Priors PAC-Bayes avec covariance pleine qui dépendent de la distribution source

Mémoire

Mathieu Alain

Maîtrise en informatique - avec mémoire
Maître ès sciences (M. Sc.)

Québec, Canada

Priors PAC-Bayes avec covariance pleine qui dépendent de la distribution source

Mémoire

Mathieu Alain

Sous la direction de:

François Laviolette, directeur de recherche
Pascal Germain, codirecteur de recherche

Résumé

L'ambition du présent mémoire est la présentation d'un ensemble de principes appelés la théorie PAC-Bayes. L'approche offre des garanties de type PAC aux algorithmes d'apprentissage bayésiens généralisés. Le mémoire traite essentiellement des cas où la distribution prior dépend des données.

Le mémoire est divisé en trois chapitres. Le premier chapitre détaille les notions de base en apprentissage automatique. Il s'agit d'idées nécessaires à la bonne compréhension des deux chapitres subséquents. Le deuxième chapitre présente et discute de la théorie PAC-Bayes. Finalement, le troisième chapitre aborde l'idée d'une garantie PAC-Bayes où le prior dépend des données.

Il y a deux contributions principales. La première contribution est une formulation analytique du risque empirique espéré pour les distributions elliptiques. La seconde contribution est une extension du travail de Parrado-Hernández et al. (34). En effet, il s'agit du développement d'une garantie PAC-Bayes avec un prior espérance non sphérique.

Abstract

The ambition of this thesis is to present a set of principles called the PAC-Bayes theory. The approach provides PAC-like guarantees for generalised Bayesian learning algorithms. This thesis deals essentially with cases where the prior distribution is data dependent.

The paper is divided into three chapters. The first chapter details the core concepts of machine learning. These are ideas that are necessary for a good understanding of the two subsequent chapters. The second chapter presents and discusses the PAC-Bayes theory. Finally, the third chapter addresses the idea of a PAC-Bayes guarantee where the prior depend on the data.

There are two main contributions. The first contribution is an analytical formulation of the empirical expected risk for elliptical distributions. The second contribution is an extension of the work of Parrado-Hernández et al. (34). Indeed, it is the development of a PAC-Bayes guarantee with a non-spherical prior expectation.

Table des matières

Remerciements

Le passage du premier au deuxième cycle est souvent tortueux, mais j'ai eu la chance d'être bien accompagné. Tout d'abord, je tiens à remercier Mario Marchand et Luc Lamontagne d'avoir accepté d'être mes évaluateurs. Je tiens également à remercier mes deux directeurs de mémoire, François Laviolette et Pascal Germain. Malheureusement, François nous a quitté récemment, mais j'espère garder la bonne humeur et la rigueur intellectuelle qu'il a toujours mises de l'avant. Un grand merci à Pascal d'avoir accepté de m'encadrer et surtout de m'avoir toujours encouragé et aidé, malgré mes fréquents hauts et bas. Il va sans dire que je ne serais pas là où je suis aujourd'hui sans lui. Par ailleurs, je dois également mentionner Jules Desharnais qui m'a donné mon baptême de la recherche à l'été 2016 et qui m'a ensuite généreusement emmené à une conférence en France l'année suivante. Tout au long de mon séjour à l'Université Laval, j'ai eu le plaisir de rencontrer de nombreux professeurs qui ont eu un impact positif sur mon développement : Thierry Eude, Nicolas Doyon, Pascal Tesson, Robert Guénette, Bernard Hodgson, José Urquiza, André Fortin, Robert Bergevin, Nadia Tawbi et plusieurs autres. Je tiens à remercier mes amis et collègues, Gabriel, Nicolas, Alexandre, Frédérik, Fan, Jean-Samuel, Maxime, Gaël, Alex, Jonathan, Ulysse, Philippe, Philippe, Jean-François, David, Pierre-Louis, Mazid, Yann, Loïc et plusieurs autres. Je tiens particulièrement à remercier ma famille, Daniel, Diane, Mila et Myriam pour toujours avoir été là pour moi. Enfin, mes derniers remerciements reviennent à Maëva qui m'a soutenu pendant toutes ces années.

Introduction

L'**intelligence artificielle** est une idée étonnamment ancienne qui remonte aux automates imaginés dans la mythologie grecque. Un automate était en principe constitué de bronze, ou d'un autre métal précieux, et il était animé d'une volonté propre. Le rêve moderne de l'intelligence artificielle est de simuler l'intelligence humaine au moyen de machines de calcul. Les tâches nécessitant traditionnellement des facultés humaines pourraient être automatisées. En dehors de l'automatisation, il existe également d'autres raisons qui justifient l'étude de l'intelligence artificielle. Tout d'abord, les connaissances acquises pourraient contribuer à une meilleure compréhension de la cognition humaine. Ensuite, elles pourraient aboutir à la création d'une intelligence artificielle **forte**, c'est-à-dire une machine consciente raisonnant librement et indépendamment de ses créateurs.

Le présent mémoire se concentre sur une branche récente, mais dominante, de l'intelligence artificielle, appelée **apprentissage automatique**. Le terme a été utilisé pour la première fois en 1959 par l'informaticien Arthur Samuel alors qu'il travaillait au sein de la multinationale américaine IBM. L'ambition de l'apprentissage automatique est de permettre à un ordinateur d'accomplir des tâches sans avoir à les programmer explicitement. La pratique de cette science est pourtant loin de conduire à la naissance d'une entité virtuelle belliqueuse. Il s'agit plutôt d'une discipline née de l'union de la statistique et de l'informatique. Le début du 20e siècle a connu l'émergence de la statistique moderne et le développement de nombreuses méthodes statistiques. Néanmoins, une grande partie des techniques étaient longtemps restées impraticables, car elles nécessitaient, en autres choses, une puissance de calcul inexistante à l'époque. Les premiers principes des ordinateurs modernes ont été établis en 1936 par le mathématicien Alan Turing et ils ont rapidement été matérialisés grâce à l'arrivée des transistors une décennie plus tard. Dès lors, une course à la puissance de calcul, prédite par les célèbres lois empiriques de Moore, a été amorcée. Le bilan est une accessibilité croissante à des ressources computationnelles de plus en plus performantes. L'une des autres raisons du décalage entre l'apparition de méthodes statistiques et la mise en application était le manque de **données**. La majorité des méthodes d'apprentissage automatique nécessitent une grande quantité de données pour être opérationnelles. La démocratisation des nouvelles technologies a plongé le monde dans l'ère des données **massives** où les individus sont à la fois consommateurs et créateurs de données numériques. L'abondance de données couplées à des ordinateurs puissants a permis à

l'apprentissage automatique une progression fulgurante et la redécouverte d'idées auparavant délaissées. En effet, il y a de nombreux concepts présentés comme nouveaux qui sont en fait des idées qui ont été remises en lumière, testées et puis améliorées. La croissance de l'apprentissage automatique a traversé des hauts et des bas, mais aujourd'hui, elle est bien présente et plus déterminée que jamais. Les prouesses spectaculaires de l'apprentissage **profond**, une classe d'**algorithmes** d'apprentissage automatique, ont récemment suscité un grand intérêt. Les algorithmes d'apprentissage sont partout, dans les moteurs de recherche, les puces des téléphones, les microprocesseurs équipant les ordinateurs et même sur les routes grâce aux voitures à conduite assistée.

Un algorithme est souvent présenté comme étant analogue à une recette de cuisine. Il s'agit formellement d'une séquence d'instructions qu'une machine à calculs doit suivre pour réussir une certaine tâche. Dans un même esprit, un algorithme d'apprentissage est simplement une procédure d'optimisation traitant des données et créant ensuite un programme informatique, appelé **modèle** d'apprentissage. Par exemple, les algorithmes d'apprentissage profond engendrent les célèbres **réseaux de neurones**. En fonction du contexte, un modèle est utilisé pour classer des images, identifier des visages, traduire des textes dans une langue étrangère, etc. Les possibilités semblent presque infinies. L'optimisation menant à un modèle, souvent appelée **entraînement**, est basée sur des mécanismes statistiques. L'apprentissage automatique est un problème d'optimisation **imparfait**, c'est-à-dire que des hypothèses fortes, appelées **biais inductifs**, sont nécessaires et même alors, une solution exacte n'existe pas systématiquement. L'élaboration d'un algorithme d'apprentissage témoigne d'un biais inductif hérité des créateurs de l'algorithme. Le biais nécessaire à l'entraînement d'un modèle. En effet, sans un biais inductif, le problème d'optimisation serait trop difficile et l'algorithme ne serait pas en mesure de choisir un modèle. En d'autres mots, l'espace des modèles serait trop complexe. L'entraînement est une approximation s'améliorant progressivement grâce à l'information contenue dans les données. Un modèle est donc sujet à commettre des erreurs. Par exemple, il peut réussir une tâche à plusieurs reprises et ensuite échouer sans fournir d'explications tangibles. L'apprentissage automatique est victime de vives critiques au sujet du manque de garanties théorique. L'engouement initial a été freiné dans plusieurs domaines d'application. Les industries aérospatiales et médicales ont besoin de garanties solides. La défaillance d'un modèle dans le cadre d'une application critique est susceptible de générer des coûts faramineux ou pire encore, des pertes humaines. L'incertitude autour des performances des modèles représente l'une des notions les plus importantes à comprendre et à contrôler. La fiabilité d'un modèle d'apprentissage dépend de la justesse des biais inductifs de même que de la qualité et de la quantité de données. La notion de **probabilité** est un cadre naturel pour quantifier cette incertitude. Il existe deux types d'incertitude : l'incertitude **aléatorique** et l'incertitude **épistémique**. La première émane des données et elle est irréductible, c'est-à-dire que des données additionnelles ne la réduisent pas. La seconde apparaît en raison d'un manque de connaissance au sujet du problème. En d'autres mots, les biais inductifs considérés engendrent

de mauvais modèles. Les méthodes d'apprentissage **bayésiennes** permettent l'encapsulation de l'incertitude épistémique dans des distributions de probabilité. Toutefois, l'approche est délicate à mettre en place, car elle exige une modélisation statistique explicite de la distribution de probabilité qui a généré les données, appelée distribution **source**. La difficulté ou même l'impossibilité de la modélisation d'un phénomène statistique avec une grande précision est bien connue. Le célèbre aphorisme énoncé en 1976 par George Box l'illustre parfaitement : « All models are wrong, but some are useful » (Tous les modèles sont incorrects, mais certains sont utiles). L'approche bayésienne est élégante, mais elle semble être dans une impasse du point de vue pratique. Par ailleurs, elle est souvent gourmande en ressources computationnelles. Les méthodes **fréquentistes** n'expriment pas directement l'incertitude épistémique, mais l'assouplissement au niveau de la modélisation statistique permet l'interprétation de l'incertitude épistémique comme une incertitude aléatorique. L'objet du mémoire n'est pas une discussion des avantages ou même simplement des différences entre les praticiens bayésiens et fréquentistes. Le débat est encore bien vivant et suscite parfois des échanges un peu trop passionnés. Néanmoins, comme en témoigne le prochain paragraphe, il est ponctué de terrains d'entente.

Une stratégie fréquentiste prometteuse, mais à saveur bayésienne, est la théorie **PAC-Bayes**. L'approche s'appuie sur la notion d'algorithmes **probablement approximativement corrects** (PAC) proposée en 1984 par l'informaticien Leslie Valiant. La théorie PAC cherche une garantie que le modèle retenu par un algorithme d'apprentissage a très probablement une forte aptitude de **généralisation**. En d'autres mots, le modèle est bon sur des données qui n'ont pas été utilisées pendant l'entraînement, mais tout de même générées par la même distribution source. Par ailleurs, la distribution source n'a pas besoin d'être connue et en pratique, elle est inconnue. Il s'agit d'un critère de réussite intéressant : un modèle possédant une garantie PAC est décrit comme ayant bien appris la tâche demandée, par opposition à apprise par cœur. La théorie PAC-Bayes applique la théorie PAC à des agrégations de modèles, c'est-à-dire un ensemble pondéré de modèles. La formalisation de cette idée est réalisée par deux types de distributions de probabilités : la distribution **a priori** et la distribution **a posteriori**. En accord avec la philosophie bayésienne, un prior impose une hiérarchie sur un ensemble de modèles, c'est-à-dire qu'il favorise, avant l'entraînement, certains modèles plus que d'autres. Par exemple, il peut être utilisé pour encoder des connaissances provenant d'une expertise tierce. Ensuite, l'algorithme d'apprentissage, à l'aide d'un prior et des données, construit un posterior qui attribue à chaque modèle une pondération d'être le bon, c'est-à-dire de bien simuler la distribution source. L'un des traits caractéristiques d'une garantie PAC-Bayes est la **mesure de complexité** entre le posterior et le prior. L'action de choisir un prior régularise la complexité du posterior. La mesure de complexité est souvent la célèbre divergence de **Kullback-Leibler**, mais des mesures plus générales sont aussi possibles. La portée des garanties PAC-Bayes repose en grande partie sur le choix d'un prior. Un mauvais prior a pour effet de dévier le posterior dans la même direction. Hélas, le plus souvent il est difficile de trouver un

prior qui s'adapte à des situations spécifiques. Par conséquent, les priors sont souvent simples et de nature générique. Néanmoins, dans plusieurs situations, il a été démontré qu'un prior générique conduit à des garanties peu intéressantes. En un sens, une garantie intéressante n'est possible que si un prior imite le posterior de sorte que la mesure de complexité soit petite. L'un des principaux inconvénients est qu'un prior doit être déterminé indépendamment des données utilisées pendant l'entraînement. Récemment, de nombreuses stratégies ont été mises en place pour échapper à la contrainte. Le but est toujours de recourir à des données pour guider le choix d'un prior. Toutefois, toutes les approches possèdent aussi des désavantages. L'objectif ultime du mémoire est la présentation et l'extension d'une méthode où le prior dépend de la distribution source.

Chapitre 1

Notions de base

1.1 Apprentissage automatique

L'apprentissage automatique (ou *machine learning* en anglais) est une branche de l'intelligence artificielle en pleine effervescence. Toutefois, il convient d'être précis : elle ne vise pas à doter un être mécanique d'une volonté propre, mais plutôt à répondre à des questions concrètes en puisant dans l'immense royaume de la statistique. La psychologie cognitive et les neurosciences ont successivement inspiré l'apprentissage automatique. Néanmoins, elles ne constituent pas la force motrice derrière le développement des algorithmes d'apprentissage. Une analogie intéressante est l'évolution technique de l'aéronautique : les premiers appareils ont peut-être été inspirés par la forme des ailes des oiseaux, mais les méthodes modernes en sont désormais éloignées. L'apprentissage automatique est assis sur les principes de la statistique **inductive**, par opposition aux méthodes **déductives** fondées sur des ensembles de règles prédéfinies. Les systèmes à base de règles prédéfinies (45) étaient auparavant dominants dans la pratique de l'intelligence artificielle, mais ils sont considérablement limités. En effet, ils sont incapables de recourir à l'induction, c'est-à-dire utiliser des données pour tirer des conclusions générales. L'hypothèse la plus fondamentale de l'apprentissage automatique est que les données contiennent des informations précieuses pour l'induction. En principe, plus il y a de données et meilleure est l'induction. L'objectif de l'apprentissage automatique est le développement d'algorithmes d'apprentissage automatique. Un algorithme d'apprentissage exploite des données et puis entraîne un modèle d'apprentissage capable d'effectuer la tâche souhaitée. L'utilité des algorithmes d'apprentissage n'est plus à démontrer. Ils exécutent désormais de nombreuses tâches, de l'identification des meilleurs indices boursiers à l'accélération de la découverte de médicaments. Les machines ont déjà prouvé qu'ils étaient en mesure de surmonter des problèmes complexes, comme la multiplication ou l'inversion de matrices de très grande taille. Il n'y a aucun doute : les ordinateurs dépassent de loin les aptitudes humaines face à cette catégorie de questions. Néanmoins, il est amusant de constater que plusieurs tâches peu complexes, dont un enfant est capable, représentent encore des défis majeurs pour les chercheurs en apprentissage

automatique. Malgré l'avènement de l'apprentissage profond (18), l'identification de visages, la reconnaissance de la parole et la description textuelle d'images sont encore des problèmes difficiles. Ils sont peut-être simples à exprimer de manière informelle, mais ils sont compliqués à être interprétés par un ordinateur. Par exemple, un enfant peut rapidement apprendre à distinguer un chat d'un chien en regardant une illustration. Le même enfant sera ensuite capable de les distinguer dans différents contextes : à l'intérieur d'une maison, devant un parc, dans une ruelle sombre, etc. En revanche, un algorithme d'apprentissage aura besoin d'une grande quantité d'images sous différents angles et éclairages. Le comble de l'ironie : le modèle sera tout de même en proie aux attaques **antagonistes**, c'est-à-dire aux images retouchées d'une manière imperceptible pour un humain dans le but d'induire le modèle en erreur (2). La situation souligne une très grande dissemblance entre le fonctionnement d'un algorithme d'apprentissage et d'un cerveau humain. De même, une interrogation presque philosophique reste à savoir si une image générée par un modèle, comme un réseau **antagoniste génératif** (19), peut être considérée comme de l'art. En résumé, le défi de l'apprentissage automatique consiste à réussir des tâches qui semblent simples à première vue, mais qui s'avérant difficiles à formaliser.

