# Introduction au Domaine de Recherche : Compromis entre performance statistique et autres critères

Jaime Roquero Gimenez

## Table des matières



Cette Introduction au Domaine de Recherche est basée sur mes travaux de mémoire de Master 2 au Laboratoire de Probabilités et Modèles Aléatoires de UPMC Paris 6 sous la direction du Pr. Gérard Biau et du mémoire de Master 2 Mathématiques, Vision, Apprentissage de l'École Normale Supérieure de Cachan sous la direction du Pr. Arnak Dalalyan.

# 1 Introduction

Les avancées technologiques du début du XXIème siècle en informatique ont permis de développer des méthodes de collecte et analyse de données en quantité massive, ensemble appelé communément "Big Data". En sciences sociales, économie, journalisme ou marketing, l'analyse et exploitation de quantités très importantes de données permet une utilisation plus efficace des ressources, permet de découvrir des phénomènes inattendus, et plus récemment d'enseigner à un ordinateur à prendre des décisions en apprenant sur les données qu'on lui fournit, entre autres choses. Aujourd'hui, plus on a de données, mieux on peut faire.

Cependant, d'autres problèmes ont surgi avec ces nouvelles technologies. Avoir un volume extraordinaire de données est en effet très utile pour faire de l'inférence ou de l'apprentissage, mais on se heurte vite à des défis techniques du fait d'une puissance de calcul parfois limitée, ou à des problématiques issues d'un débat social comme peut être celle du respect de la vie privée.

Dans la suite, nous considérons des problèmes de statistique paramétrique où un équilibre entre performance statistique et d'autres critères peut être caractérisé.

# 2 Équilibre entre performance statistique et vitesse de calcul

Malgré la puissance de calcul croissante, Le volume de données qu'on est capables de collecter augmente bien plus rapidement que la puissance de calcul disponible : par exemple, les volumes de données croissants en biologie [1] se heurtent à un futur où la loi de Moore sur l'augmentation de la capacité de calcul est remise en question. Ainsi apparaît le besoin de tenir en compte la complexité de calcul dans la création d'estimateurs statistiques.

## 2.1 Le problème statistique adapté à des relaxations convexes

Un problème classique en statistique est le débruitage et de nombreuses méthodes sont connues dans le cas de grande dimension, dont les travaux de D. Donoho sur le *soft thresholding* [2] que nous verrons ensuite appliqué pour des matrices bruitées.

Nous considérons un modèle simple de débruitage dans $\mathbb{R}^d$ où nous observons un échantillon i.i.d. $(X_1, \dots, X_n)$ de $n \geq 1$ réalisations d'un vecteur aléatoire $X \sim \mathcal{N}(x^*, \sigma^2 I_d)$ que

l'on supposera Gaussien, de variance connue. Le but à partir de cet échantillon est de retrouver la valeur $x^*$, sachant qu'elle appartient à un ensemble $\mathcal{S}$ connu, l'ensemble des signaux possibles que l'on considère convexe compact.

La méthode de débruitage qu'on utilise par la suite consiste simplement à considérer $\overline{X}_n = \frac{1}{n}\sum_i X_i$, moyenne empirique de l'échantillon, et ensuite projeter

$$\hat{X}_n(\mathcal{C}) = \underset{X\in\mathbb{R}^d}{\arg\min} \|\overline{X_n} - X\|_2 \qquad \text{tel que } X \in \mathcal{C}$$

sur un ensemble convexe $\mathcal{C}$ contenant notre ensemble de signaux. On appelle $\hat{X}_n(\mathcal{C})$ estimateur de notre paramètre $x^*$. Pour un ensemble $\mathcal{C}$ donné, on s'intéresse à deux critères de performance de notre méthode de débruitage :

– la performance statistique, mesurée à partir de l'erreur quadratique :

$$\mathcal{R}_n\big(\hat{X}_n\big) = \mathbb{E}\Big[\|\hat{X}_n - x^*\|^2_{L^2(\mathbb{R}^d)}\Big]$$

Cette quantité mesure l'écart moyen mesurée à partir de la distance $L^2$ de notre estimateur par rapport à la vraie valeur du paramètre. D'autres fonctions de perte pourraient être considérées (autres que la distance $L^2$) [3], mais celle-ci a l'avantage d'être en lien étroit avec des notions de géométrie adaptées aux relaxations convexes.

