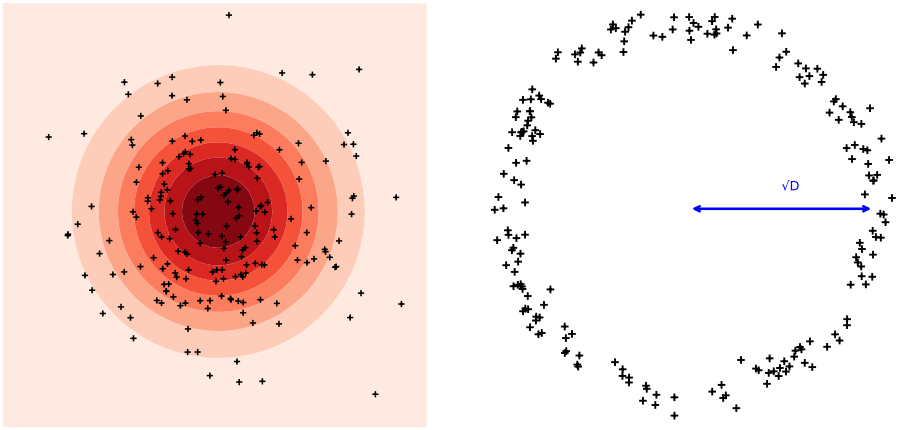

FIGURE 1.1 – Représentation d'une gaussienne $\mathcal{N}_D(0, \underline{\underline{\boldsymbol{id}}}_D)$ en petite dimension (à gauche) et en grande dimension (à droite). En petite dimension, les échantillons se placent majoritairement près de l'espérance de la gaussienne, tandis qu'en grande dimension, elles se placent majoritairement sur une pellicule sphérique de rayon $\sqrt{D}$

La moyenne de $\|X_D\|^2$ varie en D alors que l'écart-type varie en $\sqrt{D}$: quand D grandit, l'intervalle de fluctuation typique de $\|X_D\|^2$ devient négligeable par rapport à l'espérance (pour $D = 10000$, $\|X_D\|^2$ se situe avec grande probabilité dans $[9900, 10100]$). Là où en petite dimension, la norme ne compte que la contribution de quelques coordonnées qui peuvent beaucoup fluctuer, en grande dimension, l'addition de très nombreuses contributions permet de stabiliser la norme. On observe donc informellement que $\|X_D\|$ vaut presque toujours $\sqrt{D}$ (et on pourra le vérifer expérimentalement avec le script `S01_norm_high_dimensional_gaussian.py`). Ceci s'écrit formellement avec une majoration de la probabilité de grande déviation du type suivant (avec $C, c > 0$ deux constantes) :

$$\forall t > 0,\ \mathbb{P}\left(\left|\|X_D\| - \sqrt{n}\right| \geq t\right) \leq C \exp\left(-ct^2\right)$$

Ceci signifie que, même si cela paraît très déroutant compte tenu de nos intuitions en petite dimension, l'écrasante majorité de la masse d'une distribution gaussienne en grande dimension est située sur une pellicule sphérique autour de son centre et que le centre de la gaussienne est très peu probable, même s'il s'agit bien du point qui maximise la densité ! Il faut s'imaginer qu'il y a tellement d'espace en grande dimension que la pellicule sphérique devient énorme et que sa masse cumulée s'approche de 1 sous une distribution gaussienne. On peut s'imaginer informellement un comportement ressemblant à l'image de droite de la figure 1.1.

Il faut comprendre deux phénomènes dans les espaces de grande dimension : d'abord qu'il y a beaucoup d'espace. Deux points pris dans un espace de grande dimension ont toutes les chances de se situer très loin les uns des autres (au sens par exemple de la norme $\|\cdot\|_2$) car la distance est la somme des contributions des distances dans toutes les dimensions. Ensuite que le régime $N \fallingdotseq_\alpha D$ ($N \to +\infty, D \to +\infty, \frac{N}{D} \to \alpha$, voir la partie notations) n'est pas aussi anodin qu'il n'y paraît. Dans un tel régime où N compte le nombre de points et D compte la dimension de chaque point, il n'y aura *jamais* assez de points par rapport à la taille de l'espace. Par exemple, on ne pourra jamais recouvrir la sphère $\mathbb{S}^{D-1}$ avec N points dans le régime $N \fallingdotseq_5 D$. En dimension 2, $\mathbb{S}^1$ est le cercle unité et l'on dispose de $2 \times 5 = 10$ points, ce qui est peu pour recouvrir un segment courbe de longueur 2π. Comme $\alpha = 5$ est fixé, si l'on augmente N avec l'espoir de pouvoir mieux recouvrir la sphère, la dimension augmente aussi en même temps et l'on nous demande de recouvrir la sphère unité $\mathbb{S}^2$ avec 15 points, ce qui est bien trop peu encore une fois. Même pour N grand, on est toujours à court de données pour échantillonner correctement la sphère, ce qui fait ressortir le bruit des données (i.e. que les objets que les données sont censées approcher le seront toujours imprécisément et qu'il est impossible d'affiner en ajoutant des points).

1.2 Apprentissage automatique en grande dimension

Comme nous venons de le voir, la grande dimension fournit un cadre géométrique assez déroutant dans lequel les algorithmes d'apprentissage statistique classiques fonctionnent en réalité mal ou parfois ne fonctionnent pas. Avant d'en venir à l'apprentissage semi-supervisé, qui a constitué le cadre de ce stage, nous allons prendre deux exemples issus des travaux récents. Notons que dans tous ces travaux, le régime $N \asymp_\alpha D$ prend expérimentalement vite sens, ce même pour des valeurs de N et D relativement faibles, de l'ordre de 10 à 100.

Estimation de matrices de covariance. L'estimation de matrices de covariance à partir de données est une tâche centrale en apprentissage car elle permet de reconstruire les modèles probabilistes sous-jacents pour effectuer des tâches de prévision ou de classification. Par exemple, prenons $\underline{\underline{\sigma}} \in \mathbb{R}^{D\times D}$ ainsi que $\underline{x_1}, ..., \underline{x_N}$ des vecteurs tirés selon $\mathcal{N}_D\left(\underline{0}, \underline{\underline{\sigma}}\right)$. L'estimateur classique, obtenu par maximum de vraisemblance, est donné par

$$\widehat{\underline{\underline{\sigma}}} \triangleq \frac{1}{N}\sum_{i=1}^{N} \underline{x_i}\,\underline{x_i}^T$$

Si cet estimateur est consistant dans le cadre statistique classique où l'on suppose D fixé et où $N \longrightarrow +\infty$, ce n'est en fait plus le cas dans le régime où $N \asymp_\alpha D$ où la norme $\|\widehat{\underline{\underline{\sigma}}} - \underline{\underline{\sigma}}\|$ ne tend plus vers 0 car la distribution limite des valeurs propres fait intervenir par exemple la distribution de Marchenko-Pastur [11]. Une étude poussée de l'estimateur de la matrice de covariance en grande dimension est menée par X. Mestre dans [12], où des estimateurs consistants en grande dimension sont également explicités.

Partitionnement spectral. Considérons des points $\underline{x_1}, ..., \underline{x_N}$ dans un espace de grande dimension D et supposons qu'ils se partitionnent en deux classes équiprobables (ils ont été générés autour de deux centres suivant une loi normale). Sous quelques hypothèses raisonnables sur les moyennes et les covariances, il existe $\tau > 0$ tel que presque sûrement,

$$\max_{1\leq i\neq j\leq N} \left| \frac{1}{D}\left\|\underline{x_i} - \underline{x_j}\right\|^2 - \tau \right| \underset{N\asymp_\alpha D}{\longrightarrow} 0$$

Ceci signifie que les distances deux à deux convergent toutes vers une même valeur τ quelles que soient les classes d'appartenance des points. La distance entre deux points du même cluster sera asymptotiquement semblable à la distance entre deux points de clusters différents, ce qui est contraire à ce à quoi l'on s'attend en petite dimension ! Ceci fragilise les algorithmes classiques d'apprentissage qui se basent sur la similarité des échantillons pour apprendre les relations entre les données, en particulier le partitionnement spectral. Couillet et Benaych-Georges analysent dans [1] le comportement de cet algorithme en grande dimension : si sa performance s'amenuise à cause du fait que les distances se ressemblent, il est possible de pousser les calculs et de faire réapparaître les centres des classes dans des termes d'ordre supérieur dans des développements asymptotiques.

Apprentissage semi-supervisé Comme on le voit, et c'est un point central des études de la chaire "LargeDATA" dans lequel mon travail s'est inscrit, les algorithmes d'apprentissage se comportent de manière peu intuitive en grande dimension mais il est possible de les corriger pour améliorer de façon conséquente leurs performances. Durant son travail de doctorat avec Romain Couillet, Xiaoyi Mai s'est en particulier intéressée à l'apprentissage semi-supervisé en grande dimension. Contrairement à l'apprentissage supervisé, où l'on donne en entrée à l'algorithme les données et les étiquettes correspondantes (c'est-à-dire la bonne réponse pour chaque donnée), le paradigme de l'apprentissage semi-supervisé consiste à nourrir l'algorithme à la fois de données étiquetées et de données non-étiquetées (on ne connaît pas leurs étiquettes). L'objectif est que celui-ci se serve des données non étiquetées, en particulier de leur géométrie et de leur structure, pour améliorer son apprentissage. Comme les données étiquetées coûtent cher et sont longues à produire, ce paradigme d'appren-

tissage a le potentiel d'améliorer conséquemment les performances des algorithmes à peu de frais, à condition de confirmer la capacité à apprendre des données sans étiquette.

Dans [9], Mai et Couillet ont montré que les algorithmes classiques d'apprentissage semi-supervisé avaient de très mauvaises performances en grande dimension, et que la présence de données non-étiquetées dégradait parfois même les performances de ces algorithmes (par rapport à l'utilisation de la portion des données qui était étiquetée uniquement) ! Plus récemment, les mêmes auteurs ont proposé dans [10] une méthode intitulée « similarités centrées » permettant de tirer parti efficacement des données non-étiquetées en grande dimension, grâce à la théorie des matrices aléatoires. Cet article est l'un des points de départ de mon travail.

1.3 Limites fondamentales, physique et information

En parallèle du travail de Xiaoyi Mai que nous venons de citer, la communauté des chercheurs travaillant au croisement de la physique, de la théorie de l'information et de l'informatique a mis au point une méthode d'analyse des performances d'algorithmes d'apprentissage automatique basée sur des outils issus de la physique statistique. En 2019, Lelarge et Miolane ont calculé dans [7] le risque bayésien minimal pour un problème de classification semi-supervisée sur des mélanges de gaussiennes en grande dimension. Le risque est une mesure de la performance d'un algorithme de classification : plus celui-ci est faible, plus l'algorithme est performant. Les résultats de ce travail peuvent être directement transposés au cadre utilisé par Mai et Couillet et ces derniers ont calculé le risque obtenu avec l'algorithme des similarités centrées pour le comparer au risque bayésien optimal obtenu par Lelarge et Miolane, ce qui donne lieu au graphe de la figure 1.2 (que l'on pourra reproduire avec le script `S02_comparison_xmrc_mclm.py`) qui représente le risque en fonction de η, la proportion de données étiquetées. On observe que les deux courbes coïncident pour $\eta = 1$ mais se décollent légèrement ailleurs.

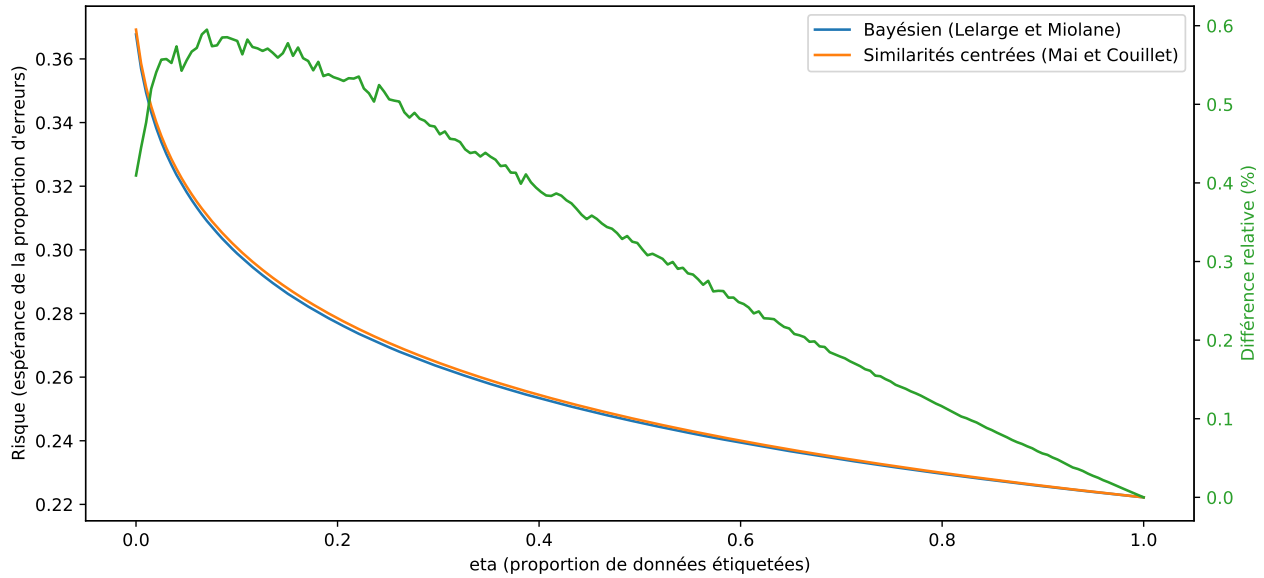

FIGURE 1.2 – Risque bayésien théorique et risque théorique obtenu avec les similarités centrées (ainsi que leur différence relative), en fonction de η, la proportion de données étiquetées.

La comparaison des courbes de ce graphe donne des résultats très encourageants mais interroge : on y voit que le risque de l'algorithme développé à l'aide de la théorie des matrices aléatoires tutoie le risque bayésien optimal au point de coïncider lorsqu'il n'y a que des données étiquetées. Ailleurs, les deux courbes, bien que proches, ne coïncident pas. Comment expliquer cette différence et comment l'exploiter pour améliorer les algorithmes d'apprentissage, si une telle amélioration est possible ? Telle était la problématique ambitieuse de mon stage à mon arrivée en avril 2020.

Les travaux ayant mené au résultat de Lelarge et Miolane proviennent de communautés scientifiques historiquement plus centrées sur la théorie de l'information et la physique, et sont généralement menés avec des outils plus ou moins éloignés de ceux utilisés en statistiques et en apprentissage. Un outil très particulier est fréquemment utilisé dans ces communautés : la méthode des répliques. Il s'agit d'une méthode de calcul développée dans les années 1980 par Marc Mézard, Giorgio Parisi et Miguel Virasoro pour l'étude des verres de spins. Elle permet, au prix de passages non-rigoureux, d'obtenir des expressions simples et élégantes pour les fonctions de partition et les énergies libres de systèmes complexes. On en trouvera une présentation au chapitre 8 de [14]. Hormis certains cas particuliers que nous verrons dans la dernière partie de ce manuscrit, les résultats obtenus par cette méthode n'ont jamais été prouvés mathématiquement ; ils sont même faux dans certains cas. En règle générale, les résultats sont conjecturés grâce à la méthodes des répliques, avant d'être validés ou non par des méthodes mathématiquement rigoureuses.

La méthode des répliques peut être appliquée à la théorie de l'information et permettre de conjecturer des résultats sur les algorithmes de traitement du signal et d'apprentissage. Déjà en 2005, Guo et Verdú ont appliqué la méthode dans [4] pour déduire des propriétés sur la performance d'algorithmes de détection multi-utilisateur dans des systèmes CDMA (Code Division Multiple Access), utilisés pour les réseaux mobiles. Plus récemment, Miolane a prouvé la méthode des répliques dans le cas particulier de l'estimation de matrices de petit rang bruitées [13], ce qui permet d'en déduire des propriétés sur les performances optimales des algorithmes d'estimation associés. En 2019, une autre équipe, américaine cette fois-ci, a prouvé dans [17] l'exactitude des résultats obtenus par la méthode des répliques dans le cas de l'estimation d'une entrée vectorielle transformée de façon linéaire par une matrice à entrées gaussiennes et bruitée.

1.4 Organisation du stage et contributions

Ce stage s'est ainsi situé à la croisée entre géométrie en grande dimension et théorie de l'information que nous avons essayé de combiner pour les appliquer à l'apprentissage semi-supervisé. Assez rapidement, nous nous sommes rendu compte que le fossé entre les travaux des physiciens-théoriciens de l'information et des statisticiens n'était pas négligeable tant les outils et méthodes utilisés sont différents. Pour bien comprendre les résultats fondamentaux de Lelarge et Miolane [7], nous avons donc décidé de procéder par étapes pour comprendre d'où viennent les phénomènes de simplification et de concentration. Nous avons commencé par poser rigoureusement les quantités d'intérêt (risque et classifieur bayésiens) pour le modèle semi-supervisé utilisé par Lelarge et Miolane dans le chapitre 2. Une fois les objets définis et les formules générales du classifieur de Bayes rappelées, nous avons mené à bien dans le chapitre 3 une analyse complète de la performance dans le cas supervisé qui est, par sa nature non-combinatoire, parfaitement maîtrisé. Nous nous sommes ensuite appuyés sur cette analyse pour tenter de généraliser l'étude au cas semi-supervisé au chapitre 4, sans utiliser la méthode des répliques et sans toutefois parvenir au bout. Ceci nous a permis de comprendre les difficultés et l'intérêt de la méthode employée par les équipes mentionnées à la partie 1.3. Au chapitre 5, nous avons effectué une lecture comparative de deux articles énonçant des résultats basés sur la méthode des répliques afin d'en saisir les points communs et d'en déduire les idées sous-jacentes principales.

Chapitre 2

Classifieur et risque bayésiens pour des données gaussiennes

Résumé du chapitre

Le modèle étudié lors de ce stage est un mélange simple de deux gaussiennes. Pour chaque point, on souhaite faire retrouver automatiquement à un algorithme quelle gaussienne l'a généré. On évalue la performance d'un classifieur par le risque, qui est l'espérance de la proportion des mauvaises classifications. On définit le classifieur bayésien, qui est le classifieur minimisant le risque, puis le risque bayésien qui est la valeur minimale du risque. L'objectif de l'étude est d'analyser le comportement du risque bayésien sur notre modèle, dans le régime $N \backsimeq_\alpha D$.

Dans ce chapitre, nous allons poser le modèle qui servira de cadre à l'étude des chapitres ultérieurs, tout en rappelant au besoin des notions de statistiques bayésiennes qui seront utiles pour mesurer la performance des algorithmes étudiés.

2.1 Classifieur et risque bayésiens pour l'apprentissage supervisé

Durant tout ce stage, nous nous sommes intéressés au problème de la classification binaire. L'algorithmicien observe les caractéristiques de deux classes d'objets différentes et son but est de deviner la classe de l'objet seulement à travers une ou plusieurs observations. L'exemple classique est la reconnaissance d'un chien ou d'un chat sur une image en couleurs. Dans ce cas, on possède seulement une image (par exemple 128×128 pixels sur 3 canaux donc 49 152 observations) et l'on souhaite déterminer de façon procédurale de quel animal il s'agit. Dans le cas de la classification supervisée ou semi-supervisée, l'algorithme à concevoir peut s'aider d'exemples, c'est-à-dire que l'on donne à l'algorithme des exemples d'images de chiens et de chats avec la solution (s'il s'agit effectivement d'un chat ou d'un chien). Le travail de l'algorithmicien est alors principalement un travail de statisticien : il faut construire un estimateur de la quantité recherchée (la classe de l'objet) qui est une fonction uniquement des observations (dans notre cas les 49 152 valeurs des pixels) et des exemples. La détermination de cet estimateur est appelée la phase d'apprentissage.

Formalisons les notions que nous venons d'introduire. On se donne $\mathcal{V} = \{-1, +1\}$ l'ensemble des classes, que l'on appelle aussi étiquettes (ou *labels*), par exemple -1 pour les chats et $+1$ pour les chiens. On se donne ensuite $\mathcal{Y}$ l'ensemble des observations possibles : dans notre exemple des images $128 \times 128 \times 3$ donc $\mathcal{Y} \subset \{0, ..., 255\}^{128\times128\times3}$. Pour entraîner l'algorithme (lors de la phase d'apprentissage), on lui donne un échantillon constitué de N exemples $\underline{\underline{\boldsymbol{y}}} = (\underline{\boldsymbol{y}_1}, ..., \underline{\boldsymbol{y}_N}) \in \mathcal{Y}^N$ avec leurs étiquettes $\underline{v} = (v_1, ..., v_N) \in \mathcal{V}^N$. Le but d'un algorithme de classification binaire est de fournir de manière automatique une fonction

$$f_{\underline{\underline{\boldsymbol{y}}},\underline{v}} : \mathcal{Y} \mapsto \mathcal{V}$$

qui soit optimale en un certain critère qui mesure l'erreur de classification. Il existe plusieurs manières de mesurer cette erreur de classification : notamment le risque empirique (qui se fonde uniquement sur le jeu de données d'apprentissage) ou le risque théorique (si l'on connaît par exemple le modèle probabiliste qui génère les données).
Le risque empirique est par exemple donné par

$$\epsilon_{\underline{\underline{y}},\underline{v}}(f) = \sum_{i=1}^{N} \mathbb{1}_{f(\underline{y}_i)\neq v_i}$$

tandis que le risque théorique a posteriori s'écrit (avec $\underline{\boldsymbol{Y}}_+, V_+$ un nouvel échantillon et une nouvelle étiquette générés indépendemment du jeu de données d'apprentissage et de la même manière)

$$\ell_{\underline{\underline{y}},\underline{v}}(f) = \mathbb{P}\left(f(\underline{\boldsymbol{Y}}_+) \neq V_+ \,\middle|\, \underline{\underline{\boldsymbol{Y}}} = \underline{\underline{y}}, \underline{\boldsymbol{V}} = \underline{v}\right) \tag{2.1}$$

Dans l'étude que nous effectuons ici, le jeu de données est aléatoire, tiré selon une loi connue. La fonction $f_{\underline{\underline{y}},\underline{v}}$ est donc aussi aléatoire et l'on cherche à minimiser la probabilité qu'une observation soit mal prévue par la fonction $f_{\underline{\underline{y}},\underline{v}}$. Ceci permet de définir le classifieur bayésien :

Définition 2.1.1: Classifieur bayésien

$$f^*_{\underline{\underline{y}},\underline{v}} = \operatorname*{argmin}_f \ell_{\underline{\underline{y}},\underline{v}}(f) \tag{2.2}$$

La fonction $f^*_{\underline{\underline{\boldsymbol{Y}}},\underline{\boldsymbol{V}}}$ dépend du jeu de données, qui est aléatoire. En moyennant le risque théorique a posteriori obtenu avec le classifieur bayésien, on définit le risque bayésien.

