Introduction au domaine de recherche

La Concentration de la Mesure comme cadre naturel à l'analyse de données avec entraînement

Cosme Louart - Promotion 2013

Table des matières

Preambule

Le projet initial de mon stage était de réussir à prévoir les performances d'un réseau de neurone simple à l'aide d'outils de matrices aléatoires. Il semblait irréaliste de considérer directement les réseaux de neurones à plusieurs couches qui s'entraînent par "backpropagation", c'est-à-dire grâce à une descente de gradient qui rectifie petit à petit les erreurs couche par couche en partant de la dernière qui répertorie les écarts entre les sorties et les labels de l'ensemble d'entrainement. Nous avons décidé plus modestement de tenter de comprendre dans un

premier temps l'apport de la non linéarité dans les fonctions de transfert de chaque neurone. Le réseau de neurone en question est ce qui s'est fait appeler un "ELM" (extreme learning machine... cf. [HZS06]) : une donnée entrante $x \in \mathbb{R}^q$ est multipliée par une matrice fixée choisie aléatoirement $W \in \mathcal{M}_{p,q}$ puis transformée entrée par entrée par une fonction non linéaire $\sigma : \mathbb{R} \to \mathbb{R}$ qui caractérise la non linéarité des fonctions de transfert. On obtient un vecteur qu'on note $\sigma(Wx) \in \mathbb{R}^p$ et qui contient les sorties des p neurones du réseau, ce qui s'appelle en anglais les "features" des données et qui pourrait se traduire par les "caractéristiques". On peut les voir comme des points de vue différents sur les données. Un des enjeux d'un réseau de neurone est de diversifier au maximum les features pour avoir le plus de chance de trouver des neurones pertinents parmi les p neurones disponibles. Dans notre cas, p tend vers l'infini. Le tout est finalement multiplié par un vecteur $\beta \in \mathbb{R}^p$ sur lequel se fera l'entraînement pour donner en sortie un scalaire $\beta^T \sigma(Wx)$ le plus proche possible du label caractéristique de la classe du vecteur x (1 ou -1 dans le cas d'une classification binaire par exemple). L'entraînement se fait sur $n \sim \infty$ données rangées dans une matrice $X \in \mathcal{M}_{q,n}$ et labellisée suivant un vecteur $Y \in \mathbb{R}^n$ de 1 et de -1. Il consiste à minimiser le problème quadratique :

$$\frac{1}{n} \left\| \sigma(WX)^T \beta - Y \right\|^2 + \gamma \left\| \beta \right\|^2 \ , \quad \beta \in \mathbb{R}^p$$

où σ agit comme on l'a dit entrée par entrée et γ est un paramètre de régularisation à choisir pour limiter le sur-apprentissage (l'écueil qui consiste à choisir un vecteur β très efficace sur l'ensemble d'entraînement mais inadapté à toute autre donnée). La solution du problème se trouve facilement et étant donné une nouvelle donné entrante x, la sortie du réseau de neurone sera

$$S = \frac{1}{n} \sigma(Wx) \Sigma (\Sigma^T \Sigma / n + \gamma I_n)^{-1} Y \quad \text{où} \quad \Sigma = \sigma(WX)$$

La formule est relativement simple, et les connaisseurs reconnaissent l'apparition d'une résolvante sous la forme $Q = (\Sigma^T \Sigma / n + \gamma I_n)^{-1}$.

Dans un premier temps nous avons considéré cette sortie comme une variable aléatoire dépendant des tirages initiaux de W (cf. [LLC17]). Elle suivait alors une distribution gaussienne autour d'une moyenne proche de 1 ou bien de -1 lorsque l'on faisait varier les tirages de W, tout autre quantité fixée par ailleurs. Des outils de matrices aléatoire nous ont permis de trouver heuristiquement un estimateur déterministe de la moyenne et de la variance de la sortie et d'en déduire les performances du réseau de neurone. Ces quantités s'expriment au travers de la matrice $\mathbb{E}\left[\Sigma^T \Sigma / n\right] \in \mathcal{M}_n$, où l'espérance est calculée sur les variation de W.

Cependant, les hypothèses adoptées ne nous permettaient pas de démontrer la validité de ces estimations. Jusque là nos hypothèses portaient essentiellement sur la matrice W qui devait vérifier certaines hypothèses de concentration ensuite transmises à la sortie grâce à des hypothèses de bornitude de la matrice des données X et de lipschitzianité de la fonction d'activation σ. Il fallait de plus

supposer que les lignes de W, qui représentent les différents neurones, étaient indépendantes. Mais tout cela était insuffisant et même inadapté à notre problème puisque ça ne nous permettait même pas d'expliquer le fait que la sortie était concentrée autour de 1 ou -1. La difficulté s'explique par le fait que s'il est aisé de voir que la matrice $\Sigma(\Sigma^T\Sigma/n + \gamma I_n)^{-1}/\sqrt{n}$ et le vecteur $\|Y\|/\sqrt{n}$ sont borné (cf.Lemme 2), on ne comprend pas bien pourquoi la sortie serait bornée étant donnée que le vecteur $\sigma(Wx)$ a généralement une norme d'ordre $\sqrt{n}$.

Ce qu'il fallait comprendre, c'est que si la sortie ne diverge pas après l'entraînement du vecteur β, c'est que les données testée ne sortent pas de nulle part. Suivant leur classe, elles suivent une des deux distributions d'où ont été extraites les données d'entrainement contenues dans la matrice X. La sortie peut en effet s'écrire :

$$S = \frac{1}{n}\sigma(Wx)(\Sigma\Sigma^T/n + \gamma I_n)^{-1}\Sigma Y$$

et étant donné qu'il est possible de borner $\frac{1}{n}\Sigma^T(\Sigma\Sigma^T/n+\gamma I_n)^{-1}\Sigma$ (cf Lemme 2), on peut montrer que la moyenne du vecteur $\sigma(Wx)(\Sigma\Sigma^T/n + \gamma I_n)^{-1}\Sigma/\sqrt{n}$ est elle aussi bornée.