– le temps de calcul. Pour obtenir la moyenne empirique on réalise $nd$ opérations, et ensuite on note $f_{\mathcal{C}}(d)$ le nombre d'opérations nécessaires à projeter sur le convexe $\mathcal{C}$, d'où une complexité de calcul égale à $nd + f_{\mathcal{C}}(d)$.

En effet, la dimension $d$ joue un rôle essentiel dans le nombre d'opérations nécessaires pour obtenir les projections. Dans la suite, à chaque fois on déterminera une taille minimale $n$ de l'échantillon (en fonction de $d$) pour vérifier un critère sur le risque, et ensuite on regardera la complexité de calcul pour obtenir l'estimateur (à nouveau, en fonction de $d$).

Si on se fixe un seuil limite d'erreur statistique acceptable (par exemple, $\mathcal{R}_n(\hat{X}_n) \leq 1$), en fonction du nombre de données disponibles on cherche un ensemble $\mathcal{C}$ qui réduise au mieux le temps de calcul. On verra que parfois, il est plus intéressant de se défausser d'une partie des données pour accélérer le temps de calcul.

## 2.2 Un exemple sur des matrices signées

L'amélioration du temps de calcul en augmentant la taille de l'échantillon a été étudiée préalablement dans [4] avec des résultats où l'amélioration du temps de calcul se faisait grâce à une augmentation considérable de la taille d'échantillon. Le résultat de [5] montre par contre que les méthodes de relaxation convexe ont juste besoin dans certains cas d'une augmentation de la taille de l'échantillon d'un facteur constant.

Si on considère l'ensemble des matrices signées de rang 1 comme notre espace de signal :

$$\mathcal{S} = \left\{ \mathbf{x}\mathbf{x}^t,\ \mathbf{x} \in \{-1,+1\}^{\sqrt{d}} \right\} \subset \mathbb{R}^d$$

on cherche à débruiter une matrice donnée $\overline{X}_n$ en projetant celle-ci sur les convexes suivants :

- $\mathcal{C}_3$ : l'enveloppe convexe de $\mathcal{S}$, appelée *cut polytope*.
- $\mathcal{C}_2$ : l'ensemble des matrices de corrélation : $\left\{ X \in \mathcal{S}_d^+(\mathbb{R}), X_{ii} = 1\ \forall i \in \{1,...,d\} \right\}$
- $\mathcal{C}_1$ : la boule unité renormalisée de la norme matricielle de la trace.

On a ainsi $\mathcal{S} \subset \mathcal{C}_3 \subset \mathcal{C}_2 \subset \mathcal{C}_1$. À partir de considérations géométriques faisant intervenir des outils tels que la largeur Gaussienne des cônes tangents aux convexes $\mathcal{C}_i$ en les points extrémaux, on obtient les valeurs suivantes pour les tailles d'échantillon nécessaires à majorer le risque par 1 : $c_i\sqrt{d}$ pour une projection sur $\mathcal{C}_i$ pour chaque $i = 1, 2, 3$, avec $c_1 \geq c_2 \geq c_3$. D'autre part, des résultats en informatique nous donnent les valeurs de $f_{\mathcal{C}_i}(d)$ pour chaque convexe :

- le *cut polytope* ayant un nombre exponentiel de faces en la dimension, le temps de calcul de la projection est super-polynomial.
- à partir de méthodes d'optimisation convexe avec point intérieur, on obtient un temps de calcul pour l'ensemble des matrices de corrélation en $\mathcal{O}(d^{9/8})$.
- en décomposant selon les valeurs singulières et ensuite par troncature, on obtient pour le convexe $\mathcal{C}_1$ un temps de calcul de la projection en $\mathcal{O}(d^{3/2})$

Cet exemple montre comment en disposant d'un échantillon plus grand on arrive à réduire considérablement le temps de calcul tout en gardant une garantie de performance statistique. Finalement, si on considère la boule unité par rapport à la norme $L^2$, le temps de calcul de la projection est extrèmement rapide, en $\mathcal{O}(\sqrt{d})$, et la taille de l'échantillon nécessaire pour garantir un risque inférieur à 1 est supérieure à $c_1\sqrt{d}$. Cependant le temps de calcul pour obtenir le vecteur de moyenne empirique étant $\mathcal{O}(d^{3/2})$, on n'obtient pas de gain substantiel de temps et il est plus intéressant dans ce cas de ne pas tenir en compte les données

supplémentaires par rapport à celles qu'on aurait utilisé dans le cas du convexe $\mathcal{C}_1$.