Définition 2.1.2: Risque bayésien

On appelle risque bayésien le risque minimal, qui est obtenu avec le classifieur bayésien. Comme minimiser l'espérance suivante revient à minimiser l'intégrande pour chaque valeur du jeu de données d'apprentissage, on peut l'écrire :

$$R^* = \min_f \mathbb{P}\left(V_+ \neq f(\underline{\boldsymbol{Y}}_+)\right) = \min_f \mathbb{E}\left[\ell_{\underline{\underline{\boldsymbol{Y}}},\underline{\boldsymbol{V}}}(f)\right] = \mathbb{E}\left[\ell_{\underline{\underline{\boldsymbol{Y}}},\underline{\boldsymbol{V}}}\left(f^*_{\underline{\underline{\boldsymbol{Y}}},\underline{\boldsymbol{V}}}\right)\right] \tag{2.3}$$

2.2 Mélange de données gaussiennes étudié

Dans la suite de ce document, on s'intéressera au modèle suivant, qui est également celui considéré par Lelarge et Miolane dans [7]. L'originalité de l'étude que nous menons ici ne réside pas dans le modèle génératif utilisé, qui est très simple et probablement l'un des mieux connus en statistiques : nous étudions ce modèle dans le contexte particulier de l'**apprentissage semi-supervisé** en **grande dimension**.

- On fixe 3 constantes strictement positives α, η, σ ainsi que $N, D > 0$ (le modèle a vocation à être étudié dans le régime $N \asymp_\alpha D$)
- On tire un vecteur aléatoire $\underline{\boldsymbol{U}} \in \mathbb{R}^D$ selon une loi $p_{\underline{\boldsymbol{U}}}$ soit uniforme sur la sphère $\mathbb{S}^{D-1}$ soit selon une loi normale $\mathcal{N}_D\left(\underline{\mathbf{0}}, \frac{1}{D}\underline{\underline{\boldsymbol{id}}}_D\right)$. Ce vecteur permet de définir une direction dans l'espace.
- On tire N variables aléatoires de Rademacher $\underline{\boldsymbol{V}} = (V_1, ..., V_N)$ indépendantes, indépendemment de $\underline{\boldsymbol{U}}$. Ces variables représentent la classe ± 1 de chacun des N points du jeu de données d'apprentissage.
- Indépendemment de $\underline{\boldsymbol{U}}$ et de $\underline{\boldsymbol{V}}$, on tire $N \times D$ variables aléatoires $\underline{\underline{\boldsymbol{Z}}} = (\underline{\boldsymbol{Z}}_1, ..., \underline{\boldsymbol{Z}}_N)$ indépendantes et identiquement distribuées selon une loi normale $\mathcal{N}_D\left(\underline{\mathbf{0}}, \underline{\underline{\boldsymbol{id}}}_D\right)$. Ces variables aléatoires bruitent les points $V_1\underline{\boldsymbol{U}}, ..., V_N\underline{\boldsymbol{U}}$.

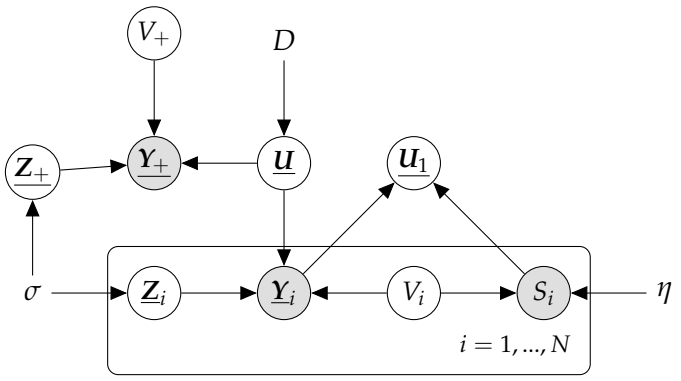

FIGURE 2.1 – Réseau bayésien représentant le modèle étudié

— Indépendemment de $\underline{\boldsymbol{U}}, \underline{V}$ et $\underline{\underline{Z}}$, on tire N variables aléatoires $\underline{B} = (B_1, ..., B_N)$ indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre η. Ces variables aléatoires servent à déterminer, pour chaque point, si son étiquette est connue ($B_i = 1$, avec une probabilité η) ou non.
— On note enfin $\underline{\underline{Y}} = (\underline{Y}_1, ..., \underline{Y}_N) = \underline{\boldsymbol{U}}\underline{V}^T + \sigma\underline{\underline{Z}}$ et $\underline{S} = \underline{B} \otimes \underline{V}$ où $\otimes$ représente le produit terme à terme.

On peut représenter le tout de manière succinte avec le modèle graphique de la figure 2.1, sur lequel on a également introduit, un point de test $\underline{Y}_+$ et son étiquette V_+, générés selon le même processus que les points d'apprentissage ; ainsi que $\underline{\boldsymbol{U}}_1$, une variable aléatoire tirée selon la distribution a posteriori $p(\underline{\boldsymbol{U}}|\underline{\underline{Y}}, \underline{S})$ (pour la notation, se reporter au paragraphe sur les répliques, à la section 5.1).

Intuitivement, on peut s'imaginer deux gaussiennes centrées en $+\underline{\boldsymbol{U}}$ et $-\underline{\boldsymbol{U}}$ avec à peu près le même nombre de points, comme à la figure 2.2. L'objectif est alors, pour un nouveau point, de déterminer à partir de quelle gaussienne il a été généré. Dans le cas de l'apprentissage semi-supervisé, on cherche à construire le meilleur estimateur $f(\underline{Y}_+)$ de l'étiquette V_+ d'un point $\underline{Y}_+$ uniquement à partir des observations $\underline{\underline{Y}}$ et $\underline{S}$. Ce problème est en général difficile mais les résultats de Lelarge et Miolane suggèrent qu'il suffit d'estimer la moyenne a posteriori $\mathbb{E}\left[\underline{\boldsymbol{U}} \mid \underline{\underline{Y}}, \underline{S}\right]$ (aussi estimateur d'erreur quadratique moyenne minimale pour $\underline{\boldsymbol{U}}$). En utilisant cet estimateur a posteriori à la place de $\underline{\boldsymbol{U}}$, on peut obtenir un classifieur optimal simple au sens du risque théorique (2.1).

Notons que sauf mention explicite du contraire, on considèrera ici (contrairement à Lelarge et Miolane) que $\underline{\boldsymbol{U}} \sim \mathcal{N}_D\left(0, \frac{1}{D}\underline{\underline{\boldsymbol{id}}}_D\right)$ au lieu d'une distribution uniforme sur la sphère unité. Les deux se ressemblent asymptotiquement mais il n'est pas prouvé que les deux distributions mènent au même résultat asymptotique.

2.3 Classifieur et risque bayésiens du modèle étudié

Dans cette partie, nous allons commencer par rappeler des résultats classiques d'apprentissage statistique appliqués à notre problème avant de calculer le classifieur et le risque de Bayes définis aux équations (2.2) et (2.3). Fixons un jeu de données d'apprentissage $(\underline{\underline{\boldsymbol{y}}}, \underline{\underline{\boldsymbol{s}}}) \in \mathbb{R}^{N \times D} \times \{-1, 0, +1\}^N$ et notons que l'on peut remplacer les V par des S dans les formules de la section 2.1. Un résultat classique de calcul bayésien (se reporter à [5] si besoin pour les bases de l'estimation statistique) montre que la règle de décision de Bayes suivante donne le classifieur optimal pour le risque $\ell_{\underline{\underline{\boldsymbol{y}}}, \underline{\underline{\boldsymbol{s}}}}$ défini précédemment à l'équation (2.1).

$$f^*_{\underline{\underline{\boldsymbol{y}}}, \underline{\underline{\boldsymbol{s}}}}(\underline{\boldsymbol{y}}_+) = \begin{cases} +1 & \text{si } \mathbb{P}(V_+ = +1 | \underline{Y}_+ = \underline{\boldsymbol{y}}_+, \underline{\underline{Y}} = \underline{\underline{\boldsymbol{y}}}, \underline{S} = \underline{\boldsymbol{s}}) > \mathbb{P}(V_+ = -1 | \underline{Y}_+ = \underline{\boldsymbol{y}}_+, \underline{\underline{Y}} = \underline{\underline{\boldsymbol{y}}}, \underline{S} = \underline{\boldsymbol{s}}) \\ -1 & \text{sinon} \end{cases}$$

$$= \text{signe}\left(\log\left(\frac{\mathbb{P}(V_+ = +1 | \underline{Y}_+ = \underline{\boldsymbol{y}}_+, \underline{\underline{Y}} = \underline{\underline{\boldsymbol{y}}}, \underline{S} = \underline{\boldsymbol{s}})}{\mathbb{P}(V_+ = -1 | \underline{Y}_+ = \underline{\boldsymbol{y}}_+, \underline{\underline{Y}} = \underline{\underline{\boldsymbol{y}}}, \underline{S} = \underline{\boldsymbol{s}})}\right)\right)$$

FIGURE 2.2 – Mélange de deux gaussiennes de covariance scalaire en dimension 2. Les points verts correspondent aux points sans étiquette, étudiés dans le cas général semi-supervisé.

Dans notre cas particulier, on peut encore réécrire, en utilisant l'indépendance de V_+ avec $\underline{\underline{\boldsymbol{Y}}}$ et $\underline{\underline{S}}$, puis que $\mathbb{P}(V_+ = +1) = \mathbb{P}(V_+ = -1) = \frac{1}{2}$

$$f^*_{\underline{\underline{y}},\underline{\underline{s}}}(\underline{y_+}) = \text{signe}\left(\log\left(\frac{p(\underline{y_+}|V_+ = +1, \underline{\underline{\boldsymbol{Y}}} = \underline{\underline{y}}, \underline{\underline{S}} = \underline{\underline{s}})}{p(\underline{y_+}|V_+ = -1, \underline{\underline{\boldsymbol{Y}}} = \underline{\underline{y}}, \underline{\underline{S}} = \underline{\underline{s}})}\right)\right) \tag{2.4}$$

En utilisant la formule des probabilités totales avec la variable $\underline{\boldsymbol{U}}$ puis en marginalisant sur $\underline{\boldsymbol{Y}_+}$ on a encore (il s'agit d'une reformulation de la première équation de la page 8 de [7]) :

$$\begin{aligned} f^*_{\underline{\underline{y}},\underline{\underline{s}}}(\underline{y_+}) &= \text{signe}\left(\log\left(\frac{\int p(\underline{y_+}|V_+ = +1, \underline{\boldsymbol{U}} = \underline{u})\, \mathrm{d}p(\underline{u}|\underline{\underline{\boldsymbol{Y}}} = \underline{\underline{y}}, \underline{\underline{S}} = \underline{\underline{s}})}{\int p(\underline{y_+}|V_+ = -1, \underline{\boldsymbol{U}} = \underline{u})\, \mathrm{d}p(\underline{u}|\underline{\underline{\boldsymbol{Y}}} = \underline{\underline{y}}, \underline{\underline{S}} = \underline{\underline{s}})}\right)\right) \\ &= \text{signe}\left(\log\left(\frac{\int \exp\left(-\frac{1}{2\sigma^2}\left\|\underline{y_+} - \underline{u}\right\|^2\right) \mathrm{d}p(\underline{u}|\underline{\underline{\boldsymbol{Y}}} = \underline{\underline{y}}, \underline{\underline{S}} = \underline{\underline{s}})}{\int \exp\left(-\frac{1}{2\sigma^2}\left\|\underline{y_+} + \underline{u}\right\|^2\right) \mathrm{d}p(\underline{u}|\underline{\underline{\boldsymbol{Y}}} = \underline{\underline{y}}, \underline{\underline{S}} = \underline{\underline{s}})}\right)\right) \end{aligned} \tag{2.5}$$

Il est en fait remarquable que la solution de ce problème par nature très combinatoire se simplifie asymptotiquement de façon conséquente. L'article [7] montre en effet que dans le régime $N \asymp_\alpha D$, les équations (2.5) et (2.3) se simplifient (de façon informelle) ainsi (il suffit asymptotiquement de remplacer $\underline{\boldsymbol{U}}$ par son espérance conditionnellement au jeu de données d'apprentissage puis de vérifier que le point à tester est dans le même sens) :

$$f^*_{\underline{\underline{y}},\underline{\underline{s}}}(\underline{y_+}) \approx \text{signe}\left(\underline{y_+}^T \mathbb{E}\left[\underline{\boldsymbol{U}} \,\middle|\, \underline{\underline{\boldsymbol{Y}}} = \underline{\underline{y}}, \underline{\underline{S}} = \underline{\underline{s}}\right]\right) \tag{2.5'}$$

$$R^* \approx 1 - \Phi\left(\frac{\sqrt{q^*}}{\sigma}\right) \tag{2.3'}$$

où q^* est la solution d'une équation à point fixe dépendant uniquement de α, σ, η.

Finalement, le but de notre travail consiste à comprendre pour quelles raisons on peut en quelque sorte remplacer asymptotiquement

$$\int \exp\left(-\frac{1}{2\sigma^2}\left\|\underline{\boldsymbol{y}_+} - v_+\underline{\boldsymbol{u}}\right\|^2\right) \mathrm{d}p(\underline{\boldsymbol{u}}|\underline{\underline{\boldsymbol{y}}}, \underline{\boldsymbol{s}})$$

par

$$\exp\left(-\frac{1}{2\sigma^2}\left\|\underline{\boldsymbol{y}_+} - v_+ \int \underline{\boldsymbol{u}}\, \mathrm{d}p(\underline{\boldsymbol{u}}|\underline{\underline{\boldsymbol{y}}}, \underline{\boldsymbol{s}})\right\|^2\right)$$

puis pourquoi $\left\|\mathbb{E}\left[\underline{\boldsymbol{U}} \,\middle|\, \underline{\underline{\boldsymbol{Y}}} = \underline{\boldsymbol{y}}, \underline{\boldsymbol{S}} = \underline{\boldsymbol{s}}\right]\right\|^2$ ainsi que $\underline{\boldsymbol{U}}^T\mathbb{E}\left[\underline{\boldsymbol{U}} \,\middle|\, \underline{\underline{\boldsymbol{Y}}} = \underline{\boldsymbol{y}}, \underline{\boldsymbol{S}} = \underline{\boldsymbol{s}}\right]$ tendent en un certain sens vers une même constante q^* déterministe. Le cœur du travail a donc pour but de caractériser la distribution a posteriori de $\underline{\boldsymbol{U}}$ sachant le jeu de données d'apprentissage.

Chapitre 3

Étude du cas supervisé

Résumé du chapitre

Dans le cas supervisé, la combinatoire disparaît car on peut remplacer les variables latentes $(S_1, ..., S_N)$ par $(V_1, ..., V_N)$, puisque l'on connaît toutes les étiquettes. Il est dès lors possible d'expliciter le classifieur et le risque bayésiens à N et D fixés puis d'étudier le comportement asymptotique de ce dernier. Celui-ci converge vers la valeur fournie par Lelarge et Miolane, ce qui constitue une première manifestation du phénomène de concentration de la mesure. Si le classifieur de Bayes observé est fondamentalement aléatoire, le risque (qui en est une observation unidimensionnelle) converge.

Dans cette partie, nous allons étudier le risque de Bayes associé au cadre mis en place au chapitre 2, dans le régime asymptotique $N \doteq_\alpha D$ où N et D croissent et où leur rapport tend vers $\alpha > 0$ fini. Le cas supervisé nous permettra de mieux comprendre les phénomènes de régularisation qui se produisent dans le cas semi-supervisé qui en sera une généralisation. Nous mènerons l'étude à N, D fixés et calculerons explicitement le classifieur et le risque de Bayes avant de nous intéresser à leur comportement asymptotique. Comme on l'a vu à l'équation (2.4), on cherche à calculer la distribution de $\underline{Y}_+$ sachant le jeu de données d'apprentissage $\underline{\underline{y}}, \underline{s}$ et une étiquette $v_+ = \pm 1$. Les calculs menant à l'équation (2.5) montrent qu'il suffit de déterminer la distribution de probabilité de $\underline{U}$ sachant $\underline{Y} = \underline{y}, \underline{S} = \underline{s}$. Il suffit ici de remplacer les occurrences de $\underline{S}$ par $\underline{V}$, ce qui permet d'observer directement les variables $\underline{V}$ qui sont sinon latentes dans le modèle plus général semi-supervisé.

3.1 Explicitation du classifieur et du risque bayésiens

Dans ce paragraphe, nous allons expliciter la distribution *a posteriori* de $\underline{U}$ sachant $\underline{\underline{Y}}, \underline{S}$ dans le cas supervisé afin d'en déduire ensuite des propriétés que nous tenterons de généraliser et d'expliquer dans le cas semi-supervisé. Commençons par définir deux quantités fonctions du jeu de données d'apprentissage.

Définition 3.1.1: Estimateurs de $\underline{U}$ (cas supervisé)

$$\underline{u}_{\mathrm{ML}}(\underline{v}, \underline{\underline{y}}) = \frac{1}{N} \sum_{i=1}^{N} v_i \underline{y}_i$$

$$\underline{u}_{\mathrm{MSE}}(\underline{v}, \underline{\underline{y}}) = \frac{1}{N + \sigma^2 D} \sum_{i=1}^{N} v_i \underline{y}_i$$

Ces deux quantités sont en fait respectivement l'estimateur du maximum de vraisemblance et l'estimateur par moindres carrés moyens de $\underline{U}$, d'où leurs noms (MSE signifie *mean squared error* en

anglais). On ne démontrera pas que $\underline{\boldsymbol{u}}_{\mathrm{ML}}(\underline{\boldsymbol{v}}, \underline{\underline{\boldsymbol{y}}})$ est l'estimateur du maximum de vraisemblance. Pour $\underline{\boldsymbol{u}}_{\mathrm{MSE}}(\underline{\boldsymbol{v}}, \underline{\underline{\boldsymbol{y}}})$, cela découle de la forme de la loi (3.1) ci-dessous car l'espérance conditionnelle est de façon classique l'estimateur MSE (pour s'en convaincre on pourra regarder le calcul fait en début de section 5.1). Intéressons-nous désormais à la loi a posteriori de $\underline{\boldsymbol{U}}|\underline{\boldsymbol{Y}}, \underline{\underline{\boldsymbol{S}}}$ puis déduisons-en la loi conditionnelle de $\underline{\boldsymbol{Y}}_+|V_+, \underline{\boldsymbol{Y}}, \underline{\underline{\boldsymbol{V}}}$ afin de calculer le classifieur défini à l'équation (2.4).

Lemme 3.1.2: Lois conditionnelles (cas supervisé)

Soit $\underline{\boldsymbol{u}} \in \mathbb{R}^D$, $v_+ \in \{-1, +1\}$, $\underline{\underline{\boldsymbol{y}}} \in \mathbb{R}^{D\times N}$ et $\underline{\boldsymbol{v}} \in \{-1, +1\}^N$.
Alors on a

$$p(\underline{\boldsymbol{u}}|\underline{\underline{\boldsymbol{y}}}, \underline{\boldsymbol{v}}) = \mathcal{N}_D\left(\underline{\boldsymbol{u}}; \underline{\boldsymbol{u}}_{\mathrm{MSE}}(\underline{\underline{\boldsymbol{y}}}, \underline{\boldsymbol{v}}), \frac{\sigma^2}{N+\sigma^2 D}\underline{\underline{\boldsymbol{id}}}_D\right) \tag{3.1}$$

$$p(\underline{\boldsymbol{y}}_+|v_+, \underline{\underline{\boldsymbol{y}}}, \underline{\boldsymbol{v}}) = \mathcal{N}_D\left(\underline{\boldsymbol{y}}_+; v_+\, \underline{\boldsymbol{u}}_{\mathrm{MSE}}(\underline{\boldsymbol{v}}, \underline{\underline{\boldsymbol{y}}}), \sigma^2\frac{N+1+\sigma^2 D}{N+\sigma^2 D}\underline{\underline{\boldsymbol{id}}}_D\right) \tag{3.2}$$

Démonstration. Pour la loi conditionnelle de $\underline{\boldsymbol{U}}|\underline{\boldsymbol{Y}}, \underline{\underline{\boldsymbol{V}}}$, comme $\underline{\boldsymbol{U}}$ et $\underline{\underline{\boldsymbol{V}}}$ sont indépendantes :

$$\begin{aligned}
p(\underline{\boldsymbol{u}}|\underline{\underline{\boldsymbol{y}}}, \underline{\boldsymbol{v}}) &\propto p(\underline{\underline{\boldsymbol{y}}}|\underline{\boldsymbol{u}}, \underline{\boldsymbol{v}})p(\underline{\boldsymbol{u}}) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^N (\underline{\boldsymbol{y}}_i - v_i\underline{\boldsymbol{u}})^T(\underline{\boldsymbol{y}}_i - v_i\underline{\boldsymbol{u}})\right)\right)\exp\left(-\frac{D}{2}\underline{\boldsymbol{u}}^T\underline{\boldsymbol{u}}\right) \\
&\propto \exp\left(\frac{1}{\sigma^2}\underline{\boldsymbol{u}}^T\sum_{i=1}^N v_i\underline{\boldsymbol{y}}_i - \frac{N+\sigma^2 D}{2\sigma^2}\underline{\boldsymbol{u}}^T\underline{\boldsymbol{u}}\right) \\
&\propto \exp\left(-\frac{N+\sigma^2 D}{2\sigma^2}\left\|\underline{\boldsymbol{u}} - \underline{\boldsymbol{u}}_{\mathrm{MSE}}(\underline{\underline{\boldsymbol{y}}}, \underline{\boldsymbol{v}})\right\|^2\right)
\end{aligned}$$

On exploite ensuite ce dernier résultat pour déduire la loi conditionnelle de $\underline{\boldsymbol{Y}}_+|V_+, \underline{\boldsymbol{Y}}, \underline{\underline{\boldsymbol{V}}}$:

$$\begin{aligned}
p(\underline{\boldsymbol{y}}_+|v_+, \underline{\underline{\boldsymbol{y}}}, \underline{\boldsymbol{v}}) &= \int_{\underline{\boldsymbol{u}}} \mathrm{d}p(\underline{\boldsymbol{y}}_+, \underline{\boldsymbol{u}}|v_+, \underline{\underline{\boldsymbol{y}}}, \underline{\boldsymbol{v}}) \\
&= \int_{\underline{\boldsymbol{u}}} p(\underline{\boldsymbol{y}}_+|\underline{\boldsymbol{u}}, v_+, \underline{\underline{\boldsymbol{y}}}, \underline{\boldsymbol{v}})\mathrm{d}p(\underline{\boldsymbol{u}}|v_+, \underline{\underline{\boldsymbol{y}}}, \underline{\boldsymbol{v}}) \\
&= \int_{\underline{\boldsymbol{u}}} p(\underline{\boldsymbol{y}}_+|\underline{\boldsymbol{u}}, v_+)\mathrm{d}p(\underline{\boldsymbol{u}}|\underline{\underline{\boldsymbol{y}}}, \underline{\boldsymbol{v}}) \\
&\propto \int_{\underline{\boldsymbol{u}}} \exp\left(-\frac{1}{2\sigma^2}(\underline{\boldsymbol{y}}_+ - v_+\underline{\boldsymbol{u}})^T(\underline{\boldsymbol{y}}_+ - v_+\underline{\boldsymbol{u}})\right)\exp\left(\frac{1}{\sigma^2}\underline{\boldsymbol{u}}^T\sum_{i=1}^N v_i\underline{\boldsymbol{y}}_i - \frac{N+\sigma^2 D}{2\sigma^2}\underline{\boldsymbol{u}}^T\underline{\boldsymbol{u}}\right)\mathrm{d}\underline{\boldsymbol{u}} \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\underline{\boldsymbol{y}}_+{}^T\underline{\boldsymbol{y}}_+\right)\int_{\underline{\boldsymbol{u}}} \exp\left(-\frac{N+1+\sigma^2 D}{2\sigma^2}\underline{\boldsymbol{u}}^T\underline{\boldsymbol{u}} + \frac{1}{\sigma^2}\underline{\boldsymbol{u}}^T\left(\sum_{i=1}^N v_i\underline{\boldsymbol{y}}_i + v_+\underline{\boldsymbol{y}}_+\right)\right)\mathrm{d}\underline{\boldsymbol{u}} \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\underline{\boldsymbol{y}}_+{}^T\underline{\boldsymbol{y}}_+\right)\exp\left(\frac{1}{2\sigma^2(N+1+\sigma^2 D)}\left(\underline{\boldsymbol{y}}_+{}^T\underline{\boldsymbol{y}}_+ + 2\underline{\boldsymbol{y}}_+{}^T v_+\sum_{i=1}^N v_i\underline{\boldsymbol{y}}_i\right)\right) \\
&\propto \exp\left(-\frac{N+\sigma^2 D}{2\sigma^2(N+1+\sigma^2 D)}\left\|\underline{\boldsymbol{y}}_+ - v_+\underline{\boldsymbol{u}}_{\mathrm{MSE}}(\underline{\underline{\boldsymbol{y}}}, \underline{\boldsymbol{v}})\right\|^2\right)
\end{aligned}$$

□

La détermination de la loi (3.2) permet de déduire dans le cas gaussien à la fois le classifieur optimal (à N et D fixés) et la performance de ce classifieur. En effet, cette loi montre que les deux distributions à séparer dans la formule (2.4) ne sont autres que deux gaussiennes de centres opposés et de même variance proportionnelle à l'identité. Il suffit alors de tracer la médiatrice du segment reliant les centres des classes pour obtenir le classifieur. La performance, quant à elle, est directement

reliée à la longueur de ce segment : plus il est long, plus les estimateurs $\underline{u}_{\text{MSE}}(\underline{\underline{Y}}, \underline{V})$ sont certains. Moins l'algorithme est "sûr" de son estimation, plus il aura tendance à délivrer des estimateurs de très faible norme, le cas limite étant bien entendu le vecteur nul qui n'apportera aucune information.