Paradigmes d'apprentissage. Un paradigme d'apprentissage est un ensemble de tâches partageant des biais inductifs similaires. Il existe trois principaux paradigmes : l'apprentissage **supervisé**, l'apprentissage **non supervisé** et l'apprentissage par **renforcement**. Les différents paradigmes se reconnaissent notamment à la nature des données fournies à l'algorithme d'apprentissage. En apprentissage supervisé, une donnée est une paire, appelée **exemple**, comprenant un vecteur de **caractéristiques** et une **étiquette**. Une caractéristique est tout simplement une propriété mesurable, comme l'intensité d'une couleur, un poids, une hauteur, le prix d'un objet, etc. Une étiquette est à bien des égards identique à une caractéristique, mais une bonne étiquette dépend idéalement d'une ou plusieurs caractéristiques. En fonction des besoins, une caractéristique peut devenir une étiquette et inversement. Le choix des caractéristiques considérées est un biais inductif important. Par exemple, est-ce que la couleur d'une voiture est un indicateur important du taux d'accidents ? Les compagnies d'assurance semblent le croire. Le lien entre une étiquette et les caractéristiques est une relation de corrélation et pas nécessairement de causalité (35). Le constat précédent est un élément crucial pour mieux comprendre les biais **algorithmiques**, c'est-à-dire les biais inductifs à l'origine d'inégalités ou de discriminations (31). Les tâches en apprentissage supervisé sont prédictives, c'est-à-dire que le modèle est entraîné à recevoir un vecteur de caractéristiques et à retourner la bonne étiquette. Dans le présent mémoire, le paradigme utilisé est l'apprentissage supervisé. En apprentissage non supervisé, les données ne sont pas étiquetées et l'algorithme d'apprentissage est amené à découvrir des régularités parmi les vecteurs de caractéristiques. Par exemple, suite à l'observation d'images de chiens et de chats, un algorithme d'apprentissage pourrait reconnaître qu'il existe deux groupes de vecteurs de caractéristiques. L'apprentissage par ren-

forcement est un peu différent des deux paradigmes précédents. Un **agent** doit apprendre une séquence d'**actions** basée sur les expériences vécues dans un environnement dans le but de maximiser une récompense au cours du temps. Par exemple, un agent pourrait être un robot explorant un labyrinthe à la recherche d'une sortie. Les actions sont les mouvements possibles dans le labyrinthe et si le robot se rapproche d'une sortie, il obtient un grand degré de récompenses. Le but est de déterminer la séquence de déplacements nécessaires pour atteindre une sortie.

1.2 Paradigme supervisé

L'apprentissage supervisé est un paradigme d'apprentissage regroupant un ensemble de tâches intuitives et très couramment utilisées. En effet, l'approche supervisée imite de manière générale la relation qui existe entre un étudiant et un enseignant.

1. L'enseignant pose des questions à l'étudiant (des vecteurs de caractéristiques sans étiquettes sont fournis à l'algorithme).
2. L'étudiant répond du mieux qu'il peut (l'algorithme propose un modèle qui étiquette les vecteurs de caractéristiques).
3. L'enseignant corrige les réponses de l'étudiant (les prédictions du modèle sont comparées aux véritables étiquettes).
4. L'étudiant apprend de ses erreurs (l'algorithme modifie le modèle en conséquence).

Entraînement. L'entraînement d'un algorithme d'apprentissage peut être réalisé de multiples façons. Les exemples peuvent être envoyés à l'algorithme en une seule fois ou par petits lots. Ils peuvent être normalisés, c'est-à-dire que les vecteurs de caractéristiques sont modifiés pour faciliter l'optimisation. Un certain nombre de pratiques d'entraînement répandues n'ont pas de fondement théorique, mais seulement une utilisation motivée par des considérations empiriques. L'explication ou même simplement l'énumération des différentes stratégies d'entraînement ne coïncident pas avec les objectifs du mémoire. Le lecteur intéressé est invité à consulter les ressources appropriées sur le sujet.

Il existe deux grands types de problèmes supervisés : la **classification** et la **régression**. La classification suppose que les étiquettes sont des **classes**, aussi appelés **catégories**, prenant des valeurs discrètes. Le problème de classification est un problème de **discrimination**, c'est-à-dire qu'il consiste à trier les vecteurs de caractéristiques dans différentes classes. Par exemple, l'identification de caractères manuscrits à partir d'images est un problème de classification : il s'agit d'associer un caractère à un ensemble ordonné de pixels. En classification, un modèle est appelé **classifieur**. La régression présume que les étiquettes sont placées sur un continuum de valeurs réelles. Par exemple, la prédiction de la valeur d'une voiture est généralement fonction

du kilométrage, de son âge, du pourcentage de l'usure des pièces, etc. En régression, un modèle est appelé **régresseur**.

Dans le présent mémoire, la tâche de prédiction est la classification **binaire**. Le problème revient à décider si un vecteur de caractéristiques appartient à la classe -1 ou à la classe 1. Il arrive parfois d'avoir plutôt les classes 0 et 1. Le choix n'est pas important en théorie, mais en pratique, il a souvent des conséquences mineures sur l'implémentation de l'algorithme d'apprentissage. La classification binaire est un problème intéressant, car il est relativement simple et très intuitif. Un bon exemple est la détection des pourriels : est-ce qu'un courriel est un pourriel ou non ? Malgré le caractère en apparence limité, il existe quelques façons d'étendre la classification binaire à la classification **multiclasse**, c'est-à-dire la classification avec plus de deux classes. Un autre avantage de la classification binaire est qu'elle est généralement peu coûteuse en ressources de calcul.

1.2.1 Classification binaire

Soit $\mathscr{X} \subseteq \mathbb{R}^n$ un espace de caractéristiques et $\mathscr{Y} = \{-1, 1\}$ un espace de classes binaires. Soit $\mathscr{Z} = \mathscr{X} \times \mathscr{Y}$ l'espace des exemples et $\mathscr{M}(\mathscr{Z})$ l'ensemble de toutes les distributions de probabilités sur $\mathscr{Z}$. Soit $\mathscr{D} \in \mathscr{M}(\mathscr{Z})$ une distribution source **inconnue**. Soit $\mathcal{S} \sim \mathscr{D}^m$ un **échantillon** de m exemples générés par $\mathscr{D}$. Il y a deux hypothèses fondamentales :

1. Il existe une distribution source $\mathscr{D}$ et elle ne varie pas.
2. La distribution $\mathscr{D}$ n'a pas de mémoire des exemples qu'elle a déjà générés.

Les hypothèses précédentes sont regroupées sous l'ombrelle de l'hypothèse des variables aléatoires **indépendantes** et **identiquement distribuées** (iid). En pratique, il s'agit d'une hypothèse difficile à satisfaire et même juste à vérifier. Un bon exemple est un échantillon de cas médicaux : un cas produit par un médecin est une paire composée d'un patient et d'un diagnostic. La première hypothèse n'est pas respectée, car le médecin possède un état physique et émotionnel qui affecte son jugement au cours de la journée. La seconde hypothèse n'est pas non plus respectée, car un cas médical complété est possiblement utile pour en traiter un autre. Il est évident que l'hypothèse iid n'est pas appropriée dans l'étude des séries temporelles et correspond à une approximation de la majorité des scénarios où elle est utilisée. Il existe quelques travaux qui tente une relaxation d'une ou des deux hypothèses précédentes, mais ils dépassent les intérêts du présent mémoire.

Classifieurs linéaires. Soit $\Omega \subseteq \mathbb{R}^n$ un espace de **poids**. Un classifieur linéaire paramétré par un vecteur de poids $\mathbf{w} \in \Omega$ est une fonction $\mathrm{c}_{\mathbf{w}} : \mathscr{X} \to \mathscr{Y}$ définie comme $\mathrm{c}_{\mathbf{w}}(\mathbf{x}) = \mathrm{sgn}(\mathbf{w}'\mathbf{x})$.

Définition 1. Pour tout nombre réel a, la fonction **signe** est définie comme

$$\operatorname{sgn}(a) = \begin{cases} -1 & \text{si } a \leq 0, \\ 1 & \text{si } a > 0. \end{cases}$$

La prédiction $\mathrm{c}_{\mathbf{w}}(\mathbf{x})$ est le signe de la combinaison linéaire entre $\mathbf{w}$ et $\mathbf{x}$, c'est-à-dire le signe du produit scalaire $\mathbf{w}'\mathbf{x}$. Intuitivement, si les vecteurs $\mathbf{w}$ et $\mathbf{x}$ sont orientés dans la même direction, alors le produit $\mathbf{w}'\mathbf{x}$ est plus grand que 0 et la classe prédite est 1. De manière similaire, si $\mathbf{w}$ et $\mathbf{x}$ sont orientés dans des directions opposées, alors $\mathbf{w}'\mathbf{x}$ est plus petit que 0 et la classe prédite est -1. Il est également possible d'interpréter l'hyperplan $\mathbf{w}'\mathbf{x} = 0$ comme une frontière de **décision** dans $\mathscr{X}$, c'est-à-dire que le côté où se situe $\mathbf{x}$ détermine la classe qui lui est attribuée. L'ensemble des classifieurs linéaires binaires de l'espace $\mathscr{X}$ correspond à l'ensemble $\mathscr{C} = \{\mathrm{c}_{\mathbf{w}} : \mathbf{x} \mapsto \operatorname{sgn}(\mathbf{w}'\mathbf{x}) \,|\, \mathbf{x} \in \mathscr{X}, \mathbf{w} \in \Omega\}$. L'ensemble $\mathscr{C}$ est aussi appelé espace des **classifieurs**. La composition de l'espace des classifieurs est un biais inductif important.

Marges. La marge **fonctionnelle** d'un classifieur $\mathrm{c}_{\mathbf{w}}$ calculée sur un exemple $\mathbf{z} = (\mathbf{x}, y)$ est simplement la quantité $y\mathbf{w}'\mathbf{x}$. La marge indique si le vecteur $\mathbf{x}$ est mal classé par $\mathrm{c}_{\mathbf{w}}$ (elle est négative) ou bien classé (elle est positive). Toutefois, elle ne fournie pas de renseignements sur la distance dans l'espace $\mathscr{X}$ entre $\mathbf{x}$ et l'hyperplan $\mathbf{w}'\mathbf{x} = 0$. En d'autres mots, seul le signe de $y\mathbf{w}'\mathbf{x}$ compte, mais la valeur n'est pas interprétable. La marge **géométrique** de $\mathrm{c}_{\mathbf{w}}$ calculée sur $\mathbf{z}$ est définie comme $\|\mathbf{w}\|^{-1} y\mathbf{w}'\mathbf{x}$. Il s'agit simplement de la marge fonctionnelle normalisée par la norme euclidienne du vecteur $\mathbf{w}$. La valeur de la marge géométrique quantifie la distance la plus courte dans $\mathscr{X}$ entre $\mathbf{x}$ et $\mathbf{w}'\mathbf{x} = 0$. En d'autres mots, il s'agit de la distance Euclidienne entre $\mathbf{x}$ et l'hyperplan $\mathbf{w}'\mathbf{x} = 0$. La notion de marges géométriques a inspirée les fonctions objectives de plusieurs algorithmes de classification. Par exemple, le but de l'algorithme des **séparateurs à vastes marges** (SVM) (10) est de déterminer une marge géométrique **maximale**, c'est-à-dire un vecteur de poids dans Ω qui maximise en moyenne la marge géométrique des exemples de l'échantillon $\mathscr{S}$.

Noyaux. Les classifieurs linéaires ont l'avantage d'être simples et lorsqu'ils sont couplés à une fonction **noyau**, ils deviennent très flexibles. Il arrive que les vecteurs de caractéristiques ne soient pas **linéairement séparable**, c'est-à-dire qu'aucun hyperplan dans l'espace $\mathscr{X}$ peut classer correctement les vecteurs de caractéristiques. L'expressivité d'un classifieur linéaire peut être améliorée en projetant les vecteurs de caractéristiques dans un espace de caractéristiques $\mathscr{H} \subseteq \mathbb{R}^k$ de dimension supérieure ou même infinie. En effet, il est fréquent qu'une bonne frontière de décision soit plus facile à obtenir dans un espace de dimension supérieure. L'espace de poids devient alors $\Omega \subseteq \mathbb{R}^k$.

Le problème dual de Lagrange maximisant la marge géométrique fait intervenir des produits scalaires entre des paires de vecteurs de caractéristiques. Évidemment, la dimension k est

potentiellement infinie et les produits scalaires dans $\mathscr{K}$ sont difficiles ou tout simplement pas calculables. Par conséquent, les calculs dans $\mathscr{K}$ sont implicites, c'est-à-dire, réalisés par l'intermédiaire d'une fonction noyau mesurant la similarité entre toutes les paires de vecteurs dans $\mathscr{X}$. Une fonction noyau $\kappa : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ est définie comme

$$\kappa(\mathbf{x}, \mathbf{x}^\star) = \phi(\mathbf{x})'\phi(\mathbf{x}^\star),$$

où $\phi : \mathscr{X} \to \mathscr{K}$ est une projection non linéaire. L'**astuce du noyau** permet d'éviter de calculer explicitement la fonction noyau, c'est-à-dire le produit scalaire $\phi(\mathbf{x})'\phi(\mathbf{x}^\star)$. Il existe de nombreuses fonctions noyaux telles que la fonction polynomiale de degré $d \in \mathbb{N}$ définie comme

$$\kappa(\mathbf{x}, \mathbf{x}^\star) = (\mathbf{x}'\mathbf{x}^\star)^d$$

ou encore la fonction gaussienne à base radiale de forme $\gamma > 0$ définie comme

$$\kappa(\mathbf{x}, \mathbf{x}^\star) = \exp(-\gamma\|\mathbf{x} - \mathbf{x}^\star\|^2).$$

Les algorithmes d'apprentissage utilisant des fonctions noyaux sont souvent appelés méthodes à noyau. Ils ont longtemps représenté l'état de l'art et malgré la présence des algorithmes d'apprentissage profond, ils restent encore très populaires. L'algorithme SVM introduit en 1992 par Bernhard Boser, Isabelle Guyon et Vladimir Vapnik (6) est probablement la méthode à noyau la plus connue. L'un des principaux inconvénients des approches à noyau est la **malédiction de la dimension** : la complexité du problème d'optimisation de la marge maximale augmente linéairement avec le nombre de dimensions n. Par conséquent, les méthodes à noyau ne sont pas à privilégier si le nombre d'exemples m est beaucoup plus petit que n. Les références aux fonctions noyaux seront omises dans la suite du mémoire, mais seulement par souci de concision. En réalité, l'approche des noyaux fonctionne très bien avec les résultats subséquents, mais encombrerait inutilement la notation.