## 2.3 La construction d'ensembles par relaxation convexe

Le problème repose donc sur la construction d'ensembles $\mathcal{C}$ convexes compacts. On applique des méthodes de relaxation convexe pour obtenir une suite de convexes se resserrant autour de l'ensemble des signaux :

$$\mathcal{C}_1 \supset \mathcal{C}_2 \supset ... \supset \mathcal{C}_i \supset ... \supset \mathcal{S}$$

En considérant deux cônes convexes bien particuliers, $\mathbb{R}^d_+$ et l'ensemble des matrices symétriques positives $\mathcal{S}_n^+(\mathbb{R}^d)$, on obtient deux classes de convexes à partir de sections linéaires de ces cônes. C'est sur ces familles de convexes que les techniques de relaxation convexe se sont développées, basées sur le principe de *lift-and-project* : un convexe en dimension $d$ est représenté comme la projection d'un convexe en dimension $d' >> d$ dont l'écriture est moins complexe. On parle alors de représentations liftées LP (*Linear programming*) et SDP (*Semi-definite programming*)

H. D. Sherali et W. P. Adams dans [6] proposent une méthode de relaxation convexe pour des polytopes $\mathcal{S}$ dont les points extrémaux sont dans $\{0,1\}^d$ et qui génère $d$ polytopes représentations LP de $\mathcal{S}$. J. B. Lasserre [7] reprend ce principe pour construire les relaxations convexes à partir de représentations liftées SDP de convexes.

Emboîter des convexes permet d'avoir des inégalités directes sur les erreurs statistiques (pour $i \leq j$, on observe que le risque quadratique vérifie $\mathcal{R}_n(\hat{R}_n(\mathcal{C}_j)) \leq \mathcal{R}_n(\hat{R}_n(\mathcal{C}_i))$), et les méthodes mentionnées simplifient l'écriture du convexe en dimension supérieure, rendant ainsi plus rapide le calcul de la projection. Un équilibre apparaît alors clairement.

Un des objectifs que j'ai poursuivi était d'obtenir des garanties sur les temps de calcul en fonction du degré de la relaxation de Sherali-Adams dans le but d'obtenir un critère paramétrique pour choisir un degré de relaxation optimal dans un contexte de débruitage de signaux à valeurs binaires.

## 2.4 Méthode de Sherali-Adams

On considère le problème linéaire 0/1 défini comme suit : nous voulons déterminer l'ensemble de vecteurs dans $\mathbb{R}^d$ à valeurs binaires satisfaisant certaines inégalités linéaires. Étant

donnés certains coefficients on cherche à caractériser l'ensemble :

$$\mathcal{S}_0 = \Big\{ x \in \mathbb{R}^d,\ \sum_{j=1}^{d} \alpha_{ij} x_j \geq \beta_i \ \forall i \in \{1, \dots, A\} \\ x \in \{0,1\}^d \Big\}$$

Il est possible de définir des contraintes linéaires à partir d'égalités en considérant deux contraintes avec des inégalités.

Pour résoudre ce problème on s'intéresse aux approximations de l'ensemble $\mathcal{S} = conv\, \mathcal{S}_0$. C'est le plus petit ensemble convexe au sens de l'inclusion qui contient tous les éléments de notre ensemble de signaux. Les relaxations convexes que nous construirons par la suite tenteront d'approcher au mieux cet ensemble $\mathcal{S}$.