Proposition 3.1.3: Classifieur bayésien conditionnel (cas supervisé)

Soit N, D des entiers strictement positifs quelconques et $(\underline{\underline{y}}, \underline{v})$ un jeu de données d'apprentissage complètement étiqueté. Le classifieur optimal au sens de $\ell_{\underline{\underline{y}}, \underline{v}}$ est alors :

$$f^*_{\underline{\underline{y}}, \underline{v}}(\underline{y_+}) = \text{signe}\left(\underline{y_+}^T \underline{u}_{\text{MSE}}(\underline{\underline{y}}, \underline{v})\right) \tag{3.3}$$

Démonstration. On injecte l'équation (3.2) dans la formule du classifieur de Bayes (2.4) en remplaçant les occurences de $\underline{S}$ par $\underline{V}$:

$$\begin{aligned} f^*_{\underline{\underline{y}}, \underline{v}}(\underline{y_+}) &= \text{signe}\left(\log\left(\frac{p(\underline{y_+}|V_+ = +1, \underline{\underline{Y}} = \underline{\underline{y}}, \underline{V} = \underline{v})}{p(\underline{y_+}|V_+ = -1, \underline{\underline{Y}} = \underline{\underline{y}}, \underline{V} = \underline{v})}\right)\right) \\ &= \text{signe}\left(\frac{N + \sigma^2 D}{2\sigma^2(N + 1 + \sigma^2 D)}\left(\left\|\underline{y_+} + \underline{u}_{\text{MSE}}(\underline{\underline{y}}, \underline{v})\right\|^2 - \left\|\underline{y_+} - \underline{u}_{\text{MSE}}(\underline{\underline{y}}, \underline{v})\right\|^2\right)\right) \\ &= \text{signe}\left(\underline{y_+}^T \underline{u}_{\text{MSE}}(\underline{\underline{y}}, \underline{v})\right) \end{aligned}$$

□

Remarque. Remarquons que la proposition 3.1.3 correspond exactement à ce que l'on cherche à montrer asymptotiquement dans le cas général semi-supervisé dans l'équation (2.5′) puisqu'en effet l'estimateur par moindres carrés coïncide par un résultat classique d'estimation avec l'espérance conditionnelle.

Proposition 3.1.4: Risque bayésien conditionnel (cas supervisé)

Dans le cas supervisé, le risque bayésien (ie. le risque obtenu avec le classifieur (3.3) de la proposition précédente) conditionné à $\underline{\underline{Y}} = \underline{\underline{y}}, \underline{V} = \underline{v}$ est donné par

$$\ell_{\underline{\underline{y}}, \underline{v}}(f^*_{\underline{\underline{y}}, \underline{v}}) = 1 - \Phi\left(\frac{\left\|\underline{u}_{\text{MSE}}(\underline{\underline{y}}, \underline{v})\right\|}{\sigma}\sqrt{\frac{N + \sigma^2 D}{N + 1 + \sigma^2 D}}\right) \tag{3.4}$$

Démonstration.

$$\begin{aligned} \ell_{\underline{\underline{Y}}, \underline{V}}(f^*_{\underline{\underline{Y}}, \underline{V}}) &= \mathbb{P}\left(V_+ \neq \text{signe}\left(f^*_{\underline{\underline{Y}}, \underline{V}}(\underline{Y_+})\right) \,\middle|\, \underline{\underline{Y}}, \underline{V}\right) \\ &= \mathbb{P}\left(V_+ \underline{Y_+}^T \underline{u}_{\text{MSE}}(\underline{\underline{Y}}, \underline{V}) < 0 \,\middle|\, \underline{\underline{Y}}, \underline{V}\right) \\ &= \mathbb{P}\left(\underline{u}_{\text{MSE}}(\underline{\underline{Y}}, \underline{V})^T \left(\underline{U} + \sigma V_+ \underline{Z_+}\right) < 0 \,\middle|\, \underline{\underline{Y}}, \underline{V}\right) \end{aligned}$$

Comme $\underline{Z_+}$ et V_+ sont indépendants de $\underline{\underline{Y}}$ et $\underline{V}$, on peut étudier leur loi sans conditionnement. La loi de $V_+ \underline{Z_+}$ est la même que celle de la variable aléatoire $\underline{Z_+}$, ce qui fait que l'on peut remplacer la première par la seconde dans l'expression ci-dessus.

$$\ell_{\underline{\underline{Y}}, \underline{V}}(f^*_{\underline{\underline{Y}}, \underline{V}}) = \mathbb{P}\left(\underline{u}_{\text{MSE}}(\underline{\underline{Y}}, \underline{V})^T \left(\underline{U} + \sigma \underline{Z_+}\right) < 0 \,\middle|\, \underline{\underline{Y}}, \underline{V}\right)$$

Conditionnellement à $\underline{\underline{Y}}, \underline{V}$, les deux variables $\underline{Z_+}$ et $\underline{U}$ sont deux gaussiennes indépendantes. La loi $\underline{U}|\underline{\underline{Y}}, \underline{V}$ est donnée à l'équation (3.1) et la loi $\underline{Z_+}|\underline{\underline{Y}}, \underline{V}$ est la même que la loi marginale de $\underline{Z_+}$ car $\underline{Z_+}$ est indépendante du jeu de données d'apprentissage. On a donc

$$\begin{aligned}
\ell_{\underline{\underline{Y}},\underline{V}}(f^*_{\underline{\underline{Y}},\underline{V}}) &= \mathbb{P}\left(\underline{u}_{\mathrm{MSE}}(\underline{\underline{Y}},\underline{V})^T \mathcal{N}_D\left(\underline{u}_{\mathrm{MSE}}(\underline{\underline{Y}},\underline{V}), \frac{\sigma^2}{N+\sigma^2 D}\underline{\underline{id}}_D\right) + \sigma \underline{u}_{\mathrm{MSE}}(\underline{\underline{Y}},\underline{V})^T \mathcal{N}_D\left(0, \underline{\underline{id}}_D\right) < 0 \,\middle|\, \underline{\underline{Y}},\underline{V}\right) \\
&= \mathbb{P}\left(\left\|\underline{u}_{\mathrm{MSE}}(\underline{\underline{Y}},\underline{V})\right\|^2 + \sigma\left\|\underline{u}_{\mathrm{MSE}}(\underline{\underline{Y}},\underline{V})\right\|\left(\mathcal{N}\left(0, \frac{1}{N+\sigma^2 D}\right) + \mathcal{N}(0,1)\right) < 0 \,\middle|\, \underline{\underline{Y}},\underline{V}\right) \\
&= \mathbb{P}\left(\left\|\underline{u}_{\mathrm{MSE}}(\underline{\underline{Y}},\underline{V})\right\| + \sigma\mathcal{N}\left(0, \frac{N+1+\sigma^2 D}{N+\sigma^2 D}\right) < 0 \,\middle|\, \underline{\underline{Y}},\underline{V}\right) \\
&= \mathbb{P}\left(\mathcal{N}(0,1) < -\frac{\left\|\underline{u}_{\mathrm{MSE}}(\underline{\underline{Y}},\underline{V})\right\|}{\sigma}\sqrt{\frac{N+\sigma^2 D}{N+1+\sigma^2 D}} \,\middle|\, \underline{\underline{Y}},\underline{V}\right) \\
&= 1 - \Phi\left(\frac{\left\|\underline{u}_{\mathrm{MSE}}(\underline{\underline{Y}},\underline{V})\right\|}{\sigma}\sqrt{\frac{N+\sigma^2 D}{N+1+\sigma^2 D}}\right)
\end{aligned}$$

Ceci prouve le résultat. □

Remarque. Dans cette expression, on remarque que le terme sous la racine importe très peu car il ressemble très rapidement à 1. Le risque dépend cependant encore du jeu de données d'apprentissage $\underline{\underline{y}}, \underline{v}$ via $\underline{u}_{\mathrm{MSE}}(\underline{\underline{Y}},\underline{V})$ qui est une variable aléatoire. En grande dimension, il se passe cependant un phénomène déroutant au premier abord : bien que $\underline{u}_{\mathrm{MSE}}(\underline{\underline{Y}},\underline{V})$ soit fondamentalement aléatoire, sa norme est essentiellement constante : comme on le verra plus tard, il s'agit d'une première manifestation du phénomène de concentration de la mesure déjà évoqué à la section 1.1.

3.2 Concentration et asymptotique du risque bayésien

Dans la section précédente, nous avons remarqué qu'il était possible, dans le cas supervisé, de calculer exactement le classifieur optimal au sens du risque $\ell_{\underline{\underline{y}},\underline{v}}$ puis le risque associé. Nous allons maintenant étudier le comportement du risque dans le régime asymptotique $N \asymp_\alpha D$. Pour continuer, nous aurons besoin du lemme suivant :

Lemme 3.2.1: Asymptotique de l'espérance d'une loi du χ

Soit $(X_D)_{D\in\mathbb{N}^*}$ une suite de variables aléatoires réelles suivant chacune une loi du χ à D degrés de liberté : $X_D \sim \chi_D$ Alors, lorsque $D \to +\infty$,

$$\mathbb{E}[X_D] = \sqrt{D}\left(1 - \frac{1}{4D} + o\left(\frac{1}{D}\right)\right)$$

et en particulier

$$D - \mathbb{E}[X_D]^2 \to \frac{1}{2}$$

Démonstration. L'espérance d'une loi du χ à D degrés de liberté est donnée par

$$\mathbb{E}[X_D] = \sqrt{2}\frac{\Gamma\left(\frac{D+1}{2}\right)}{\Gamma\left(\frac{D}{2}\right)}$$

On a par ailleurs les développements suivants pour $x \to +\infty$:

$$\Gamma(x) = \sqrt{2\pi}x^{x-\frac{1}{2}}e^{-x}\left(1 + \frac{1}{12x} + o\left(\frac{1}{x}\right)\right)$$

$$\Gamma\left(x+\frac{1}{2}\right) = \sqrt{2\pi}x^x e^{-x}\left(1-\frac{1}{24x}+o\left(\frac{1}{x}\right)\right)$$

Par conséquent

$$\begin{aligned}\mathbb{E}\left[X_D\right] &= \sqrt{D}\frac{1-\frac{1}{12D}+o\left(\frac{1}{D}\right)}{1+\frac{1}{6D}+o\left(\frac{1}{D}\right)}\\ &= \sqrt{D}\left(1-\frac{1}{12D}+o\left(\frac{1}{D}\right)\right)\left(1-\frac{1}{6D}+o\left(\frac{1}{D}\right)\right)\\ &= \sqrt{D}\left(1-\frac{1}{4D}+o\left(\frac{1}{D}\right)\right)\end{aligned}$$

On a alors en particulier

$$\mathbb{E}\left[X_D\right]^2 = D\left(1-\frac{1}{2D}+o\left(\frac{1}{D}\right)\right)$$

et donc

$$D-\mathbb{E}\left[X_D\right]^2 = \frac{1}{2}+o\left(1\right)$$

□

Proposition 3.2.2: Convergence de la norme de l'estimateur $\underline{\boldsymbol{U}}_{\text{MSE}}$

Notons $\underline{\boldsymbol{U}}_{\text{MSE},N}$ la variable aléatoire $\underline{\boldsymbol{U}}_{\text{MSE},N} = \frac{1}{N+\sigma^2 D}\sum_{i=1}^N V_i \underline{\boldsymbol{Y}}_i$ (on fait apparaître explicitement la dépendance en (N,D) via l'indice N)
Alors dans le régime $N \backsimeq_\alpha D$, $\left\|\underline{\boldsymbol{U}}_{\text{MSE},N}\right\|$ se concentre autour de sa moyenne

$$\forall \varepsilon > 0, \mathbb{P}\left(\left|\left\|\underline{\boldsymbol{U}}_{\text{MSE},N}\right\| - \mathbb{E}\left[\left\|\underline{\boldsymbol{U}}_{\text{MSE},N}\right\|\right]\right| \geq \varepsilon\right) \longrightarrow 0$$

et on a même convergence en probabilité de la suite de variables aléatoires $\left(\left\|\underline{\boldsymbol{U}}_{\text{MSE},N}\right\|\right)_N$ vers la limite de la suite des moyennes : $\sqrt{\frac{\alpha}{\alpha+\sigma^2}}$

$$\forall \varepsilon > 0, \mathbb{P}\left(\left|\left\|\underline{\boldsymbol{U}}_{\text{MSE},N}\right\| - \sqrt{\frac{\alpha}{\alpha+\sigma^2}}\right| \geq \varepsilon\right) \longrightarrow 0$$

On pourra observer ce comportement à l'aide du script `S03_concentration_norm_SL.py`.

Démonstration.

$$\begin{aligned}\underline{\boldsymbol{U}}_{\text{MSE},N} &= \frac{1}{N+\sigma^2 D}\sum_{i=1}^N V_i(V_i\underline{\boldsymbol{U}}+\sigma\underline{\boldsymbol{Z}}_i)\\ &= \frac{1}{N+\sigma^2 D}\left(N\underline{\boldsymbol{U}}+\sigma\sum_{i=1}^N V_i\underline{\boldsymbol{Z}}_i\right)\end{aligned}$$

Or, les variables aléatoires $\underline{\boldsymbol{A}}_i \triangleq V_i\underline{\boldsymbol{Z}}_i$ (pour $i \in [N]$) suivent des lois normales multivariés $\mathcal{N}_D(\underline{\boldsymbol{0}}, \underline{\underline{\boldsymbol{id}}}_D)$ toutes indépendantes entre elles (car les V_i et les $\underline{\boldsymbol{Z}}_i$ sont toutes indépendantes). Par conséquent, la variable aléatoire $\underline{\boldsymbol{A}} \triangleq \sum_{i=1}^N V_i\underline{\boldsymbol{Z}}_i$ suit une loi normale $\mathcal{N}_D\left(\underline{\boldsymbol{0}}, N\underline{\underline{\boldsymbol{id}}}_D\right)$. Finalement, $\underline{\boldsymbol{U}}_{\text{MSE}}$ suit la même loi que

$$\underline{\boldsymbol{U}}_{\text{MSE},N} \overset{\mathcal{D}}{=} \frac{1}{N+\sigma^2 D}(N\underline{\boldsymbol{U}}+\sigma\underline{\boldsymbol{A}})$$

Or, comme $\underline{\boldsymbol{U}}$ est indépendante des V_i et des $\underline{\boldsymbol{Z}}_i$, $\underline{\boldsymbol{U}}$ reste indépendante de $\underline{\boldsymbol{A}}$. En sommant les variances de ces deux gaussiennes :

$$\underline{\boldsymbol{U}}_{\mathrm{MSE},N} \overset{\mathcal{D}}{=} \mathcal{N}_D\left(\underline{\mathbf{0}}, \frac{N}{D(N+\sigma^2 D)}\underline{\underline{\boldsymbol{id}}}_D\right) \tag{3.5}$$

Ceci nous mène donc à :

$$\left\|\underline{\boldsymbol{U}}_{\mathrm{MSE},N}\right\| \overset{\mathcal{D}}{=} \sqrt{\frac{N}{D(N+\sigma^2 D)}} X_D \tag{3.6}$$

où $(X_D)_D$ est une suite de variables aléatoires suivant une loi du χ à D degrés de liberté (souvenons-nous que dans le régime $N \doteq_\alpha D$, N et D croissent asymptotiquement).

Nous allons désormais procéder en deux temps et étudier la convergence de l'espérance de $\left\|\underline{\boldsymbol{U}}_{\mathrm{MSE},N}\right\|$ puis la concentration de $\left\|\underline{\boldsymbol{U}}_{\mathrm{MSE},N}\right\|$ autour de son espérance avant de revenir au résultat énoncé :

Convergence de l'espérance de la norme On peut écrire l'espérance de la norme de la façon suivante :

$$\mathbb{E}\left[\left\|\underline{\boldsymbol{U}}_{\mathrm{MSE},N}\right\|\right] = \sqrt{\frac{N}{D(N+\sigma^2 D)}}\mathbb{E}\left[X_D\right]$$

D'après le lemme (3.2.1), cette espérance est équivalente à

$$\mathbb{E}\left[\left\|\underline{\boldsymbol{U}}_{\mathrm{MSE},N}\right\|\right] \sim \sqrt{\frac{N}{N+\sigma^2 D}} \longrightarrow \sqrt{\frac{\alpha}{\alpha+\sigma^2}}$$

Concentration de la norme autour de son espérance On étudie maintenant la probabilité de déviation de $\left\|\underline{\boldsymbol{U}}_{\mathrm{MSE},N}\right\|$ à sa moyenne pour montrer qu'elle converge vers 0 : soit $\varepsilon > 0$

$$\begin{aligned}
\mathbb{P}\left(\left|\left\|\underline{\boldsymbol{U}}_{\mathrm{MSE},N}\right\| - \mathbb{E}\left[\left\|\underline{\boldsymbol{U}}_{\mathrm{MSE},N}\right\|\right]\right| > \varepsilon\right) &= \mathbb{P}\left(\left|X_D - \mathbb{E}\left[X_D\right]\right| \geq \varepsilon\sqrt{\frac{D(N+\sigma^2 D)}{N}}\right) \\
&\leq \frac{N(D - \mathbb{E}\left[X_D\right]^2)}{\varepsilon^2 D(N+\sigma^2 D)} \\
&\sim \frac{\alpha}{2\varepsilon^2(\alpha+\sigma^2)}\frac{1}{D} \\
&\longrightarrow 0
\end{aligned}$$

où l'inégalité est obtenue grâce à Bienaymé-Tchebychev et où l'équivalent de la dernière étape a été obtenu grâce au lemme 3.2.1.

Convergence en probabilité On termine en utilisant l'inégalité triangulaire suivante :

$$\left|\left\|\underline{\boldsymbol{U}}_{\mathrm{MSE},N}\right\| - \sqrt{\frac{\alpha}{\alpha+\sigma^2}}\right| \leq \left|\left\|\underline{\boldsymbol{U}}_{\mathrm{MSE},N}\right\| - \mathbb{E}\left[\left\|\underline{\boldsymbol{U}}_{\mathrm{MSE},N}\right\|\right]\right| + \left|\mathbb{E}\left[\left\|\underline{\boldsymbol{U}}_{\mathrm{MSE},N}\right\|\right] - \sqrt{\frac{\alpha}{\alpha+\sigma^2}}\right|$$

On a donc

$$\begin{aligned}
\mathbb{P}\left(\left|\left\|\underline{\boldsymbol{U}}_{\mathrm{MSE},N}\right\| - \sqrt{\frac{\alpha}{\alpha+\sigma^2}}\right| \geq \varepsilon\right) \leq\ & \mathbb{P}\left(\left|\left\|\underline{\boldsymbol{U}}_{\mathrm{MSE},N}\right\| - \mathbb{E}\left[\left\|\underline{\boldsymbol{U}}_{\mathrm{MSE},N}\right\|\right]\right| \geq \frac{\varepsilon}{2}\right) \\
&+ \mathbb{P}\left(\left|\mathbb{E}\left[\left\|\underline{\boldsymbol{U}}_{\mathrm{MSE},N}\right\|\right] - \sqrt{\frac{\alpha}{\alpha+\sigma^2}}\right| \geq \frac{\varepsilon}{2}\right)
\end{aligned}$$

Comme les deux termes de droite tendent vers 0, la convergence en probabilité est prouvée. □

Proposition 3.2.3: Risque bayésien asymptotique (cas supervisé)

Dans le régime $N \doteq_\alpha D$, le risque bayésien optimal R_N^* tend vers $1 - \Phi\left(\frac{\sqrt{\alpha}}{\sigma\sqrt{\alpha+\sigma^2}}\right)$
On rappelle que Φ, définie dans les notations, est une queue de gaussienne, et on pourra observer la concentration du risque bayésien à l'aide du script `S04_risk_SL.py`.