Risque. Une fois un classifieur $\mathrm{c}_\mathbf{w}$ en main, la suite logique est l'évaluation de ses performances, c'est-à-dire la qualité de ses prédictions. La fonction de **perte** quantifie la différence entre la classe prédite par $\mathrm{c}_\mathbf{w}$ et la vraie classe. En discrimination binaire, la fonction de perte est souvent la fonction **indicatrice** $\mathrm{I} : \mathscr{X} \times \mathscr{Y} \to \{0, 1\}$ retournant 1 si la classe prédite est identique à la vraie classe et 0 sinon. Elle est aussi appelée perte **zéro-un**. Le **risque théorique** binaire et le risque **empirique** binaire associé à $\mathrm{c}_\mathbf{w}$ et à la perte zéro-un sont définis respectivement comme

$$R_d(\mathrm{c}_\mathbf{w}) = \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathscr{D}} \mathrm{I}(\mathrm{c}_\mathbf{w}(\mathbf{x}) \neq y) \qquad \text{et} \qquad R_s(\mathrm{c}_\mathbf{w}) = \frac{1}{m}\sum_{\mathbf{z}\in\mathcal{S}} \mathrm{I}(\mathrm{c}_\mathbf{w}(\mathbf{x}) \neq y). \tag{1.1}$$

Le risque correspond à la proportion de fois que le classifieur $\mathrm{c}_\mathbf{w}$ commet une erreur. L'objectif de l'apprentissage automatique est la minimisation de la fonctionnelle R_d, mais il est impossible d'y parvenir, du moins directement, dans la mesure où la distribution $\mathscr{D}$ est inconnue. Il

est alors tentant de minimiser la fonctionelle R_s. L'idée n'est pas totalement incongrue : il est bien vrai que R_s est le substitut empirique immédiat de R_d. Néanmoins, cette stratégie appelée **minimisation du risque empirique** (ERM) amène à un phénomène appelé le **sur-apprentissage**.

Phénomènes d'apprentissage. Un phénomène d'apprentissage est tout d'abord une observation empirique. Il y a deux principaux phénomènes : le sur-apprentissage et le **sous-apprentissage**. Le sur-apprentissage d'un classifieur $\mathrm{c}_{\mathbf{w}}$ est un phénomène caractérisé par un risque $R_s(\mathrm{c}_{\mathbf{w}})$ très faible ou même nul, mais un risque $R_d(\mathrm{c}_{\mathbf{w}})$ élevé. En d'autres mots, $\mathrm{c}_{\mathbf{w}}$ a sur-appris s'il a mémorisé dans le moindre détail les exemples de l'échantillon $\mathcal{S}$. Le sur-apprentissage devient problématique dès que $\mathrm{c}_{\mathbf{w}}$ est confronté à des exemples différents, mais générés par la même distribution $\mathcal{D}$. La situation est comparable à celle d'un étudiant qui a mémorisé l'examen d'une année précédente, mais qui est déstabilisé face à un autre examen comportant des questions à peine différentes. Une explication possible est que les biais inductifs de l'algorithme d'apprentissage n'étaient pas suffisamment forts, c'est-à-dire que la **capacité** de $\mathrm{c}_{\mathbf{w}}$ n'était pas suffisamment circonscrite. La capacité contrôle la facilité à apprendre par cœur un échantillon. En classification, la capacité augmente généralement avec la complexité des frontières de décision produites par $\mathrm{c}_{\mathbf{w}}$. L'autre phénomène d'apprentissage notable est le sous-apprentissage. Un classifieur $\mathrm{c}_{\mathbf{w}}$ a sous-appris s'il possède un grand $R_s(\mathrm{c}_{\mathbf{w}})$ et $R_d(\mathrm{c}_{\mathbf{w}})$. Le phénomène est souvent imputable à un ensemble $\mathcal{S}$ trop petit ou à des biais inductifs trop forts, c'est-à-dire que la capacité de $\mathrm{c}_{\mathbf{w}}$ était trop petite.

La grande difficulté de l'apprentissage automatique consiste à naviguer entre les régimes de sous-apprentissage et de sur-apprentissage. L'interprétation courante, mais quelquefois remise en question, considère que le véritable apprentissage se situe quelque part entre les deux. La problématique est souvent appelée le compromis **biais-variance** et représente un thème récurrent en apprentissage automatique. Un principe bien connu est le théorème du *no free lunch* qui énonce qu'il n'existe pas un algorithme d'apprentissage universel susceptible de répondre à tous les problèmes d'apprentissage. En d'autres termes, il n'y a pas de biais inductifs universels : un biais efficace face à un problème précis est inévitablement un handicap face à un autre.

Double descente. La formulation du compromis biais-variance est intuitive, mais elle ne fait plus l'unanimité. Notamment, les réseaux de neurones semblent enfreindre le compromis en disposant d'une capacité énorme sans connaître de régime de sur-apprentissage. Un phénomène appelé double descente est parfois observé. L'augmentation progressive de la capacité se caractérise par une descente du risque empirique, suivie d'une remontée assimilable à un sur-apprentissage et puis, contre toute attente, d'une nouvelle descente (32).

Billet de loterie. Une autre observation intéressante au sujet des réseaux de neurones est l'hypothèse du billet de loterie (13). L'hypothèse affirme qu'un réseau de neurones initialisé de manière aléatoire contient des sous-réseaux, appelés **billets gagnants** pouvant au moins égaler les performances du réseau complet. La conjecture offre une perspective combinatoire intéressante : un grand réseau de neurones est préférable, car il a plus de chances de contenir un billet gagnant qu'un petit réseau. Il s'agit actuellement d'un terrain de recherche fertile générant de nombreuses questions intrigantes.

1.3 Généralisation

La généralisation est l'une des notions les plus fondamentales en apprentissage automatique : il s'agit de l'objectif de tous les modèles. Un classifieur $c_{\mathbf{w}}$ a généralisé si la différence $|R_d(c_{\mathbf{w}}) - R_s(c_{\mathbf{w}})|$ est petite. En d'autres mots, il n'a pas sous-appris ou sur-appris. Il existe différentes façons d'encourager la généralisation. L'une des plus simples consiste à faciliter la recherche dans l'espace des classifieurs $\mathscr{C}$ en réduisant la taille de l'espace de poids Ω.

1.3.1 Sélection de modèles.

La sélection de modèles est un ensemble de principes et de techniques utilisées afin de sélectionner un modèle possédant de bonnes propriétés de généralisation. La capacité d'un modèle est l'un des facteurs cruciaux déterminant le potentiel de généralisation. Un modèle trop capable a tendance à sur-apprendre, tandis qu'un modèle pas suffisamment capable a plutôt le réflexe de sous-apprendre. Il n'est pas facile de savoir quel est le bon niveau de capacité pour un problème d'apprentissage donné, c'est-à-dire le seuil où un modèle ne sur-apprend pas, mais ne sous-apprend pas. En effet, comme la distribution source $\mathscr{D}$ est inconnue, il est impossible de calculer le risque $R_d(c_{\mathbf{w}})$. L'une des façons d'estimer les performances d'un modèle consiste à recourir aux techniques de **validations**.

Validation non croisée. La stratégie est la division de l'échantillon $\mathscr{S}$ en deux parties : l'ensemble d'entraînement et l'ensemble de validation. En règle générale, l'ensemble d'entraînement est plus grand que l'ensemble de validation. L'ensemble d'entraînement est transmis à l'algorithme et utilisé pour entraîner le modèle. L'ensemble de validation est ensuite utilisé pour valider les performances du modèle. En effet, il est inapproprié de juger les performances d'un modèle avec les mêmes données utilisées pour l'entraînement. Néanmoins, il faut comprendre que le risque empirique calculé sur l'ensemble de validation ne donne aucune garantie d'être proche du risque théorique.

Validation croisée. La validation croisée est un moyen de tirer le meilleur parti possible des exemples de l'échantillon $\mathscr{S}$. Elle consiste à diviser l'échantillon en $N \in \mathbb{N}$ parties. L'algorithme est entraîné sur $N - 1$ parties et utilise la partie restante pour la validation, c'est-à-dire

calculer un risque empirique. L'opération est ensuite répétée avec une autre partie de validation différente. Finalement, une moyenne des N pertes empiriques est calculée. L'inconvénient principal de cette méthode est le coût de calcul élevé. Encore une fois, il n'y a aucune garantie que la différence $|R_d(c_{\mathbf{w}}) - R_s(c_{\mathbf{w}})|$ soit petite.

Critères d'information. Une autre manière de faire de la sélection de modèles est grâce à un critère de sélection, aussi appelé critère d'information. Les critères d'information représentent des mesures de la qualité de généralisation d'un modèle. Il existe plusieurs critères tels que le critère d'**information Bayésien** (BIC)(37) et le critère d'**information d'Akaike** (AIC) (1). Les critères fonctionnent généralement bien dans un cadre spécifique et beaucoup moins bien autrement.

1.3.2 Régularisation.

Le principe de la régularisation consiste à pénaliser la capacité d'un modèle. En d'autres mots, un terme de pénalité est ajouté à la procédure ERM. Il s'agit d'une déclinaison du célèbre **rasoir d'Ockham** stipulant qu'il faut écarter les explications complexes et préconiser une solution simple. En d'autres mots, pour pour un ensemble de prédicteurs ayant un risque empirique égal, il est préférable de choisir le plus simple. Le rasoir n'affirme pas que l'explication la plus simple est nécessairement la bonne, mais plutôt qu'elle doit être envisagée attentivement. La vision précédente prend beaucoup de sens dans le cadre de la statistique bayésienne où la régularisation s'apparente au choix d'un prior sur l'espace des classifieurs. L'approche de régularisation peut être utilisée en conjonction d'autres méthodes telles que la sélection de modèles.

Chapitre 2

Théorie PAC-Bayes

Depuis les deux dernières décennies, la théorie PAC-Bayes s'est avérée être un instrument riche et efficace permettant de trouver des garanties théoriques en apprentissage automatique. D'abord initiée par John Shawe-Taylor et David McAllester (41; 30), elle a également été développée de manière indépendante par Olivier Catoni à la fin du 20e siècle (9). L'approche PAC-Bayes a fait ses preuves à de nombreuses reprises, notamment dans l'étude des méthodes à noyaux telles que le fameux algorithme SVM (27). La théorie PAC-Bayes est un ensemble d'outils fréquentistes offrant des garanties de généralisation PAC (44) aux algorithmes d'apprentissage **bayésiens généralisés**. Un algorithme bayésien généralisé est tout simplement un algorithme d'apprentissage qui produit un posterior sur l'espace des hypothèses au lieu d'un seul modèle. La qualification de généralisé découle du fait qu'il ne nécessite pas une modélisation statistique complète, c'est-à-dire qu'il utilise une fonction de perte et non une fonction de vraisemblance. Il s'agit d'une approche fréquentiste, mais fortement influencée par des idées bayésiennes. Un survol des développements récents est disponible, lire Guedj (20). La notion d'agrégation de modèles, c'est-à-dire l'étude simultanée d'un ensemble de modèles, est appelée apprentissage **esembliste**. Un bon exemple est la famille des algorithmes de *boosting* (14; 36). Le *boosting* est basé sur le principe que la combinaison de plusieurs modèles à faible capacité produit une meilleure généralisation qu'un seul modèle à forte capacité. Une analogie possible est le système démocratique, sauf que les votes sont pondérés. Il existe deux types de garanties PAC-Bayes : les garanties **empiriques** et les garanties **oracles**. Les garanties oracles renseignent sur la vitesse de convergence, c'est-à-dire à quelle vitesse les méthodes PAC-Bayes deviennent précises avec plus de données. Néanmoins, elles ne peuvent pas être calculées, car elles reposent sur des quantités inaccessibles. En comparaison, les garanties empiriques peuvent être calculées et donc optimisées. Dans le présent mémoire, seules les garanties empiriques sont étudiées.

2.1 Bornes PAC-Bayes

Les garanties PAC-Bayes présentées dans le présent chapitre se rapportent toutes à la classification linéaire binaire. Néanmoins, la théorie PAC-Bayes s'applique aussi à plusieurs autres problèmes supervisés, non supervisés (38) et même par renforcement (12). La notation du chapitre précédent est reprise, avec un seul changement : la référence à un classifieur est simplifié à son vecteur de poids. Par exemple, un classifieur $c_{\mathbf{w}}$ est désigné seulement par le vecteur $\mathbf{w}$.

Soit $\mathscr{M}(\mathscr{C})$ l'ensemble de toutes les distributions de probabilités sur l'espace des hypothèses $\mathscr{C}$. Soit $\mathscr{P} \in \mathscr{M}(\mathscr{C})$ un prior et $\mathscr{Q} \in \mathscr{M}(\mathscr{C})$ un posterior. La distribution $\mathscr{Q}$ est déterminée après l'observation de l'échantillon $\mathcal{S}$. Dans un sens, $\mathscr{Q}$ est aléatoire, tandis que $\mathscr{P}$ est déterministe. Il y a trois éléments déterminants dans une garantie PAC-Bayes, le risque **théorique espéré**, le risque **empirique espéré** et une mesure de complexité. Le risque théorique espéré et le risque empirique espéré associés à $\mathscr{C}$ et $\mathscr{Q}$ sont respectivement définis comme

$$R_d(\mathscr{Q}) = \mathop{\mathbb{E}}_{\mathbf{w}\sim\mathscr{Q}} R_d(\mathbf{w}) \qquad \text{et} \qquad R_s(\mathscr{Q}) = \mathop{\mathbb{E}}_{\mathbf{w}\sim\mathscr{Q}} R_s(\mathbf{w}).$$

Les deux définitions précédentes contiennent un abus de notation. En effet, par souci de simplicité, les termes R_d et R_s sont surchargés (voir les définitions 1.1). Le risque théorique espéré et le risque empirique espéré sont respectivement parfois appelés risque de **Gibbs** et risque de Gibbs empirique. La divergence de Kullback-Leibler (DKL) est une mesure de dissimilarité entre deux distributions de probabilités et elle parfois appelée **entropie relative**. La DKL apparaît dans de nombreux contextes en statistique et en théorie de l'information. Il existe des travaux sur des garanties PAC-Bayes utilisant d'autres mesures de complexité différentes. L'utilisation de mesures de la famille des divergences f permet une relaxation de l'hypothèse des variables indépendentes et identiquement distribuées (iid) (3). Néanmoins, dans le présent mémoire, seule la DKL est considérée.

La forme d'une garantie PAC-Bayes est une borne supérieure sur le risque théorique espéré. Le théorème général suivant rassemble plusieurs bornes PAC-Bayes connues.