On est donc arrivé naturellement à la conclusion qu'il fallait adopter une hypothèse de concentration sur les données et que cette hypothèse devait même devenir l'hypothèse de base de n'importe quel réseau de neurone incluant une phase d'entraînement dans son logiciel.

Dans un premier temps nous développerons une approche didactique de la théorie de la concentration de la mesure et présenterons en particulier son champ d'application. Nous appliquerons ensuite les outils développés au cas fondamental de la distribution spectrale d'une matrice de covariance empirique. Enfin nous exposerons brièvement l'exemple de la régression de Ridge sur lequel nous avons récemment travaillé et qui s'apparente étroitement au cas des ELM.

1 La concentration de la mesure

Les spécialistes de la concentration de la mesure s'accordent à dire que cette théorie fut initialement introduite par Milman dans les années 70. Il reprenait les travaux de Levy sur la sphère $\mathbb{S}^n$ pour étudier la géométrie asymptotique des espaces de Banach. Cette notion est en effet essentiellement géométrique puisqu'elle a à voir avec l'inégalité isopérimétrique (ie, la minoration à volume fixée de la surface d'une variété. dans R^p, la sphère optimise cette inégalité. Dans une variété Riemannienne, la concentration peut se traduire comme une borne inférieure donnée à la courbure de Ricci comme exposé par Gromov dans [Gro79]). Nous contournerons cependant ici cette approche structurelle pour garder une vision strictement probabiliste qui se trouve suffisante pour nos besoins. Nous pouvons signaler le succès qu'a reçu la théorie en France dans les années 80 et 90 avec les mathématiciens Pisier et Maurey qui nous ont donné en particulier

une démonstration très synthétique de la concentration des vecteurs Gaussien (c'est la preuve du théorème 1), Talagrand [Tal95] qui a étudié la concentration dans les espaces produits (c'est le théorème 2) et Ledoux [Led01] qui a développé le lien avec l'entropie en traduisant des propriétés de concentration sous forme d'inégalités de Sobolev. La concentration de la mesure apparaît comme une notion transversale assez puissante pour formaliser certains phénomènes caractéristiques des grandes dimensions. L'idée de base qui ressort de la plupart des papiers sur la question est qu'une fonction à valeur réelle dépendant de beaucoup de paramètres de manière équilibrée est quasiment constante (sur un espace concentré...).

Bien que ce soit en grande dimension que la théorie trouve toute son sens, nous avons fait le choix de développer une approche originale qui part du cas des variable aléatoires pour ensuite la généraliser au cas des vecteurs aléatoires. Nous exposerons ainsi les outils probabilistes avant les théorèmes de concentration pour persuader le lecteur avant de le convaincre.

1.1 Concentration d'une variable aléatoire

La concentration d'une variable aléatoire peut se voir dans un premier temps comme un contrôle sur ses variations.

Définition 1. *Étant donnés une variable aléatoire $Z \in \mathbb{R}$ et une fonction $\alpha : \mathbb{R} \to \mathbb{R}$, nous dirons que Z est* α-concentrée, *et nous noterons $Z \propto \alpha$ si et seulement si pour tout Z', copie indépendante de Z :*

$$\forall t > 0 \ : \ \mathbb{P}\left(|Z - Z'| \geq t\right) \leq \alpha(t) \tag{1}$$

Remarque 1. *La concentration dans $\mathbb{R}$ peut aussi se traduire comme une concentration autour d'une médiane (cf. [Led01], cor 1.5.). Si $\mathbb{P}(Z \geq m_Z) \geq 1/2$ et $\mathbb{P}(Z \leq m_Z) \geq 1/2$, l'équation (1) implique :*

$$\forall t > 0 \ : \ \mathbb{P}\left(|Z - m_Z| \geq t\right) \leq 2\alpha(t)$$

Et c'est même une équivalence moyennant l'apparition d'un facteur $1/2$ devant l'argument de α.

On peut déjà noter trois opérations simples, stables pour la concentration :

Proposition 1. *Si $Z_1 \propto \alpha$, $Z_2 \propto \beta$ et $f : \mathbb{R} \to \mathbb{R}$ est λ-lipschitzienne, alors :*
- $f(Z) \propto \alpha(\cdot/\lambda)$
- $Z_1 + Z_2 \propto \alpha(\cdot/2) + \beta(\cdot/2)$
- $\max(Z_1, Z_2) \propto \alpha(\cdot/2) + \beta(\cdot/2)$

Le lecteur pourrait être tenté d'étudier le cas d'une somme ou d'un maximum calculé sur n termes avec n possiblement très grand, il espérerait alors pouvoir obtenir une concentration intéressante proportionnelle à $\log(n)$. On peut en

réalité avoir beaucoup mieux comme nous le verrons dans la sous-section 1.2 sur la concentration des vecteurs aléatoires qui est le bon cadre pour traiter ces grandes dimensions.

Pour le produit, la concentration n'est pas aussi évidente, il faudrait supposer de plus que les deux variables sont bornées. Mais nous préférons passer tout de suite au cas plus précis mais aussi plus riche qui nous occupera tout le reste du papier : le cas de la concentration normale. Elle est caractéristique de nombreuses distributions et apparaîtra dans nos exemples par le biais du Théorème 1 pour le cas de vecteurs gaussiens et du Théorème 2 pour le cas de vecteurs à entrées indépendantes et bornées.