Démonstration. En utilisant successivement la formule de l'espérance totale puis la proposition (3.1.4), on obtient

$$\begin{aligned} R_N^* &= \mathbb{E}\left[\mathbb{P}\left(V_+ \neq f^*_{\underline{\underline{Y}},\underline{V}}(\underline{Y_+}) \;\middle|\; \underline{\underline{Y}},\underline{V}\right)\right] \\ &= 1 - \mathbb{E}\left[\Phi\left(\frac{\left\|\underline{\boldsymbol{u}_{\mathrm{MSE},N}}\right\|}{\sigma}\sqrt{\frac{N+\sigma^2 D}{N+1+\sigma^2 D}}\right)\right] \end{aligned}$$

On se convainc sans peine que la proposition 3.2.2 fonctionne aussi en remplaçant $\left\|\underline{\boldsymbol{u}_{\mathrm{MSE},N}}\right\|$ dans la formule (3.6) par

$$\left\|\underline{\boldsymbol{u}_{\mathrm{MSE},N}}\right\|\sqrt{\frac{N+\sigma^2 D}{N+1+\sigma^2 D}} \overset{\mathcal{D}}{=} \sqrt{\frac{N}{D(N+1+\sigma^2 D)}}X_D$$

On a donc convergence en probabilité de cette quantité, et en particulier en loi vers $\sqrt{\frac{\alpha}{\alpha+\sigma^2}}$
On termine en notant que la fonction $t \mapsto \Phi\left(\frac{t}{\sigma}\right)$ est continue et bornée, si bien que par caractérisation de la convergence en loi,

$$R_N^* \longrightarrow 1 - \Phi\left(\frac{\sqrt{\alpha}}{\sigma\sqrt{\alpha+\sigma^2}}\right)$$

□

Remarque. Comme on l'avait déjà noté à la remarque de la page 20, le phénomène que l'on observe ici est très intimement lié à celui de la concentration de la mesure, dont il est un cas particulier simple. La variable aléatoire $\underline{\boldsymbol{u}_{\mathrm{MSE},N}}$ est un objet par essence non-déterministe : comme on l'a vu à l'équation (3.5), il suit une loi normale. On pourrait même avoir l'impression que, puisque sa variance est en $\frac{N}{D(N+\sigma^2 D)}\underline{\underline{\boldsymbol{id}}}_D$, cet objet va se rapprocher du vecteur nul mais ce serait oublier que dans le régime $N \doteq_\alpha D$, la dimension de l'espace dans lequel vivent ces lois normales augmente aussi. Si les coordonnées des vecteurs aléatoires tirés diminuent asymptotiquement, le nombre de coordonnées augmente et cette augmentation contribue largement au fait que $\underline{\boldsymbol{u}_{\mathrm{MSE},N}}$ comporte asymptotiquement toujours de l'information significative.

Asymptotiquement donc, le vecteur aléatoire $\underline{\boldsymbol{u}_{\mathrm{MSE},N}}$ voit ses coordonnées décroître mais reste fondamentalement aléatoire, sans aucune direction de l'espace préférée. Cependant, sa norme tend vers une valeur constante ! Plus généralement, un vecteur aléatoire concentré sera un vecteur aléatoire dont l'évaluation par certaines fonctionnelles (ie. des fonctions renvoyant un réel) est asymptotiquement essentiellement constante : la probabilité de déviation par rapport à une valeur limite décroît très rapidement. Pour plus d'information, on pourra se référer à [6] ou [19], ou encore aux travaux récents de Louart et Couillet [8] pour des applications en apprentissage automatique.

Dans notre cas, si le classifieur optimal est aléatoire (si $\underline{\boldsymbol{U}}$ varie, l'estimateur associé aussi), ce n'est pas le cas de la performance optimale qui est asymptotiquement indépendante de l'aléa et possède une expression très simple.

Chapitre 4

Étude du cas semi-supervisé

Résumé du chapitre

Dans le cas semi-supervisé, les calculs sont plus complexes que précédemment et mènent soit à des difficultés combinatoires, soit à des difficultés liées au calcul d'intégrales en grande dimension. Pour poursuivre les calculs, on propose une piste pour obtenir un classifieur bayésien approché. Si l'on ne trouve pas immédiatement un algorithme efficace, il est possible de l'adapter avec un paramètre de recentrage que l'on choisit empiriquement. Le classifieur se rapproche alors de l'algorithme proposé par Xiaoyi Mai dans [10] mais n'atteint évidemment pas la performance optimale prédite par Lelarge et Miolane. Des études ultérieures pourront affiner les simplifications qui ont dû être faites pour aboutir à cet algorithme.

Les calculs qui ont été menés ci-dessus dans le cas supervisé au chapitre précédent ont cet avantage qu'ils peuvent être menés explicitement jusqu'au bout et que l'on peut ensuite étudier le comportement asymptotique des quantités associées sans faire d'approximation en amont. Cette démarche se révèle infructueuse dans le cas semi-supervisé car les étiquettes inconnues compliquent grandement le problème en le rendant fondamentalement combinatoire. Dans le cas supervisé, nous nous étions intéressés à l'estimation du vecteur $\underline{\boldsymbol{U}}$ en apportant la meilleure information possible grâce à la statistique $\underline{\underline{\boldsymbol{Y}}}\underline{\boldsymbol{V}} = \sum_{i=1}^{N} V_i \underline{\boldsymbol{Y}_i}$. Finalement, aux signes V_i près (on peut les négliger car on les connaît tous, et dès que l'un est négatif, on peut remettre le vecteur tiré « dans le bon sens »), notre problème revenait à estimer le vecteur D-dimensionnel $\underline{\boldsymbol{U}}$ à partir de N copies bruitées de ce même vecteur. Notons que la statistique $\underline{\underline{\boldsymbol{Y}}}\underline{\boldsymbol{V}}$ de $(\underline{\boldsymbol{V}}, \underline{\underline{\boldsymbol{Y}}})$ est exhaustive pour le paramètre $\underline{\boldsymbol{U}}$ puisque l'on a la factorisation de Fisher-Neyman suivante :

$$\begin{aligned} p(\underline{\underline{\boldsymbol{y}}}, \underline{\boldsymbol{v}} | \underline{\boldsymbol{u}}) &\propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(\underline{\boldsymbol{y}_i} - v_i \underline{\boldsymbol{u}} \right)^T \left(\underline{\boldsymbol{y}_i} - v_i \underline{\boldsymbol{u}} \right) \right) \\ &= \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} \underline{\boldsymbol{y}_i}^T \underline{\boldsymbol{y}_i} \right) \exp\left(\frac{1}{\sigma^2} \underline{\boldsymbol{u}}^T \underline{\underline{\boldsymbol{y}}} \underline{\boldsymbol{v}} - \frac{1}{2\sigma^2} \underline{\boldsymbol{u}}^T \underline{\boldsymbol{u}} \right) \end{aligned}$$

Plaçons-nous maintenant dans le cadre fictif où l'on connaîtrait l'oracle $\underline{\boldsymbol{U}}$ et l'on souhaiterait estimer $\underline{\boldsymbol{V}}$. Il suffit alors d'estimer chaque V_i en prenant le signe du produit scalaire $\underline{\boldsymbol{U}}^T \underline{\boldsymbol{Y}_i}$, autrement dit trouver si $\underline{\boldsymbol{Y}_i}$ est orienté dans le même sens que $\underline{\boldsymbol{U}}$. $\underline{\underline{\boldsymbol{Y}}}^T \underline{\boldsymbol{U}}$ est d'ailleurs une statistique exhaustive pour $\underline{\boldsymbol{V}}$ puisque l'on a la factorisation de Fisher-Neyman suivante :

$$p(\underline{\underline{\boldsymbol{y}}}, \underline{\boldsymbol{u}} | \underline{\boldsymbol{v}}) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} \underline{\boldsymbol{y}_i}^T \underline{\boldsymbol{y}_i} - \frac{N + \sigma^2}{2\sigma^2} \underline{\boldsymbol{u}}^T \underline{\boldsymbol{u}} \right) \exp\left(\frac{1}{\sigma^2} \underline{\boldsymbol{v}}^T \underline{\underline{\boldsymbol{y}}}^T \underline{\boldsymbol{u}} \right)$$

Que se passe-t-il alors lorsque l'on ne connaît ni $\underline{\boldsymbol{U}}$ ni $\underline{\boldsymbol{V}}$, ou bien dans le cas semi-supervisé, lorsque l'on ne connaît que partiellement $\underline{\boldsymbol{V}}$? Dans ce cas, notre problème se ramène à un problème d'estimation d'une matrice de rang 1 à laquelle on a ajouté un bruit additif gaussien dont toutes les composantes sont indépendantes :

$$\underline{\underline{\boldsymbol{Y}}} = \underline{\boldsymbol{U}}\,\underline{\boldsymbol{V}}^T + \sigma\underline{\underline{\boldsymbol{Z}}}$$

Plusieurs approches sont envisageables pour aborder ce problème. La plus naturelle est peut-être celle de poursuivre les calculs tant que l'on obtient des formes explicites (notamment pour la distribution de $\underline{\boldsymbol{U}}|\underline{\underline{\boldsymbol{Y}}}, \underline{\boldsymbol{S}}$) avant de chercher des simplifications ou approximations éventuelles.

4.1 Loi a posteriori du vecteur engendrant les classes

Dans le cas supervisé, la loi a posteriori de $\underline{\boldsymbol{U}}|\underline{\underline{\boldsymbol{Y}}}, \underline{\boldsymbol{V}}$ avait une expression simple comme en témoigne la formule (3.1). Dans le cas semi-supervisé, ce n'est plus le cas et les paragraphes qui suivent ont pour but de l'expliciter. Commençons par définir l'ensemble suivant :

Définition 4.1.1

Soit $\underline{s} = (s_1, ..., s_N) \in \{-1, 0, +1\}^N$. On définit $\mathcal{V}(\underline{s})$ comme étant le sous-ensemble de $\{-1, +1\}^N$ contenant tous les éléments $\underline{v} = (v_1, ..., v_N) \in \{-1, +1\}^N$ vérifiant

$$\forall i \in \{1, ..., N\}, s_i \neq 0 \Rightarrow v_i = s_i$$

On note également $N_u = |\{i \in [N] | s_i = 0\}|$ et $N_l = |\{i \in [N] | s_i \neq 0\}|$ le nombre d'éléments respectivement non-étiquetés et étiquetés du jeu de données d'apprentissage.
Enfin, on note $\underline{\underline{\boldsymbol{y}_\mathbf{u}}}$ la sous-matrice de $\underline{\underline{\boldsymbol{y}}}$ contenant uniquement les colonnes de $\underline{\underline{\boldsymbol{y}}}$ qui ne sont pas étiquetées. Elle est donc de taille $D \times N_u$.

Exemple. Prenons $N = 3$ et $\underline{s} = (+1, -1, 0)$. Alors :

$$\mathcal{V}(\underline{s}) = \{(+1, -1, +1), (+1, -1, -1)\}$$

Dit autrement, l'ensemble $\mathcal{V}(\underline{s})$ correspond à toutes les réalisations possibles de $\underline{\boldsymbol{V}}$ sachant $\underline{\boldsymbol{S}} = \underline{s}$: on complète simplement toutes les étiquettes inconnues en prenant tous les -1 ou +1 possibles.

Définition 4.1.2

On définit le poids suivant sur les éléments de $\underline{v} \in \mathcal{V}(\underline{s})$:

$$w(\underline{v}) \triangleq \exp\left(\frac{1}{2\sigma^2(N+\sigma^2 D)} \underline{v}^T \underline{\underline{\boldsymbol{y}}}^T \underline{\underline{\boldsymbol{y}}}\,\underline{v}\right) = \exp\left(\frac{1}{2\sigma^2(N+\sigma^2 D)} \left\| \sum_{i=1}^{N} v_i \underline{\boldsymbol{y}_i} \right\|^2\right)$$

Ce poids permet de définir une mesure de probabilité discrète (après normalisation) sur l'ensemble fini $\mathcal{V}(\underline{s})$, ou de façon équivalente sur $\{-1, +1\}^{N_u}$ en concaténant les N_l étiquettes connues :

$$w'(\underline{v_\mathrm{u}}) = \exp\left(\frac{1}{2\sigma^2(N+\sigma^2 D)} \left(2\underline{s}^T \underline{\underline{\boldsymbol{y}}}^T \underline{\underline{\boldsymbol{y}_\mathbf{u}}}\,\underline{v_\mathrm{u}} + \underline{v_\mathrm{u}}^T \underline{\underline{\boldsymbol{y}_\mathbf{u}}}^T \underline{\underline{\boldsymbol{y}_\mathbf{u}}}\,\underline{v_\mathrm{u}}\right)\right)$$

Le lemme qui suit est l'analogue du lemme 3.1.2 dans le cas semi-supervisé :

Lemme 4.1.3: Lois conditionnelles (cas semi-supervisé)

Soit $\underline{\underline{\boldsymbol{y}}} \in \mathbb{R}^{N\times D}$, $\underline{\boldsymbol{s}} \in \{-1,0,1\}^N$ un jeu de données d'apprentissage constitué de N points dans l'espace de dimension D et de N étiquettes masquées ou non.
On a alors

$$p(\underline{\boldsymbol{u}}|\underline{\underline{\boldsymbol{y}}},\underline{\boldsymbol{s}}) = \frac{1}{Z}\sum_{\underline{\boldsymbol{v}}\in\mathcal{V}(\underline{\boldsymbol{s}})} w(\underline{\boldsymbol{v}})\mathcal{N}_D\left(\underline{\boldsymbol{u}}; \frac{1}{N+\sigma^2 D}\underline{\underline{\boldsymbol{y}}}\underline{\boldsymbol{v}}, \frac{\sigma^2}{N+\sigma^2 D}\underline{\underline{\boldsymbol{id}}}_D\right) \tag{4.1}$$

$$= \frac{1}{Z'}\mathcal{N}_D\left(\underline{\boldsymbol{u}}; \frac{1}{N+\sigma^2 D}\underline{\underline{\boldsymbol{y}}}\underline{\boldsymbol{s}}, \frac{\sigma^2}{N+\sigma^2 D}\underline{\underline{\boldsymbol{id}}}_D\right)\prod_{i=1}^{N_u}\cosh\left(\frac{1}{\sigma^2}\underline{\boldsymbol{u}}^T\underline{\boldsymbol{y}_{u,i}}\right) \tag{4.2}$$

Par ailleurs,

$$p(\underline{\boldsymbol{y}_+}|v_+,\underline{\underline{\boldsymbol{y}}},\underline{\boldsymbol{s}}) = \frac{1}{Z''}\sum_{\underline{\boldsymbol{v}}\in\mathcal{V}(\underline{\boldsymbol{s}})} w(\underline{\boldsymbol{v}})\mathcal{N}_D\left(\underline{\boldsymbol{y}_+}; \frac{v_+}{N+\sigma^2 D}\underline{\underline{\boldsymbol{y}}}\underline{\boldsymbol{v}}, \sigma^2\frac{N+1+\sigma^2 D}{N+\sigma^2 D}\underline{\underline{\boldsymbol{id}}}_D\right) \tag{4.3}$$

Z, Z' et Z'' sont des constantes de normalisation.

Démonstration. Dans cette démonstration, on utilisera beaucoup le signe $\propto$ qui signifie, rappelons-le, que l'on ne calcule pas la constante de normalisation de la densité de probabilité, car il suffira d'intégrer le résultat final obtenu. On l'emploie dès qu'une constante multiplicative ne dépend pas de la quantité mesurée par la densité en cours de calcul. En utilisant la formule de Bayes puis la formule des probabilités totales sur la variable $\underline{\boldsymbol{V}}$, on écrit ($\mathrm{d}\mu$ est une mesure discrète uniforme sur $\{-1,+1\}^N$) :

$$p(\underline{\boldsymbol{u}}|\underline{\underline{\boldsymbol{y}}},\underline{\boldsymbol{s}}) \propto p(\underline{\boldsymbol{u}},\underline{\underline{\boldsymbol{y}}}|\underline{\boldsymbol{s}}) = \int_{\underline{\boldsymbol{v}}} p(\underline{\boldsymbol{u}},\underline{\boldsymbol{v}},\underline{\underline{\boldsymbol{y}}}|\underline{\boldsymbol{s}})\mathrm{d}\mu(\underline{\boldsymbol{v}}) = \int_{\underline{\boldsymbol{v}}} \underbrace{p(\underline{\underline{\boldsymbol{y}}}|\underline{\boldsymbol{u}},\underline{\boldsymbol{v}})}_{(a)}\underbrace{p(\underline{\boldsymbol{u}})}_{(b)}\mathrm{d}\underbrace{p(\underline{\boldsymbol{v}}|\underline{\boldsymbol{s}})}_{(c)} \tag{4.4}$$

Pour aboutir à (a), on a utilisé que la variable aléatoire $\underline{\underline{\boldsymbol{Y}}}$ est indépendante de $\underline{\boldsymbol{S}}$ conditionnellement à $\underline{\boldsymbol{V}}$. On sépare ensuite les termes (b) et (c) en utilisant que $\underline{\boldsymbol{U}}$ et $\underline{\boldsymbol{V}}$ sont indépendants conditionnellement à $\underline{\boldsymbol{S}}$. On peut enfin omettre le conditionnement par rapport à $\underline{\boldsymbol{S}}$ dans (b) car $\underline{\boldsymbol{U}}$ est indépendant de $\underline{\boldsymbol{S}}$.
On décompose maintenant $\underline{\boldsymbol{v}} = (v_1,...,v_N)$, $\underline{\underline{\boldsymbol{y}}} = (\underline{\boldsymbol{y}_1},...,\underline{\boldsymbol{y}_N})$ et on utilise que les N variables aléatoires associées $(\underline{\boldsymbol{Y}_1},...,\underline{\boldsymbol{Y}_N})$ sont indépendantes conditionnellement à $\underline{\boldsymbol{U}}$ et que chacune ne dépend que de son étiquette. Comme les variables aléatoires $\underline{\boldsymbol{Y}_i}$ sont normales centrées en $V_i\underline{\boldsymbol{U}}$ conditionnellement à $(V_i,\underline{\boldsymbol{U}})$, on a en réinjectant leurs densités (ainsi que celle de $\underline{\boldsymbol{U}}$) :

$$\begin{aligned} p(\underline{\boldsymbol{u}},\underline{\boldsymbol{v}},\underline{\underline{\boldsymbol{y}}}|\underline{\boldsymbol{s}}) &= \prod_{i=1}^{N} p(\underline{\boldsymbol{y}_i}|\underline{\boldsymbol{u}},v_i)p(\underline{\boldsymbol{u}})p(\underline{\boldsymbol{v}}|\underline{\boldsymbol{s}}) \\ &\propto \exp\left(\sum_{i=1}^{N}\left(-\frac{1}{2\sigma^2}(\underline{\boldsymbol{y}_i}-v_i\underline{\boldsymbol{u}})^T(\underline{\boldsymbol{y}_i}-v_i\underline{\boldsymbol{u}})\right) - \frac{D}{2}\underline{\boldsymbol{u}}^T\underline{\boldsymbol{u}}\right)p(\underline{\boldsymbol{v}}|\underline{\boldsymbol{s}}) \end{aligned}$$

Dans l'équation (4.4), les termes qui nous intéressent sont ceux en $\underline{\boldsymbol{u}}$ et en $\underline{\boldsymbol{v}}$. Ceux qui en sont indépendants peuvent être inclus dans la constante Z.
En développant les produits scalaires, en omettant les termes ne dépendant pas de $\underline{\boldsymbol{v}}$ ni de $\underline{\boldsymbol{u}}$ et en utilisant que $\forall\underline{\boldsymbol{v}},\forall i, v_i^2 = 1$ on obtient :

$$
\begin{aligned}
p(\underline{\boldsymbol{u}}, \underline{\boldsymbol{v}}, \underline{\underline{\boldsymbol{y}}} | \underline{\boldsymbol{s}}) &\propto \exp\left(\sum_{i=1}^{N}\left(-\frac{1}{2\sigma^2}(-2v_i \underline{\boldsymbol{y}_i}^T \underline{\boldsymbol{u}} + \underline{\boldsymbol{u}}^T \underline{\boldsymbol{u}})\right) - \frac{D}{2}\underline{\boldsymbol{u}}^T \underline{\boldsymbol{u}}\right) p(\underline{\boldsymbol{v}}|\underline{\boldsymbol{s}}) \\
&\propto \exp\left(-\frac{N+\sigma^2 D}{2\sigma^2}\underline{\boldsymbol{u}}^T \underline{\boldsymbol{u}}\right) \exp\left(\frac{1}{\sigma^2}\underline{\boldsymbol{u}}^T \underline{\underline{\boldsymbol{y}}}\,\underline{\boldsymbol{v}}\right) p(\underline{\boldsymbol{v}}|\underline{\boldsymbol{s}}) \\
&= \exp\left(-\frac{N+\sigma^2 D}{2\sigma^2}\left(\underline{\boldsymbol{u}}^T \underline{\boldsymbol{u}} - \frac{2}{N+\sigma^2 D}\underline{\boldsymbol{u}}^T \underline{\underline{\boldsymbol{y}}}\,\underline{\boldsymbol{v}}\right)\right) p(\underline{\boldsymbol{v}}|\underline{\boldsymbol{s}}) \qquad (4.5) \\
&\propto \exp\left(\frac{N+\sigma^2 D}{2\sigma^2}\underline{\boldsymbol{v}}^T \underline{\underline{\boldsymbol{y}}}^T \underline{\underline{\boldsymbol{y}}}\,\underline{\boldsymbol{v}}\right) \mathcal{N}_D\left(\underline{\boldsymbol{u}}; \frac{1}{N+\sigma^2 D}\underline{\underline{\boldsymbol{y}}}\,\underline{\boldsymbol{v}}, \frac{\sigma^2}{N+\sigma^2 D}\underline{\underline{\boldsymbol{id}}}_D\right) p(\underline{\boldsymbol{v}}|\underline{\boldsymbol{s}}) \qquad (4.6)
\end{aligned}
$$

La mesure de probabilité $\mathbb{P}(\underline{\boldsymbol{v}}|\underline{\boldsymbol{s}})$ est en fait une mesure de probabilité discrète car la loi a priori de $\underline{\boldsymbol{V}}$ contraint $\underline{\boldsymbol{v}}$ à des valeurs dans l'ensemble $\{-1,+1\}^N$. Conditionnellement à $\underline{\boldsymbol{S}} = \underline{\boldsymbol{s}}$, cette mesure associe même une probabilité constante sur toutes les réalisations de $\underline{\boldsymbol{V}}$ compatibles avec $\underline{\boldsymbol{s}}$, à savoir les éléments de $\mathcal{V}(\underline{\boldsymbol{s}})$. En reprenant l'exemple de la définition 4.1.1,

$$
\mathbb{P}(\underline{\boldsymbol{V}} = (+1,-1,+1)|\underline{\boldsymbol{S}} = (+1,-1,0)) = \mathbb{P}(\underline{\boldsymbol{V}} = (+1,-1,-1)|\underline{\boldsymbol{S}} = (+1,-1,0)) = \frac{1}{2}
$$

tandis que toutes les autres réalisations possibles de $\underline{\boldsymbol{V}}$ auront une probabilité a posteriori valant 0. Finalement, en partant de la forme (4.6), la probabilité recherchée à l'équation (4.4) peut se réécrire

$$
p(\underline{\boldsymbol{u}}|\underline{\underline{\boldsymbol{y}}}, \underline{\boldsymbol{s}}) \propto \sum_{\underline{\boldsymbol{v}} \in \mathcal{V}(\underline{\boldsymbol{s}})} w(\underline{\boldsymbol{v}}) \mathcal{N}_D\left(\underline{\boldsymbol{u}}; \frac{1}{N+\sigma^2 D}\underline{\underline{\boldsymbol{y}}}\,\underline{\boldsymbol{v}}, \frac{\sigma^2}{N+\sigma^2 D}\underline{\underline{\boldsymbol{id}}}_D\right)
$$