Théorème 1 (Bégin et al. (5))**.** *Soit $\Delta : [0,1]^2 \to \mathbb{R}$ une fonction convexe. Pour toute distribution source $\mathscr{D} \in \mathscr{M}(\mathscr{X})$ et un échantillon $\mathcal{S}$ tiré dans $\mathscr{D}^m$, pour tout espace des hypothèses $\mathscr{C}$, pour tout prior $\mathscr{P} \in \mathscr{M}(\mathscr{C})$, pour tout $\delta \in (0,1]$, avec probabilité au moins $1-\delta$, pour tout posterior $\mathscr{Q} \in \mathscr{M}(\mathscr{C})$,*

$$\Delta(R_s(\mathscr{Q}), R_d(\mathscr{Q})) \leq \frac{1}{n}\left(\mathrm{KL}\{\mathscr{Q}\|\mathscr{P}\} + \ln\frac{I_\Delta(m)}{\delta}\right),$$

où $I_\Delta(m) = \sup_{r\in[0,1]}\left(\sum_{k=0}^{m}\binom{m}{k} r^k(1-r)^{m-k}\exp\left\{m\Delta\left(\frac{k}{m}, r\right)\right\}\right)$ *et* $\mathrm{KL}\{\mathscr{Q}\|\mathscr{P}\} = \int \log\frac{d\mathscr{Q}}{d\mathscr{P}} d\mathscr{Q}$.

Une borne PAC-Bayes est dite **étanche** lorsque la quantité de droite est petite puisque cela implique que $\Delta(R_s(\mathscr{Q}), R_d(\mathscr{Q}))$ est aussi petite.

2.1.1 Avantages et limitations

L'un des avantages évidents de la théorie PAC-Bayes est qu'elle ne nécessite pas la partition de l'échantillon en ensembles d'entraînement et de validation. En effet, tous les exemples peuvent être conservés pour calculer la borne PAC-Bayes. Un autre avantage provient du caractère empirique des bornes, c'est-à-dire que les bornes sont calculables. Il est envisageable de les prendre comme fonctions objectives et de les minimiser, c'est-à-dire de choisir un posterior qui minimise le côté droit du théorème 1. Il est donc possible de concevoir de nouveaux algorithmes d'apprentissage basés sur des fonctions objectives à saveur PAC-Bayes. L'une des façons de minimiser une borne PAC-Bayes est l'utilisation d'une méthode de descente en gradient. Par exemple, l'algorithme PAC-Bayes *Gradient Descent* (PBGD) est un algorithme permettant la minimisation par descente en gradient d'une fonction objective donnée par la théorie PAC-Bayes (16). L'approche PAC-Bayes fournit également les outils permettant l'explication de pourquoi certains algorithmes fonctionnent mieux que d'autres. Il s'avère que plusieurs algorithmes d'apprentissage populaires tels que l'algorithme SVM minimisent une fonction objective semblable à certaines déclinaisons de la borne PAC-Bayes.

La théorie PAC-Bayes suscite un intérêt croissant, mais elle présente actuellement quelques inconvénients. Les bornes classiques supposent que les exemples sont iid, c'est-à-dire que l'hypothèse iid est satisfaite. Évidemment, dans la plupart des situations, l'hypothèse n'est pas respectée. De plus, en pratique, certaines bornes ne sont pas étanches. En effet, un mauvais choix de prior accentue la valeur de la DKL.

2.1.2 Bayes et PAC-Bayes

La statistique bayésienne et la théorie PAC-Bayes partagent à la fois des similitudes et des différences. Les deux approches travaillent avec les notions de prior et de posterior. Toutefois, il faut faire attention, car il ne s'agit pas des mêmes objets. En effet, le prior PAC-Bayes ne se limite pas nécessairement à encapsuler les connaissances d'une expertise externe. Il s'agit d'une manière générale de mettre en place une structure sur l'espace des classifieurs. En théorie PAC-Bayes, le principe de la mise à jour bayésienne est repris, mais modifié pour que le posterior dépende non pas d'une fonction de vraisemblance, mais d'une fonction de perte. Le stratagème permet d'éviter une modélisation explicite en faisant davantage confiance aux données. Les méthodes PAC-Bayes sont très axées sur les données et parfois considérées comme une généralisation des stratégies bayésiennes. Néanmoins, il faut souligner que depuis plusieurs années, la statistique bayésienne connaît de nombreux assouplissements. Par exemple, certains praticiens bayésiens se modifient la fonction de vraisemblance. De plus, dans le domaine de la statistique non paramétrique, il existe une riche littérature sur les vraisemblances **fractionnaires**, c'est-à-dire les fonctions de vraisemblance élevées à une puissance fractionnaire (17). La valeur de la fraction contrôle la contribution des données dans le choix d'un posterior. La technique précédente est utile en **misspécification**, c'est-à-dire lorsque la modélisation

n'est pas parfaitement fidèle à la réalité. Il s'agit d'une problématique qui ne se pose pas dans le cadre PAC-Bayes. Il existe également une littérature sur les approches bayésiennes empiriques, c'est-à-dire où le prior est estimé à partir des données. Néanmoins, malgré une flexibilité accrue, la statistique bayésienne reste tributaire d'une modélisation explicite. Les garanties offertes par les méthodes bayésiennes sont asymptotiques et valides pour une modélisation spécifique. Les bornes PAC-Bayes sont généralement valides pour toute distribution source et pour tout choix de prior et posterior. Par ailleurs, elles fonctionnent en échantillon fini. Même s'il est intéressant d'avoir une idée du comportement asymptotique d'une garantie, il est impossible, en pratique, d'avoir une quantité infinie d'exemples. Finalement, la théorie PAC-Bayes et la statistique bayésienne ont en commun l'échantillonnage de distributions de probabilité, potentiellement complexes, en utilisant des outils numériques tels que la méthode de Monte-Carlo. Une exception est présentée à la section 2.3.

2.2 Exemples de bornes

Le choix d'une fonction convexe Δ au théorème 1 conduit au choix d'une borne PAC-Bayes. La borne de McAllester est l'une des toutes premières bornes PAC-Bayes empiriques (30) et possède l'atout d'être particulièrement simple. La borne de McAllester est obtenue en posant $\Delta(R_d(\mathcal{Q}), R_s(\mathcal{Q})) = 2(R_d(\mathcal{Q}) - R_s(\mathcal{Q}))^2$,

$$R_d(\mathcal{Q}) \leq R_s(\mathcal{Q}) + \sqrt{\frac{1}{2m}\left(\mathrm{KL}\{\mathcal{Q}\|\mathcal{P}\} + \log\tfrac{2\sqrt{m}}{\delta}\right)}.$$

La prochaine borne est appelée la borne de Langford et Seeger (26). Elle est obtenue en posant $\Delta(R_d(\mathcal{Q}), R_s(\mathcal{Q})) = \mathrm{kl}\{R_d(\mathcal{Q})\|R_s(\mathcal{Q})\}$,

$$\mathrm{kl}\{R_s(\mathcal{Q})\|R_d(\mathcal{Q})\} \leq \frac{1}{m}\left(\mathrm{KL}\{\mathcal{Q}\|\mathcal{P}\} + \log\tfrac{2\sqrt{m}}{\delta}\right). \tag{2.1}$$

Le terme $\mathrm{kl}\{R_d(\mathcal{Q})\|R_s(\mathcal{Q})\}$ est simplement la DKL entre deux distributions de Bernouilli de probabilités $R_d(\mathcal{Q})$ et $R_s(\mathcal{Q})$, aussi appelée **petite** DKL,

$$\mathrm{kl}\{R_d(\mathcal{Q})\|R_s(\mathcal{Q})\} = R_d(\mathcal{Q})\ln\frac{R_d(\mathcal{Q})}{R_s(\mathcal{Q})} + (1 - R_d(\mathcal{Q}))\ln\frac{1 - R_d(\mathcal{Q})}{1 - R_s(\mathcal{Q})}.$$

La borne de Langford et Seeger est toujours au moins aussi étanche que la borne de McAllester. En effet, il est possible de montrer que $\mathrm{kl}\{R_d(\mathcal{Q})\|R_s(\mathcal{Q})\} \leq 2(R_d(\mathcal{Q}) - R_s(\mathcal{Q}))^2$. En particulier, si le risque $R_s(\mathcal{Q}) = 0$, alors la borne de Langford et Seeger décroit en $\mathcal{O}(\frac{1}{m})$, tandis que la borne de McAllester décroit en $\mathcal{O}(\frac{1}{\sqrt{m}})$. Dans le présent mémoire, la borne utilisée est la borne de Langford et Seeger. Néanmoins, les résultats s'adaptent sans aucun problème à d'autres bornes.

2.3 Risque empirique elliptique

Le risque empirique espéré est une quantité importante exprimant la qualité de prédiction des classifieurs de l'espace $\mathscr{C}$ pondéré par le posterior $\mathcal{Q}$. En classification linéaire binaire, il est

possible de réécrire le risque empirique espéré de la façon suivante (15),

$$R_s(\mathcal{Q}) = \frac{1}{2} - \frac{1}{2}\frac{1}{m}\sum_{\mathbf{z}\in\mathcal{S}} y \mathop{\mathbb{E}}_{\mathbf{w}} \operatorname{sgn}(\mathbf{w}'\mathbf{x}). \tag{2.2}$$

Dans la sous-section, le contexte est suffisamment claire pour écrire $\mathbf{w}$ et $\mathbf{z}$ sous les espérances au lieu de $\mathbf{w} \sim \mathcal{Q}$ et $\mathbf{z} \sim \mathcal{S}$. L'étape clé de l'équivalence précédente est le constat que pour tout $\mathbf{z} \in \mathcal{Z}$ et $\mathbf{w} \in \Omega$,

$$\mathrm{I}(\mathrm{c}_{\mathbf{w}}(\mathbf{x}) \neq y) = \frac{1}{2} - \frac{y}{2}\operatorname{sgn}(\mathbf{w}'\mathbf{x}).$$

En général, la seconde espérance (sur $\mathbf{w}$) de l'équation 2.2 est approximée par une méthode de Monte-Carlo, mais il s'agit d'une procédure potentiellement lourde en calculs (section 5 de (20)). Toutefois, il a été démontré que si $\mathcal{Q}$ est une distribution normale **sphérique**, c'est-à-dire que la matrice de covariance est la matrice identité, alors une forme analytique existe (27). L'une des contributions du mémoire est de trouver une forme analytique si $\mathcal{Q}$ appartient à la famille des distributions elliptiques (8). De plus, une forme analytique est aussi trouvée si la fonction signe est remplacée par la fonction **maximum**, connue par la communauté de l'apprentissage profond sous le nom anglais de *rectified linear unit* (ReLU).

Définition 2. Pour tout nombre réel a, la fonction maximum est définie comme

$$\max(0, a) = \begin{cases} 0 & \text{si } a \leq 0, \\ a & \text{si } a > 0. \end{cases}$$

La famille des distributions elliptiques englobe de nombreuses distributions de probabilité célèbres, telles que la distribution de Student et la distribution logistique. Notamment, elle inclut aussi la distribution normale généralisée (22) contenant la distribution normale et la distribution de Laplace. Les distributions elliptiques sont fréquemment étudiées dans des domaines de la finance quantitative (21) comme l'actuariat.

Les deux lemmes suivants seront réutilisés pour la démonstration des deux théorèmes principaux. Les preuves sont relativement classiques et se retrouve dans le livre *Multivariate Statistical Simulation* (23).

Soit $\boldsymbol{\mu}$ un vecteur réel de dimension $n \times 1$, $\boldsymbol{\Sigma}$ une matrice réelle définie positive de dimension $n \times n$ et g une fonction réelle. La distribution elliptique de dimension n de paramètres $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ et g est dénotée par $\mathscr{E}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$. La densité de probabilité de la distribution elliptique $\mathscr{E}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ est définie comme

$$Ng((\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}))$$

où N est simplement une constante de normalisation.

Lemme 1. *Si* $\mathbf{w} \sim \mathscr{E}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$*, alors*

$$\mathbf{w}'\mathbf{x} \sim \mathscr{E}_1(\boldsymbol{\mu}'\mathbf{x}, \mathbf{x}'\boldsymbol{\Sigma}\mathbf{x}, g).$$

Lemme 2. *Si* $\mathbf{w} \sim \mathscr{E}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$*, alors*

$$\frac{\mathbf{w}'\mathbf{x} - \boldsymbol{\mu}'\mathbf{x}}{\sqrt{\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x}}} \sim \mathscr{E}_1(0, 1, g).$$

Le théorème suivant exprime l'espérance et la variance de $\operatorname{sgn}(\mathbf{w}'\mathbf{x})$ sous une forme analytique simple qui ne dépend que de la fonction cumulative de la distribution elliptique standard $\mathscr{E}_1(0, 1, g)$. Le théorème a l'avantage de contourner un échantillonnage numérique possiblement fastidieux. Il s'agit d'une contribution du mémoire permettant l'extension des résultats existants à la famille des distributions elliptiques.

Théorème 2. *Si* $\mathbf{w} \sim \mathscr{E}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$*, alors pour tout* $\mathbf{x} \in \mathbb{R}$*,*

$$\mathop{\mathbb{E}}_{\mathbf{w}} \operatorname{sgn}(\mathbf{w}'\mathbf{x}) = 1 - 2\Phi(r) \quad \textit{et} \quad \mathop{\mathbb{V}}_{\mathbf{w}} \operatorname{sgn}(\mathbf{w}'\mathbf{x}) = 1 - (1 - 2\Phi(r))^2,$$

où Φ *est la fonction cumulative de la distribution elliptique standard et* $r = -\frac{\boldsymbol{\mu}'\mathbf{x}}{\sqrt{\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x}}}$*.*

Démonstration. Par le lemme 1 (première ligne), le théorème de l'espérance totale (deuxième ligne) et le lemme 2 (quatrième ligne),

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{w}} \operatorname{sgn}(\mathbf{w}'\mathbf{x}) &= \mathop{\mathbb{E}}_{\mathbf{w}'\mathbf{x}} \operatorname{sgn}(\mathbf{w}'\mathbf{x}) \\
&= \mathbb{P}(\mathbf{w}'\mathbf{x} > 0) \mathop{\mathbb{E}}_{\mathbf{w}|\mathbf{w}'\mathbf{x} \le 0} \operatorname{sgn}(\mathbf{w}'\mathbf{x}) + \mathbb{P}(\mathbf{w}'\mathbf{x} \le 0) \mathop{\mathbb{E}}_{\mathbf{w}|\mathbf{w}'\mathbf{x} \le 0} \operatorname{sgn}(\mathbf{w}'\mathbf{x}) \\
&= \mathbb{P}(\mathbf{w}'\mathbf{x} > 0) - \mathbb{P}(\mathbf{w}'\mathbf{x} \le 0) = 1 - 2\,\mathbb{P}(\mathbf{w}'\mathbf{x} \le 0). \\
&= 1 - 2\,\mathbb{P}\left(\frac{\mathbf{w}'\mathbf{x} - \boldsymbol{\mu}'\mathbf{x}}{\sqrt{\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x}}} \le r\right) = 1 - 2\Phi(r), \qquad (2.3)
\end{aligned}$$

où Φ est la fonction cumulative de la distribution elliptique standard et $r = -\frac{\boldsymbol{\mu}'\mathbf{x}}{\sqrt{\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x}}}$. Par la définition de la variance (première ligne), le théorème de l'espérance totale (deuxième ligne) et l'équation 2.3 (quatrième ligne),

$$\begin{aligned}
\mathop{\mathbb{V}}_{\mathbf{w}} \operatorname{sgn}(\mathbf{w}'\mathbf{x}) &= \mathop{\mathbb{E}}_{\mathbf{w}} \operatorname{sgn}^2(\mathbf{w}'\mathbf{x}) - \left(\mathop{\mathbb{E}}_{\mathbf{w}} \operatorname{sgn}(\mathbf{w}'\mathbf{x})\right)^2 \\
&= \mathbb{P}(\mathbf{w}'\mathbf{x} > 0) \mathop{\mathbb{E}}_{\mathbf{w}|\mathbf{w}'\mathbf{x} \le 0} \operatorname{sgn}^2(\mathbf{w}'\mathbf{x}) + \mathbb{P}(\mathbf{w}'\mathbf{x} \le 0) \mathop{\mathbb{E}}_{\mathbf{w}|\mathbf{w}'\mathbf{x} \le 0} \operatorname{sgn}^2(\mathbf{w}'\mathbf{x}) - \left(\mathop{\mathbb{E}}_{\mathbf{w}} \operatorname{sgn}(\mathbf{w}'\mathbf{x})\right)^2 \\
&= \mathbb{P}(\mathbf{w}'\mathbf{x} > 0) + \mathbb{P}(\mathbf{w}'\mathbf{x} \le 0) - \left(\mathop{\mathbb{E}}_{\mathbf{w}} \operatorname{sgn}(\mathbf{w}'\mathbf{x})\right)^2 \\
&= 1 - \left(\mathop{\mathbb{E}}_{\mathbf{w}} \operatorname{sgn}(\mathbf{w}'\mathbf{x})\right)^2 \\
&= 1 - (1 - 2\Phi(r))^2.
\end{aligned}$$

□

Le théorème suivant est analogue au théorème 2, mais en remplaçant la fonction signe par la fonction maximum.