Définition 2. *Nous dirons qu'une variable aléatoire est* normalement concentrée *avec un* paramètre de queue σ *(il est de même ordre que l'écart type) et un* paramètre de tête $C \geq 1$ *autour du* pivot $a \in \mathbb{R}$ *si :*

$$\forall t > 0 \ : \ \mathbb{P}\left(|Z - a| \geq t\right) \leq Ce^{-(t/\sigma)^2}$$

On notera alors $Z \in a \pm Ce^{-(\frac{\cdot}{\sigma})^2}$*, ou plus simplement* $Z \in Ce^{-(\frac{\cdot}{\sigma})^2}$ *si le pivot* a *n'a pas besoin d'être notifié.*

Cette définition rejoint bien la définition de concentration définie plus haut car dans le cas de la concentration normale, la moyenne comme la médiane jouent toutes deux le rôle de *pivots* de la concentration :

Proposition 2 (cf. [Led01], prop 1.8.)**.** *Soit* Z*, une variable aléatoire,* $C \geq 1$*,* $c > 0$ *et* $a \in \mathbb{R}$*, si* $Z \in a \pm Ce^{-c\cdot^2}$*, alors pour toute médiane* m_Z *de* Z*,* $Z \in m_Z \pm 2C^2e^{-c\cdot^2/2}$ *et* $Z \in \mathbb{E}Z \pm Ce^{C^2/2}e^{-c\cdot^2/2}$*.*

Remarque 2. *Notons que d'après la Proposition 2 et la Remarque 1 :*

$$Z \in a \pm Ce^{-c\cdot^2} \quad \Rightarrow \quad Z \propto 2C^2e^{-c\cdot^2/4} \quad \Rightarrow \quad Z \in m_Z \pm 4C^2e^{-c\cdot^2/4}$$

Lorsque l'on monte en dimension le paramètre de queue se fera appeler *diamètre observable* et contiendra alors l'information utile de l'inégalité de concentration. C'est lui qui caractérisera la concentration en fonction de la dimension. Le paramètre de tête joue quand à lui un rôle strictement technique, gardons en effet à l'esprit que l'exponentielle décroit très rapidement et que quoi qu'il arrive $\mathbb{P}\left(|Z - \mathbb{E}Z| \geq t\right) \leq 1$ avec égalité quand $t = 0$.

La concentration normale est relativement respectueuse pour le produit, il suffit qu'un des deux termes du produit soit borné.

Proposition 3. *Etant donnés deux variables aléatoires* $Z_2, Z_2 \in \mathbb{R}$ *et des constantes* $C_1, C_2 \geq 1$*,* $\sigma_1, \sigma_2, K_1 > 0$ *et* $b \in \mathbb{R}$*, si on suppose que que* $Z_1 \leq K_1$*, alors on a l'implication :*

$$Z_1 \in C_1e^{-(\frac{\cdot}{\sigma_1})^2} \ \ \& \ \ Z_2 \in b \pm C_2e^{-(\frac{\cdot}{\sigma_2})^2} \quad \Longrightarrow \quad Z_1Z_2 \in (C_1 + C_2)e^{-\left(\frac{\cdot}{\sigma}\right)^2},$$

avec $\sigma = 2\max\left(\sigma_1\left|b\right|,\ \sigma_2K_1\right)$.

Démonstration. Saoit a, un pivot de Z_1 et b, un pivot de Z_2. Il suffit de borner :

$$\begin{aligned}\mathbb{P}\left(|Z_1Z_2 - ab| \geq t\right) &= \mathbb{P}\left(|Z_1\left(Z_2 - b\right) + \left(Z_1 - a\right)b| \geq t\right) \\ &= \mathbb{P}\left(|Z_1|\,|Z_2 - b| \geq t/2\ \right) + \mathbb{P}\left(|Z_1 - a|\,|b| \geq t/2\right)\end{aligned}$$

□

Comme on va le voir dans la proposition suivante, la concentration normale offre des facilités calculatoires substantielles puisqu'elle se prête naturellement aux inégalités de Hölder pour les majorations des espérances par exemple. Elle peut en effet se traduire comme une inégalité sur les moments centrés :

Proposition 4 ([Led01], prop 1.10.)**.** *Soit Z une variable aléatoire et $C \geq 1$ et $c > 0$, on a :*

$$Z \in a \pm Ce^{-(\frac{\cdot}{\sigma})^2} \quad \Rightarrow \quad \mathbb{E}\left[|Z - a|^r\right] \leq \frac{C\sigma^r r^{\frac{r}{2}}}{2^{\frac{r}{2}}}, r > 0 \quad \Rightarrow \quad Z \in a \pm Ce^{-\frac{\cdot^2}{e\sigma^2}}$$

Comme on peut le montrer il est en réalité suffisant d'avoir une inégalité sur les moments pairs pour en déduire une concentration normale. Mais la condition nécessaire $r > 0$ est plus large et tout aussi attendue.

Démonstration. La condition suffisante se démontre avec une inégalité de Markov bien choisie et une optimisation sur r, la condition nécessaire est une conséquence de Fubini :

$$\mathbb{E}\left[|Z - a|^r\right] = \int_{t=0}^{\infty} \mathbb{P}\left(|Z - a|^r \geq t\right) dt$$

et on conclue grâce à une intégration par partie ou bien un changement de variable. □

Un autre type de concentration va apparaître lorsqu'on s'intéressera au carré de variable normalement concentrée, il s'agit de la concentration exponentielle :

$$Z \in Ce^{-t/\sigma} \qquad \Longleftrightarrow \qquad \forall t > 0\ :\ \mathbb{P}\left(|Z - \mathbb{E}Z| \geq t\right) \leq Ce^{-t/\sigma}$$

La concentration exponentielle peut elle aussi se traduire en terme d'inégalités sur les moments centrés d'ordre $r > 0$, cette fois majorés par $C(\frac{\sigma r}{2})^r$.

Nous donnons enfin la proposition qui nous a été très utile dans notre étude de la matrice de covariance empirique et qui nous permettra de démontrer simplement les concentrations des quantités X^TAX lorsque A est une matrice de norme spectrale bornée et X est un vecteur aléatoire concentré.