Ceci démontre la première partie de la proposition (équation 4.1). Montrons maintenant (4.2). On repart de la forme (4.5) et on utilise la même remarque que ci-dessus pour réécrire l'intégrale en une somme discrète avant de séparer la somme à l'intérieur de l'exponentielle en deux :

$$
\begin{aligned}
p(\underline{\boldsymbol{u}}|\underline{\underline{\boldsymbol{y}}}, \underline{\boldsymbol{s}}) &\propto \exp\left(-\frac{N+\sigma^2 D}{2\sigma^2}\left(\underline{\boldsymbol{u}}^T \underline{\boldsymbol{u}} - 2\underline{\boldsymbol{u}}^T \frac{1}{N+\sigma^2 D}\underline{\underline{\boldsymbol{y}}}\,\underline{\boldsymbol{s}}\right)\right) \sum_{\underline{\boldsymbol{v}_{\mathrm{u}}} \in \{-1,+1\}^{N_u}} \exp\left(\frac{1}{\sigma^2}\underline{\boldsymbol{u}}^T \underline{\underline{\boldsymbol{y}_{\mathrm{u}}}}\,\underline{\boldsymbol{v}_{\mathrm{u}}}\right) \qquad (4.7) \\
&\propto \mathcal{N}_D\left(\underline{\boldsymbol{u}}; \frac{1}{N+\sigma^2 D}\underline{\underline{\boldsymbol{y}}}\,\underline{\boldsymbol{s}}, \frac{\sigma^2}{N+\sigma^2 D}\underline{\underline{\boldsymbol{id}}}_D\right) \underbrace{\sum_{\underline{\boldsymbol{v}_{\mathrm{u}}} \in \{-1,+1\}^{N_u}} \prod_{i=1}^{N_u} \exp\left(\frac{v_{\mathrm{u},i}}{\sigma^2}\underline{\boldsymbol{u}}^T \underline{\boldsymbol{y}_{u,i}}\right)}_{\Delta}
\end{aligned}
$$

Le facteur de droite (Δ) peut en fait se simplifier en un produit de cosinus hyperboliques grâce au raisonnement par récurrence suivant :

$$
\begin{aligned}
\Delta &= \sum_{v_{\mathrm{u},2},\ldots,v_{\mathrm{u},N_u} \in \{-1,1\}^{N_u-1}} \prod_{i=2}^{N_u} \exp\left(\frac{v_{\mathrm{u},i}}{\sigma^2}\underline{\boldsymbol{u}}^T \underline{\boldsymbol{y}_{u,i}}\right)\left(\exp\left(\frac{1}{\sigma^2}\underline{\boldsymbol{u}}^T \underline{\boldsymbol{y}_{u,1}}\right) + \exp\left(-\frac{1}{\sigma^2}\underline{\boldsymbol{u}}^T \underline{\boldsymbol{y}_{u,1}}\right)\right) \\
&= 2\cosh\left(\frac{1}{\sigma^2}\underline{\boldsymbol{u}}^T \underline{\boldsymbol{y}_{u,1}}\right) \sum_{v_{\mathrm{u},2},\ldots,v_{\mathrm{u},N_u} \in \{-1,1\}^{N_u-1}} \prod_{i=2}^{N_u} \exp\left(\frac{v_{\mathrm{u},i}}{\sigma^2}\underline{\boldsymbol{u}}^T \underline{\boldsymbol{y}_{u,i}}\right) \\
&\ldots \\
&= 2^{N_u} \prod_{i=1}^{N_u} \cosh\left(\frac{1}{\sigma^2}\underline{\boldsymbol{u}}^T \underline{\boldsymbol{y}_{u,i}}\right)
\end{aligned}
$$

Ceci prouve la formule (4.2). Pour la loi (4.3), les calculs sont analogues à ceux menant à l'équation (3.2), en utilisant cette fois-ci la densité (4.5).

$$
\begin{aligned}
p(\underline{\underline{\boldsymbol{y}}}_+|v_+,\underline{\underline{\boldsymbol{y}}},\underline{\boldsymbol{s}}) &= \int_{\underline{\boldsymbol{u}}} p(\underline{\underline{\boldsymbol{y}}}_+|\underline{\boldsymbol{u}},v_+)\,\mathrm{d}p(\underline{\boldsymbol{u}}|\underline{\underline{\boldsymbol{y}}},\underline{\boldsymbol{s}}) \qquad (4.8)\\
&\propto \int_{\underline{\boldsymbol{u}}} \exp\left(-\frac{\left\|\underline{\underline{\boldsymbol{y}}}_+ - v_+\underline{\boldsymbol{u}}\right\|^2}{2\sigma^2}\right) \int_{\underline{\boldsymbol{v}}} \exp\left(-\frac{N+\sigma^2 D}{2\sigma^2}\left(\underline{\boldsymbol{u}}^T\underline{\boldsymbol{u}} - \frac{2}{N+\sigma^2 D}\underline{\boldsymbol{u}}^T\underline{\underline{\boldsymbol{y}}}\,\underline{\boldsymbol{v}}\right)\right)\mathrm{d}p(\underline{\boldsymbol{v}}|\underline{\boldsymbol{s}})\,\mathrm{d}\underline{\boldsymbol{u}}\\
&\propto \sum_{v\in\mathcal{V}(\underline{\boldsymbol{s}})} \int_{\underline{\boldsymbol{u}}} \exp\left(-\frac{\left\|\underline{\underline{\boldsymbol{y}}}_+ - v_+\underline{\boldsymbol{u}}\right\|^2}{2\sigma^2}\right) \exp\left(-\frac{N+\sigma^2 D}{2\sigma^2}\left(\underline{\boldsymbol{u}}^T\underline{\boldsymbol{u}} - 2\underline{\boldsymbol{u}}^T\frac{1}{N+\sigma^2 D}\sum_{i=1}^{N} v_i\underline{\underline{\boldsymbol{y}}}_i\right)\right)\mathrm{d}\underline{\boldsymbol{u}}
\end{aligned}
$$

On intègre selon $\underline{\boldsymbol{u}}$ et on aboutit à la forme proposée. □

Remarque. La forme (4.1) de la distribution a posteriori a l'inconvénient d'avoir un très grand nombre de termes, de l'ordre de $2^{(1-\eta)N}$. La combinatoire de ce problème peut facilement exploser, alors qu'avec la forme (4.2), on a seulement un produit d'environ $(1-\eta)N$ facteurs.

Remarque. On remarque que dans le cas où l'apprentissage est supervisé ($s_i \neq 0\ \forall i$ et $\mathcal{V}(\underline{\boldsymbol{s}}) = \{\underline{\boldsymbol{s}}\}$), la formule (4.1) se réduit à la formule (3.1) puisque dans ce cas, $\underline{\boldsymbol{U}}_{\mathrm{MSE}} = \frac{1}{N+\sigma^2 D}\underline{\underline{\boldsymbol{Y}}}\,\underline{\boldsymbol{V}}$. Dans le cas semi-supervisé, les différentes densités normales correspondant aux éléments de $\mathcal{V}(\underline{\boldsymbol{s}})$ sont pondérées par le facteur $w(\underline{\boldsymbol{v}})$ qui favorise les grandes normes de $\underline{\underline{\boldsymbol{y}}}\,\underline{\boldsymbol{v}}$ donc en quelque sorte les vecteurs qui « sortent du lot » et différencient un maximum les deux classes.

On déduit de la densité (4.1) le corollaire suivant :

Corollaire 4.1.4

On appelle $\underline{\boldsymbol{\Omega}}(\underline{\boldsymbol{s}})$ une variable aléatoire discrète prenant ses valeurs dans $\mathcal{V}(\underline{\boldsymbol{s}})$ selon la loi de probabilité définie par $w(\underline{\boldsymbol{v}})$ dans la définition 4.1.2. On appelle $\underline{\boldsymbol{Z}}$ une variable aléatoire suivant la loi $\underline{\boldsymbol{Z}} \sim \mathcal{N}_D\left(0, \underline{\underline{\boldsymbol{id}}}_D\right)$, indépendante de $\underline{\boldsymbol{\Omega}}(\underline{\boldsymbol{s}})$. Alors $\underline{\boldsymbol{U}}|\underline{\underline{\boldsymbol{Y}}},\underline{\boldsymbol{S}}$ suit la même loi que

$$
\{\underline{\boldsymbol{U}}|\underline{\underline{\boldsymbol{Y}}} = \underline{\underline{\boldsymbol{y}}},\underline{\boldsymbol{S}} = \underline{\boldsymbol{s}}\} \overset{\mathcal{D}}{=} \frac{1}{N+\sigma^2 D}\underline{\underline{\boldsymbol{y}}}\,\underline{\boldsymbol{\Omega}}(\underline{\boldsymbol{s}}) + \frac{\sigma}{\sqrt{N+\sigma^2 D}}\underline{\boldsymbol{Z}}
$$

Bien que nous ayons réussi à les mener jusqu'ici, les calculs ne permettent pas d'aboutir à une expression simple du classifieur bayésien comme dans le cas supervisé, au moyen par exemple de la formule (2.4). Il est même difficile d'avoir une expression de $\mathbb{E}\left[\underline{\boldsymbol{U}}|\underline{\underline{\boldsymbol{Y}}},\underline{\boldsymbol{S}}\right]$ car les deux formules explicites que nous avons de la loi $\underline{\boldsymbol{U}}|\underline{\underline{\boldsymbol{Y}}},\underline{\boldsymbol{S}}$ sont soit fortement combinatoires soit difficiles à intégrer en grande dimension. Des simplifications vont nécessairement devoir être menées pour aboutir à un résultat exploitable.

4.2 Classification semi-supervisée bayésienne approchée

Pour l'étude qui suit, on définit les quantités suivantes :

Définition 4.2.1

On appelle $\underline{\underline{Q_u}}$ la résolvante de la matrice $\underline{\underline{y_u}}\,\underline{\underline{y_u}}^T$ (qui est, à normalisation près, la matrice de covariance empirique des données non-étiquetées)

$$\begin{aligned}\underline{\underline{Q_u}} : \mathbb{R} &\longrightarrow \mathbb{R}^{D\times D}\\ t &\longmapsto \left(\underline{\underline{y_u}}\,\underline{\underline{y_u}}^T - t\,\underline{\underline{id}}_D\right)^{-1}\end{aligned}$$

On définit ensuite $\underline{\underline{S_u}}$ la matrice suivante :

$$\underline{\underline{S_u}} \triangleq -\frac{1}{\sigma^2}\left(\underline{\underline{Q_u}}\left(\sigma^2(N+\sigma^2 D)\right)\right)^{-1} = (N+\sigma^2 D)\underline{\underline{id}}_D - \frac{1}{\sigma^2}\underline{\underline{y_u}}\,\underline{\underline{y_u}}^T \tag{4.9}$$

Approximation 4.2.2

Sous certaines hypothèses qu'il faudrait clarifier et quantifier dans un futur travail, la loi $\underline{U}|\underline{Y},\underline{S}$ s'approche par une loi normale

$$p(\underline{u}|\underline{\underline{y}},\underline{s}) \approx \mathcal{N}_D\left(\underline{\underline{S_u}}^{-1}\underline{\underline{y}}\,\underline{s}, \frac{1}{\sigma^2}\underline{\underline{S_u}}\right) \tag{4.10}$$

À condition que $\underline{\underline{id}}_D + \underline{\underline{S_u}} > 0$ (matrice symétrique définie positive, nous reviendrons à cette hypothèse à la section 4.3), on a alors

$$p(\underline{y_+}|v_+,\underline{\underline{y}},\underline{s}) \approx \mathcal{N}_D\left(\underline{y_+}; v_+\underline{\underline{S_u}}^{-1}\underline{\underline{y}}\,\underline{s}, \underline{\underline{id}}_D + \underline{\underline{S_u}}^{-1}\right) \tag{4.11}$$

et on en déduit le classifieur semi-supervisé approché suivant :

$$f_{\underline{\underline{y}},\underline{s}}(\underline{y_+}) = \text{signe}\left(\underline{y_+}^T\left((N+1+\sigma^2 D)\underline{\underline{id}}_D - \frac{1}{\sigma^2}\underline{\underline{y_u}}\,\underline{\underline{y_u}}^T\right)^{-1}\underline{\underline{y}}\,\underline{s}\right) \tag{4.12}$$

Dans les deux paragraphes qui suivent, nous allons par deux façons différentes aboutir à l'approximation (4.10). Dans le troisième paragraphe, nous montrerons l'approximation (4.11) et nous aboutirons au classifieur semi-supervisé approché.

4.2.1 Approximation par une intégrale continue sur une sphère

On commence par reprendre la forme de l'équation (4.7)

$$p(\underline{u}|\underline{\underline{y}},\underline{s}) \propto \exp\left(-\frac{N+\sigma^2 D}{2\sigma^2}\left(\underline{u}^T\underline{u} - 2\underline{u}^T\frac{1}{N+\sigma^2 D}\underline{\underline{y}}\,\underline{s}\right)\right)\sum_{\underline{v_u}\in\{-1,+1\}^{N_u}}\exp\left(\frac{1}{\sigma^2}\underline{u}^T\underline{\underline{y_u}}\,\underline{v_u}\right)$$

Approximation 1 : on suppose sans le justifier que la somme sur l'hypercube $\{-1,+1\}^{N_u}$ est une bonne approximation discrète d'une intégrale sur la sphère $\sqrt{N_u}\mathbb{S}^{N_u-1}$ (autrement dit, on assimile la mesure uniforme discrète sur $\{-1,+1\}^{N_u}$ à la mesure uniforme sur la sphère $\mathrm{d}\mu$). Ceci mène à la forme suivante :

$$p(\underline{u}|\underline{\underline{y}},\underline{s}) \propto \exp\left(-\frac{N+\sigma^2 D}{2\sigma^2}\underline{u}^T\underline{u} + \frac{1}{\sigma^2}\underline{u}^T\underline{\underline{y}}\,\underline{s}\right)\int_{\underline{v_u}\in\sqrt{N_0}\mathbb{S}^{N_0-1}}\exp\left(\frac{1}{\sigma^2}(\underline{\underline{y_u}}^T\underline{u})^T\underline{v_u}\right)\,\mathrm{d}\mu(\underline{v_u})$$

On se place maintenant dans une base orthonormée $(\underline{e_1}, ..., \underline{e_{N_u}})$ de $\mathbb{R}^{N_u}$ où $\underline{e_1} = \frac{\underline{\underline{y_u}}^T \underline{u}}{\|\underline{\underline{y_u}}^T \underline{u}\|}$, dans laquelle tout vecteur $\underline{v_u} \in \mathbb{R}^{N_u}$ se décompose alors ainsi :

$$\underline{v_u} = \frac{(\underline{\underline{y_u}}^T \underline{u})^T \underline{v_u}}{\|\underline{\underline{y_u}}^T \underline{u}\|} \underline{e_1} + v_{u,2}\underline{e_2} + \ldots + v_{u,N_u}\underline{e_{N_u}}$$

de sorte que

$$p(\underline{u}|\underline{\underline{y}}, \underline{s}) \propto \exp\left(-\frac{N+\sigma^2 D}{2\sigma^2}\underline{u}^T\underline{u} + \frac{1}{\sigma^2}\underline{u}^T\underline{\underline{y}}\underline{s}\right) \int_{\underline{v_u} \in \sqrt{N_0}\mathbb{S}^{N_0-1}} \exp\left(\frac{\|\underline{\underline{y_u}}^T \underline{u}\| v_{u,1}}{\sigma^2}\right) \mathrm{d}\mu(\underline{v_u})$$

On peut réduire cette intégrale à une intégrale à une seule dimension en calculant la loi marginale d'une coordonnée du vecteur $\underline{v_u}$ uniformément distribué sur la sphère [18] :

$$p(\underline{u}|\underline{\underline{y}}, \underline{s}) \propto \exp\left(-\frac{N+\sigma^2 D}{2\sigma^2}\underline{u}^T\underline{u} + \frac{1}{\sigma^2}\underline{u}^T\underline{\underline{y}}\underline{s}\right) \int_{-\sqrt{N_0}}^{\sqrt{N_0}} \exp\left(\frac{\|\underline{\underline{y_u}}^T \underline{u}\| t}{\sigma^2}\right) \left(1 - \frac{t^2}{N_0}\right)^{\frac{N_0-1}{2}-1} \mathrm{d}t$$

Approximation 2 : en grande dimension, la loi d'une marginale de $\underline{v_u}$ s'approche d'une distribution normale [18] (pour le voir, il suffit d'effectuer un développement limité dans la limite $N_0 \to +\infty$), ce qui permet de remplacer informellement :

$$p(\underline{u}|\underline{\underline{y}}, \underline{s}) \propto \exp\left(-\frac{N+\sigma^2 D}{2\sigma^2}\underline{u}^T\underline{u} + \frac{1}{\sigma^2}\underline{u}^T\underline{\underline{y}}\underline{s}\right) \int_{-\sqrt{N_0}}^{\sqrt{N_0}} \exp\left(\frac{\|\underline{\underline{y_u}}^T \underline{u}\| t}{\sigma^2}\right) \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t$$

Approximation 3 : on termine le calcul en modifiant le domaine d'intégration pour l'étendre à $\mathbb{R}$ tout entier :

$$\begin{aligned} p(\underline{u}|\underline{\underline{y}}, \underline{s}) &\propto \exp\left(-\frac{N+\sigma^2 D}{2\sigma^2}\underline{u}^T\underline{u} + \frac{1}{\sigma^2}\underline{u}^T\underline{\underline{y}}\underline{s}\right) \int_{-\infty}^{+\infty} \exp\left(\frac{\|\underline{\underline{y_u}}^T \underline{u}\| t}{\sigma^2}\right) \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t \\ &\propto \exp\left(-\frac{N+\sigma^2 D}{2\sigma^2}\underline{u}^T\underline{u} + \frac{1}{\sigma^2}\underline{u}^T\underline{\underline{y}}\underline{s}\right) \exp\left(\frac{1}{2\sigma^4}\underline{u}^T\underline{\underline{y_u}}\,\underline{\underline{y_u}}^T\underline{u}\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\underline{u}^T\underline{\underline{S_u}}\underline{u} + \frac{1}{\sigma^2}\underline{u}^T\underline{\underline{y}}\underline{s}\right) \end{aligned}$$

Ceci conduit bien à l'équation (4.10).

4.2.2 Approximation par une densité gaussienne

Pour ce calcul, on commence par repartir encore une fois de la forme (4.7)

$$p(\underline{u}|\underline{\underline{y}}, \underline{s}) \propto \exp\left(-\frac{N+\sigma^2 D}{2\sigma^2}\left(\underline{u}^T\underline{u} - 2\underline{u}^T\frac{1}{N+\sigma^2 D}\underline{\underline{y}}\underline{s}\right)\right) \sum_{\underline{v_u} \in \{-1,+1\}^{N_u}} \exp\left(\frac{1}{\sigma^2}\underline{u}^T\underline{\underline{y_u}}\underline{v_u}\right)$$

L'idée est alors de supposer qu'à $\underline{\underline{y}}$ et $\underline{s}$ fixé, la somme de droite est en fait un tirage de Monte-Carlo de la variable aléatoire suivante (fonction de la variable aléatoire $\underline{V_u} \sim p_{\underline{V_u}|\underline{S}=\underline{s}}$ qui est en fait une loi uniforme sur $\{-1,+1\}^{N_u}$)

$$\underline{\underline{A}}(\underline{\underline{y_u}}, \underline{V_u}) \triangleq \underline{\underline{y_u}}\underline{V_u}$$

que l'on suppose suivre une loi normale (car il s'agit d'une somme de N_u variables aléatoires indépendantes) dont on connaît les deux premiers moments par le calcul.

Espérance et variance de $\underline{A}(\underline{\underline{y_u}}, \underline{V_u})$: Fixons $N_u \in \mathbb{N}$, $\underline{\underline{y_u}} \in \mathbb{R}^{N_u \times D}$, et déterminons l'espérance et la variance de la variable aléatoire $\underline{A}(\underline{\underline{y_u}}, \underline{V_u})$ où $\underline{V_u}$ suit la loi uniforme sur $\{-1, +1\}^{N_u}$. Notons que selon cette loi, les (V_i) sont tous indépendants entre eux, ce qui donne les égalités suivantes :

$$\mathbb{E}\left[\underline{A}(\underline{\underline{y_u}}, \underline{V_u})\right] = \sum_{i=1}^{N} \mathbb{E}\left[V_{u,i}\right] \underline{y_{u,i}} = \underline{0}$$
$$\mathbb{V}\left[\underline{A}(\underline{\underline{y_u}}, \underline{V_u})\right] = \sum_{i=1}^{N} \mathbb{V}\left[V_{u,i}\right] \underline{y_{u,i}}\ \underline{y_{u,i}}^T = \underline{\underline{y_u}}\ \underline{\underline{y_u}}^T \geq 0$$

Approximation par une gaussienne : On suppose dans ce qui suit que lorsque l'on parcourt $\{-1, +1\}^{N_u}$, $\underline{A}$ se comporte comme une gaussienne car il s'agit d'une somme de N variables aléatoires indépendantes (théorème central limite). On caractérise cette loi normale grâce à son espérance et sa variance calculées ci-dessus

$$\underline{A}(\underline{\underline{y_u}}, \underline{V_u}) \sim \mathcal{N}_D\left(\underline{0}, \underline{\underline{y_u}}\ \underline{\underline{y_u}}^T\right)$$

Simplification de la loi de $\underline{U}, \underline{Y}, \underline{S}$: Avec l'hypothèse simplificatrice précédente, on calcule

$$\begin{aligned} p(\underline{u}|\underline{\underline{y}}, \underline{s}) &\propto \exp\left(-\frac{N+\sigma^2 D}{2\sigma^2}\underline{u}^T\underline{u} + \frac{1}{\sigma^2}\underline{u}^T\underline{\underline{y}}\,\underline{s}\right) \int_{\underline{a}} \exp\left(\frac{1}{\sigma^2}\underline{u}^T\underline{a} - \frac{1}{2}\underline{a}^T\left(\underline{\underline{y_u}}\ \underline{\underline{y_u}}^T\right)^{-1}\underline{a}\right) \mathrm{d}\underline{a} \\ &\propto \exp\left(-\frac{N+\sigma^2 D}{2\sigma^2}\underline{u}^T\underline{u} + \frac{1}{\sigma^2}\underline{u}^T\underline{\underline{y}}\,\underline{s}\right) \exp\left(\frac{1}{2\sigma^4}\underline{u}^T\underline{\underline{y_u}}\ \underline{\underline{y_u}}^T\underline{u}\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\underline{u}^T\underline{\underline{S_u}}\,\underline{u} + \frac{1}{\sigma^2}\underline{u}^T\underline{\underline{y}}\,\underline{s}\right) \end{aligned}$$

Ceci conduit également bien à (4.10).