Théorème 3. *Si* $\mathbf{w} \sim \mathscr{E}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, *alors pour tout* $\mathbf{x} \in \mathbb{R}$,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{w}} \max(0, \mathbf{w}'\mathbf{x}) = (1 - \Phi(r)) \mathop{\mathbb{E}}_{\mathbf{w}|\mathbf{w}'\mathbf{x}\leq 0} \mathbf{w}'\mathbf{x}, \\
&\mathop{\mathbb{V}}_{\mathbf{w}} \max(0, \mathbf{w}'\mathbf{x}) = (1 - \Phi(r)) \Big(\mathop{\mathbb{E}}_{\mathbf{w}|\mathbf{w}'\mathbf{x}\leq 0} (\mathbf{w}'\mathbf{x})^2 - \mathop{\mathbb{E}}_{\mathbf{w}|\mathbf{w}'\mathbf{x}\leq 0} (\mathbf{w}'\mathbf{x}) \Big).
\end{aligned}$$

où Φ *est la fonction cumulative de la distribution elliptique standard et* $r = -\frac{\boldsymbol{\mu}'\mathbf{x}}{\sqrt{\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x}}}$.

Démonstration. Par le lemme 1 et le théorème de l'espérance totale,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{w}} \max(0, \mathbf{w}'\mathbf{x}) &= \mathop{\mathbb{E}}_{\mathbf{w}'\mathbf{x}} \max(0, \mathbf{w}'\mathbf{x}) \\
&= \mathbb{P}(\mathbf{w}'\mathbf{x} > 0) \mathop{\mathbb{E}}_{\mathbf{w}|\mathbf{w}'\mathbf{x}\leq 0} \max(0, \mathbf{w}'\mathbf{x}) + \mathbb{P}(\mathbf{w}'\mathbf{x} \leq 0) \mathop{\mathbb{E}}_{\mathbf{w}|\mathbf{w}'\mathbf{x}\leq 0} \max(0, \mathbf{w}'\mathbf{x}) \\
&= \mathbb{P}(\mathbf{w}'\mathbf{x} > 0) \mathop{\mathbb{E}}_{\mathbf{w}|\mathbf{w}'\mathbf{x}\leq 0} \mathbf{w}'\mathbf{x} \\
&= (1 - \mathbb{P}(\mathbf{w}'\mathbf{x} \leq 0)) \mathop{\mathbb{E}}_{\mathbf{w}|\mathbf{w}'\mathbf{x}\leq 0} \mathbf{w}'\mathbf{x}. \\
&= \left(1 - \mathbb{P}\left(\frac{\mathbf{w}'\mathbf{x} - \boldsymbol{\mu}'\mathbf{x}}{\sqrt{\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x}}} \leq r\right)\right) \mathop{\mathbb{E}}_{\mathbf{w}|\mathbf{w}'\mathbf{x}\leq 0} \mathbf{w}'\mathbf{x} = (1 - \Phi(r)) \mathop{\mathbb{E}}_{\mathbf{w}|\mathbf{w}'\mathbf{x}\leq 0} \mathbf{w}'\mathbf{x}, \quad (2.4)
\end{aligned}$$

où Φ est la fonction cumulative de la distribution elliptique standard et $r = -\frac{\boldsymbol{\mu}'\mathbf{x}}{\sqrt{\mathbf{x}^t\boldsymbol{\Sigma}\mathbf{x}}}$. Par la définition de la variance (première ligne), le théorème de l'espérance totale (deuxième ligne) et l'équation 2.4 (quatrième ligne),

$$\begin{aligned}
\mathop{\mathbb{V}}_{\mathbf{w}} \max(0, \mathbf{w}'\mathbf{x}) &= \mathop{\mathbb{E}}_{\mathbf{w}} \max^2(0, \mathbf{w}'\mathbf{x}) - \big(\mathop{\mathbb{E}}_{\mathbf{w}} \max(0, \mathbf{w}'\mathbf{x}) \big)^2 \\
&= \mathbb{P}(\mathbf{w}'\mathbf{x} > 0) \mathop{\mathbb{E}}_{\mathbf{w}|\mathbf{w}'\mathbf{x}\leq 0} \max^2(0, \mathbf{w}'\mathbf{x}) + \mathbb{P}(\mathbf{w}'\mathbf{x} \leq 0) \mathop{\mathbb{E}}_{\mathbf{w}|\mathbf{w}'\mathbf{x}\leq 0} \max^2(0, \mathbf{w}'\mathbf{x}) - \big(\mathop{\mathbb{E}}_{\mathbf{w}} \max(0, \mathbf{w}'\mathbf{x}) \big)^2 \\
&= \mathbb{P}(\mathbf{w}'\mathbf{x} > 0) \mathop{\mathbb{E}}_{\mathbf{w}|\mathbf{w}'\mathbf{x}\leq 0} (\mathbf{w}'\mathbf{x})^2 - \big(\mathop{\mathbb{E}}_{\mathbf{w}} \max(0, \mathbf{w}'\mathbf{x}) \big)^2 \\
&= (1 - \mathbb{P}(\mathbf{w}'\mathbf{x} > 0)) \mathop{\mathbb{E}}_{\mathbf{w}|\mathbf{w}'\mathbf{x}\leq 0} (\mathbf{w}'\mathbf{x})^2 - \big(\mathop{\mathbb{E}}_{\mathbf{w}} \max(0, \mathbf{w}'\mathbf{x}) \big)^2 \\
&= (1 - \Phi(r)) \Big(\mathop{\mathbb{E}}_{\mathbf{w}|\mathbf{w}'\mathbf{x}\leq 0} (\mathbf{w}'\mathbf{x})^2 - \mathop{\mathbb{E}}_{\mathbf{w}|\mathbf{w}'\mathbf{x}\leq 0} \mathbf{w}'\mathbf{x} \Big).
\end{aligned}$$

□

L'espérance $\mathbb{E}_{\mathbf{w}|\mathbf{w}'\mathbf{x}\leq 0} \mathbf{w}'\mathbf{x}$ du théorème 3 est simplement l'espérance d'une distribution elliptique tronquée pour les valeurs $\mathbf{w}'\mathbf{x}$ strictement positives. Une forme analytique de l'espérance est étudiée dans Landsman and Valdez (25).

Définition 3. Pour tout nombre réel z, la fonction d'**erreur** est définie comme

$$\operatorname{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a \exp\{-t^2\}\, dt.$$

Le corollaire suivant est spécifique aux distributions normales et il a déjà été rapporté dans certains travaux (26; 16).

Corollaire 1. *Si* $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, alors*

$$\mathop{\mathbb{E}}_{\mathbf{w}} \operatorname{sgn}(\mathbf{w}'\mathbf{x}) = \operatorname{erf}\left(\frac{\boldsymbol{\mu}'\mathbf{x}}{\sqrt{2\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x}}}\right).$$

Démonstration. Par le théorème 2 et l'expression de la fonction cumulative de la distribution normale standard, □

Corollaire 2. *Si* $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, alors*

$$\mathop{\mathbb{E}}_{\mathbf{w}} \max(0, \mathbf{w}'\mathbf{x}) = \frac{1}{2}\boldsymbol{\mu}'\mathbf{x}\left(1 - \operatorname{erf}\left(\frac{r}{\sqrt{2}}\right)\right) + \sqrt{\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x}}\,\phi(r),$$

où ϕ *est la fonction de densité de la distribution normale standard et* $r = -\frac{\boldsymbol{\mu}'\mathbf{x}}{\sqrt{\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x}}}$.

Démonstration. Par l'équation 3 et l'espérance d'une distribution normale tronquée (25),

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{w}} \max(0, \mathbf{w}'\mathbf{x}) &= (1 - \Phi(r))\left(\boldsymbol{\mu}'\mathbf{x} + \sqrt{\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x}}\frac{\phi(r)}{1 - \Phi(r)}\right) \\
&= (1 - \Phi(r))\boldsymbol{\mu}'\mathbf{x} + \sqrt{\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x}}\phi(r) \\
&= \left(1 - \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{r}{\sqrt{2}}\right)\right)\right) + \sqrt{\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x}}\phi(r).
\end{aligned}$$

□

Le théorème 2 permet d'obtenir une forme analytique de l'équivalence 2.2. Par exemple, pour une distribution normale, en utilisant le corollaire 1,

$$R_s(\mathcal{Q}) = \frac{1}{2} - \frac{1}{2}\sum_{\mathbf{z} \in \mathcal{S}} y \operatorname{erf}\left(\frac{\boldsymbol{\mu}'\mathbf{x}}{\sqrt{2\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x}}}\right). \tag{2.5}$$

Les résultats précédents s'adaptent facilement aux distributions de **mélange**, c'est-à-dire aux distributions de probabilités comprenant une somme pondérée de plusieurs distributions.

Travaux futurs. La limitation du théorème 2 est que la divergence de Kullback-Leibler entre deux distributions elliptiques ne semble pas avoir de forme analytique ou du moins simple. Néanmoins, il existe une forme pour deux distributions normales. Par conséquent, l'intérêt du théorème 2 dans une borne PAC-Bayes est réduit, car un échantillonnage de la DKL est tout de même nécessaire. Une question possible : est-ce qu'il existe une divergence s'exprimant de manière analytique pour deux distributions elliptiques ? Une autre avenue intéressante serait de trouver une équivalence analogue à l'équation 2.2, mais pour la fonction max. Le théorème 3 permettrait alors de calculer une forme analytique du risque empirique espérée. La fonction max est fréquemment utilisée comme fonction d'activation dans les réseaux de neurones.

Chapitre 3

Priors PAC-Bayes

Le prior et les données sont deux éléments très importants qui déterminent le posterior d'une borne PAC-Bayes. Le choix d'un prior ne quantifie pas seulement l'information disponible au sujet de l'espace des classifieurs, mais procure surtout une structure aux classifieurs. L'utilisation d'un prior PAC-Bayes capture la notion de chance (ou *luckiness* en anglais) proposée par Shawe-Taylor et al. (42). Un prior est chanceux si, sans aucun indice provenant des données, la valeur de la divergence de Kullback-Leibler est petite, c'est-à-dire que le prior est proche du posterior. Toutefois, sans surprise, il se trouve qu'être chanceux est difficile. Par conséquent, plusieurs bornes PAC-Bayes ne sont pas étanches en raison d'une valeur de DKL trop grande. Il apparaît alors nécessaire d'avoir des priors **informatifs**, c'est-à-dire informés par les données. La théorie PAC-Bayes, comme l'illustre le chapitre précédent, privilégie les données à une modélisation explicite. Le choix d'un prior grâce aux données n'est donc pas une proposition déraisonnable. Malheureusement, l'utilisation directe des données utilisées par le posterior est impossible. L'optimisation d'un prior est aussi connue sous le nom de **localisation** (9). Il y a deux approches principales pour traiter la question des priors localisées :

1. La dépendance aux données. L'optimisation d'un prior sur une partie des données n'ayant pas été utilisée par le posterior est possible (4). L'autre option est l'utilisation de toutes les données, mais l'imposition d'une condition de stabilité comme par exemple que le prior doit être stable aux petites perturbations dans les données (11).
2. La dépendance à la distribution source (28; 34; 29). La construction d'un prior reposant sur la distribution source. Le présent mémoire explore cette approche.

3.1 Dépendance à la source

La mesure de complexité joue un rôle primordial dans l'étanchéité d'une borne PAC-Bayes : un prior mal choisi ou trop générique amène à des valeurs de la DKL élevées. En effet, le DKL pénalise les grandes différences entre le prior et le posterior. Le posterior est décidé par

l'algorithme d'apprentissage, mais il est également influencé par le choix de prior par le biais de la DKL. La principale restriction dans la sélection d'un prior est l'impossibilité de recourir aux données utilisées par le posterior, sauf si certaines conditions sont imposées. Une astuce est la création d'un prior qui dépend de la distribution source $\mathscr{D}$. De cette façon, il est considéré comme fixé avant que les données ne soient observées. Évidemment, $\mathscr{D}$ est inconnue et donc la DKL n'est pas calculable. L'analyse suivante consiste à établir une borne supérieure sur la DKL. Encore une fois, le présent chapitre porte sur la classification linéaire binaire et reprend la notation des chapitres précédents.

Prior espérance. L'idée de Parrado-Hernández et al. (34) est la construction d'un prior espérance $\mathscr{P}_{\mathrm{e}}$ qui dépend de la distribution $\mathscr{D}$. Il s'agit d'une distribution normale sphérique et centrée sur le **centre de masse** $\mathbf{v}_d$. Le vecteur $\mathbf{v}_d$ représente la moyenne d'une variable aléatoire dépendant de $\mathscr{D}$. En particulier, le travail de Parrado-Hernández et al. (34) se concentre sur la variable aléatoire $\mathbf{V} = \mathrm{Y}\mathbf{X}$. La moyenne du centre de masse $\mathbf{v}_d$ et la moyenne du centre de masse empirique $\mathbf{v}_s$ sont respectivement définies comme

$$\mathbf{v}_d = \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathscr{D}} y\mathbf{x} \qquad \text{et} \qquad \mathbf{v}_s = \sum_{\mathbf{z}\sim\mathcal{S}} y\mathbf{x}\,. \tag{3.1}$$

Si les vecteurs de caractéristiques sont distribués en deux groupes de classes différentes autour de l'origine de l'espace $\mathscr{X}$ et que les classes sont équilibrées, alors le vecteur $\mathbf{v}_d$ est perpendiculaire à un hyperplan entre la moyenne des vecteurs de classe positive et la moyenne des vecteurs de classe négative. Le prior espérance est compatible avec différentes bornes PAC-Bayes, mais dans le présent mémoire, il est formulé avec la borne de Langford et Seeger. Par l'équation 2.1, avec probabilité au moins $1-\delta$, pour tout posterior $\mathscr{Q}_{\mathrm{e}}$ de la forme $\mathscr{N}(\mathbf{v}_q, \mathrm{I})$ avec $\mathbf{v}_q \in \mathscr{X}$,

$$\begin{aligned}\mathrm{kl}\{R_s(\mathscr{Q}_{\mathrm{e}}) \| R_d(\mathscr{Q}_{\mathrm{e}})\} &\leq \frac{1}{m}\left(\mathrm{KL}\{\mathscr{Q}_{\mathrm{e}} \| \mathscr{P}_{\mathrm{e}}\} + \log\frac{2\sqrt{m}}{\delta}\right) \\ &= \frac{1}{m}\left(\frac{1}{2}\|\mathbf{v}_q - \mathbf{v}_d\|^2 + \log\frac{2\sqrt{m}}{\delta}\right).\end{aligned} \tag{3.2}$$

La mesure de complexité est problématique, car elle n'est pas calculable. La DKL entre deux distributions normales sphériques est simplement la norme euclidienne entre la différence des vecteurs $\mathbf{v}_d$ et $\mathbf{v}_q$. Évidemment, elle n'est pas non plus calculable. La stratégie consiste à la borner supérieurement par des quantités calculables. Par l'inégalité de Cauchy-Schwarz,

$$\begin{aligned}\|\mathbf{v}_d - \mathbf{v}_q\|^2 &= \|\mathbf{v}_d - \mathbf{v}_s + \mathbf{v}_s - \mathbf{v}_q\|^2 \\ &\leq \|\mathbf{v}_d - \mathbf{v}_s\|^2 + \|\mathbf{v}_s - \mathbf{v}_q\|^2 + 2\|\mathbf{v}_d - \mathbf{v}_s\|\|\mathbf{v}_s - \mathbf{v}_q\| \\ &= \left(\|\mathbf{v}_d - \mathbf{v}_s\| + \|\mathbf{v}_s - \mathbf{v}_q\|\right)^2.\end{aligned} \tag{3.3}$$

La difficulté revient alors à borner $\|\mathbf{v}_d - \mathbf{v}_s\|$. Heureusement, $\mathbf{v}_d$ et $\mathbf{v}_s$ ne sont pas des termes quelconques. L'un est une approximation empirique de l'autre et il est possible de se servir

d'une inégalité de **concentration** pour borner $\|\mathbf{v}_d - \mathbf{v}_s\|$. Une excellente référence sur les inégalités de concentration est le livre *Concentration Inequalities : A Nonasymptotic Theory of Independence* (7).