Résultat 1. *Étant donné deux variables aléatoire $Z_1, Z_2 \in \mathbb{R}$, quatres parametre $C_1, C_2 \geq 1$ et $\sigma_1, \sigma_2 > 0$ et deux pivots $a, b \in \mathbb{R}$, on a l'implication :*

$$\begin{cases} Z_1 \in a \pm C_1 e^{-(\frac{\cdot}{\sigma_1})^2} \\ Z_2 \in b \pm C_2 e^{-(\frac{\cdot}{\sigma_2})^2} \end{cases} \quad \Longrightarrow \quad Z_1Z_2 \in ab \pm (C_1 + C_2)\left(e^{-\frac{\cdot}{\sigma}} + e^{-(\frac{\cdot}{\sigma'})^2}\right)$$

avec $\sigma = 3\max(\sigma_1^2, \sigma_2^2)$ et $\sigma' = 3\max(\sigma_2|a|, \sigma_1|b|)$

On préfère prendre la même concentration pour Z_1 et Z_2 pour simplifier l'expression de la concentration de $Z_1 Z_2$. Du reste, cette proposition sera essenciellement utilisée pour démontrer la concentration de carrés comme $X^T A X = \|\sqrt{A}X\|$ si A est symétrique positive.

Démonstration. Tout repose sur l'identité algébrique donnée pour deux pivots $a, b \in \mathbb{R}$:

$$Z_1 Z_2 - ab = (Z_1 - a)(Z_2 - b) + a(Z_2 - b) + b(Z_1 - a)$$

On s'en sort ensuite comme dans la démonstration de la proposition 3 (avec l'inégalité $\mathbb{P}(|Z_1|\,|Z_2| \geq t) \leq \mathbb{P}(|Z_1| \geq \sqrt{t}) + \mathbb{P}(|Z_2| \geq \sqrt{t})$). □

On peut aussi déduire la concentration de n'importe quelle puissance d'une variable concentrée Z :

Résultat 2. *Etant donné une puissance $m \in \mathbb{N}$ et une variable normalement conentrée Z verifiant $Z \in a \pm Ce^{-(\frac{\cdot}{\sigma})^2}$, où $C \geq 1$ et $c > 0$, on a la concentration :*

$$Z^m \in a^m \pm Ce^{-(\frac{\cdot}{\sigma_2})^2} + Ce^{-(\frac{\cdot}{\sigma_m})^{\frac{2}{m}}} \qquad \textit{avec :} \qquad \begin{cases} \sigma_2 = 2^m |a|^m \sigma \\ \sigma_m = \sqrt{2}\sigma^m \end{cases}$$

Démonstration. C'est une conséquence du binome de Newton :

$$Z^m = (Z - a + a)^m = a^m + a^m \sum_{i=1}^{m} \binom{m}{i} \left(\frac{Z-a}{a}\right)^i.$$

En étudiant séparément les cas $|\frac{Z-a}{a}| \leq 1$ et $|\frac{Z-a}{a}| \geq 1$, on on conclue grace à :

$$|Z^m - a^m| \leq (2|a|)^m \left(\left|\frac{Z-a}{a}\right| + \left|\frac{Z-a}{a}\right|^m\right).$$

□

1.2 Concentration d'un vecteur aléatoire

Le passage de la dimension 1 aux dimensions supérieures est crucial car c'est ici que la théorie prend tout son sens. L'erreur serait de remplacer la valeur absolue de la définition 1 par une norme. Alors la concentration resterait la *concentration autour de quelque chose*, un vecteur aléatoire serait concentré lorsque ses tirages s'agglutinent. C'est donc déjà se priver du cas des vecteurs gaussiens qui en grande dimension se répartissent autour d'une sphère.

Notons γ, la loi gaussienne de moyenne nulle et de variance unitaire, et $Z \sim \gamma^{\otimes n}$, le vecteur gaussien canonique à n entrées indépendantes. Lorsque $n = 1$ (et même $n = 2$), les tirages de Z semblent se concentrer autour de 0. Cependant, quand n augmente, on voit très vite que Z s'éloigne de zéro pour parcourir un

espace proche de la sphère de rayon $\sqrt{n}$: $\sqrt{n}\mathbb{S}^{n-1}$. On dit que le diamètre de la distribution $\gamma^{\otimes n}$ est d'ordre $\sqrt{n}$. Mais ce qui nous intéresse, et ce qu'avait relevé Levy vers 1917 en étudiant la distribution uniforme sur la sphère, c'est le *diamètre observable*, c'est à dire très pratiquement, l'évaluation que ferait un observateur (humainement limité à la dimension 2 voir 3) de la distribution. Et bien il aurait l'impression de voir un amas de taille humaine, c'est à dire d'ordre 1.

Avant de démontrer cet effet, essayons de formaliser ce que l'on entend par "diamètre observable". C'est ce qui s'appelait plus haut "paramètre de queue" :

Définition 3. *Étant donné un espace vectoriel normé $(E, \|\cdot\|)$, nous dirons qu'un vecteur aléatoire $Z \in E$ est normalement concentré avec un* diamètre observable σ *et paramètre de tête C si pour toute fonction $f : E \to \mathbb{R}$, 1-Lipschitzienne, $f(Z) \in \mathbb{E}f(Z) \pm Ce^{-(\frac{\cdot}{\sigma})^2}$. On notera alors $Z \in \mathbb{E} \pm Ce^{-(\frac{\cdot}{\sigma})^2}$.*

On peut s'intéresser à la concaténation de deux vecteurs aléatoires suivant le produit de leurs espaces vectoriels respectifs. C'est l'indépendance qu'on avait évoqué dans la sous section précédente, comme promis, la proposition qui suit nous donne en particulier la concentration du produit de deux variables aléatoires indépendantes.