4.2.3 Classifieur approché

Supposons dans la suite qu'on ait en effet

$$p(\underline{u}|\underline{\underline{y}}, \underline{s}) \propto \exp\left(-\frac{1}{2\sigma^2}\underline{u}^T\underline{\underline{S_u}}\,\underline{u} + \frac{1}{\sigma^2}\underline{u}^T\underline{\underline{y}}\,\underline{s}\right)$$

et déduisons-en la loi $p(\underline{y_+}|v_+, \underline{\underline{y}}, \underline{s})$ grâce à la formule (4.8) :

$$p(\underline{y_+}|v_+, \underline{\underline{y}}, \underline{s}) \propto \exp\left(-\frac{1}{2\sigma^2}\underline{y_+}^T\underline{y_+}\right) \int_{\underline{u}} \exp\left(-\frac{1}{2\sigma^2}\left[\underline{u}^T\left(\underline{\underline{id}}_D + \underline{\underline{S_u}}\right)\underline{u} - 2\underline{u}^T\left(v_+\underline{y_+} + \underline{\underline{y}}\,\underline{s}\right)\right]\right) \mathrm{d}\underline{u}$$

À supposer que $(\underline{\underline{id}}_D + \underline{\underline{S}}_u) > 0$ (c'est-à-dire que toutes ses valeurs propres sont strictement positives), on peut intégrer la densité normale apparaissant sous l'intégrale et obtenir (la dernière égalité s'obtient par exemple avec l'identité de Woodbury)

$$\begin{aligned} p(\underline{y_+}|v_+, \underline{\underline{y}}, \underline{s}) &\propto \exp\left(-\frac{1}{2\sigma^2}\left[\underline{y_+}^T\left(\underline{\underline{id}}_D - \left(\underline{\underline{id}}_D + \underline{\underline{S_u}}\right)^{-1}\right)\underline{y_+} - 2v_+\underline{y_+}^T\left(\underline{\underline{id}}_D + \underline{\underline{S_u}}\right)^{-1}\underline{\underline{y}}\,\underline{s}\right]\right) \\ &\propto \mathcal{N}_D\left(\underline{y_+}; v_+(\underline{\underline{S_u}})^{-1}\underline{\underline{y}}\,\underline{s}, \left(\underline{\underline{id}}_D - \left(\underline{\underline{id}}_D + \underline{\underline{S_u}}\right)^{-1}\right)^{-1}\right) \\ &= \mathcal{N}_D\left(\underline{y_+}; v_+(\underline{\underline{S_u}})^{-1}\underline{\underline{y}}\,\underline{s}, \underline{\underline{id}}_D + \underline{\underline{S_u}}^{-1}\right) \end{aligned}$$

Dans ce cas, le classifieur bayésien défini à l'équation (2.4) est

$$\underline{y_+} \mapsto \text{signe}\left(\underline{y_+}^T\left(\underline{\underline{id}}_D + \underline{\underline{S_u}}\right)^{-1}\underline{\underline{y}}\,\underline{s}\right) = \text{signe}\left(\underline{y_+}^T\left((N+1+\sigma^2 D)\underline{\underline{id}}_D - \frac{1}{\sigma^2}\underline{\underline{y_u}}\ \underline{\underline{y_u}}^T\right)^{-1}\underline{\underline{y}}\,\underline{s}\right)$$

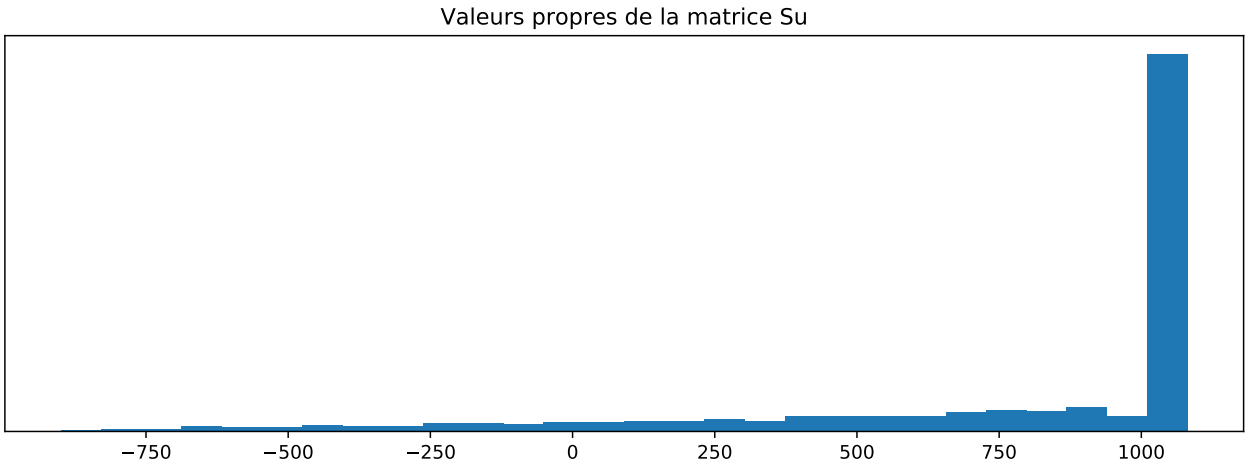

FIGURE 4.1 – Histogramme des valeurs propres de la matrice $\underline{\underline{S_u}}$ pour $N = 400$, $D = 800$, $\sigma = \sqrt{0.9}$, $\eta = 0.1$

4.3 Valeurs propres de la matrice résolvante

Nous avons vu à la partie précédente que, pour aboutir à l'approximation (4.11), il était nécessaire que la matrice $\underline{\underline{\boldsymbol{id}}}_D + \underline{\underline{S_u}}$, qui est une très proche variante de la matrice $\underline{\underline{S_u}}$ définie à l'équation (4.9), ait des valeurs propres strictement positives. Malheureusement, des simulations (script `S05_eiganvalues.py`) suggèrent que ces matrices ont un nombre non-négligeable de valeurs propres négatives, comme on peut le voir à la figure 4.1. Il semble donc que les approximations et inégalités qui ont dû être faites dans les paragraphes précédents soient trop fortes et qu'un travail ultérieur plus poussé doive être mené à bien pour aboutir. Cependant, et aussi étonnant que cela puisse paraître compte tenu de l'invalidité empirique de l'hypothèse, le classifieur approché trouvé à l'équation (4.12) semble fournir une piste prometteuse pour recentrer l'estimateur naïf de $\underline{\boldsymbol{U}}$ donné par $\underline{\underline{\boldsymbol{ys}}} = \sum_{i=1}^{N} s_i \underline{\boldsymbol{y}_i}$. Nous allons le voir dans le paragraphe qui suit.

4.4 Performance de la classification semi-supervisée bayésienne approchée

Le classifieur approché trouvé à l'équation (4.12) suggère de recentrer l'estimateur obtenu naturellement en moyennant les échantillons en le multipliant à gauche par la matrice suivante, qui est une variante redimensionnée de la matrice résolvante des données non-étiquetées, prise en un point particulier :

$$(\underline{\underline{\boldsymbol{id}}}_D + \underline{\underline{S_u}})^{-1} = -\sigma^2 \underline{\underline{Q_u}} \left(\sigma^2(N + 1 + \sigma^2 D)\right)$$

On se propose d'étudier une famille de classifieurs bayésiens approchés en ajoutant un paramètre de recentrage λ variant entre 0 et 1 pour tester différentes intensités de la transformation (lorsque $\lambda \to 0$, la résolvante ressemble à la matrice identité (pas de transformation appliquée à l'estimateur supervisé) alors que lorsque $\lambda \to 1$, le recentrage est important) :

$$[0, 1] \ni \lambda \mapsto -\underline{\underline{Q_u}} \left(\frac{\sigma^2(N + 1 + \sigma^2 D)}{\lambda} \right) \sum_{i=1}^{N} s_i \underline{\boldsymbol{y}_i} \tag{4.13}$$

On trace à la figure 4.2 la performance du classifieur approché en changeant la valeur de λ. Le script `S07_approximate_classifier.py` permet de reproduire l'expérience. On observe que lorsque la valeur du recentrage augmente, la performance du classifieur commence par s'améliorer (le risque diminue) par rapport à celle du classifieur obtenu uniquement avec les points étiquetés. Cependant, à

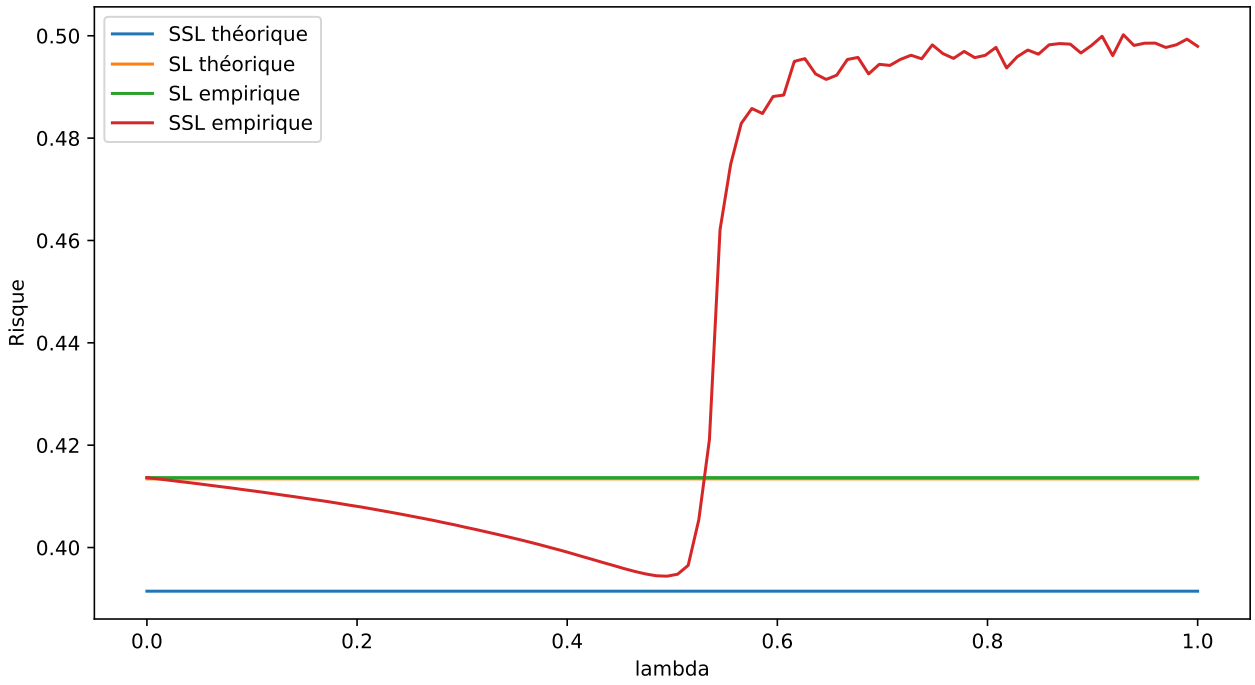

FIGURE 4.2 – Performance obtenue pour différentes valeurs du recentrage λ, avec $N = 400, D = 800, \sigma = \sqrt{0.9}, \eta = 0.1$, moyenne empirique sur 200 simulations. On compare la performance avec celle du classifieur supervisé, pour lequel on supprime toutes les données non-étiquetées. Le risque diminue progressivement en augmentant l'intensité du recentrage, puis on observe une transition de phase entre 0 et 1.

partir d'une valeur de λ située entre 0 et 1, la performance se dégrade et finit par donner un classifieur totalement inefficace (50% de mauvaises classifications). Ce comportement est étonnant au premier abord car la formule obtenue plus haut suggère de prendre $\lambda = 1$ et devrait permettre d'obtenir la performance optimale bayésienne. Il faut toutefois se souvenir que la formule découle d'une série d'approximations. Cette figure fait alors sens : on ne parvient pas à atteindre la performance optimale prédite par Lelarge et Miolane car on n'a pas trouvé le classifieur optimal. D'où vient alors la brutale dégradation des performances au milieu de l'intervalle $[0, 1]$? Pour le comprendre, on peut réécrire l'équation (4.13) de façon plus explicite (à un facteur scalaire positif près sans intérêt car le classifieur requiert seulement le signe du produit scalaire suivant) :

$$[0,1] \ni \lambda \mapsto \left(\sigma^2(N+1+\sigma^2 D)\underline{\underline{\boldsymbol{id}}}_D - \lambda \underline{\underline{\boldsymbol{y_u}}}\,\underline{\underline{\boldsymbol{y_u}}}^T\right)^{-1} \sum_{i=1}^{N} s_i \underline{\boldsymbol{y}}_i$$

En traçant l'histogramme des valeurs propres de $\lambda \underline{\underline{\boldsymbol{y_u}}}\,\underline{\underline{\boldsymbol{y_u}}}^T$ (figure 4.3, script `S06_resolvent_bulk.py`) et en les comparant à $\sigma^2(N+1+\sigma^2 D)$, on comprend alors l'origine de la transition de phase : elle a pour origine le fait qu'à partir d'une certaine valeur de dilatation de $\underline{\underline{\boldsymbol{y_u}}}\,\underline{\underline{\boldsymbol{y_u}}}^T$, la valeur propre $\sigma^2(N+1+\sigma^2 D)$ entre dans le bulk des valeurs propres de $\underline{\underline{\boldsymbol{y_u}}}\,\underline{\underline{\boldsymbol{y_u}}}^T$ correspondant à une distribution de Marchenko-Pastur. Si l'on étudie le bon modèle de matrice aléatoire $(\underline{\underline{\boldsymbol{U}}} + \sigma\underline{\underline{\boldsymbol{Z}}})(\underline{\underline{\boldsymbol{U}}} + \sigma\underline{\underline{\boldsymbol{Z}}})^T$, il doit être possible de déterminer explicitement la valeur de λ qui donne lieu à la transition de phase.

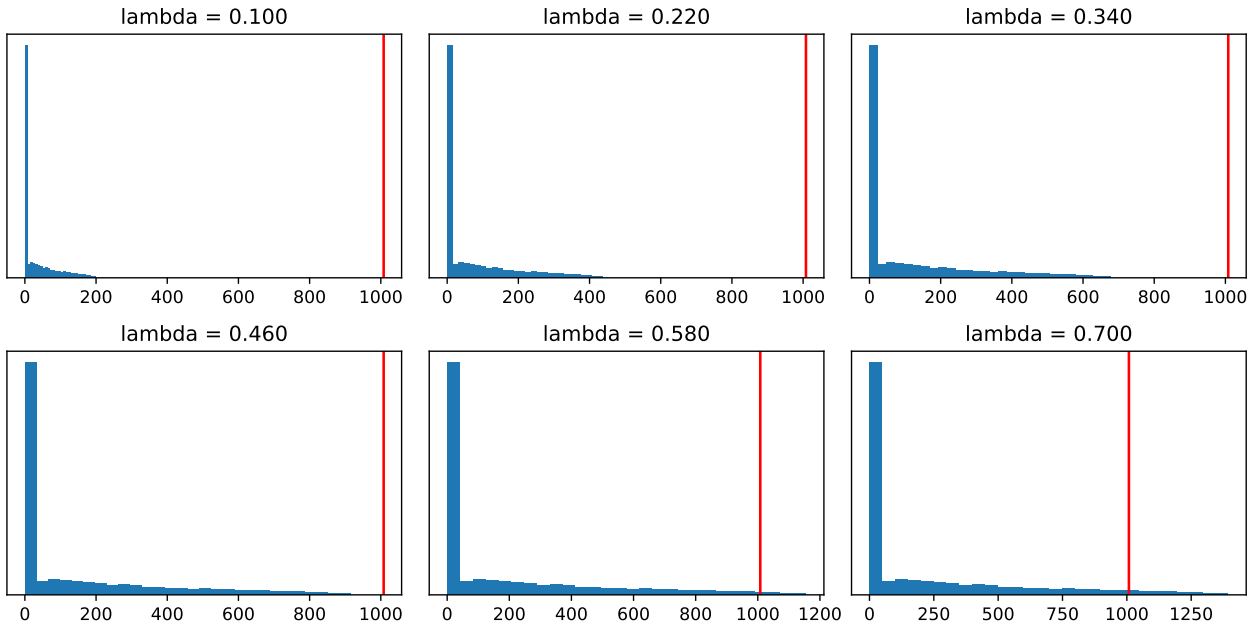

FIGURE 4.3 – Histogrammes des valeurs propres de $\lambda \underline{\underline{\boldsymbol{y_u}}}\ \underline{\underline{\boldsymbol{y_u}}}^T$ pour différentes valeurs de la dilatation λ (il s'agit du paramètre de recentrage défini plus haut), $N = 400, D = 800, \sigma = \sqrt{0.9}, \eta = 0.1$. La ligne rouge représente l'unique valeur propre (de multiplicité D) de $\sigma^2(N+1+\sigma^2 D)\underline{\underline{\boldsymbol{id}}}_D$. Plus λ est grand, plus le bulk des valeurs propres de $\lambda \underline{\underline{\boldsymbol{y_u}}}\ \underline{\underline{\boldsymbol{y_u}}}^T$ avance jusqu'à rencontrer les valeurs propres de la matrice identité.

Remarque (Lien avec Mai/Couillet (2020))**.** Dans [10], les auteurs aboutissent à une formule très similaire à la nôtre pour estimer les étiquettes des points non-étiquetés (équation numérotée 12 dans [10]) :

$$\left(\alpha \underline{\underline{\boldsymbol{id}}}_{N_u} - \underline{\underline{\boldsymbol{W}_{\mathrm{u,u}}}}\right)^{-1} \underline{\underline{\boldsymbol{W}_{\mathrm{u,l}}}}\ \underline{\boldsymbol{f}_{\mathrm{l}}} \quad , \quad \alpha > 0$$

Dans ce cas, $\underline{\boldsymbol{f}_l}$ donne les étiquettes des points connus et la matrice $\underline{\underline{\boldsymbol{W}}}$ est une matrice de similarités de taille $N \times N$ entre les N points 2 à 2. Elle est donc symétrique et peut être découpée en 3 blocs : similarités entre les points étiquetés $\underline{\underline{\boldsymbol{W}_{\mathrm{l,l}}}}$, similarités entre les points non-étiquetés $\underline{\underline{\boldsymbol{W}_{\mathrm{u,u}}}}$ et similarités entre les deux types de points $\underline{\underline{\boldsymbol{W}_{\mathrm{u,l}}}}$. Une étude est menée dans l'article pour déterminer la valeur de α optimale.

Chapitre 5

Approche informationnelle

Résumé du chapitre

L'étude menée au chapitre précédent suggère que des éléments nous échappent encore pour bien comprendre le résultat fondamental prouvé par Lelarge et Miolane dans [7]. Nous proposons donc de synthétiser des éléments et résultats récents utilisant le formalisme de la théorie de l'information et de la méthode des répliques. Nous commençons par présenter quelques résultats sur les canaux gaussiens scalaires qui sont en fait fondamentaux et ce même pour des problèmes d'estimation muiltivariés. En effet, sous certaines conditions, ceux-ci peuvent se ramener à des problèmes à une dimension, parfois couplés, dont l'amplitude du bruit est à déterminer : c'est ce qui a été prouvé mathématiquement dans deux travaux récents dont nous extrayons dans la suite quelques résultats choisis avant de les comparer et de proposer des pistes pour de futurs travaux.

5.1 Canaux gaussiens unidimensionnels

Les calculs d'estimation effectués dans les références que nous allons citer dans la suite nécessitent d'introduire (ou rappeler) certaines notions de théorie de l'information. Ce paragraphe est une synthèse de différents éléments de [13, 16] que nous avons repris, uniformisés et étendus. Pour plus de détails sur les canaux gaussiens et pour une présentation beaucoup moins spécifique mais plus complète, on se référera au chapitre 9 de [2].

5.1.1 Définitions

Soit X une variable aléatoire réelle de variance finie, suivant la loi P_X et appelée « signal », Z une variable aléatoire réelle indépendante de X, suivant une loi normale centrée réduite et appelée « bruit » et $\gamma > 0$ un réel appelé « rapport de puissances » (si X est de variance 1, son inverse est appelé "snr" en anglais, pour "signal-to-noise ratio"). Dans un canal gaussien, on observe le signal amplifié par γ et auquel on ajoute le bruit gaussien additif Z :

$$Y = \sqrt{\gamma}X + Z \tag{5.1}$$

On se pose la question suivante : en moyenne, quelle est la meilleure façon d'estimer X à partir de son observation Y et quelle sera la performance associée ? Les réponses pourront varier en fonction de la distribution d'entrée P_X ainsi que du réel γ. Si l'on choisit comme critère l'erreur quadratique moyenne entre le tirage de X et sa reconstruction, on cherche donc une fonction $f : \mathbb{R} \mapsto \mathbb{R}$ telle que

$$f = \underset{f:\mathbb{R}\mapsto\mathbb{R}}{\operatorname{argmin}} \, \mathbb{E}\left[(X - f(Y))^2\right]$$

En réécrivant le critère à minimiser, on a

$$\begin{aligned}\mathbb{E}\left[(X-f(Y))^2\right] &= \mathbb{E}\left[\mathbb{E}\left[(X-f(Y))^2 \,\middle|\, Y\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[X^2 \mid Y\right] - 2f(Y)\mathbb{E}\left[X \mid Y\right] + f(Y)^2\right]\end{aligned}$$

Pour minimiser l'espérance, il faut et il suffit de minimiser chaque terme en Y, donc il faut choisir $f(y)$ qui minimise $\mathbb{E}\left[X^2 \mid Y=y\right] - 2f(y)\mathbb{E}\left[X \mid Y=y\right] + f(y)^2$. On définit l'estimateur des moindres carrés de X à partir de l'observation $Y=y$ par

$$X_{\text{MSE}} \triangleq \mathbb{E}\left[X \mid Y\right]$$

Puisque nous avons identifié l'estimateur optimal de X au sens de l'erreur quadratique moyenne, nous pouvons désormais calculer l'erreur associée, qui donne donc l'erreur minimale que l'on peut obtenir lors de l'estimation de X à partir de son observation Y via un canal gaussien de paramètres P_X et γ :

Définition 5.1.1: Erreur quadratique moyenne minimale

$$\text{mmse}_X(\gamma) \triangleq \mathbb{E}\left[(X-\mathbb{E}\left[X \mid Y\right])^2\right] \tag{5.2}$$

Comme on va le voir dans les paragraphes qui suivent, l'erreur quadratique moyenne minimale est intrinsèquement liée à l'information d'un canal que l'on définit ainsi :

Définition 5.1.2: Information mutuelle

On définit l'information mutuelle d'un canal comme étant l'information mutuelle entre son entrée et sa sortie :

$$I_X(\gamma) \triangleq I(X;\sqrt{\gamma}X+Z) = \mathbb{E}\left[\log\left(\frac{p(X|Y)}{p(X)}\right)\right] \tag{5.3}$$

Rappelons que l'information mutuelle mesure la quantité d'information que la sortie contient sur l'entrée et qu'elle s'écrit comme la divergence de Kullback-Leibler entre la distribution jointe de (X,Y) et le produit des deux distributions marginales de X et de Y.

Explicitons maintenant la distribution a posteriori de X sachant Y (qui intervient dans l'information), ce qui permet de définir la fonction de partition $\mathcal{Z}(\gamma,Y)$ pour normaliser la probabilité :

$$p_{X|Y}(x|y) = \frac{1}{\mathcal{Z}(\gamma,y)} p_X(x) \exp\left(-\frac{1}{2}\gamma x^2 + \sqrt{\gamma}xy\right) \tag{5.4}$$

Définition 5.1.3: Fonction de partition

On appelle fonction de partition de la probabilité a posteriori la fonction suivante :

$$\mathcal{Z}(\gamma,y) \triangleq \int_x p_X(x) \exp\left(-\frac{1}{2}\gamma x^2 + \sqrt{\gamma}xy\right) \mathrm{d}x \tag{5.5}$$

La fonction de partition permet de définir l'énergie libre du canal monodimensionnel de la façon suivante, analogue à la définition de la quantité ψ_{P_X} définie à l'annexe A.1 de [13].

