Apprentissage statistique et grande dimension
Notes de cours 2023

C. Levrard. Ces notes doivent beaucoup au cours de Statistiques en grande dimension de C. Giraud, et d'Apprentissage statistique de S. Arlot.

Chapitre 1

Point de vue général sur l'apprentissage

1.1 Régression/classification

Le point de vue général en apprentissage statistique est celui de la *prédiction* : à partir d'une variable explicative X à valeurs dans $\mathcal{X}$, on veut prédire une sortie Y, ou variable à expliquer, à valeurs dans $\mathcal{Y}$. Si la variable à expliquer Y est binaire ($\mathcal{Y} \simeq \{0,1\}$) on parle de *classification*, si Y est à valeurs continues ($\mathcal{Y} \simeq \mathbb{R}$ par exemple), on parlera plutôt de *régression*.

Exemple 1.1.

- Si on cherche à prédire, à partir d'une image (modélisée comme un vecteur X à valeurs dans $\mathbb{R}^D$, où D est le nombre de pixels) si elle représente un chat ou non, c'est un problème de *classification* (variable réponse Y en Oui/Non, ou 0/1).
- Si à partir de paramètres socio-économiques X on cherche à prédire la durée de vie en bonne santé d'un individu Y, c'est un problème de *régression*.

Remarque 1.2 *: Classification multi-classes et régression vectorielle.* On peut étendre légèrement ces deux notions : si $\mathcal{Y} \simeq \{1, \dots, K\}$ (K classes au lieu de 2), on peut parler de *classification multi-classes*. Si $\mathcal{Y} \simeq \mathbb{R}^p$, on peut parler de *régression vectorielle*.

Ces problèmes étant souvent traités via des techniques adaptées de la classification binaire ou de la régression réelle, on ne présentera que ces deux derniers dans ce cours.

Dans ces deux cas, un *prédicteur* est une fonction $f : \mathcal{X} \to \mathcal{Y}$, c'est à dire une machine à prédire. Dans un problème de régression on parle de *régresseur*, dans un problème de classification c'est un *classifieur*.

Lorsque l'on parle *d'apprentissage statistique*, le côté "statistique" vient du fait que l'on suppose que les variables (X, Y) suivent une certaine loi (que l'on notera $P_{(X,Y)}$ la plupart du temps, ou juste P lorsque c'est transparent). C'est là où la partie "modélisation" rentre en jeu :

Exemple 1.3 *: Anti-spam.* Le but de la tâche que l'on se donne est de construire un anti-spam, c'est à dire un prédicteur f qui prend en entrée différentes quantités mesurées sur un mail en entrée et ressort un label (0 pour pas spam, 1 pour spam).

Par exemple, on peut mesurer $X = (N_1, N_2)$, où N_1 est le nombre de fois où votre nom apparaît et N_2 le nombre de caractères non latins. Un modèle fruste est de supposer $X \sim \mathcal{P}(\theta_1) \otimes \mathcal{P}(\theta_2)$, et $Y \mid X \sim \mathcal{B}(p(\theta_1, \theta_2))$, où p est décroissante en la première variable et croissante en la seconde.

Une fois le modèle posé, on peut faire comme en stats classiques et essayer d'estimer θ_1, θ_2 et p. Cela dit ce n'est pas le but de l'apprentissage statistique : ce qui nous intéresse est de construire un classifieur qui se trompe le moins possible. Cela peut passer par l'estimation des paramètres d'un modèle, mais ce n'est pas obligatoire (et des fois même pas possible).

Le but est alors de construire $f : \mathcal{X} \to \mathcal{Y}$ tel que $f(X) \approx Y$ avec grosse proba. Reste à définir ce qu'on entend par $f(X) \approx Y$.

1.2 Pertes, risques et prédicteurs optimaux

Pour définir la qualité d'un prédicteur qui, pour l'entrée x prédit $f(x)$, il faut commencer par définir une notion de proximité entre $f(x)$ et y.