Proposition 5 ([Led01], prop 1.15.)**.** *Soient $Z_1 \in E$ et $Z_2 \in F$, deux vecteurs normalement concentrés avec des diamètres observables inférieurs à σ alors le vecteur aléatoire $(Z_1, Z_2) \in E \times F$ est lui aussi normalement concentré avec un diamètre observable inférieur à 2σ*

Cette proposition se montre avec la transformée de Laplace, grâce au lemme suivant qui découle de l'inégalité de Markov et d'une optimisation sur λ :

Lemme 1 ([Led01], prop 1.9, 1.14.)**.** *Soit f, une variable aléatoire de moyenne nulle (par exemple $f(Z) - \mathbb{E}f(Z)$). Si pour tout $\lambda \in \mathbb{R}$, on a $\mathbb{E}e^{\lambda f} \leq e^{\lambda^2\sigma^2/4}$, alors $f \in 0 \pm 2Ce^{-(\frac{\cdot}{\sigma})^2}$. Réciproquement, si $f \in 0 \pm e^{-(\cdot/\sigma)^2}$, alors $\forall \lambda \in \mathbb{R}$: $\mathbb{E}\left[e^{\lambda f}\right] \leq 2e^{\frac{\sigma^2\lambda^2}{2}}$.*

Démonstration de la proposition 5. Etant donnés $\lambda \in \mathbb{R}$ et une fonction $f : E \otimes F \to \mathbb{R}$ 1-lipschitzienne telle que $\mathbb{E}f(Z_1, Z_2) = 0$, on a :

$$\begin{aligned}\mathbb{E}e^{\lambda f(Z_1,Z_2)} &= \mathbb{E}\left[e^{\lambda\mathbb{E}[f(Z_1,Z_2)|Z_2]}\mathbb{E}\left[e^{\lambda(f(Z_1,Z_2)-\mathbb{E}[f(Z_1,Z_2)|Z_2])} \mid Z_2\right]\right] \\ &\leq \mathbb{E}\left[e^{\lambda\mathbb{E}[f(Z_1,Z_2)|Z_2]}2e^{\frac{\lambda^2\sigma}{2}}\right] \;\leq\; 4e^{\sigma\lambda^2},\end{aligned}$$

car $\mathbb{E}[f(Z_1, Z_2) \mid Z_2]$ est lipschitzien et de moyenne nulle par hypothèse. □

Nous voyons dans la proposition 5 que le produit des mesures nous fait augmenter à chaque étape le diamètre observable d'un facteur 2, elle ne peut donc pas nous permettre de monter en dimension puisqu'on aboutirait à un diamètre

observable de 2^n largement sur évalué. Avec les mêmes méthodes utilisée plus intelligemment, on peut montrer qu'il est majoré par n. Cependant pour les besoins de notre étude il nous fallait un diamètre observable d'ordre $O(1)$ et c'est ce qu'on retrouve pour les transformations lipschitzienne d'un vecteur gaussien :

Théorème 1 ([Led01], cor 2.6, prop 2.18)**.** $Z \sim \gamma^{\otimes n} \implies Z \in 2\alpha_{\mathcal{N}}(\cdot/\sqrt{2})$.

Ce théorème peut se démontrer géométriquement à partir de l'inégalité isopérimétrique. Elle nous donne la concentration de Lévy sur la sphère et ensuite on reprend un argument de Poincaré qui construit la mesure gaussienne $\gamma^{\otimes n}$ comme la limite quand m tend vers ∞ des projections sur $\mathbb{R}^n$ des mesures uniformes sur $\mathbb{S}^m$. On peut aussi reprendre l'argument astucieux de Pisier et Maurey qui n'obtiennent pas des constantes aussi fines mais démontrent le résultats très rapidement.

Démonstration. (cf. [Tao11], theo 1.1.12.) On veut utiliser le lemme 1, on part de l'inégalité étrange :

$$\mathbb{E}e^{-tf(Z)} \leq \mathbb{E}e^{-t(f(Z)-f(Z'))} \qquad \text{d'après Jensen car } \mathbb{E}f(Z) = 0,$$

où Z est une copie indépendante de Z. Il faut ensuite écrire la différence $f(Z) - f(Z')$ comme un taux de variation d'un $df = \frac{d}{d\theta} f(Z_\theta)$ où $Z_\theta = \cos\theta Z + \sin\theta Z'$(on suppose f différentiable). On se ramène à la majoration recherchée grâce à l'inégalité de Jensen et on conclue grâce à l'hypothèse sur f en remarquant que $df = \nabla f(Z_\theta) Z'_\theta$ est un vecteur gaussien de variance unitaire. □

Le champ d'application de ce théorème concerne toutes les transformations Lipschitziennes de vecteurs gaussien c'est à dire des modifications respectueuses de la sphère qui se permettent possiblement quelques ablations mais pas de dissociation. On aurait aimé pouvoir traiter des cas de distributions discontinues. Talagrand nous offre un résultat de ce type un peut plus faible que la concentration normale mais toutefois suffisant encore une fois pour nos besoins.

Théorème 2 ([Tal95] Theo 4.1.1)**.** *Soit $Z \in [0,1]^n$, un vecteur aléatoire avec des entrées* indépendantes *(et* bornées*), pour toute fonction $f : [0,1]^n$ 1-Lipschitzienne et* convexe*, $f(Z) \in 4e^{-(\frac{\cdot}{2})^2}$.*

Rappelons que l'on cherche à appliquer ces théorème à des ensembles de données qu'on retrouve en pratique. Pour des ensembles d'images, le caractère borné des entrées ne pose pas de problème c'est en revanche l'indépendance qui devient un obstacle quasiment infranchissable, par exemple si l'on a affaire à des image de chiffres. Le théorème 1 autorisait un élargissement de ses hypothèses à toute les transformations Lipschitzienne d'un vecteur gaussien. Ici à cause de l'hypothèse de convexité, on doit se borner aux transformations affines qui englobe tout de même certains cas où les entrées sont indépendantes.