Définition 5.1.4: Énergie libre

$$\psi_X(\gamma) \triangleq \mathbb{E}\left[\log\left(\mathcal{Z}(\gamma,Y)\right)\right] \tag{5.6}$$

On définit enfin la notion de réplique et la fonction de coïncidence (traduction libre du terme anglais *overlap*) comme étant l'espérance du produit entre X et une de ses répliques :

Définition 5.1.5: Réplique

On appelle dans la suite réplique de X toute variable aléatoire X_i tirée selon la loi a posteriori $p_{X|Y}$:

$$p_{X_i|Y}(x|y) = p_{X|Y}(x|y) \tag{5.7}$$

On représente deux répliques sur le réseau bayésien de la figure 5.1 symbolisant un canal gaussien. Notons que conditionnellement à Y, deux répliques sont indépendantes deux à deux et qu'il en est de même pour X et une de ses répliques.

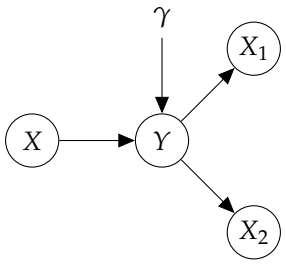

FIGURE 5.1 – Réseau bayésien pour un canal gaussien et deux répliques

Définition 5.1.6: Coïncidence

On définit la coïncidence comme suit :

$$F_X(\gamma) \triangleq \mathbb{E}\left[XX_1\right]$$

On montre que F_X peut aussi s'écrire de façon équivalente :

$$F_X(\gamma) = \mathbb{E}\left[XX_1\right] = \mathbb{E}\left[X\mathbb{E}\left[X \mid Y\right]\right] = \mathbb{E}\left[X_1X_2\right] = \mathbb{E}\left[\mathbb{E}\left[X \mid Y\right]^2\right] \tag{5.8}$$

On prouve dans la démonstration qui suit l'égalité entre les quatre expressions ci-dessus (il s'agit d'une méthode de calcul clé pour comprendre les travaux de Miolane, basée sur une identité probabiliste aussi dénommée identité de Nishimori dans [13]) :

Démonstration. Notons que les passages surmontés d'un numéro sont justifiés à la fin de la démonstration pour éviter de rompre le fil des calculs.

$$\begin{aligned}
\mathbb{E}\left[X\mathbb{E}\left[X \mid Y\right]\right] &= \int_{a,c} a\,\mathbb{E}\left[X \mid Y = c\right]\, \mathrm{d}\mathbb{P}_{X,Y}(a,c) \\
&= \int_{a,b,c} ab\, \mathrm{d}\mathbb{P}_{X|Y}(b,c)\mathrm{d}\mathbb{P}_{X,Y}(a,c) \\
&\overset{(1)}{=} \int_{a,b,c} ab\, \mathrm{d}\mathbb{P}_{X_1|Y}(b,c)\mathrm{d}\mathbb{P}_{X,Y}(a,c) \\
&\overset{(2)}{=} \int_{a,b,c} ab\, \mathrm{d}\mathbb{P}_{X_1|X,Y}(b|a,c)\mathrm{d}\mathbb{P}_{X,Y}(a,c) \\
&\overset{(3)}{=} \int_{a,b,c} ab\, \mathrm{d}\mathbb{P}_{X,X_1,Y}(a,b,c) \\
&= \mathbb{E}\left[XX_1\right]
\end{aligned}$$

En repartant de l'avant-dernière expression (3),

$$\begin{aligned}
\mathbb{E}\left[XX_1\right] &= \int_{a,b,c} ab \, \mathrm{d}\mathbb{P}_{X,X_1,Y}(a,b,c) \\
&= \int_{a,b,c} ab \, \mathrm{d}\mathbb{P}_{X,X_1|Y}(a,b|c)\mathrm{d}\mathbb{P}_Y(c) \\
&\overset{(4)}{=} \int_{a,b,c} ab \, \mathrm{d}\mathbb{P}_{X|Y}(a|c)\mathrm{d}\mathbb{P}_{X_1|Y}(b|c)\mathrm{d}\mathbb{P}_Y(c) \\
&\overset{(5)}{=} \int_{a,b,c} ab \, \mathrm{d}\mathbb{P}_{X_2|Y}(a|c)\mathrm{d}\mathbb{P}_{X_1|Y}(b|c)\mathrm{d}\mathbb{P}_Y(c) \\
&\overset{(6)}{=} \int_{a,b,c} ab \, \mathrm{d}\mathbb{P}_{X_2,X_1|Y}(a,b|c)\mathrm{d}\mathbb{P}_Y(c) \\
&= \int_{a,b,c} ab \, \mathrm{d}\mathbb{P}_{X_2,X_1,Y}(a,b,c) \\
&= \mathbb{E}\left[X_1X_2\right]
\end{aligned}$$

En repartant de l'expression (4),

$$\begin{aligned}
\mathbb{E}\left[XX_1\right] &\overset{(7)}{=} \int_{a,b,c} ab \, \mathrm{d}\mathbb{P}_{X|Y}(a|c)\mathrm{d}\mathbb{P}_{X|Y}(b|c)\mathrm{d}\mathbb{P}_Y(c) \\
&= \int_c \mathbb{E}\left[X \mid Y=c\right]^2 \, \mathrm{d}\mathbb{P}_Y(c) \\
&= \mathbb{E}\left[\mathbb{E}\left[X \mid Y\right]^2\right]
\end{aligned}$$

On justifie les différents passages ainsi :
- (1), (5) et (7) grâce à la formule (5.7) qui définit les répliques X_1 et X_2
- (2) et (4) par indépendance de X_1 et X sachant Y (chaîne de Markov)
- (6) par indépendance de X_1 et X_2 sachant Y (cause commune)

□

5.1.2 Propriétés

Les grandeurs que nous avons définies au paragraphe précédent sont toutes liées par des relations algébriques ou intégrales simples que nous allons expliciter ici.

Les deux relations suivantes sont presque immédiates et on en déduit qu'information et énergie libre sont deux quantités presque équivalentes, et qu'il en est de même pour la MMSE et la coïncidence.

Proposition 5.1.7: Relations I-ψ et F-MMSE

$$I_X(\gamma) = -\psi_X(\gamma) + \frac{1}{2}\gamma\mathbb{E}\left[X^2\right] \tag{5.9}$$

$$\mathrm{mmse}_X(\gamma) = -F_X(\gamma) + \mathbb{E}\left[X^2\right] \tag{5.10}$$

Il existe maintenant des relations intégrales entre l'information et la MMSE d'une part, et de façon équivalente entre énergie libre et coïncidence.

Proposition 5.1.8: Relations I-MMSE et F-ψ

$$\frac{d}{d\gamma} I_X(\gamma) = \frac{1}{2}\mathrm{mmse}_X(\gamma) \tag{5.11}$$

$$\frac{d}{d\gamma} \psi_X(\gamma) = \frac{1}{2} F_X(\gamma) \tag{5.12}$$

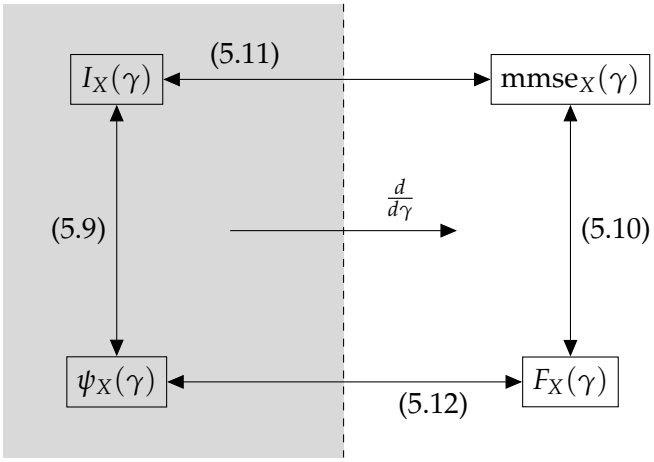

FIGURE 5.2 – Illustration synthétique des propriétés du paragraphe 5.1.2

L'identité I-MMSE, mise en évidence en 2005 par Guo, Shamai et Verdu dans [3] et dérivée de l'identité de de Bruijn (théorème 17.7.2 de [2]) est fréquemment utilisée pour calculer la MMSE de canaux plus complexes où X n'est pas une variable gaussienne (l'estimateur X_{MSE} n'est alors pas calculable directement ou alors compliqué). On la retrouve par exemple à la page 11 de [7], ou encore à la page 11 de [16].

Démonstration. Commençons par réécrire l'énergie libre :

$$\begin{aligned}
\psi_X(\gamma) &= \mathbb{E}\left[\log(\mathcal{Z}(\gamma, Y))\right] \\
&= \mathbb{E}\left[\log \int_x p_X(x) \exp\left(-\frac{1}{2}\gamma x^2 + \sqrt{\gamma} x Y\right) \mathrm{d}x\right] \\
&= \mathbb{E}\left[\log \int_x p_X(x) \exp\left(-\frac{1}{2}\gamma x^2 + \gamma x X + \sqrt{\gamma} x Z\right) \mathrm{d}x\right]
\end{aligned}$$

On dérive alors par rapport à γ sous les deux espérances, ce qui est possible sur l'intervalle $(0, +\infty)$ car on peut dominer deux fois sur tout segment $[a, b], 0 < a < b < +\infty$ par une fonction intégrable (une fois pour l'espérance conditionnelle à l'intérieur puis une fois pour l'espérance totale à l'extérieur).

$$\begin{aligned}
\frac{d}{d\gamma}\psi_X(\gamma) &= \mathbb{E}\left[\mathop{\mathbb{E}}_{X_1 \sim p_{X|Y}}\left[-\frac{1}{2}X_1^2 + X X_1 + \frac{1}{2\sqrt{\gamma}} X_1 Z\right]\right] \\
&= \frac{1}{2}\mathbb{E}\left[-\mathbb{E}\left[X^2 \mid Y\right] + X\mathbb{E}\left[X \mid Y\right] + X\mathbb{E}\left[X \mid Y\right] + \frac{1}{\sqrt{\gamma}} Z\mathbb{E}\left[X \mid Y\right]\right] \\
&= \frac{1}{2}\mathbb{E}\left[-\mathbb{E}\left[X^2 \mid Y\right] + X\mathbb{E}\left[X \mid Y\right] + \frac{1}{\sqrt{\gamma}}\mathbb{E}\left[XY \mid Y\right]\right] \\
&= \frac{1}{2}F_X(\gamma) + \frac{1}{2}\mathbb{E}\left[-X^2 + \frac{1}{\sqrt{\gamma}} XY\right] \\
&= \frac{1}{2}F_X(\gamma)
\end{aligned}$$

□

Les propriétés de ce paragraphe sont résumées graphiquement à la figure 5.2.

Remarque. Miolane montre des propriétés plus complètes sur la coïncidence et l'énergie libre, dans le cadre plus général d'un signal vectoriel et d'un rapport signal sur bruit matriciel (symétrique défini positif). On pourra retrouver les détails à l'annexe A.1 de [13] et les relier aux travaux de Palomar et Verdu [15].

5.1.3 Exemples de canaux gaussiens

Cette section donne trois exemples de canaux gaussiens pour des distributions d'entrée choisies utiles pour l'application ultérieure de ces notions à l'apprentissage semi-supervisé. Elle sert à la fois d'illustration permettant de se familiariser avec les notions introduites et de formulaire pour la suite.

5.1.3.1 Signal gaussien centré réduit

On considère dans ce paragraphe un signal d'entrée $U \sim \mathcal{N}(0,1)$ dans le canal gaussien suivant ($Z \sim \mathcal{N}(0,1)$ indépendant de U) :

$$Y = \sqrt{\gamma}U + Z$$

On calcule très simplement U_{MSE} de la façon suivante, utilisée par Lelarge et Miolane dans [7] : l'espérance conditionnelle peut aussi être interprétée comme la projection au sens du produit scalaire $\langle U \mid Y \rangle = \mathbb{E}[UY]$ de U sur la droite vectorielle engendrée par Y. En normalisant Y par $\|Y\|$ pour obtenir un vecteur unitaire, on calcule :

$$U_{\text{MSE}} = \mathbb{E}[U \mid Y] = \frac{\mathbb{E}[UY]}{\mathbb{E}[Y^2]}Y = \frac{\mathbb{E}\left[\sqrt{\gamma}U^2 + UZ\right]}{\mathbb{E}\left[\gamma U^2 + 2\sqrt{\gamma}UZ + Z^2\right]}Y = \frac{\sqrt{\gamma}}{1+\gamma}Y \tag{5.13}$$

Ceci fournit

$$\text{mmse}_U(\gamma) = \mathbb{E}\left[\left(\frac{1}{1+\gamma}U - \frac{\sqrt{\gamma}}{1+\gamma}Z\right)^2\right] = \frac{1}{1+\gamma} \tag{5.14}$$

Pour la suite, on aura besoin de calculer la distribution a posteriori de $U|Y$ selon la formule (5.4)

$$p(u|y) = \mathcal{N}\left(u; \frac{\sqrt{\gamma}}{1+\gamma}y, \frac{1}{1+\gamma}\right) \tag{5.15}$$

On peut dès lors calculer l'information mutuelle selon la formule (5.3), ce qui donne

$$I_U(\gamma) = \frac{1}{2}\log(1+\gamma) \tag{5.16}$$

On déduit maintenant de (5.15) la fonction de partition selon la définition de la formule (5.5) :

$$\mathcal{Z}(\gamma, y) = \frac{1}{\sqrt{1+\gamma}}\exp\left(\frac{\gamma}{2(1+\gamma)}y^2\right) \tag{5.17}$$

puis l'énergie libre grâce à la définition (5.6)

$$\psi_U(\gamma) = \frac{\gamma}{2} - \frac{1}{2}\log(1+\gamma) \tag{5.18}$$

et enfin la coïncidence avec la définition (5.8) et en exploitant (5.13) :

$$F_U(\gamma) = \frac{\gamma}{1+\gamma} \tag{5.19}$$

On peut remarquer par le calcul, pour se familiariser avec elles, que les relations I-MMSE, F-MMSE, F-ψ et I-ψ sont vérifiées.

5.1.3.2 Signal Rademacher (Bernoulli centré réduit)

On considère désormais $V = 2B - 1$ où B est une variable aléatoire de Bernoulli de paramètre $1/2$: V est ainsi centrée, de variance unité et prend ses valeurs dans $\{-1, 1\}$. Avec les notations des canaux gaussiens et en utilisant les propriétés trouvées précédemment, on a :

$$\mathcal{Z}(\gamma, Y) = \exp\left(-\frac{1}{2}\gamma\right)\cosh\left(\sqrt{\gamma}Y\right) \tag{5.20}$$

$$\psi_V(\gamma) = -\frac{1}{2}\gamma + \mathbb{E}\left[\log\cosh\left(\sqrt{\gamma}Z + \gamma\right)\right] \tag{5.21}$$

$$I_V(\gamma) = \gamma - \mathbb{E}\left[\log\cosh\left(\sqrt{\gamma}Z + \gamma V\right)\right] = \gamma - \mathbb{E}\left[\log\cosh\left(\sqrt{\gamma}Z + \gamma\right)\right] \tag{5.22}$$

$$F_V(\gamma) = \mathbb{E}\left[\tanh\left(\sqrt{\gamma}Z + \gamma\right)\right] \tag{5.23}$$

$$\mathrm{mmse}_V(\gamma) = 1 - \mathbb{E}\left[\tanh\left(\sqrt{\gamma}Z + \gamma\right)\right] \tag{5.24}$$

Démonstration. Les passages non-triviaux de ce formulaire sont :

$$\mathbb{E}\left[\log\cosh\left(\sqrt{\gamma}Z + \gamma V\right)\right] = \mathbb{E}\left[\log\cosh\left(\sqrt{\gamma}Z + \gamma\right)\right] \tag{5.22'}$$

$$\frac{d}{d\gamma}\mathbb{E}\left[\log\cosh\left(\sqrt{\gamma}Z + \gamma\right)\right] = \frac{1}{2}\left(1 + \mathbb{E}\left[\tanh\left(\sqrt{\gamma}Z + \gamma\right)\right]\right) \tag{5.23'}$$

La première égalité s'obtient comme suit, en utilisant successivement que Z et $-Z$ ont même loi, puis la parité de la fonction cosh :

$$\begin{aligned}
\mathbb{E}\left[\log\cosh\left(\sqrt{\gamma}Z + \gamma V\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\log\cosh\left(\sqrt{\gamma}Z + \gamma V\right) \mid V\right]\right] \\
&= \frac{1}{2}\left(\mathbb{E}\left[\log\cosh\left(\sqrt{\gamma}Z - \gamma\right)\right] + \mathbb{E}\left[\log\cosh\left(\sqrt{\gamma}Z + \gamma\right)\right]\right) \\
&= \frac{1}{2}\left(\mathbb{E}\left[\log\cosh\left(-\sqrt{\gamma}Z - \gamma\right)\right] + \mathbb{E}\left[\log\cosh\left(\sqrt{\gamma}Z + \gamma\right)\right]\right) \\
&= \mathbb{E}\left[\log\cosh\left(\sqrt{\gamma}Z + \gamma\right)\right]
\end{aligned}$$

Pour la deuxième égalité, on réécrit l'espérance comme une intégrale :

$$\mathbb{E}\left[\log\cosh\left(\sqrt{\gamma}Z + \gamma\right)\right] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \log\cosh\left(\sqrt{\gamma}z + \gamma\right)\exp\left(-\frac{z^2}{2}\right)\,\mathrm{d}z$$

La fonction h suivante est définie sur $(-\infty, +\infty) \times (0, +\infty)$

$$h(z, \gamma) = \log\cosh\left(\sqrt{\gamma}z + \gamma\right)\exp\left(-\frac{z^2}{2}\right)$$

et a pour dérivée partielle par rapport à γ

$$\frac{\partial h}{\partial \gamma}(z, \gamma) = \tanh\left(\sqrt{\gamma}z + \gamma\right)\left(1 + \frac{z}{2\sqrt{\gamma}}\right)\exp\left(-\frac{z^2}{2}\right)$$

On domine sur tout intervalle de la forme $[a, +\infty), a > 0$:

$$\left|\frac{\partial h}{\partial \gamma}(z, \gamma)\right| \leq \left|1 + \frac{z}{2\sqrt{a}}\right|\exp\left(-\frac{z^2}{2}\right)$$

qui est une fonction intégrable. Par conséquent, on peut intervertir dérivée et intégrale, de sorte que

$$\begin{aligned}
\frac{d}{d\gamma}\mathbb{E}\left[\log\cosh\left(\sqrt{\gamma}Z + \gamma\right)\right] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \tanh\left(\sqrt{\gamma}z + \gamma\right)\left(1 + \frac{z}{2\sqrt{\gamma}}\right)\exp\left(-\frac{z^2}{2}\right)\,\mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \tanh\left(\sqrt{\gamma}z + \gamma\right)\exp\left(-\frac{z^2}{2}\right)\,\mathrm{d}z \\
&\quad + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \tanh\left(\sqrt{\gamma}z + \gamma\right)\frac{z}{2\sqrt{\gamma}}\exp\left(-\frac{z^2}{2}\right)\,\mathrm{d}z
\end{aligned}$$

En intégrant par parties le deuxième terme :

$$\begin{aligned}\frac{d}{d\gamma}\mathbb{E}\left[\log\cosh\left(\sqrt{\gamma}Z+\gamma\right)\right] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\tanh\left(\sqrt{\gamma}z+\gamma\right)\exp\left(-\frac{z^2}{2}\right)\,\mathrm{d}z \\ &\quad+\frac{1}{2\sqrt{2\pi}}\int_{-\infty}^{+\infty}\tanh'\left(\sqrt{\gamma}z+\gamma\right)\exp\left(-\frac{z^2}{2}\right)\,\mathrm{d}z \\ &= \frac{1}{2}\left(1+\mathbb{E}\left[\tanh\left(\sqrt{\gamma}Z+\gamma\right)\right]\right) \\ &\quad+\frac{1}{2\sqrt{2\pi}}\underbrace{\int_{-\infty}^{+\infty}\left(\tanh\left(\sqrt{\gamma}z+\gamma\right)-\tanh^2\left(\sqrt{\gamma}z+\gamma\right)\right)\exp\left(-\frac{z^2}{2}\right)\,\mathrm{d}z}_{\triangleq A}\end{aligned}$$

On change de variables $t \triangleq z+\sqrt{\gamma}$ dans l'intégrale A :

$$\begin{aligned}A &= \int_{-\infty}^{+\infty}\tanh\left(\sqrt{\gamma}t\right)\left(1-\tanh\left(\sqrt{\gamma}t\right)\right)\exp\left(-\frac{(t-\sqrt{\gamma})^2}{2}\right)\,\mathrm{d}t \\ &= 2\int_{-\infty}^{+\infty}\frac{\tanh\left(\sqrt{\gamma}t\right)}{\cosh\left(\sqrt{\gamma}t\right)}\exp\left(-\frac{t^2+\gamma}{2}\right)\,\mathrm{d}t\end{aligned}$$

Le facteur en tanh dans l'intégrande est impair, tandis que les autres sont pairs. L'intégrande est donc impaire. Comme l'intégrale converge absolument (on peut borner les tanh par 1), $A=0$, ce qui prouve l'égalité annoncée. □

5.1.3.3 Signal déterministe

Supposons que X est déterministe et notons-le a. On a alors immédiatement :

$$\mathcal{Z}(\gamma,Y) = \exp\left(-\frac{1}{2}\gamma a^2+\sqrt{\gamma}aY\right) \tag{5.25}$$

$$I_a(\gamma) = 0 \tag{5.26}$$

$$\psi_a(\gamma) = \frac{1}{2}\gamma a^2 \tag{5.27}$$

$$F_a(\gamma) = a^2 \tag{5.28}$$

$$\mathrm{mmse}_a(\gamma) = 0 \tag{5.29}$$

5.2 Information et estimation en grande dimension

Dans cette partie qui clora le rapport, nous allons effectuer une lecture informée de deux articles scientifiques déduisant des propriétés asymptotiques de problèmes d'estimation en grande dimension grâce à la méthode des répliques appliquée à la théorie de l'information. Il convient de préciser que ces deux travaux relativement généraux ont été prouvés mathématiquement, ce qui est une avancée considérable par rapport aux conjectures utilisant la méthode des répliques sans justification. Nous en arriverons ensuite aux points communs entre ces travaux. Nous espérons que ceux-ci permettront à de futures études d'aller plus loin dans la compréhension des concepts utilisés.