Definition 1.4 : Pertes et risques

Une fonction de perte c est une fonction

$$c : \begin{cases} \mathcal{Y} \times \mathcal{Y} & \to & \mathbb{R}^+ \\ (y_1, y_2) & \mapsto & c(y_1, y_2). \end{cases}$$

La fonction de risque R associée est définie par

$$R_P : \begin{cases} \mathcal{Y}^{\mathcal{X}} & \to & \mathbb{R}^+ \\ f & \mapsto & E_{(X,Y)} c(f(X), Y), \end{cases}$$

où $E_{(X,Y)}$ désigne l'espérance sous la loi $P_{(X,Y)}$.

Il est usuel de prendre pour fonction de perte une distance dans l'espace $\mathcal{Y}$. Un minimum est de demander quelque chose ressemblant à de la séparation, c'est à dire $c(y, y') = 0 \Leftrightarrow y = y'$.

Le risque associé à une fonction de perte c est donc une manière de mesurer la performance d'un prédicteur. Il dépend uniquement de la fonction de perte choisie et de la loi $P_{(X,Y)}$, d'où la notation R_P, qui sera souvent abrégée en R (mais il faut garder en tête que ça dépend de P). Le plus petit risque atteignable est appelé *risque de Bayes*, défini par

$$R_P^* = \inf_{f \in \mathcal{Y}^X} R_P(f),$$

et dépend uniquement de la fonction de perte ℓ et de la loi $P_{(X,Y)}$. De manière générale, le risque de Bayes peut prendre n'importe quelle valeur entre 0 (prédiction parfaite de Y via X) et $+\infty$ (fonction de perte non intégrable).

Dès lors, un prédicteur *optimal* f^* est un prédicteur vérifiant

$$f^* \in \arg\min_{f \in \mathcal{Y}^{\mathcal{X}}} R_P(f).$$

Un tel prédicteur est aussi appelé *prédicteur de Bayes*, n'existe pas forcément et quand bien même n'est pas nécessairement unique. La forme d'un prédicteur de Bayes dépend là encore de c et de la loi $P_{(X,Y)}$ uniquement.

Pour un prédicteur f quelconque, on peut aussi définir *l'excès de risque* $\ell_P(f, f^*)$ (avec un abus de notation) en comparant son risque au risque de Bayes, c'est à dire

$$\ell_P(f, f^*) = R_P(f) - R_P^*.$$

1.2.1 Pertes classiques en régression

Ici $\mathcal{Y} = \mathbb{R}$, on cherche à prédire une valeur réelle.

Perte quadratique

La perte la plus classique en régression est la perte quadratique, c'est à dire $c(y, y') = (y - y')^2$. Dans ce cas, les prédicteurs de Bayes sont donnés par la *fonction de régression* $\eta(X) = E(Y \mid X)$, lorsque bien défini.

Proposition 1.5 : Risque quadratique - Prédicteur de Bayes

On suppose que $E(|Y| \mid X) < +\infty$ P_X p.s., et $E((Y - \eta(X))^2) < +\infty$. On a alors

1. *η est un prédicteur de Bayes.*
2. *Si f est un prédicteur de Bayes, $f = \eta$ P_X p.s..*
3. *Le risque de Bayes vaut*

$$R_P^* = E_{(X,Y)}(Y - \eta(X))^2 = E(\mathrm{Var}(Y \mid X)).$$

4. *Si f est un prédicteur, son excès de risque vaut*

$$\ell_P(f, f^*) = \|f - \eta\|^2_{L_2(P_X)}.$$

Démonstration. La condition $E(|Y| \mid X) < +\infty$ assure que $\eta(X)$ est bien définie (P_X-p.s.). Soit f un régresseur, on a alors

$$\begin{aligned} E((Y - f(X))^2) &= E\left(\mathbb{E}\left((Y - f(X))^2 \mid X\right)\right) \\ &= E\left(\mathrm{Var}(Y \mid X)\right) + E\left(f(X) - \eta(X)\right)^2 \\ &= E(Y - \eta(X))^2 + E\left(f(X) - \eta(X)\right)^2, \end{aligned}$$

ce dont on déduit les quatre points. □

On peut remarquer que dans le cas où $E((Y - \eta(X))^2) = \infty$ n'importe quel régresseur est de Bayes. Enfin, dans les modèles de régression où on suppose $Y = f^*(X) + \varepsilon$, où $E(\varepsilon \mid X) = 0$, f^* correspond à la fonction de régression, et la condition minimale pour que tout fonctionne devient $E(\varepsilon^2) < +\infty$.