La relaxation convexe la plus simple consiste à simplement éliminer la contrainte binaire. On définit le polyhèdre $\mathcal{R}_0$ par :

$$\mathcal{R}_0 = \Big\{ x \in \mathbb{R}^d,\ \sum_{j=1}^{d} \alpha_{ij} x_j \geq \beta_i \ \forall i \in \{1, \dots, A\},\ x \in [0,1]^d \Big\} \tag{1}$$

La méthode de Sherali-Adams construit une suite de polyhèdres convexes $\mathcal{R}_p$ pour $p \in \{0, \dots, d\}$ telle que les polyhèdres se resserrent autour de $\mathcal{S}$ lorsque le degré $p$ de la relaxation augmente.

Ainsi on obtient une suite telle que :

$$\mathcal{R}_0 \supset \mathcal{R}_1 \supset \cdots \supset \mathcal{R}_d \supset \mathcal{S}$$

En fait, la relaxation de Sherali-Adams est exacte au sens où lorsque $p = d$ on a $\mathcal{S} = \mathcal{R}_d$. Notre but étant de trouver un compromis entre la performance statistique et la performance de calcul, on observe que pour des degrés élevés la performance statistique est bonne vu que le convexe se colle bien à notre convexe $\mathcal{S}_0$, mais en revanche la définition des ensembles $\mathcal{R}_p$ pour $p$ faible font appel à une écriture plus parcimonieuse en termes du nombre d'inégalités, permettant ainsi un calcul plus rapide de la projection d'un point sur ces convexes.

On donne maintenant une définition constructive de la méthode de Sherali-Adams : cette relaxation est basée sur l'idée du *lift-and-project* : en partant de l'ensemble $\mathcal{R}_0$, on ajoute des contraintes selon le degré $p$ de la relaxation, puis on procède à une linéarisation qui

définit un convexe dans une dimension supérieure. Ensuite, en projetant sur les premières $d$ coordonnées on aboutit à la relaxation en question.

On définit les ensembles $\mathcal{C}_p \subset \mathbb{R}^d \times \Omega^p$ for $p \in \{0, \ldots, d\}$ et $\mathcal{R}_p = \mathrm{Proj}_{\mathbb{R}^d} \mathcal{C}_p$, où $\Omega^p$ est un certain espace vectoriel indexé par le degré de la relaxation.

Nous définissons maintenant les contraintes à ajouter pour chaque $p$. On définit un ensemble de polynômes à $d$ variables appelé l'ensemble des *facteurs de degré p*. On définit une paire admissible de sous-ensembles $(I, J)$ de $\{1, \ldots, d\}$ par $I, J \subset \{1, \ldots, d\}$ tels que $I \cap J = \emptyset$, et $|I \cup J| = p$, et on définit le facteur $F(I, J)$ de degré $p$ par le polynôme :

$$F(I, J)(x) = \prod_{k \in I} x_k \prod_{l \in J} (1 - x_l)$$

En évaluant en un facteur de degré quelconque un point de notre ensemble initial de signaux $\mathcal{S}_0$ on obtient une valeur positive. Ainsi en multipliant une inégalité dans la définition de notre ensemble $\mathcal{R}_0$ par un facteur quelconque on obtient une contrainte (non-linéaire) qui est satisfaite par tous les éléments de $\mathcal{S}_0$.

On définit alors l'ensemble :

$$\tilde{\mathcal{C}_p} = \{x \in \mathbb{R}^d, \forall (I, J) \text{ paire admissible d'ordre p, } F(I, J)(x) \sum_{j=1}^{d} \alpha_j x_j \geq F(I, J)(x)\, \beta\} \quad (2)$$

et on le transforme en développant le polynôme $F(I, J)(x) \sum_{j=1}^{d} \alpha_j x_j$ puis en simplifiant tout facteur $x_k^2 = x_k$ (cette égalité étant valable en tout point de notre ensemble de signaux). On définit de nouvelles variables pour linéariser les monômes : pour $\prod_{k \in A} x_k, A \subset \{1, \ldots, d\}$ on le remplace dans l'inégalité par la nouvelle variable $y_A$. Nous définissons par $f(I, J)(x, y)$ l'expression linéarisée de $F(I, J)(x)$ définie sur $\mathbb{R}^d \times \Omega^p$ au lieu de $\mathbb{R}^d$.