Enfin on va faire avec ce qu'on a et l'hypothèse de concentration normale pourra être adoptée comme une hypothèse qu'un physicien (ou un économiste...) admet

sur le système qu'il étudie. Nous travaillerons donc suivant les deux configuration exposées par ces théorèmes fondamentaux.

L'enjeu est alors de savoir si les fonctions que nous allons rencontrer seront bien convexes en plus d'être Lipschitziennes. C'est bien le cas comme on va le voir. Signalons tout d'abord la concentration des produits scalaires $u^T Z$ lorsque $\|u\| = 1$, (cela nous donne une concentration en paramètre de queue proportionnel à $\sqrt{n}$ pour la somme des Z_i). Grâce au Résultat 1, nous obtenons aussi très facilement une version plus aboutie de l'inégalité de Hanson-Wright :

Résultat 3. *Etant donnés un vecteur aléatoire $Z \in E$ et une application linéaire u telle que $\|u\| \leq 1$, si Z vérifie les hypothèse des théorèmes 1 ou 2 avec un diamètre observable σ et un paramètre de tête C, alors :*

$$\begin{cases} \|u(Z)\| \in Ce^{-(\frac{\cdot}{\sigma})^2} \\ \|u(Z)\|^2 \in 2Ce^{-\frac{\cdot}{3\sigma}} + 2Ce^{-(\frac{\cdot}{9\sigma})^2} \end{cases}$$

En particulier si $Z \in \mathbb{R}^p$ et $u(Z) = \sqrt{A}$ avec $A \in \mathcal{M}^p$ symétrique positive de norme 1, on retrouve la concentration de $Z^T AZ$ que nous recroiserons.

2 Applications

En étudiant le problème de la convergence de la sortie de l'ELM qui a motivé nos premières recherches, on s'est rendu compte que la seule hypothèse utile était l'indépendance des colonnes de $\Sigma = \sigma(WX)$ et sa concentration suivant le Théorème 1 ou le théorème 2. Il est alors possible de simplifier le problème en ne prenant pas en compte les transformation initiales caractérisée par la multiplication par W et l'application σ. On en vient alors à réaliser une simple régression de Ridge sur la matrice des données X qui sera présentée dans la sous section 2.2. Présentons d'abord plus en détail la matrice X étudiée.

Donnons nous k distributions $\mu_1, \cdots \mu_k$ sur $\mathbb{R}^p$ et n vecteurs aléatoires $\{x_1, \cdots x_n\}$ suivant chacun une de ces k distributions qui représentent les k classes de notre modèle. On a affaire à un mélange de données qu'on voudrait identifier selon leur classe. On se place dans le cas où n et p sont très grand et de même ordre ($n = O(p)$ et $p = O(n)$). Le nombre de classe k est supposé négligeable devant n et p, nous le noterons donc $O(1)$. On suppose de plus que la matrice $X = (x_1, \cdots x_n)$, qui regroupe toutes les données rangées en colonne, vérifie les hypothèses du théorème 1 ou du théorème 2 en tant que vecteur de $\mathcal{M}_{p,n}$ muni de la norme de Frobenius $\|\cdot\|_F$ définie par $\|M\|_F^2 = \sum m_{i,j}^2$ (c'est une norme euclidienne). On notera la norme spectrale plus simplement $\|\cdot\|$.

Nous allons d'abord étudier le spectre de la matrice de covariance empirique $C_X = XX^T/n$ qui doit révéler l'existence de diverses classes. Nous présenterons ensuite la techniques de régression de Ridge qui est à l'origine des ELM.

2.1 Distribution spectrale d'une covariance empirique

La convergence quand $n \to \infty$ de la distribution spectrale $\mu_C = \frac{1}{p}\sum_{w\in \mathrm{Sp}(C)} \delta_w$ de C peut être vue comme une conséquence de la convergence de sa transformée de Stieltjes g_C qui est une fonction holomorphe définie hors du spectre de C :

$$g_C(z) = \int \frac{1}{w-z} d\mu_C(w) = \frac{1}{p} \operatorname{Tr} Q_C(-z),$$

où la matrice $Q_C(z)$ est appelée résolvante de C et est égale à $(C + zI_p)^{-1}$. Il existe une identification entre distribution et transformée de Stieltjes qu'on ne peut raisonnablement pas développer ici. Néanmoins il est possible de justifier le succès de cette approche grâce au lemme qui suit. La résolvante, contrairement à la covariance, est bornée en norme spectrale, indépendamment de n et p.

Lemme 2. *Pour toute matrice symétrique positive $D \in \mathcal{M}_p$ et tout $R \in \mathcal{M}_{p,n}$:*

$$\|Q_D(z)\| \leq \frac{1}{|z|} \qquad \|Q_D(z)D\| \leq 1 \qquad \left\|Q_{C_R}(z)R/\sqrt{n}\right\| \leq \frac{1}{\sqrt{|z|}}$$

Dorénavant toutes ces quantités seront pour nous des $O(1)$ et $Q_C(z)$ sera noté plus simplement Q. Ce contrôle sur Q nous permet (parce que les fonctions en jeu sont convexes) de montrer la concentration de plusieurs fonctionnelles de Q en fonction de leur paramètre de lipschitzianité.

Résultat 4. *Il existe deux paramètres $C, c = O(1)$ tels que pour tout vecteur $u \in \mathbb{R}^p$ et $A \in \mathcal{M}_p$ symétrique positive, vérifiant $\|u\|, \|A\| \leq 1$ on a :*

$$\begin{cases} \|u\| \leq 1 & \Longrightarrow \quad u^T Q u \in e^{-cn\,\cdot^2} \\ \|A\| \leq 1 & \Longrightarrow \quad \operatorname{Tr} AQ \in e^{-c\,\cdot^2} \\ \operatorname{Tr} A \leq p & \Longrightarrow \quad \operatorname{Tr} AQ \in e^{-\frac{c}{p}\,\cdot^2} \end{cases}$$

Donc en particulier, la transformée de Stieltjes est bien normalement concentrée avec un paramètre de queue d'ordre $O(1/n)$: reste à lui obtenir un estimateur.