Le prochain lemme borne $\|\mathbf{v}_d - \mathbf{v}_s\|$ en utilisant une généralisation de la célèbre inégalité de Hoeffding, c'est-à-dire l'inégalité de McDiarmid. L'inégalité a recours à une notion de stabilité. Une fonction satisfait la condition des **différences bornées** si la valeur ne change pas trop lorsqu'un des arguments est modifié. Formellement, une fonction f prenant en argument des vecteurs aléatoires indépendants $\mathbf{X}_1, \ldots, \mathbf{X}_m \in \mathcal{X}$ satisfait la condition des différences bornées si pour tout $1 \leq k \leq m$, alors il existe une constante $c_k \geq 0$ telle que

$$\sup_{\mathbf{x}_1,\ldots,\mathbf{x}_m,\mathbf{x}'_k \in \mathcal{X}} |f(\mathbf{x}_1, \ldots, \mathbf{x}_k, \ldots, \mathbf{x}_m) - f(\mathbf{x}_1, \ldots, \mathbf{x}'_k, \ldots, \mathbf{x}_m)| \leq c_k.$$

Théorème 4 (Inégalité de McDiarmid)**.** *Soit f une fonction qui satisfait la condition des différences bornées. Pour tout $\epsilon > 0$,*

$$\mathbb{P}\big(f(\mathbf{x}_1, \ldots, \mathbf{x}_m) - \mathbb{E}\, f(\mathbf{x}_1, \ldots, \mathbf{x}_m) \geq \epsilon\big) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{k=1}^m c_k^2}\right).$$

Lemme 3 (Shawe-Taylor and Cristianini (40))**.** *Pour tout $\delta \in (0,1)$, avec probabilité au moins $1-\delta$,*

$$\|\mathbf{v}_d - \mathbf{v}_s\| \leq \frac{K}{\sqrt{m}}\left(2 + \sqrt{2\log\tfrac{1}{\delta}}\right),$$

où $K = \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|$.

Démonstration. Soit f une fonction définie comme

$$f_d(\mathcal{S}) = \|\mathbf{v}_d - \mathbf{v}_s\|.$$

Soit $\mathcal{S}'$ un échantillon identique à $\mathcal{S}$, sauf pour l'exemple $\mathbf{z}_k$ qui est remplacé par $\mathbf{z}'_k \in \mathcal{Z}$. Par l'inégalité du triangle inverse (deuxième ligne),

$$\begin{aligned} |f_d(\mathcal{S}) - f_d(\mathcal{S}')| &= |\|\mathbf{v}_d - \mathbf{v}_s\| - \|\mathbf{v}_d - \mathbf{v}_{s'}\|| \\ &\leq \|\mathbf{v}_s - \mathbf{v}_{s'}\| = \frac{1}{m}\|\mathbf{z}_k - \mathbf{z}'_k\| \\ &\leq \frac{2K}{m}, \end{aligned}$$

où $K = \sup_{\mathbf{z}\in\mathcal{Z}} \|y\mathbf{x}\| = \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|$. Par l'inégalité de McDiarmid, pour tout $\epsilon > 0$,

$$\mathbb{P}\Big(f_d(\mathcal{S}) - \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m} f_d(\mathcal{S}) \leq \epsilon\Big) \leq 1 - \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m (\frac{2K}{m})^2}\right) = 1 - \exp\left(\frac{-m\epsilon^2}{2K^2}\right).$$

En posant le terme de droite égal à $1-\delta$ avec $\delta \in (0,1)$, alors

$$\epsilon = \sqrt{\frac{2K^2}{m}\log\frac{1}{\delta}}.$$

Par conséquent, avec probabilité au moins $1-\delta$,

$$\|\mathbf{v}_d - \mathbf{v}_s\| - \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathscr{D}^m} \|\mathbf{v}_d - \mathbf{v}_s\| \leq \sqrt{\frac{2K^2}{m}\log\frac{1}{\delta}}. \tag{3.4}$$

Maintenant, il reste à borner supérieurement le terme $\mathbb{E}_{\mathcal{S}\sim\mathscr{D}^m} \|\mathbf{v}_d - \mathbf{v}_s\|$. Soit $\tilde{\mathcal{S}} \sim \mathscr{D}^m$ un échantillon de m exemples. En observant que

$$\mathbf{v}_d = \mathop{\mathbb{E}}_{\tilde{\mathcal{S}}\sim\mathscr{D}^m} \mathbf{v}_{\tilde{s}}$$

et par l'inégalité du triangle (troisième ligne),

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathscr{D}^m} \|\mathbf{v}_d - \mathbf{v}_s\| &= \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathscr{D}^m} \left\| \mathop{\mathbb{E}}_{\tilde{\mathcal{S}}\sim\mathscr{D}^m} (\mathbf{v}_{\tilde{s}} - \mathbf{v}_s) \right\| \\ &\leq \mathop{\mathbb{E}}_{\mathcal{S},\tilde{\mathcal{S}}\sim\mathscr{D}^m} \|\mathbf{v}_{\tilde{s}} - \mathbf{v}_s\|.\end{aligned}$$

Soit $\mathbf{B} = (B_1, \ldots, B_m) \in \{-1, 1\}^m$ un vecteur aléatoire de dimension m distribué selon la distribution de Bernouilli équilibrée $\mathscr{B}$. En observant que

$$\mathop{\mathbb{E}}_{\mathcal{S},\tilde{\mathcal{S}}\sim\mathscr{D}^m} \|\mathbf{v}_{\tilde{s}} - \mathbf{v}_s\| = \frac{1}{m} \mathop{\mathbb{E}}_{\mathcal{S},\tilde{\mathcal{S}},\mathscr{B}} \left\| \sum_{i=1}^m b_i \tilde{\mathbf{z}}_i - \sum_{i=1}^m b_i \mathbf{z}_i \right\|$$

et par l'inégalité du triangle (première ligne) et comme $\mathcal{S}$ et $\mathcal{S}'$ ont tous deux été générés par $\mathscr{D}$ (deuxième ligne),

$$\begin{aligned}\frac{1}{m} \mathop{\mathbb{E}}_{\mathcal{S},\tilde{\mathcal{S}},\mathscr{B}} \left\| \sum_{i=1}^m b_i \tilde{\mathbf{z}}_i - \sum_{i=1}^m b_i \mathbf{z}_i \right\| &\leq \frac{1}{m} \mathop{\mathbb{E}}_{\mathcal{S},\tilde{\mathcal{S}},\mathscr{B}} \left(\left\| \sum_{i=1}^m b_i \tilde{\mathbf{z}}_i \right\| + \left\| \sum_{i=1}^m b_i \mathbf{z}_i \right\| \right) \\ &= \frac{2}{m} \mathop{\mathbb{E}}_{\mathcal{S},\mathscr{B}} \left\| \sum_{i=1}^m b_i \mathbf{z}_i \right\|.\end{aligned}$$

Par l'inégalité de Jensen (première ligne) et comme les signes contraires b_i et b_j s'annulent (deuxième ligne),

$$\begin{aligned}\frac{2}{m} \mathop{\mathbb{E}}_{\mathcal{S},\mathscr{B}} \left\| \sum_{i=1}^m b_i \mathbf{z}_i \right\| &\leq \frac{2}{m} \sqrt{\mathop{\mathbb{E}}_{\mathcal{S},\mathscr{B}} \sum_{i,j=1}^m b_i b_j \mathbf{z}_i \mathbf{z}_j} \\ &= \frac{2}{m} \sqrt{\mathop{\mathbb{E}}_{\mathcal{S}} \sum_{i=1}^m \|\mathbf{z}_i\|^2} \\ &\leq \frac{2}{m} \sqrt{mK^2} = \frac{2K}{\sqrt{m}}.\end{aligned} \tag{3.5}$$

Finalement, en substituant l'équation 3.5 dans l'équation 3.4, avec probabilité $1-\delta$,

$$\|\mathbf{v}_d - \mathbf{v}_s\| - \frac{2K}{\sqrt{m}} \leq \sqrt{\frac{2K^2}{m}\log\frac{1}{\delta}}.$$

□

Par la borne 3.2 et le lemme 3, avec probabilité au moins $1-\delta$,

$$\mathrm{kl}\{R_s(\mathcal{Q}_\mathrm{e})\|R_d(\mathcal{Q}_\mathrm{e})\} \leq \frac{1}{m}\left(\frac{1}{2}\left(\|\mathbf{v}_q-\mathbf{v}_s\|+\tfrac{K}{\sqrt{m}}\left(2+\sqrt{2\log\tfrac{2}{\delta}}\right)\right)^2+\log\tfrac{4\sqrt{m}}{\delta}\right). \tag{3.6}$$

Dans un contexte de classification sur des jeux de données de UCI (33), la borne précédente s'est révélée être plus étanche que la borne classique (34). Les résultats empiriques confirment l'intuition qu'une garantie PAC-Bayes se conjugue bien avec un prior informatif.

3.2 Prior espérance avec covariance pleine

L'analyse de la section précédente (34) a encouragé d'autres travaux sur les priors localisés (29; 11). Dans le mémoire, l'un des objectifs est la création d'une borne PAC-Bayes similaire au théorème 3.6, mais encore encore plus flexible, c'est-à-dire avec un prior espérance $\mathscr{P}_\mathrm{c}$ qui n'est pas nécessairement sphérique. La matrice de covariance et la matrice de covariance empirique sont respectivement définies comme

$$\boldsymbol{\Sigma}_d = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathscr{D}}\mathbf{x}\mathbf{x}' - \mathbf{v}_d\mathbf{v}_d' \qquad \text{et} \qquad \boldsymbol{\Sigma}_s = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i' - \mathbf{v}_s\mathbf{v}_s'. \tag{3.7}$$

Le prior $\mathscr{P}_\mathrm{c}$ est de la forme $\mathscr{N}(\mathbf{v}_d, \boldsymbol{\Sigma}_d^{-1})$ et le posterior $\mathcal{Q}_\mathrm{c}$ est de la forme $\mathscr{N}(\mathbf{v}_q, \boldsymbol{\Sigma}_q^{-1})$ avec $\boldsymbol{\Sigma}_q^{-1} \in \mathscr{X}\times\mathscr{X}$. La principale différence dans la dérivation de la borne PAC-Bayes est que la DKL est plus complexe. En effet, la DKL entre deux distributions normales est définie comme

$$\mathrm{KL}\{\mathcal{Q}_\mathrm{c}\|\mathscr{P}_\mathrm{c}\} = \frac{1}{2}\Bigg(\underbrace{(\mathbf{v}_d-\mathbf{v}_q)'\boldsymbol{\Sigma}_d(\mathbf{v}_d-\mathbf{v}_q)}_{\text{(i)}} + \underbrace{\log\tfrac{\det(\boldsymbol{\Sigma}_d^{-1})}{\det(\boldsymbol{\Sigma}_q)}}_{\text{(ii)}} + \underbrace{\mathrm{tr}(\boldsymbol{\Sigma}_d\boldsymbol{\Sigma}_q)}_{\text{(iii)}} - n\Bigg). \tag{3.8}$$

Les trois termes (i), (ii), (iii) seront bornés par les lemmes suivants.

Lemme 4. (i) $= (\mathbf{v}_d-\mathbf{v}_q)'\boldsymbol{\Sigma}_d(\mathbf{v}_d-\mathbf{v}_q)$ *est borné supérieurement par*

$$\left(\|\boldsymbol{\Sigma}_d-\boldsymbol{\Sigma}_s\|+\|\boldsymbol{\Sigma}_s\|\right)\left(\|\mathbf{v}_d-\mathbf{v}_s\|+\|\mathbf{v}_s-\mathbf{v}_q\|\right)^2.$$

Démonstration. Par l'inégalité de Cauchy-Schwarz (seconde ligne),

$$\begin{aligned}(\mathbf{v}_d-\mathbf{v}_q)'\boldsymbol{\Sigma}_d(\mathbf{v}_d-\mathbf{v}_q) &= \|\mathbf{v}_d-\mathbf{v}_q\|_{\boldsymbol{\Sigma}_d}^2 = \|\mathbf{v}_d-\mathbf{v}_s+\mathbf{v}_s-\mathbf{v}_q\|_{\boldsymbol{\Sigma}_d}^2\\ &\leq \|\mathbf{v}_d-\mathbf{v}_s\|_{\boldsymbol{\Sigma}_d}^2 + \|\mathbf{v}_s-\mathbf{v}_q\|_{\boldsymbol{\Sigma}_d}^2 + 2\|\mathbf{v}_d-\mathbf{v}_s\|_{\boldsymbol{\Sigma}_d}\|\mathbf{v}_s-\mathbf{v}_q\|_{\boldsymbol{\Sigma}_d}.\end{aligned}$$

Soit $\boldsymbol{\Sigma}_d = \mathbf{L}'\mathbf{L}$ une décomposition de Cholesky. Encore une fois, par l'inégalité de Cauchy-Schwarz (second ligne),

$$\begin{aligned}\|\mathbf{v}_d-\mathbf{v}_s\|_{\boldsymbol{\Sigma}_d}^2 &= (\mathbf{v}_d-\mathbf{v}_s)'\mathbf{L}'\mathbf{L}(\mathbf{v}_d-\mathbf{v}_s) = (\mathbf{L}(\mathbf{v}_d-\mathbf{v}_s))'\mathbf{L}(\mathbf{v}_d-\mathbf{v}_s)\\ &= \|\mathbf{L}(\mathbf{v}_d-\mathbf{v}_s)\|^2 \leq \|\boldsymbol{\Sigma}_d\|\|\mathbf{v}_d-\mathbf{v}_s\|^2.\end{aligned}$$

Le même stratagème s'applique au terme $\|\mathbf{v}_s - \mathbf{v}_q\|_{\boldsymbol{\Sigma}_d}$. Finalement,

$$\|\mathbf{v}_d - \mathbf{v}_q\|^2_{\boldsymbol{\Sigma}_d} \leq \|\boldsymbol{\Sigma}_d\| \Big(\|\mathbf{v}_d - \mathbf{v}_s\|^2 + \|\mathbf{v}_s - \mathbf{v}_q\|^2 + 2\|\mathbf{v}_d - \mathbf{v}_s\| \|\mathbf{v}_s - \mathbf{v}_q\| \Big). \tag{3.9}$$

Par l'inégalité du triangle,

$$\|\mathbf{v}_d - \mathbf{v}_q\|^2_{\boldsymbol{\Sigma}_d} \leq (\|\boldsymbol{\Sigma}_d - \boldsymbol{\Sigma}_s\| + \|\boldsymbol{\Sigma}_s\|)(\|\mathbf{v}_d - \mathbf{v}_s\|_2^2 + \|\mathbf{v}_s - \mathbf{v}_q\|_2^2 + 2\|\mathbf{v}_d - \mathbf{v}_s\| \|\mathbf{v}_s - \mathbf{v}_q\|).$$

□

Il convient de remarquer que l'expression au carré dans l'équation 3.9 est la même que dans l'équation 3.3. La liberté d'avoir une matrice de covariance complète se paie par un facteur multiplicatif correspondant à la plus grande valeur singulière de la matrice $\boldsymbol{\Sigma}_d$.