Nous définissons ainsi un sous-ensemble de $\mathbb{R}^d \times \Omega^p$ à partir des contraintes linéarisées : nous obtenons pour la contrainte d'être dans $[0, 1]^d$ l'ensemble

$$Z_p = \{(x, y) \in \mathbb{R}^d \times \Omega^p,\ \forall (I, J) \text{ paire admissible d'ordre p}, f(I, J)(x, y) \geq 0\} \quad (3)$$

Finalement, en prenant l'intersection de $Z_{p+1}$ avec $\tilde{\mathcal{C}_p}$ on obtient l'ensemble suivant :

$$\mathcal{C}_p = \{(x,y) \in \mathbb{R}^d \times \Omega^p, \forall (I,J) \text{ paire admissible d'ordre p,} \\ \Big(\sum_{j\in I} \alpha_j - \beta\Big) f(I,J) + \sum_{j \in \mathbb{N}-(I\cup J)} \alpha_j f(I+j,J) \geq 0 \\ \forall (I,J) \text{ paire admissible d'ordre } p+1, f(I,J)(x,y) \geq 0\}$$

En définissant, pour $(I,J)$ paire admissible d'ordre p la forme linéaire $\mathbb{R}^d \times \Omega^p$ par :

$$\Phi_{(I,J)} : (x,y) \to \Big(\sum_{j\in I} \alpha_j - \beta\Big) f(I,J)(x,y) + \sum_{j \in \mathbb{N}-(I\cup J)} \alpha_j f(I+j,J)(x,y)$$

On obtient,

$$\mathcal{C}_p = \{(x,y) \in \mathbb{R}^d \times \Omega^p, \forall (I,J) \text{ paire admissible d'ordre p}, \Phi_{(I,J)}(x,y) \geq 0 \\ \forall (I,J) \text{ paire admissible d'ordre } p+1, f(I,J)(x,y) \geq 0\}$$

En prenant $\mathcal{R}_p = \text{Proj}_{\mathbb{R}^d} \mathcal{C}_p$ pour $p$ positif. Seul $\mathcal{R}_0$ s'écrit différemment :

$$\mathcal{R}_0 = \Big\{x \in \mathbb{R}^d, \sum_{j=1}^{d} \alpha_j x_j \geq \beta \ , \ x \in [0,1]^d\Big\} \tag{4}$$

On peut donc montrer que $\mathcal{R}_d = \mathcal{S}$, et analyser en un point de $\mathcal{S}_0$ donné quel est le nombre de faces qui arrivent sur ce point dans le convexe $\mathcal{R}_p$ d'ordre $p$, ce qui permet par la suite d'approcher la complexité en termes de calcul.

# 3 Équilibre entre performance statistique et respect de la confidentialité

Le respect de la confidentialité et la vie privée est une problématique très actuelle vu la capacité de certains acteurs à obtenir de l'information sur nous à partir de la collecte et analyse de données. Or parfois ces données cachent des informations utiles pour la société. Dans la médecine, par exemple, la collecte de données individuelles permet au statisticien de faire de l'inférence qui peut déboucher sur des traitements. Cependant chaque individu

peut être réticent à dévoiler certaines données sensibles. Il est donc intéressant de mettre en place une méthode qui permette d'apprendre *à partir des données*, et non pas *sur les données*. J. Duchi [8] s'intéresse à une forme particulière de protection de la vie privée appelée *differential privacy* : le détenteur des données ne faisant pas confiance au statisticien, il bruite l'information qu'il possède avant de la communiquer pour garantir que l'information fournie n'est pas trop importante.

## 3.1 Le modèle

De nombreux procédés d'apprentissage statistique ont recours à des techniques de descente de gradient (souvent stochastique). Dans notre situation, on cherche à trouver la vraie valeur d'un paramètre $\theta^* \in \mathbb{R}^d$ où à chaque étape de notre itération on demande au détenteur des données de communiquer la valeur du gradient d'une fonction de perte notée $l = l(\mathcal{X}, \theta)$ évaluée à partir des données privées $\mathcal{X}$ nécessaire à la procédure. On suppose que les valeurs du gradient sont bornées par une constante $C$. Le procédé de bruitage consiste à choisir, sachant la vraie valeur du gradient $X = \nabla_\theta l(\mathcal{X}, \theta)$, à générer un vecteur $Y$ selon une certaine loi qui sera fourni au statisticien : cette loi prend la forme d'un noyau de transition $Q$ conditionné par $X$ dont le support sera borné par une constante $D > C$. Ce noyau doit satisfaire comme condition ne pas être biaisé : $\mathbb{E}_Q[Y|X = x] = x$.