5.2.1 Estimation de matrices de petit rang bruitées

La publication [7] qui a servi de base à ce stage se fonde sur des résultats de Miolane (2017) [13] qui étudie le problème de l'estimation de matrices de petit rang bruitées. Comme on l'a vu dans l'introduction du chapitre 4, notre problème peut s'y ramener : la matrice de petit rang est la matrice

$\underline{\underline{\boldsymbol{U}}}\underline{\underline{\boldsymbol{V}}}^T$, elle est bruitée par une matrice à entrées gaussiennes $\sigma\underline{\underline{\boldsymbol{Z}}}$. Le modèle étudié dans [13] est quant à lui le suivant :

$$\underline{\underline{\boldsymbol{Y}}} = \sqrt{\frac{\lambda}{D}}\underline{\underline{\boldsymbol{U}}}\underline{\underline{\boldsymbol{V}}}^T + \underline{\underline{\boldsymbol{Z}}}$$

où les entrées $Z_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1), U_i \overset{\text{iid}}{\sim} P_U, V_j \overset{\text{iid}}{\sim} P_V$ avec P_U et P_V des distributions à deuxième moment fini. Celui que nous étudions peut s'y ramener en prenant (ayant tiré $\underline{\boldsymbol{S}} = \underline{\boldsymbol{s}}$ qui donne la valeur des étiquettes connues et en rappelant que $s_j = 0$ intervient avec probabilité $1-\eta$ pour chaque indice $j \in [N]$)

$$\lambda = \frac{1}{\sigma^2} \quad ; \quad U_i \sim \mathcal{N}(0,1) \quad ; \quad V_j = \begin{cases} s_j & \text{si } s_j \neq 0 \\ W_j & \text{sinon} \end{cases} \quad ; \quad W_j \sim 2\mathcal{B}\left(\frac{1}{2}\right) - 1$$

Dans son analyse, l'auteur étudie des versions généralisées des quantités que nous avons définies pour le canal gaussien scalaire, et notamment la convergence de la MMSE et de l'information mutuelle vectorielles suivantes :

$$I\left((\underline{\boldsymbol{U}},\underline{\boldsymbol{V}});\underline{\underline{\boldsymbol{Y}}}\right) \underset{N \underset{\sim}{=}_\alpha D}{\longrightarrow} \frac{\alpha\lambda}{2}\mathbb{E}\left[U^2\right]\mathbb{E}\left[V^2\right] - \sup_{(q_u,q_v)\in\Gamma(\lambda,\alpha)} \mathcal{F}\left(\lambda,\alpha,q_u,q_v\right)$$

$$\text{mmse}(\lambda) \triangleq \frac{1}{ND}\mathbb{E}\left[\left\|\underline{\underline{\boldsymbol{U}}}\underline{\underline{\boldsymbol{V}}}^T - \mathbb{E}\left[\underline{\underline{\boldsymbol{U}}}\underline{\underline{\boldsymbol{V}}}^T \,\middle|\, \underline{\underline{\boldsymbol{Y}}}\right]\right\|^2\right] \underset{N \underset{\sim}{=}_\alpha D}{\longrightarrow} \mathbb{E}\left[U^2\right]\mathbb{E}\left[V^2\right] - q_u^* q_v^*$$

où

$$\mathcal{F}\left(\lambda,\alpha,q_u,q_v\right) \triangleq \psi_U(\lambda\alpha q_v) + \alpha\psi_V(\lambda q_u) - \frac{\lambda\alpha}{2}q_u q_v \tag{5.30}$$

et

$$\Gamma(\lambda,\alpha) = \left\{(q_u,q_v) \in \mathbb{R}_+^2 \middle| q_u = F_U(\lambda\alpha q_v), q_v = F_V(\lambda q_U)\right\} \tag{5.31}$$

Il est remarquable que ces quantités d'estimation multivariées fassent intervenir des fonctions de coïncidence et énergies libres de canaux gaussiens scalaires ψ_U, ψ_V, F_U, F_V que nous pouvons réécrire explicitement grâce au formulaire de la section précédente 5.1.3. On peut réécrire les équations (5.30) et (5.31) dans notre cas particulier, et retomber sur celles du théorème 1 de [7]. On a ainsi en particulier le système suivant, qui permet de retrouver l'équation de point fixe utilisée trouvée par Lelarge et Miolane dans [7].

$$\begin{cases} q_U &= F_U\left(\frac{\alpha q_V}{\sigma^2}\right) = \frac{\alpha q_V}{\sigma^2 + \alpha q_V} \\ q_V &= F_V\left(\frac{q_U}{\sigma^2}\right) = \eta F_{s=\pm 1}\left(\frac{q_U}{\sigma^2}\right) + (1-\eta)F_W\left(\frac{q_U}{\sigma^2}\right) = \eta + (1-\eta)\mathbb{E}\left[\tanh\left(\frac{\sqrt{q_U}}{\sigma}Z + \frac{q_U}{\sigma^2}\right)\right] \end{cases} \tag{5.32}$$

5.2.2 Modèle linéaire standard

L'équipe de Galen Reeves et Henry Pfister a étudié notamment un modèle linéaire dans [17, 16].

$$\mathbb{R}^D \ni \underline{\boldsymbol{X}} = (X_1, ..., X_D) \overset{\text{i.i.d.}}{\sim} P_X \qquad \mathbb{E}\left[X_i^4\right] < +\infty$$

$$\mathbb{R}^{N\times D} \ni \underline{\underline{\boldsymbol{A}}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{1}{D}\right) \qquad \mathbb{R}^N \ni \underline{\boldsymbol{Z}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

On mesure $\underline{\boldsymbol{Y}} \in \mathbb{R}^N$ et $\underline{\underline{\boldsymbol{A}}}$ et l'objectif est de reconstruire $\underline{\boldsymbol{X}}$.

$$\underline{\boldsymbol{Y}} = \underline{\underline{\boldsymbol{A}}}\underline{\boldsymbol{X}} + \underline{\boldsymbol{Z}}$$

L'analyse des performances de ce modèle se fait alors en définissant une fonction R qui dépend de l'information mutuelle entre l'entrée et la sortie d'un canal gaussien scalaire I_X (5.3) :

$$R(\alpha,z) = I_X\left(\frac{\alpha}{1+z}\right) + \frac{\alpha}{2}\left(\log(1+z) - \frac{z}{1+z}\right)$$

Le résultat principal de Reeves et Pfister donne que dans ce contexte, à une condition supplémentaire (appelée *single-crossing property* et qui peut par exemple se vérifier graphiquement), on a les convergences suivantes :

$$\frac{1}{D} I\left(\underline{\boldsymbol{X}}; \underline{\boldsymbol{Y}} | \underline{\underline{\boldsymbol{A}}}\right) \underset{N \doteq_\alpha D}{\longrightarrow} \mathcal{I}(\alpha) \triangleq \min_{z \geq 0} R(\alpha, z) \tag{5.33}$$

$$\frac{1}{D} \text{mmse}\left(\underline{\boldsymbol{X}} | \underline{\boldsymbol{Y}}, \underline{\underline{\boldsymbol{A}}}\right) = \frac{1}{D} \mathbb{E}\left[\left\| \underline{\boldsymbol{X}} - \mathbb{E}\left[\underline{\boldsymbol{X}} \mid \underline{\boldsymbol{Y}}, \underline{\underline{\boldsymbol{A}}}\right] \right\|^2\right] \underset{N \doteq_\alpha D}{\longrightarrow} \mathcal{M}(\alpha) \triangleq \operatorname*{argmin}_{z \geq 0} R(\alpha, z) \tag{5.34}$$

Comme plusieurs valeurs de z peuvent atteindre le minimum, la fonction $\alpha \mapsto \mathcal{M}(\alpha)$ peut être discontinue. La seconde limite fonctionne aux points de continuité de $\mathcal{M}$.

Cette formulation (5.33-5.34) est équivalente à celle donnée dans [16] en utilisant le changement de variable $z(s) = \frac{\alpha}{s} - 1$

$$\mathcal{I}(\alpha) = \min_{s \geq 0} I_X(s) + \frac{\alpha}{2}\left(\log\left(\frac{\alpha}{s}\right) + \frac{s}{\alpha} - 1\right) \tag{5.33'}$$

$$\mathcal{M}(\alpha) = \text{mmse}_X(s^*(\alpha)) \tag{5.34'}$$

En effet, d'après l'équation (5.33), au point z^* qui minimise $R(\alpha, z)$ on a nécessairement $\frac{\partial R}{\partial z}(\alpha, z^*) = 0$, d'où :

$$-\frac{\alpha}{(1+z^*)^2} I_X'\left(\frac{\alpha}{1+z^*}\right) + \frac{\alpha}{2}\left(\frac{1}{1+z^*} - \frac{1}{(1+z^*)^2}\right) = 0$$

D'après la relation I-MMSE (5.11)

$$\mathcal{M}(\alpha) = z^* = \text{mmse}_X\left(\frac{\alpha}{1+z^*}\right) = \text{mmse}_X\left(s^*\right)$$

On observe qu'il existe par ailleurs une façon de calculer directement $\mathcal{M}(\alpha)$ à partir d'une équation de point fixe si l'on sait évaluer la fonction $\gamma \mapsto \text{mmse}_X(\gamma)$. Il s'agit de l'équation (40) de [16].

$$\mathcal{M}(\alpha) = \text{mmse}_X\left(\frac{\alpha}{1+\mathcal{M}(\alpha)}\right)$$

Remarque. Si le modèle linéaire standard étudié par Reeves et Pfister comporte beaucoup de similarités dans sa résolution avec celui étudié par Miolane, il n'est pas directement adapté à notre problème de classification non-supervisé car il s'agit justement d'un modèle *linéaire*. L'estimation d'une matrice de petit rang bruitée est un problème d'estimation faisant intervenir une transformation quadratique $\underline{\boldsymbol{U}}\underline{\boldsymbol{V}}^T$ de l'entrée $(\underline{\boldsymbol{U}}, \underline{\boldsymbol{V}})$.

5.2.3 Synthèse

La lecture comparée de ces deux articles qui étudient le comportement d'estimateurs multidimensionnels dans un régime $N \doteq_\alpha D$ fait ressortir deux points d'intérêt que l'on peut aussi relier à l'article précurseur [4] de Guo et Verdu qui étudie les performances d'algorithmes de détection d'utilisateurs sur des réseaux CDMA grâce à la méthode des répliques :

— les quantités étudiées (en général la MMSE et l'information) convergent vers des valeurs qui sont intrinsèquement liées à des canaux gaussiens scalaires, par exemple I_X ou encore ψ_U dans les équations ci-dessus. Le canal gaussien scalaire étant parfaitement maîtrisé, les calculs deviennent alors faisables. L'article de Guo et Verdu mentionne d'ailleurs un « principe de découplage » selon lequel les différentes coordonnées sont en fait asymptotiquement découplées et qu'elles se comportent comme s'il était possible de les estimer à l'aide d'un simple canal gaussien scalaire ;

— pour le problème d'estimation à travers le modèle linéaire standard, la dépendance en la matrice $\underline{\underline{\boldsymbol{A}}}$ disparaît asympotiquement. Ce comportement est très similaire aux remarques faites par Guo et Verdu dans la partie D de leur article (la performance de l'estimation ne dépend asymptotiquement plus de l'état du canal) ainsi qu'au phénomène de concentration de la norme de l'estimateur $\underline{\boldsymbol{U}}_{\text{MSE}}$ vu dans le cas supervisé au chapitre 3 : la norme de $\underline{\boldsymbol{U}}_{\text{MSE}}$ ne dépend asymptotiquement plus du jeu de données d'apprentissage.

5.2.4 Perspectives de recherche

De façon générale, il nous semble être judicieux pour la suite de mieux comprendre les étapes qui permettent aux deux équipes dont nous venons de citer les travaux d'aboutir à la convergence de l'information mutuelle multidimensionnelle. Cette analyse passera probablement par l'application formelle de la méthode des répliques (sans justifier rigoureusement les calculs) pour calculer l'énergie libre $\psi = \mathbb{E}[\log(Z)]$ d'un système donné. Cette méthode, détaillée au chapitre 8 de [14], repose sur l'heuristique suivante : soit $X = \log(Z)$ une variable aléatoire. Son espérance est donnée par la dérivée en 0 de sa fonction génératrice des cumulants :

$$\mathbb{E}[X] = \lim_{t \to 0} \frac{1}{t} \log\left(\mathbb{E}[\exp(tX)]\right)$$

ceci se réécrit ici de la manière suivante :

$$\mathbb{E}[\log(Z)] = \lim_{t \to 0} \frac{1}{t} \log\left(\mathbb{E}\left[Z^t\right]\right)$$

Lorsque t est entier, l'espérance $\mathbb{E}\left[Z^t\right]$ se calcule en faisant cohabiter t répliques du système. Si l'on suppose que la formule obtenue se généralise sur un voisinage de 0, on pourra alors passer à la limite et obtenir l'énergie libre.

De façon générale, une fois la limite de l'information mutuelle obtenue (elle sera vraisemblablement fonction de quantités liées à des canaux gaussiens unidimensionnels), il semble raisonnable de pouvoir trouver des quantités analogues (coïncidence et MMSE multidimensionnelles) avec des relations semblables à celles trouvées à la partie 5.1.2.

Conclusion, apports et perspectives

Bilan scientifique : résultats et contributions Lors de ce stage de recherche, nous nous sommes intéressés aux performances asymptotiques des algorithmes d'apprentissage en grande dimension, et plus particulièrement à l'apprentissage semi-supervisé sur des mélanges de gaussiennes. L'objectif initial du stage était de comprendre le décalage entre l'étude de Mai et Couillet [10] et celle de Lelarge et Miolane [7]. Si cet objectif ambitieux n'a pas été totalement atteint, nous sommes en mesure d'apporter quelques éléments de réponse et des pistes pour de futures études. Nous avons commencé par réécrire de façon probabiliste le résultat de Lelarge et Miolane dans le cas supervisé au chapitre 3, ce qui nous a permis de comprendre les liens qui pouvaient exister entre ce dernier résultat et le phénomène de concentration de la mesure. Dans un second temps au chapitre 4, nous avons tenté d'élargir cette approche au cadre semi-supervisé qui est plus général. Des difficultés combinatoires nous ont empêché de mener à bout cette étude mais nous avons tout de même pu établir l'expression d'un classifieur semi-supervisé approché en effectuant quelques hypothèses simplificatrices. Celui-ci n'atteint évidemment pas la performance optimale bayésienne mais son expression se rapproche de celle du classifieur établi par Mai et Couillet. Pour aller plus loin dans l'étude et espérer retrouver l'expression exacte de la performance asymptotique de Lelarge et Miolane, nous avons dans un dernier temps commencé l'analyse des résultats fondamentaux qui ont permis à ces derniers d'obtenir le leur. Il a pour cela été nécessaire de reformuler mathématiquement les différents objets étudiés et d'effectuer une synthèse de travaux récents à l'intersection entre physique statistique et théorie de l'information. L'analyse complète des calculs reste encore à faire et cela sera probablement l'objet de travaux futurs, qui nécessiteront de lever le dernier verrou : le calcul de la limite de l'énergie libre grâce à la méthode des répliques dont nous avons résumé la démarche dans le dernier paragraphe de ce manuscrit.

Bilan personnel et professionnel Le stage de recherche dont ce manuscrit est la synthèse n'est pas seulement un travail scientifique mais aussi une expérience humaine et professionnelle qui mérite un mot de conclusion au même titre que les résultats scientifiques.

J'ai eu l'occasion d'apprendre énormément pendant toute la durée de mon stage, en fait beaucoup plus que ce que je m'imaginais au début : j'ai revu et considérablement enrichi mes connaissances en statistiques, en estimation et en probabilités et je me suis rendu compte de l'étendue des connaissances existantes. J'ai compris qu'il était possible de passer toute une vie à apprendre sans jamais avoir fait le tour de son domaine ; j'ai donc saisi un des sens de la formation par la recherche, qui est peut-être avant tout une école de la modestie. En cela, ce stage a été très complémentaire à mon stage de recherche précédent (qui était peut-être finalement plus un travail d'ingénieur de recherche) et au stage de développement qui avait ouvert mon année de césure. J'ai appris à exploiter une bibliographie très pointue car c'était nécessaire pour travailler à l'intersection entre deux domaines de recherche avancés (apprentissage et théorie de l'information) dont je ne connaissais que le premier. Si ce stage m'a donc beaucoup appris, il n'en a pas moins été éprouvant et difficile, et ce pour deux raisons. La première est évidemment l'épidémie de Covid-19 qui l'a transformé en « téléstage » effectué depuis chez mes parents, ce qui m'a mécaniquement isolé de l'équipe et du laboratoire. La deuxième est son caractère très théorique (la grande dimension est très éprouvante pour l'intuition) et sa situation inconfortable entre deux communautés scientifiques au vocabulaire et aux techniques différents, qui a nécessité d'effectuer un important travail d'adaptation et de « traduction ».

Sur le plan professionnel, ce stage a fait évoluer mes impressions et mes attentes. Depuis deux

ans, avec mes stages de recherche et développement, j'étais presque certain de vouloir poursuivre en doctorat à la fin de ma scolarité. Le confinement m'a invité à beaucoup me questionner et mon stage a suscité de grands moments de réflexion sur mon travail, ma compréhension de celui-ci et ses impacts. J'ai eu (et ai encore) beaucoup de mal à décorréler ce qui vient du contexte anxiogène suscité par la pandémie, ce qui vient du stage très théorique et ce qui vient de la recherche. Je garderai de ce stage de précieux enseignements sur les données en grande dimension, leur comportement et les algorithmes associés - il s'agissait d'ailleurs de mon objectif initial - mais je sais que je ne poursuivrai pas sur un tel sujet théorique plus tard. La question du doctorat est cependant encore ouverte car je suis parfaitement en accord avec les valeurs du monde de la recherche, je trouve l'environnement humain extrêmement agréable et stimulant et j'aimerais pouvoir penser les problèmes de façon rigoureuse et mathématique pour pouvoir leur apporter des solutions efficaces sur-mesure en tirant parti des outils les plus adaptés. Dois-je pour autant faire de la recherche mon métier ? J'ai compris pendant ce stage que passion pour la science et travail n'avaient pas forcément à se rejoindre et je tiens à remercier mon tuteur Steeve pour la très éclairante citation attribuée à Einstein qu'il utilise dans sa signature électronique (même si nous l'interprétons tous les deux différemment) : « Science is a wonderful thing if one does not have to earn one's living at it ». Avant de me décider à reprendre ou non la recherche avec un doctorat, j'aimerais reprendre d'anciens projets, reprendre confiance en moi et m'orienter sur des projets plus concrets en science des données en apportant mes connaissances aux acteurs économiques et publics afin de les aider à mieux comprendre les données massives dont ils disposent. J'ai aussi mûri lors de ces derniers mois un intérêt pour les données satellitaires, notamment grâce au cours du master MVA donné par l'équipe du centre Borelli et l'équipe IMAGES de Télécom Paris. Ce cours a fait écho à mon premier stage de césure à Reuniwatt, lors duquel j'avais exploité des images de satellites météorologiques pour améliorer des prévisions de production photovoltaïque. Ce domaine m'intéresse beaucoup, notamment du fait des applications que peut avoir l'analyse de données géospatiales à grande échelle pour la gestion des ressources (énergétiques, naturelles, spatiales, ...). Si je ne continue pas professionnellement dans cette voie, je suis au moins certain d'y contribuer un jour sur mon temps libre.

Annexe A

Algorithmes

Pour permettre au lecteur de se faire la main avec les outils présentés dans ce manuscrit, nous proposons quelques scripts écrits dans le langage Python permettant de reproduire expérimentalement certains des comportements observés. Dans le texte, il est en général fait référence à ces scripts à l'aide d'une police spéciale, par exemple `S00_script.py`. On trouvera les scripts dans le dépôt git suivant : `https://gitlab.com/h_sdl/rmt-bounds`.

Liste des scripts

- `S01_norm_high_dimensional_gaussian.py` : script simple permettant d'observer expérimentalement la concentration de la norme d'une gaussienne en grande dimension
- `S02_comparison_xmrc_mclm.py` : script permettant de reproduire le graphe de comparaison de performances obtenu à la figure 1.2 ou encore à la figure 6 de [10]
- `S03_concentration_norm_SL.py` : script permettant d'observer le phénomène de concentration de la norme de l'estimateur MSE dans le cas supervisé
- `S04_risk_SL.py` : script permettant d'observer la phénomène de concentration du risque dans le cas supervisé
- `S05_eigenvalues.py` : script permettant d'observer la distribution des valeurs propres de la matrice $\underline{\underline{S}}_{\mathrm{u}}$ définie au chapitre 4
- `S06_resolvent_bulk.py` : script permettant d'observer le phénomène d'augmentation de l'amplitude du bulk, qui explique la transition de phase pour la performance du classifieur approché trouvé dans le cas semi-supervisé
- `S07_approximate_classifier.py` : script permettant d'observer la performance du classifieur semi-supervisé approché décrit au chapitre 4 pour différentes valeurs du recentrage λ

Liste des fichiers utilitaires On fournit aussi quelques fichiers utilitaires :

- `dataset.py` utilisé pour générer des jeux de données
- `information_toolbox.py` utilisé pour calculer certaines des quantités de l'étude informationnelle
- `rmt_toolbox.py` utilisé pour inférer les paramètres d'une distribution de Marchenko-Pastur
- `benchmark.py` utilisé pour évaluer le risque d'un ensemble de classifieurs avant de les comparer
- `classifiers.py` regroupe les classifieurs expérimentés

Bibliographie

[1] Romain Couillet and Florent Benaych-Georges. Kernel spectral clustering of large dimensional data. *Electronic journal of statistics* , 10(1) :1393–1454, 2016.

[2] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.

[3] Dongning Guo, Shlomo Shamai, and Sergio Verdu. Mutual information and minimum mean-square error in gaussian channels. *IEEE Transactions on Information Theory*, 51(4) :1261–1282, 2005.

[4] Dongning Guo and Sergio Verdu. Randomly spread CDMA : asymptotics via statistical physics. *IEEE Transactions on Information Theory*, 51(6) :1983–2010, 2005.

[5] Steven M. Kay. *Fundamentals of Statistical Signal Processing : Estimation Theory*. Prentice-Hall, Inc., USA, 1993.

[6] Michel Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2001.

[7] Marc Lelarge and Léo Miolane. Asymptotic bayes risk for gaussian mixture in a semi-supervised setting. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 639–643, 2019.

[8] Cosme Louart and Romain Couillet. Concentration of measure and large random matrices with an application to sample covariance matrices, 2018. Preprint, arXiv : 1805.08295 [math.PR].

[9] Xiaoyi Mai and Romain Couillet. A random matrix analysis and improvement of semi-supervised learning for large dimensional data. *Journal of Machine Learning Research*, 19 :79 :1–79 :27, 2018.

[10] Xiaoyi Mai and Romain Couillet. Consistent semi-supervised graph regularization for high dimensional data. Preprint, arXiv : 2006.07575 [cs.LG], 2020.

[11] Vladimir A Marčenko and Leonid A Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4) :457–483, apr 1967.

[12] Xavier Mestre. Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates. *IEEE Transactions on Information Theory*, 54(11) :5113–5129, 2008.

[13] Léo Miolane. Fundamental limits of low-rank matrix estimation : the non-symmetric case. Preprint, arXiv : 1702.00473 [math.PR], 2017.

[14] Marc Mézard and Andrea Montanari. *Information, Physics, and Computation*. Oxford University Press, Inc., USA, 2009.

[15] Daniel P. Palomar and Sergio Verdu. Gradient of mutual information in linear vector gaussian channels. *IEEE Transactions on Information Theory*, 52(1) :141–154, 2006.

[16] Galen Reeves and Henry Pfister. Understanding phase transitions via mutual information and mmse, 2019. Preprint, arXiv : 1907.02095 [cs.IT].

[17] Galen Reeves and Henry D. Pfister. The replica-symmetric prediction for random linear estimation with gaussian matrices is exact. *IEEE Transactions on Information Theory*, 65(4) :2252–2283, 2019.

[18] Marcus Spruill. Asymptotic distribution of coordinates on high dimensional spheres. *Electronic Communications in Probability*, 12 :234–247, 2007.

[19] Roman Vershynin. *High-Dimensional Probability : An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[20] Eugene Wigner. Random matrices in physics. *SIAM Review*, 9(1) :1–23, 1967.