Comme on l'avait vu au début de la sous-section 1.2 il serait absurde de chercher une matrice déterministe $\tilde{C}$ telle que $\|C - \tilde{C}\|_F$ (ou la norme spectrale) tende vers 0. Il en est de même pour la matrice aléatoire Q. On va plutôt chercher ce qu'on appelle un équivalent déterministe de Q, c'est à dire une matrice dont les *observations* sont les mêmes. On cherche par exemple une matrice $\tilde{Q}$ telle que pour toute matrice symétrique A de norme spectrale unitaire et tout vecteur unitaire u, on a presque surement :

$$u^T(Q - \tilde{Q})u = O(1) \qquad \text{et} \qquad \operatorname{Tr}(A(Q - \tilde{Q}))/p = O(1).$$

Il serait naïf de croire que la matrice $\tilde{Q}$ serait simplement $(\bar{C} + zI_p)^{-1}$ où $\bar{C}$ est l'espérance de C. En partant de la forme $\tilde{Q} = (\tilde{C} + zI_p)^{-1}$ avec $\tilde{C}$ à déterminer,

on arrive au calcul classique en matrices aléatoires :

$$\tilde{Q} - \mathbb{E}Q = \mathbb{E}\left[Q\left(XX^T/n - \tilde{C}\right)\tilde{Q}\right] = \sum_{i=1}^{n} \frac{1}{n}\mathbb{E}\left[Q(x_i x_i^T - \tilde{C})\tilde{Q}\right] \tag{2}$$

$$= \sum_{i=1}^{n} \frac{1}{n}\mathbb{E}\left[Q_{-i}\left(\frac{x_i x_i^T}{1 + x_i^T Q_{-i} x_i/n} - \tilde{C}\right)\tilde{Q}\right] - \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[Q_{-i}x_i x_i^T Q\tilde{C}\tilde{Q}\right]$$

où Q_{-i} est la matrice Q privée de la contribution de x_i, on a utilisé les identités classiques qui nous permettent de gérer les interférences entre Q et les $(x_i)_{1\leq i\leq n}$:

$$Q = Q_{-i} - \frac{1}{n}\frac{Q_{-i}x_i x_i^T Q_{-i}}{1 + x_i^T Q_{-i} x_i/n} \qquad \text{et} \qquad Qx_i = \frac{Q_{-i}x_i}{1 + x_i^T Q_{-i} x_i/n}$$

on peut montrer que le terme de droite dans la différence en (2) est négligeable devant celui de gauche qui nous encourage alors à choisir $\tilde{C} = \frac{\bar{C}}{1+\operatorname{Tr}\bar{C}\mathbb{E}Q_{-i}/n}$. Mais ce serait oublier que tous les x_i ne suivent pas la même loi, il faut en fait choisir :

$$\tilde{C} = \sum_{l=1}^{k} \frac{n_l}{n}\frac{\bar{C}_l}{1+\delta_l} \quad \text{avec} \quad \forall l \in \{1, \cdots k\} \ : \ \delta_l = \frac{1}{n}\operatorname{Tr}(\bar{C}_l \mathbb{E}Q_{-l})$$

où $\bar{C}_l$ est la covariance de la distribution μ_l et Q_{-l} est un Q_{-i} où i suit la loi μ_l. On prouve la validité de cet équivalent déterministe en particulier grâce au Résultat 3 qui nous donne la concentration de $\frac{1}{n}x_i^T Q_{-i}x_i$ autour de δ_l si x_i suit la loi μ_l. L'équivalent déterministe obtenu par ces méthodes n'est cependant pas vraiment utile en pratique car il nous faut $\mathbb{E}Q_{-l}$ qu'on ne sait pas calculer. On va donc préférer choisir les $\{\delta_l\}_{1\leq l\leq k}$ comme solution du problème à k équations :

$$\forall h \in \{1, \cdots k\} \ : \ \delta_h = \frac{1}{n}\operatorname{Tr}\left(\bar{C}_h\left(\sum_{l=1}^{k}\frac{n_l}{n}\frac{\bar{C}_l}{1+\delta_l} + zI_p\right)^{-1}\right)$$

C'est un problème de point fixe dont qu'on peut approcher numériquement par itération. On peut montrer que la solution est proche de $\frac{1}{n}\operatorname{Tr}(\bar{C}_l\mathbb{E}Q_{-l})$ à $O(1/\sqrt{n})$ près.

On a donc mis au point une manière de prévoir la distribution spectrale d'une matrice de covariance empirique.

2.2 Régression de Ridge

Cette compréhension du comportement spectrale de C et la construction de l'équivalent déterministe $\tilde{Q}$ nous donne les clefs pour comprendre les performance de la régression de Ridge et donc celle des ELM présentés dans le préambule. La régression de Ridge en question consiste à minimiser le problème :

$$\frac{1}{n}\left\|\beta X - L\right\|^2 + \gamma\left\|\beta\right\|^2 \ , \quad \beta \in \mathcal{M}_{k,p}$$

sur un ensemble de données déjà labellisées (X, Y) où $L \in \mathcal{M}_{k,n}$ est remplie de 0 et contient un unique 1 dans chaque colonne i à l'indice de la loi de x_i. On espère ensuite pouvoir évaluer la classe d'une nouvelle donnée x en prenant l'indice de la coordonnée maximale du *score* de x, $S(x) = \frac{1}{n} L X^T Q x \in \mathbb{R}^k$.