Autres fonctions de perte

Une extension directe de la perte quadratique est la perte L_q, pour $q \geq 1$, donnée par

$$c(y, y') = |y - y'|^q.$$

Toujours sous condition d'existence, un prédicteur de Bayes est alors donné par

$$X \mapsto \arg\min_u E(c(u,Y)|X) = \arg\min_u (E|u-Y|^q \mid X),$$

où au passage on peut remarquer que comme en statistique bayésienne il s'agit de minimiser un risque "condionnel" à X. Un cas particulier est celui de $q=1$, où dans ce cas un prédicteur de Bayes est donné par

$$X \mapsto \mathrm{Med}(Y|X),$$

c'est à dire la médiane de la loi $Y \mid X$ (bien définie pour Y $\mathcal{B}(\mathbb{R})$-mesurable). Un avantage de la médiane sur la moyenne est la moindre sensitivité aux erreurs : si $Y' \mid X \sim (1-\varepsilon)Y \mid X + \varepsilon\delta_{X_n}$ (distribution réponse corrompue avec bruit situé en X_n X-mesurable), alors

$$|\mathbb{E}(Y' \mid X) - \mathbb{E}(Y \mid X)| \underset{X_n \to +\infty}{\longrightarrow} +\infty,$$

tandis que $\mathrm{Med}(Y' \mid X)$ va rester à distance bornée de $\mathrm{Med}(Y \mid X)$. La norme 1 n'est qu'une des nombreuses fonctions de pertes associée à la notion de "robustesse" à de la contamination (ici de distribution source, mais cela s'appliquera pareillement lorsque l'on parlera d'échantillon d'entraînement). On peut citer par exemple (FAIRE DESSINS) :

— Perte quadratique seuillée :

$$c(y,y') = (y-y')^2 \wedge \alpha,$$

minimisée par la moyenne α-seuillée. On peut aussi seuiller n'importer quelle autre fonction de perte.

— Perte de Huber :

$$c(y,y') = (y-y')^2 \wedge (2\alpha|y-y'| - \alpha^2).$$

— Perte de Tukey-biweight :

$$c(y,y') = \left[1 - \left(1 - \frac{(y-y')^2}{\alpha^2}\right)^2\right] \wedge 1.$$

— Perte de Catoni-Giulini :

$$c(y',y) = \left(\frac{\psi(\lambda|y|)}{\lambda|y|}y - y'\right)^2,$$

avec $\psi(u) = u \wedge 1$. Cette perte n'est pas symétrique, le risque est défini par $R_P(f) = E_{(X,Y)}c(f(X),Y)$: la pondération porte sur la loi de $Y \mid X$, ce qui mène à l'estimateur de Bayes $E\left(Y\frac{\psi(\lambda|Y|)}{\lambda|Y|} \mid X\right)$ qui est optimal sous certains aspects.

L'idée de toutes ces fonctions de perte est de payer moins cher les erreurs large que pour la perte quadratique (et donc d'y être moins sensible).

Remarque 1.6. En toute généralité il n'est pas nécessaire d'avoir une fonction de perte pour définir un risque : de fait n'importe quelle fonction $R_P : \mathcal{Y}^{\mathcal{X}} \to \bar{\mathbb{R}^+}$

pourrait faire l'affaire (on pourrait aussi définir des risques et prédicteurs de Bayes). Cela étant la construction de prédicteurs de Bayes consistant à minimiser

$$u \mapsto \mathbb{E}(c(u, Y) \mid X)$$

ne tient plus, et l'interprétation d'un risque en prédiction non plus. Par exemple, si $R_P(f) := E(|\eta(X) - f(X)|)$, on ne peut pas exprimer $R_P(f)$ sous la forme $E(c(f(X), Y)) - K(P)$ (dans ce cas c'est plus un risque en *estimation* de la fonction de régression). Un des avantages de la perte quadratique est que minimiser $E((f(X) - Y)^2)$ revient à minimiser $E((f(X) - \eta(X))^2$ d'après la Proposition 1.5.