Lemme 5. (ii) $= \log \frac{\det(\boldsymbol{\Sigma}_d^{-1})}{\det(\boldsymbol{\Sigma}_q)}$ *est borné supérieurement par* $n \log(\|\boldsymbol{\Sigma}_d - \boldsymbol{\Sigma}_s\| + \|\boldsymbol{\Sigma}_s\|) - \log\det \boldsymbol{\Sigma}_q$.

Démonstration. Par le caractère défini positif de $\boldsymbol{\Sigma}_d$,

$$\begin{aligned}
\log \frac{\det(\boldsymbol{\Sigma}_d^{-1})}{\det(\boldsymbol{\Sigma}_q)} &= -\log\det(\boldsymbol{\Sigma}_d) - \log\det(\boldsymbol{\Sigma}_q) = -\log \prod_{i=1}^{n} \lambda_i(\boldsymbol{\Sigma}_d) - \log\det(\boldsymbol{\Sigma}_q) \\
&= -\sum_{i=1}^{n} \log \lambda_i(\boldsymbol{\Sigma}_d) - \log\det(\boldsymbol{\Sigma}_q) \leq -n \log \lambda_{\min}(\boldsymbol{\Sigma}_d) - \log\det(\boldsymbol{\Sigma}_q) \\
&= n \log \|\boldsymbol{\Sigma}_d\| - \log\det(\boldsymbol{\Sigma}_q) \leq n \log(\|\boldsymbol{\Sigma}_d - \boldsymbol{\Sigma}_s\| + \|\boldsymbol{\Sigma}_s\|) - \log\det \boldsymbol{\Sigma}_q.
\end{aligned}$$

□

Lemme 6. (iii) $= \operatorname{tr}(\boldsymbol{\Sigma}_d \boldsymbol{\Sigma}_q)$ *est borné supérieurement par* $(\|\boldsymbol{\Sigma}_d - \boldsymbol{\Sigma}_s\| + \|\boldsymbol{\Sigma}_s\|) \operatorname{tr}(\boldsymbol{\Sigma}_q)$.

Démonstration. Par l'inégalité de la trace de Von Neumann et le caractère défini positif de $\boldsymbol{\Sigma}_q$,

$$\begin{aligned}
\operatorname{tr}(\boldsymbol{\Sigma}_d \boldsymbol{\Sigma}_q) &\leq \sum_{i=1}^{n} \sigma_i(\boldsymbol{\Sigma}_d)\sigma_i(\boldsymbol{\Sigma}_q) \leq \sum_{i=1}^{n} \|\boldsymbol{\Sigma}_d\| \sigma_i(\boldsymbol{\Sigma}_q) = \|\boldsymbol{\Sigma}_d\| \sum_{i=1}^{n} \lambda_i(\boldsymbol{\Sigma}_q) = \|\boldsymbol{\Sigma}_d\| \operatorname{tr}(\boldsymbol{\Sigma}_q) \\
&\leq (\|\boldsymbol{\Sigma}_d - \boldsymbol{\Sigma}_s\| + \|\boldsymbol{\Sigma}_s\|) \operatorname{tr}(\boldsymbol{\Sigma}_q).
\end{aligned}$$

□

Les trois lemmes précédents possèdent tous le même terme incalculable : la matrice $\boldsymbol{\Sigma}_d$ est inconnue et il faut une borne supérieure calculable sur $\|\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_d\|$. La recherche de la borne n'a pas été directe. En effet, il existe quelques bornes pour $\|\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_d\|$, mais la majorité contiennent des constantes inconnues ou possèdent des hypothèses incompatibles avec la situation présente (43). À la suite d'une recherche infructueuse, le lemme suivant a été développé de manière originale à partir du lemme 3.

Lemme 7. *Pour tout $\delta \in (0,1)$, avec probabilité au moins $1-\delta$,*

$$\|\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_d\| \leq \tfrac{2K^2}{\sqrt{m^3}}\left((m-1)\sqrt{2\log\tfrac{1}{\delta}} + 2(m+\sqrt{m})\right).$$

La preuve du lemme précédent est en annexe. Néanmoins, il se trouve qu'un meilleur résultat existait déjà. En effet, le lemme suivant sera retenu dans la suite de l'analyse.

Lemme 8 (Shawe-Taylor and Cristianini (39))**.** *Pour tout $\delta \in (0,1)$, avec probabilité au moins $1-\delta$,*

$$\|\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_d\| \leq \tfrac{2K^2}{\sqrt{m}}\left(2 + \sqrt{2\log\tfrac{2}{\delta}}\right).$$

Finalement, tous les ingrédients sont disponibles pour énoncer la prochaine contribution du mémoire.

Théorème 5. *Pour toute distribution $\mathscr{D} \in \mathscr{M}(\mathscr{X})$ et un échantillon $\mathscr{S} \sim \mathscr{D}^m$, pour tout espace des hypothèses $\mathscr{C}$, pour tout $\mathscr{P}_{\mathrm{c}}$ de la forme $\mathscr{N}(\mathbf{v}_d, \boldsymbol{\Sigma}_d^{-1})$, pour tout $\delta \in (0,1)$, avec probabilité au moins $1-\delta$, pour tout posterior $\mathscr{Q}_{\mathrm{c}}$ de la forme $\mathscr{N}(\mathbf{v}_q, \boldsymbol{\Sigma}_q^{-1})$,*

$$\mathrm{kl}\{R_s(\mathscr{Q}_{\mathrm{c}})\|R_d(\mathscr{Q}_{\mathrm{c}})\} \leq \frac{1}{2m}\Bigg(\Big((\alpha + \|\boldsymbol{\Sigma}_s\|)\left((\beta + \|\mathbf{v}_s - \mathbf{v}_q\|)^2 + \operatorname{tr}\boldsymbol{\Sigma}_q\right) + n\log\frac{\alpha + \|\boldsymbol{\Sigma}_s\|}{\det\boldsymbol{\Sigma}_q} - n\Big) + 2\log\frac{6\sqrt{m}}{\delta}\Bigg).$$

où $\alpha = \frac{2K^2}{\sqrt{m}}\left(2 + \sqrt{2\log\frac{6}{\delta}}\right)$, $\beta = \frac{K}{\sqrt{m}}\left(2 + \sqrt{2\log\frac{3}{\delta}}\right)$ et $K = \sup_{\mathbf{x}\in\mathscr{X}}\|\mathbf{x}\|$.

Démonstration. En reprenant l'équation 3.8 et par les lemmes 4, 5 and 6,

$$\mathrm{KL}\{\mathscr{Q}_{\mathrm{c}}\|\mathscr{P}_{\mathrm{c}}\} \leq \tfrac{1}{2}\Big((\|\boldsymbol{\Sigma}_d - \boldsymbol{\Sigma}_s\| + \|\boldsymbol{\Sigma}_s\|)(\|\mathbf{v}_d - \mathbf{v}_s\| + \|\mathbf{v}_s - \mathbf{v}_q\|)^2 + n\log(\|\boldsymbol{\Sigma}_d - \boldsymbol{\Sigma}_s\| + \|\boldsymbol{\Sigma}_s\|) - \log\det\boldsymbol{\Sigma}_q + (\|\boldsymbol{\Sigma}_d - \boldsymbol{\Sigma}_s\| + \|\boldsymbol{\Sigma}_s\|)\operatorname{tr}(\boldsymbol{\Sigma}_q) - n\Big).$$

Par les lemmes 3 et 8 avec la borne de l'union, avec probabilité au moins $1 - \frac{2}{3\delta}$,

$$\mathrm{KL}\{\mathscr{Q}_{\mathrm{c}}\|\mathscr{P}_{\mathrm{c}}\} \leq \frac{1}{2}\left((\alpha + \|\boldsymbol{\Sigma}_s\|)\left((\beta + \|\mathbf{v}_s - \mathbf{v}_q\|)^2 + \operatorname{tr}\boldsymbol{\Sigma}_q\right) + n\log\frac{\alpha + \|\boldsymbol{\Sigma}_s\|}{\det\boldsymbol{\Sigma}_q} - n\right)$$

avec $\alpha = \frac{2K^2}{\sqrt{m}}\left(2 + \sqrt{2\log\frac{6}{\delta}}\right)$, $\beta = \frac{K}{\sqrt{m}}\left(2 + \sqrt{2\log\frac{3}{\delta}}\right)$ et $K = \sup_{\mathbf{x}\in\mathscr{X}}\|\mathbf{x}\|$. Finalement, par la borne de Langford et Seeger et encore une fois la borne de l'union, avec probabilité au moins $1-\delta$,

$$\mathrm{kl}\{R_s(\mathscr{Q}_{\mathrm{c}})\|R_d(\mathscr{Q}_{\mathrm{c}})\} \leq \frac{1}{m}\Bigg(\frac{1}{2}\Big((\alpha + \|\boldsymbol{\Sigma}_s\|)\left((\beta + \|\mathbf{v}_s - \mathbf{v}_q\|)^2 + \operatorname{tr}\boldsymbol{\Sigma}_q\right) + n\log\frac{\alpha + \|\boldsymbol{\Sigma}_s\|}{\det\boldsymbol{\Sigma}_q} - n\Big) + \log\frac{6\sqrt{m}}{\delta}\Bigg).$$

□

Le théorème précédent permet de prendre en compte une matrice de covariance pleine. Il s'agit d'une extension du théorème de Parrado-Hernández et al. (34) qui fonctionne pour une matrice de covariance identité. La borne 5 ajoute sans surprise quelques termes moins étanches. Néanmoins, il y a possiblement un grand gain à avoir au niveau du risque empirique espéré.

Conclusion

La notion de généralisation est primordiale en apprentissage automatique. En effet, il devient nécessaire de garantir les capacités de généralisation des modèles d'apprentissage. La théorie PAC-Bayes se propose comme une solution fréquentiste élégante, mais avec une touche bayésienne. Elle applique des garanties de type PAC aux algorithmes bayésiens généralisés. Il y a plusieurs avantages : la modélisation est flexible et il est possible d'avoir des fonctions objectives basées sur des bornes de généralisation. Néanmoins, les bornes PAC-Bayes sont souvent peu étanches en raison d'une grande dissimilarité entre le prior et le posterior. L'une des manières de réduire la mesure de complexité est l'utilisation un prior qui dépend des données. L'importance des priors informatifs a été soulevée par les travaux de Parrado-Hernández et al. (34). Il s'agit de la création d'un prior qui dépend de la distribution source. Le mémoire a montré qu'il est possible d'étendre la borne obtenue dans Parrado-Hernández et al. (34) à des priors espérances avec une matrice de covariance pleine.

Travaux futurs.

1. La suite immédiate du travail contenu dans le présent mémoire est la réalisation d'expérimentations empiriques de la borne du théorème 5. Il serait intéressant de la comparer avec la borne obtenue par (34). Est-ce qu'une matrice de covariance informée par les données peut rendre une borne PAC-Bayes plus étanche ?
2. Une piste intéressante est aussi le lien entre les bornes PAC-Bayes et les critères d'information. En effet, le BIC correspond au *unit information prior* (UIP) (24) dans l'utilisation des facteurs de Bayes. Le prior UIP est simplement une distribution normale centrée sur le vecteur de vraisemblance maximale et avec comme covariance la matrice d'information de Fisher. Quelle serait la signification d'une borne PAC-Bayes avec le UIP ?
3. Une autre extension possible est l'étude des priors espérances d'une forme différente de $\mathbf{V} = \mathrm{Y}\mathbf{X}$. La forme de $\mathbf{V}$ doit être choisie en fonction des spécificités du problème d'apprentissage et du type de modèles utilisés. L'une des difficultés est qu'un prior opère dans l'espace des hypothèses $\mathscr{C}$ et qu'il n'est pas toujours simple d'anticiper le résultat dans l'espace des données $\mathscr{X}$.

Annexe A

Annexe - Démonstration supplémentaire effectuée au cours de la maîtrise

La démonstration suivante a nécessité un effort non négligeable au cours des activités de recherche du mémoire. En effet, il n'était pas évident de savoir quelle était la bonne direction à prendre. Ce n'est qu'après être arrivé au résultat suivant que nous avons pris conscience de l'existence du lemme 8 dans la littérature. Bien que le lemme 8 permet de lier la norme de $\mathbf{\Sigma}_d$ avec un peu plus de précision, il s'agit d'un résultat très similaire au nôtre.

Preuve du lemme 7

Démonstration. Soit $\mathcal{S}'$ un échantillon identique à $\mathcal{S}$, sauf pour l'exemple $\mathbf{z}_k$ qui est remplacé par $\mathbf{z}'_k \in \mathcal{Z}$. Par l'inégalité du triangle,

$$\begin{aligned}
|\|\mathbf{\Sigma}_s - \mathbf{\Sigma}_d\| - \|\mathbf{\Sigma}_{s'} - \mathbf{\Sigma}_d\|| &\leq \|\mathbf{\Sigma}_s - \mathbf{\Sigma}_{s'}\| \\
&= \left\| \tfrac{1}{m}(\mathbf{v}_r\mathbf{v}_r^t - \mathbf{v}_r\mathbf{v}_r^t) - \tfrac{1}{m^2}(\textstyle\sum \mathbf{v} \sum \mathbf{v}^t - \sum' \mathbf{v} \sum' \mathbf{v}^t) \right\| \\
&\leq \tfrac{1}{m}\|\mathbf{v}_r\mathbf{v}_r^t - \mathbf{v}_r\mathbf{v}_r^t\| + \tfrac{1}{m^2}\|\textstyle\sum \mathbf{v} \sum \mathbf{v}^t - \sum' \mathbf{v} \sum' \mathbf{v}^t\| \\
&= \tfrac{1}{m}\|\mathbf{v}_r\mathbf{v}_r^t - \mathbf{v}_r\mathbf{v}_r^t\| + \tfrac{1}{m^2}\|\textstyle\sum \mathbf{v}\mathbf{v}_r^t - \sum' \mathbf{v}\mathbf{v}_r^t + \sum \mathbf{v}_r\mathbf{v}^t - \sum' \mathbf{v}_r\mathbf{v}^t - \mathbf{v}_r\mathbf{v}_r^t + \mathbf{v}_r\mathbf{v}_r^t\| \\
&\leq \tfrac{2K^2}{m} + \tfrac{2(m-1)K^2}{m^2} = \tfrac{4K^2}{m} - \tfrac{2K^2}{m^2} = \tfrac{2K^2(m-1)}{m^2},
\end{aligned}$$

où $K^2 = \sup_{\mathbf{z}\in\mathcal{Z}} \|\mathbf{v}\mathbf{v}^t\| = \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|^2$.