On peut alors évaluer l'erreur quadratique E_q avec la formule :

$$E_q = \mathbb{E} \left\| S(x) - l_x \right\|^2 = \left\| \mathbb{E} S(x) - l_x \right\|^2 + \mathcal{V}(S(x))$$

où $l_x \in \mathbb{R}^k$ désigne le vecteur caractéristique de la classe k_x de x (il contient un 1 au $k_x^{\text{ième}}$ indice). On voit apparaître naturellement l'espérance de $S(x)$ et sa variance $\mathcal{V}(S(x))$ qu'il faut alors évaluer. On arrive à montrer que :

$$\mathbb{E} S(x) = \sum_{l=1}^{k} \frac{n_l}{n} \frac{\bar{x} \tilde{Q} \bar{x}_l}{1 + \delta_l} + O\left(\frac{1}{\sqrt{n}}\right)$$

où $\bar{x}$ désigne l'espérance de la distribution de x et $\bar{x}_l$ désigne l'espérance de μ_l. Pour l'estimation de la variance, il faut d'abord trouver un équivalent déterministe de QAQ. Avec la notation $\nu_l = \frac{n_l}{1+\delta_l}$ introduisons les matrices :

$$\Psi = \left(\frac{\nu_h}{n} \frac{\mathbb{E} Q \bar{C}_h Q \bar{C}_l}{1 + \delta_l} \right)_{1 \leq l,h \leq k} \qquad \tilde{\Psi} = \left(\frac{\nu_h}{n} \frac{\tilde{Q} \bar{C}_h \tilde{Q} \bar{C}_l}{1 + \delta_l} \right)_{1 \leq l,h \leq k}$$

$$\Phi = \left(\frac{\nu_h \nu_g}{1 + \delta_l} \mathbb{E} \bar{x}_h^T Q \bar{C}_l Q \bar{x}_g \right)_{1 \leq l,h,g \leq k} \qquad \tilde{\Phi} = \left(\frac{\nu_h \nu_g}{1 + \delta_l} \bar{x}_h^T \tilde{Q} \bar{C}_l \tilde{Q} \bar{x}_g \right)_{1 \leq l,h,g \leq k}$$

$$\tilde{\Theta} = \left(\nu_h \nu_g \bar{x}_h \tilde{Q} \bar{x}_g \right)_{1 \leq h,g \leq k}$$

Notons que les matrices Φ et $\tilde{\Phi}$ sont des pavés, si on les multiplie à droite par des matrices rectangulaire, on obtient une nouvelle matrice pavé suivant la règle de calcul :

$$\Psi \Phi = \left\{ \sum_{m=1}^{k} \Psi_{l,m} \Phi_{m,h,g} \right\}_{1 \leq l,h,g \leq k}$$

Résultat 5. *On a l'estimation (les matrices sont de taille $O(1)$) :*

$$\Psi = (I_p - \tilde{\Psi})^{-1} \tilde{\Psi} + O\left(\frac{1}{\sqrt{n}}\right) \qquad \Phi = (I_p - \tilde{\Psi})^{-1} \tilde{\Phi} + O\left(\frac{\log n}{\sqrt{n}}\right)$$

Résultat 6. *Introduisons les matrices diagonales $D^l = Diag(\Psi_{\cdot,l})$, $D_l = Diag(\Psi_{l,\cdot})$ et $D_\delta = Diag(\frac{1}{1+\delta_l})_{1 \leq l \leq k}$. On a l'estimation :*

$$\mathbb{E}\left[S(x)\right] = \frac{\Theta_{\cdot,k(x)}}{\nu_x} + O\left(\frac{1}{\sqrt{n}}\right)$$

$$\mathbb{E}\left[S(x) S(x)^T\right] = (1 + \delta_{k(x)})(\Phi_{k(x),\cdot,\cdot} + D_l D_\delta) - \frac{\Theta D^l + D^l \Theta}{\nu_{k(x)}} + O\left(\frac{\log n}{\sqrt{n}}\right)$$

2.3 à venir...

Les résultats sur l'erreur quadratique sont assez récents, il nous reste à trouver le moyen de les analyser pour inférer si possible une technique d'optimisation du paramètre de régularisation γ de la régression de Ridge.

Par ailleurs, plutôt que l'erreur quadratique, c'est l'erreur de classification qui nous intéresse en pratique. On peut l'évaluer en supposant que le score suit une distribution normale ce qui semble être le cas. A première vue ce résultat n'est pas évident à démontrer. La $l^{\text{ième}}$ coordonnée du score peut s'écrire comme une somme $S_l(x) = \frac{1}{n}\sum_{k(i)=l} x_i Q x$ mais on ne peut pas utiliser le théorème de centrale limite car les différents termes ne sont pas indépendants à cause de la présence de la matrice Q qui dépend d'un grand nombre de variables.

Enfin nous souhaitons à terme étudier le comportement de réseaux de neurones plus complexes que les ELM présentés dans l'introduction. Nous baserons notre étude sur une hypothèse de concentration normale des données à analyser comme cela semble être nécessaire. Nous envisageons (de loin) pour l'instant deux sujets d'étude possibles, la "backpropagation" dont le succès reste toujours un mystère. Il est possible que nous devions prendre un point de vue plus géométrique interprétant la concentration avec la courbure de Ricci comme l'a fait par exemple le chercheur Yann Olivier.

Références

[Gro79] Mikhail Gromov. Paul lévy's isoperimetric inequality. *Preprint IHES*, 1979.

[HZS06] Guang-Bin Huang, Qin Yu Zhu, and Chee-Kheong Siew. Extreme learning machine : theory and applications. *Neurocomputing vol 70*, pages 489–501, 2006.

[Led01] Michel Ledoux. *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs, Number 89, 2001.

[LLC17] Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *Annals of applied Probability*, 2017.

[Tal95] Michel Talagrand. *Concentration of Measure and Isoperimetric Inequalities in product spaces*. Publications mathématiques de l'I.H.E.S., tome 81, 1995.

[Tao11] Terence Tao. Topics in random matrix theory. Department of Mathematics, UCLA, Los Angeles, 2011.