On considère comme critère pour garantir la confidentialité l'information mutuelle définie par :

$$I(P, Q) = \mathbb{E}_P[KL(Q(\cdot|X)||PQ(\cdot))]$$

où $KL$ est la divergence de Kullback-Leibler. Celle-ci est définie pour deux lois à densité $p$ et $q$ par rapport à une mesure $\nu$ par :

$$KL(p\|q) = \int p \log\left(\frac{p}{q}\right) \mathrm{d}\nu \geq 0$$

et $KL(p\|q) = 0$ si et seulement si $p = q$.

On dira que $Q$ satisfait un degré de confidentialité au niveau $I^*$ si on a pour tout $P$, $I(P, Q) \leq I^*$. Plus l'ensemble $\mathcal{D}$ est grand, plus le signal est bruité et on maintient la confidentialité. À un niveau de confidentialité désiré, on doit fixer un ensemble $\mathcal{D}$ adapté. Pour cela on cherche un Q optimal pour un $\mathcal{D}$ quelconque au sens suivant :

**Definition 3.1.** *On dit que le noyau de transition $Q^*$ satisfait la* condition optimale de

confidentialité locale *pour les ensembles* $\mathcal{C} \subset \mathcal{D} \subset \mathbb{R}^d$ *pour le niveau* $I^*$ *si*

$$\sup_P I(P, Q^*) = \inf_Q \sup_P I(P, Q) = I^*$$

*où* $P$ *est pris dans l'ensemble des probabilités sur* $\mathcal{C}$ *et* $Q$ *est un noyau de transition de* $\mathcal{C}$ *dans* $\mathcal{D}$*.*

En caractérisant la valeur de $I^*$, on pourra par la suite choisir l'ensemble $\mathcal{D}$ approprié. Souvent, $\mathcal{D}$ sera une dilatation de l'ensemble $\mathcal{C}$.

## 3.2 Noyaux optimaux et bornes minimax

**Theorem 3.1.** *Soit* $\mathcal{C} \subset \mathbb{R}^d$ *un polytope convexe compact invariant par rotation autour de ses points extrémaux, et* $\mathcal{D} = (1+\alpha)\mathcal{C}$ *avec* $\alpha > 0$*. Soit* $Q^*$ *la distribution de* $Z|X$ *qui maximise l'entropie* $H(Z|X = x)$ *telle que le support de* $Q^*$ *est dans les points extrémaux de* $\mathcal{D}$ *et que* $\mathbb{E}_Q[Z|X = x] = x$*. Alors* $Q^*$ *satisfait la condition optimale de confidentialité pour les ensembles* $\mathcal{C}$ *et* $\mathcal{D}$*.*

Ce théorème nous permet de caractériser directement les noyaux optimaux dans le cas de certains polytopes bien choisis. C'est le cas lorsque l'ensemble $\mathcal{C}$ est la boule unité pour la norme $L^1$ ou la norme $L^\infty$, pour lesquels on a ainsi des expressions explicites.

Par la suite je me suis intéressé au cas où le vecteur $X$ est s-sparse, pour un s petit, et où $\mathcal{C}$ est la boule unité pour la norme $L^1$. Le but est de caractériser un noyau optimal en se basant initialement sur le modèle non sparse, pour ensuite essayer de trouver un noyau optimal. Les bornes obtenues jusqu'à présent permettent de faire un choix de $\alpha$ pour satisfaire le niveau de confidentialité cherché. Par contre pour l'instant ces bornes ne sont pas optimales.