Par l'inégalité de McDiamird, avec probabilité $1-\delta$,

$$\|\mathbf{\Sigma}_s - \mathbf{\Sigma}_d\| - \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m} \|\mathbf{\Sigma}_s - \mathbf{\Sigma}_d\| \leq \tfrac{K^2(m-1)}{\sqrt{m^3}}\sqrt{2\log\tfrac{1}{\delta}}.$$

Soit $\tilde{\mathcal{S}}$ un échantillon de m exemples. En observant que $\mathbb{E}_{\tilde{\mathcal{S}}\sim\mathscr{D}^m} \boldsymbol{\Sigma}_{\tilde{s}} = \boldsymbol{\Sigma}_d$ et par l'inégalité du triangle,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathscr{D}^m} \|\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_d\| &= \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathscr{D}^m} \|\boldsymbol{\Sigma}_s - \mathop{\mathbb{E}}_{\tilde{\mathcal{S}}\sim\mathscr{D}^m} \boldsymbol{\Sigma}_{\tilde{s}}\| \\
&= \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathscr{D}^m} \| \mathop{\mathbb{E}}_{\tilde{\mathcal{S}}\sim\mathscr{D}^m} \{\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_{\tilde{s}}\}\| \\
&\leq \mathop{\mathbb{E}}_{\mathcal{S},\tilde{\mathcal{S}}\sim\mathscr{D}^{2m}} \|\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_{\tilde{s}}\|.
\end{aligned}$$

Soit $\mathbf{B} = (B_1, \ldots, B_m) \in \{-1, 1\}^m$ un vecteur aléatoire de dimension m distribué selon une distribution de Bernouilli équilibrée $\mathscr{B}$. Par l'inégalité du triangle,

$$\begin{aligned}
\|\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_{\tilde{s}}\| &= \left\| \frac{1}{m}(\sum_{i=1}^m b_i \mathbf{v}_i \mathbf{v}_i^t - \sum_{i=1}^m b_i \tilde{\mathbf{v}}_i \tilde{\mathbf{v}}_i^t) + \frac{1}{m^2}(\sum_{i,j=1}^m b_i \tilde{\mathbf{v}}_i \tilde{\mathbf{v}}_j^t - \sum_{i,j=1}^m b_i \mathbf{v}_i \mathbf{v}_j^t) \right\| \\
&\leq \frac{1}{m} \left\| \sum_{i=1}^m b_i \mathbf{v}_i \mathbf{v}_i^t - \sum_{i=1}^m b_i \tilde{\mathbf{v}}_i \tilde{\mathbf{v}}_i^t \right\| + \frac{1}{m^2} \left\| \sum_{i,j=1}^m b_i \tilde{\mathbf{v}}_i \tilde{\mathbf{v}}_j^t - \sum_{i,j=1}^m b_i \mathbf{v}_i \mathbf{v}_j^t \right\|.
\end{aligned}$$

En utilisant encore l'inégalité du triangle et comme $\mathcal{S}$ et $\tilde{\mathcal{S}}$ ont tous deux été générés par $\mathscr{D}$,

$$\|\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_{\tilde{s}}\| \leq \frac{2}{m} \left\| \sum_{i=1}^m b_i \mathbf{v}_i \mathbf{v}_i^t \right\| + \frac{2}{m^2} \left\| \sum_{i,j=1}^m b_i \mathbf{v}_i \mathbf{v}_j^t \right\|.$$

Finalement, $\frac{2}{m}\| \sum_{i=1}^m b_i \mathbf{v}_i \mathbf{v}_i^t\| = \frac{2}{m}\sqrt{\lambda_{\max}(\sum_{i,j=1}^m b_i b_j (\mathbf{v}_i \mathbf{v}_i^t)^t \mathbf{v}_j \mathbf{v}_j^t)}$ et $\frac{2}{m^2}\| \sum_{i,j=1}^m b_i \mathbf{v}_i \mathbf{v}_j^t\| = \frac{2}{m^2}\sqrt{\lambda_{\max}(\sum_{i,j,p,q=1}^m b_i b_p (\mathbf{v}_i \mathbf{v}_j^t)^t \mathbf{v}_p \mathbf{v}_q^t)}$. Par conséquent,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathscr{D}^m} \|\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_d\| \leq \frac{2}{m} \mathop{\mathbb{E}}_{\boldsymbol{b},\mathcal{S}\sim\mathscr{B},\mathscr{D}^m} \sqrt{\lambda_{\max}(\sum_{i,j=1}^m b_i b_j (\mathbf{v}_i \mathbf{v}_i^t)^t \mathbf{v}_j \mathbf{v}_j^t)} \\
+ \frac{2}{m^2} \mathop{\mathbb{E}}_{\boldsymbol{b},\mathcal{S}\sim\mathscr{B},\mathscr{D}^m} \sqrt{\lambda_{\max}(\sum_{i,j,p,q=1}^m b_i b_p (\mathbf{v}_i \mathbf{v}_j^t)^t \mathbf{v}_p \mathbf{v}_q^t)}.
\end{aligned}$$

Les signes contraires b_i et b_j s'annulent,

$$\begin{aligned}
\frac{2}{m} \mathop{\mathbb{E}}_{\boldsymbol{b},\mathcal{S}\sim\mathscr{B},\mathscr{D}^m} \left\{ \sqrt{\lambda_{\max}(\sum_{i,j=1}^m b_i b_j (\mathbf{v}_i \mathbf{v}_i^t)^t \mathbf{v}_j \mathbf{v}_j^t)} \right\} &= \frac{2}{m} \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathscr{D}^m} \left\{ \sqrt{\lambda_{\max}(\sum_{i=1}^m (\mathbf{v}_i \mathbf{v}_i^t)^t \mathbf{v}_i \mathbf{v}_i^t))} \right\} \\
&= \frac{2}{m} \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathscr{D}^m} \left\{ \sqrt{\sum_{i=1}^m \|\mathbf{v}_i \mathbf{v}_i^t\|^2} \right\} \\
&= \frac{2}{m} \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathscr{D}^m} \left\{ \sqrt{\sum_{i=1}^m \|\mathbf{v}_i\|^4} \right\} \\
&\leq \frac{2}{m} \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathscr{D}^m} \left\{ \sqrt{mK^4} \right\} \\
&= \frac{2K^2}{\sqrt{m}}.
\end{aligned}$$

Les signes contraires b_i et b_j s'annulent,

$$\begin{aligned}\frac{2}{m^2}\mathop{\mathbb{E}}_{\boldsymbol{b},\mathcal{S}\sim\mathscr{B},\mathscr{D}^m}\sqrt{\lambda_{\max}(\sum_{i,j,p,q=1}^{m} b_i b_p(\mathbf{v}_i\mathbf{v}_j^t)^t\mathbf{v}_p\mathbf{v}_q^t)} &= \frac{2}{m^2}\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathscr{D}^m}\sqrt{\lambda_{\max}(\sum_{i,j=1}^{m}(\mathbf{v}_i\mathbf{v}_j^t)^t\mathbf{v}_i\mathbf{v}_j^t))} \\ &= \frac{2}{m^2}\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathscr{D}^m}\sqrt{\sum_{i,j=1}^{m}\|\mathbf{v}_i\mathbf{v}_j^t\|^2} \leq \frac{2}{m^2}\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathscr{D}^m}\sqrt{m^2K^4} \\ &= \frac{2K^2}{m}.\end{aligned}$$

Enfin, avec probabilité $1-\delta$,

$$\begin{aligned}\|\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_d\| &\leq \frac{K^2(m-1)}{\sqrt{m^3}}\sqrt{2\log\frac{1}{\delta}} + \frac{2R^2}{\sqrt{m}} + \frac{2K^2}{m} \\ &= \frac{K^2}{\sqrt{m^3}}\left((m-1)\sqrt{2\log\frac{1}{\delta}} + 2(m+\sqrt{m})\right).\end{aligned}$$

□

Bibliographie

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974.

[2] N. Akhtar and A. S. Mian. Threat of adversarial attacks on deep learning in computer vision : A survey. *IEEE Access*, 6 :14410–14430, 2018. doi : 10.1109/ACCESS.2018.2807385. URL https://doi.org/10.1109/ACCESS.2018.2807385.

[3] P. Alquier and B. Guedj. Simpler pac-bayesian bounds for hostile data. *Mach. Learn.*, 107 (5) :887–902, 2018. doi : 10.1007/s10994-017-5690-0. URL https://doi.org/10.1007/s10994-017-5690-0.

[4] A. Ambroladze, E. Parrado-Hernández, and J. Shawe-Taylor. Tighter pac-bayes bounds. In B. Schölkopf, J. C. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 9–16. MIT Press, 2006.

[5] L. Bégin, P. Germain, F. Laviolette, and J. Roy. Pac-bayesian theory for transductive learning. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, volume 33 of *JMLR Workshop and Conference Proceedings*, pages 105–113. JMLR.org, 2014. URL http://proceedings.mlr.press/v33/begin14.html.

[6] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144—-152, New York, NY, USA, 1992. Association for Computing Machinery.

[7] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN 978-0-19-953525-5. doi : 10.1093/acprof:oso/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

[8] S. Cambanis, S. Huang, and G. Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3) :368–385, 1981.

[9] O. Catoni. *Pac-Bayesian Supervised Classification : The Thermodynamics of Statistical Learning*. Institute of Mathematical Statistics Lecture Notes, 2007. ISBN 978-0-94-060072-0.

[10] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3) :273–297, 1995. doi : 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018.

[11] G. K. Dziugaite and D. M. Roy. Data-dependent pac-bayes priors via differential privacy. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31 : Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8440–8450, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/9a0ee0a9e7a42d2d69b8f86b3a0756b1-Abstract.html.

[12] M. M. Fard, J. Pineau, and C. Szepesvári. Pac-bayesian policy evaluation for reinforcement learning. In F. G. Cozman and A. Pfeffer, editors, *UAI 2011, Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, July 14-17, 2011*, pages 195–202. AUAI Press, 2011. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=2218&proceeding_id=27.

[13] J. Frankle and M. Carbin. The lottery ticket hypothesis : Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

[14] Y. Freund. Boosting a weak learning algorithm by majority. In M. A. Fulk and J. Case, editors, *Proceedings of the Third Annual Workshop on Computational Learning Theory, COLT 1990, University of Rochester, Rochester, NY, USA, August 6-8, 1990*, pages 202–216. Morgan Kaufmann, 1990. URL http://dl.acm.org/citation.cfm?id=92640.

[15] P. Germain. Algorithmes d'apprentissage automatique inspirés de la théorie PAC-Bayes. Master's thesis, Université Laval, 2009. URL http://www.theses.ulaval.ca/2009/26191/.

[16] P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. Pac-bayesian learning of linear classifiers. In A. P. Danyluk, L. Bottou, and M. L. Littman, editors, *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 353–360. ACM, 2009. doi : 10.1145/1553374.1553419. URL https://doi.org/10.1145/1553374.1553419.

[17] S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017. doi : 10.1017/9781139029834.

[18] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[19] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27 : Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.

[20] B. Guedj. A primer on pac-bayesian learning. *CoRR*, abs/1901.05353, 2019. URL http://arxiv.org/abs/1901.05353.

[21] A. K. Gupta, T. Varga, and T. Bodnar. In *Elliptically Contoured Models in Statistics and Portfolio Theory*, 2013.

[22] E. Gómez, M. Gómez-Villegas, and J. Marin. A multivariate generalization of the power exponential family of distributions. *Communications in Statistics-theory and Methods - COMMUN STATIST-THEOR METHOD*, 27 :589–600, 01 1998. doi : 10.1080/03610929808832115.

[23] M. E. Johnson. Multivariate statistical simulation. In M. Lovric, editor, *International Encyclopedia of Statistical Science*, pages 930–932. Springer, 2011. doi : 10.1007/978-3-642-04898-2_39. URL https://doi.org/10.1007/978-3-642-04898-2_39.

[24] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 1995.

[25] Z. Landsman and E. Valdez. Tail conditional expectations for elliptical distributions. *North American Actuarial Journal*, 7 :55–71, 06 2003. doi : 10.1080/10920277.2003.10596118.

[26] J. Langford. Tutorial on practical prediction theory for classification. *J. Mach. Learn. Res.*, 6 :273–306, 2005. URL http://jmlr.org/papers/v6/langford05a.html.

[27] J. Langford and J. Shawe-Taylor. Pac-bayes & margins. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, pages 423–430. MIT Press, 2002.

[28] G. Lever, F. Laviolette, and J. Shawe-Taylor. Distribution-dependent pac-bayes priors. In M. Hutter, F. Stephan, V. Vovk, and T. Zeugmann, editors, *Algorithmic Learning Theory, 21st International Conference, ALT 2010, Canberra, Australia, October 6-8,*

2010. Proceedings, volume 6331 of *Lecture Notes in Computer Science*, pages 119–133. Springer, 2010. doi : 10.1007/978-3-642-16108-7_13. URL https://doi.org/10.1007/978-3-642-16108-7_13.

[29] G. Lever, F. Laviolette, and J. Shawe-Taylor. Tighter pac-bayes bounds through distribution-dependent priors. *Theor. Comput. Sci.*, 473 :4–28, 2013. doi : 10.1016/j.tcs.2012.10.013. URL https://doi.org/10.1016/j.tcs.2012.10.013.

[30] D. A. McAllester. Some pac-bayesian theorems. *Mach. Learn.*, 37(3) :355–363, 1999. doi : 10.1023/A:1007618624809. URL https://doi.org/10.1023/A:1007618624809.

[31] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6) :115 :1–115 :35, 2021. doi : 10.1145/3457607. URL https://doi.org/10.1145/3457607.

[32] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever. Deep double descent : Where bigger models and more data hurt. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=B1g5sA4twr.

[33] C. B. D. Newman and C. Merz, 1998.

[34] E. Parrado-Hernández, A. Ambroladze, J. Shawe-Taylor, and S. Sun. Pac-bayes bounds with data dependent priors. *J. Mach. Learn. Res.*, 13 :3507–3531, 2012. URL http://dl.acm.org/citation.cfm?id=2503353.

[35] J. Pearl. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, 62(3) :54–60, 2019. doi : 10.1145/3241036. URL https://doi.org/10.1145/3241036.

[36] R. E. Schapire. The strength of weak learnability. *Mach. Learn.*, 5 :197–227, 1990. doi : 10.1007/BF00116037. URL https://doi.org/10.1007/BF00116037.

[37] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 1978.

[38] Y. Seldin and N. Tishby. Pac-bayesian analysis of co-clustering and beyond. *J. Mach. Learn. Res.*, 11 :3595–3646, 2010. URL http://portal.acm.org/citation.cfm?id=1953046.

[39] J. Shawe-Taylor and N. Cristianini. Estimating the moments of a random vector with applications. 2003.

[40] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004. ISBN 9780511809682. doi : 10.1017/CBO9780511809682. URL https://kernelmethods.blogs.bristol.ac.uk/.

[41] J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a bayesian estimator. In Y. Freund and R. E. Schapire, editors, *Proceedings of the Tenth Annual Conference on Computational Learning Theory, COLT 1997, Nashville, Tennessee, USA, July 6-9, 1997*, pages 2–9. ACM, 1997. doi : 10.1145/267460.267466. URL https://doi.org/10.1145/267460.267466.

[42] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Trans. Inf. Theory*, 44(5) :1926–1940, 1998. doi : 10.1109/18.705570. URL https://doi.org/10.1109/18.705570.

[43] J. A. Tropp. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1-2) :1–230, 2015.

[44] L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11) :1134–1142, 1984. doi : 10.1145/1968.1972. URL https://doi.org/10.1145/1968.1972.

[45] S. M. Weiss and N. Indurkhya. Rule-based machine learning methods for functional prediction. *J. Artif. Intell. Res.*, 3 :383–403, 1995. doi : 10.1613/jair.199. URL https://doi.org/10.1613/jair.199.