D'autre part, ces noyaux $Q^*$ nous donnent des inégalités minimax nous montrant qu'est-ce qu'on peut attendre de mieux comme performance statistique. On définit une borne minimax $\epsilon_n$ dans un problème d'estimation de paramètre $\theta \in \Theta$ par :

$$\epsilon_n = min_{\hat{\theta}_n} max_{\theta \in \Theta} R_n(\hat{\theta}_n, \theta^*)$$

Cela revient à trouver une limite optimale pour la performance d'un estimateur donné dans notre problème : on ne peut espérer faire mieux qu'un estimateur qui atteindrait cette borne. En revenant au cas où $\mathcal{C}$ est une boule $L^\infty$, on a le résultat suivant :

**Theorem 3.2.** *On considère l'espace des signaux* $\Theta = \{\theta \in \mathbb{R}^d, \|\theta\|_1 \leq r\}$*, et que les ensembles* $\mathcal{C}$ *et* $\mathcal{D}$ *sont respectivement la boule unité pour la norme* $L^\infty$ *et une dilatation de celle-ce de rapport* $M > 1$*.*

*On a alors :*

$$\epsilon \geq \frac{rM\sqrt{\log(2d)}}{17\sqrt{n}}$$

*où* $\epsilon$ *est l'erreur minimax associée au problème avec des fonctions de perte 1-lipschitziennes.*

D'autre part, on a aussi le résultat suivant quant à la garantie optimale de confidentialité dans le cas des ensembles $\mathcal{D}$ qui sont des dilatations de la boule unité $L^\infty$.

**Theorem 3.3.** *Pour* $\mathcal{C} = \{x, \|x\|_\infty \leq 1\}$ *et* $\mathcal{D} = \{x, \|x\|_\infty \leq M\}$*, en choisissant* $Q^*$ *qui satisfait la condition optimale de confidentialité on obtient :*

$$I^* = \sup_P I(P, Q^*) = d - d.h(\frac{1}{2} - \frac{1}{2M})$$

*où* $h(x) = -x\log(x) - (1-x)\log(1-x)$

On observe ainsi clairement le compromis à faire entre le degré de confidentialité $I^*$ souhaité et la pénalisation que cela implique sur le risque minimax : en augmentant $M$ on améliore la confidentialité mais on diminue le risque optimal qu'on peut espérer.

# 4 Conclusion

Les critères à tenir en compte par le statisticien hors de la minimisation de l'erreur statistique sont bien plus nombreux que ceux décrits auparavant. Les limitations quant à la transmission de données, par exemple, est un autre critère qui pourrait conditionner une méthode d'apprentissage statistique quand la capacité du système est soumise à certaines contraintes.

C'est ainsi que les résultats évoqués précédemment sont seulement le début d'un intérêt en statistiques d'inclure dans les considérations mathématiques des critères extérieurs. D'autres modèles que le débruitage sont ouverts à des études de ce type, et probablement d'autres critères apparaîtront prochainement pour lesquels il faudra tout d'abord définir avec des outils statistiques les contraintes que l'on veut se fixer.

L'importance d'obtenir des bornes minimax dans ces contextes est donc essentielle car ces contraintes extérieures pourraient avoir des conséquences inattendues sur les garanties optimales dans certains problèmes, et qui modifient notre intuition quant aux buts à atteindre.

# Références


[1] Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data : Astronomical or Genomical ? *PLoS Biol* 13(7) : e1002195, 2015

[2] D. L. Donoho. Denoising by soft thresholding. *IEEE Transactions on Information Theory*, 52 :1289-1306, 2006

[3] T. Hastie, R. Tibshrihani, J. Friedman. Elements of Statistical Learning, *Springer Series in Statistics*, 2009

[4] R. Servedio. Computational sample complexity and attribute-efficient learning. *Journal of Computing and System Sciences*, 60 :161-178, 2000.

[5] V. Chandrasekaran, M.I.Jordan, Computational and Statistical Tradeoff via Convex Relaxations, *Proceedings of the National Academy of Sciences*, vol.10 no.13 E1181–E1190, 2013

[6] H. P. Sherali, W. P. Adams. A hierarchy of relaxations between the continuous and convex hull representation for zero-one programming problems. *SIAM Journal on Discrete Mathematics*, 3 :411-430, 1990.

[7] J. B. Lasserre, Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11 :796-817, 2001.

[8] J. C. Duchi, M. I. Jordan, M. J. Wainwright. *Privacy Aware Learning*, *Journal of the Association for Computing Machinery*, 2014