1.2.2 Pertes classiques en classification

Ici $\mathcal{Y} = \{0, 1\}$, on cherche à prédire une étiquette binaire.

Perte $0/1$

La perte la plus souvent utilisée **théoriquement** est celle définie par

$$c(y, y') = \mathbb{1}_{y \neq y'},$$

correspondant moralement à payer 1 lorsque l'on se trompe d'étiquette. Un prédicteur $f : \mathcal{X} \to \{0, 1\}$ est appelé classifieur, et son risque est donné par

$$R_P(f) = P_{(X,Y)}(f(X) \neq Y),$$

c'est à dire la probabilité qu'il se trompe sous la loi $P_{(X,Y)}$. Comme en régression la *fonction de régression*

$$\eta(X) = E(Y \mid X) = \mathbb{P}(Y = 1 \mid X),$$

va jouer un grand rôle dans la définition des risques et classifieurs de Bayes. Une borne triviale sur le risque de Bayes est donnée par

$$R_P^* \leq p \wedge (1 - p),$$

où $p = \mathbb{P}(Y = 1)$, les deux risques de droite correspondant respectivement aux risques des classifieurs stupides $f_0 \equiv 0$ et $f_1 \equiv 1$. Bien que stupide, cette première borne montre qu'il n'est pas toujours pertinent de se comparer à la valeur $1/2$ en pratique (si dans un jeu de données vous avez 98% de données d'une certaine classe et 2% de l'autre, un score de 2% ne correspond pas à une grande précision...).

On peut montrer qu'on peut remplacer $p = \mathbb{P}(Y = 1)$ par $\eta(X) = \mathbb{P}(Y = 1 \mid X)$ dans cette borne stupide pour obtenir le risque de Bayes.

Proposition 1.7 : Perte 0/1 - Classifieur de Bayes

Pour la perte 0/1, en notant $\eta(X) = \mathbb{P}(Y = 1 \mid X)$*, on a*

1. *Le classifieur* $f^*(X) = \mathbb{1}_{\eta(X) \geq 1/2}$ *est de Bayes.*
2. *Un classifieur* f *est de Bayes si et seulement si,* P_X *p.s.,*

$$\begin{cases} \eta(x) > 1/2 & \Rightarrow \quad f(x) = 1, \\ \eta(x) < 1/2 & \Rightarrow \quad f(x) = 0. \end{cases}$$

3. *Le risque de Bayes vaut*

$$R_P^* = E_X(\eta(X) \wedge (1 - \eta(X))).$$

4. *Si* f *est un classifieur,*

$$\ell_P(f, f^*) = E_X\left[|2\eta(X) - 1| \, \mathbb{1}_{f \neq f^*}\right].$$

En somme, les classifieurs de Bayes sont déterminés par la fonction de régression η : lorsque $\eta = 1/2$ on choisit le label que l'on veut, mais en dehors de cette valeur on choisit la classe majoritaire (0 si $P(Y = 0 \mid X) > P(Y = 1 \mid X)$, 1 si $P(Y = 0 \mid X) < P(Y = 1 \mid X)$. On peut abréger un peu en donnant une CNS d'optimalité de la forme

$$|2\eta(X) - 1| \mathbb{1}_{f(X) \neq \mathbb{1}_{\eta(X) \geq (1/2)}} = 0.$$

Enfin, la dernière relation sur l'excès de risque est fondamentale, et souligne le fait que les erreurs sont plus importantes sur les zones en X où le problème est le plus facile (c'est à dire $\eta(X)$ loin de 1/2).

Démonstration. Si f est un classifieur, on a

$$\begin{aligned} R_P(f) &= \mathbb{P}(f(X) \neq Y) = E_{(X,Y)}\left[\mathbb{P}(Y \neq f(X) \mid X)\right] \\ &= E_{(X,Y)}\left[\mathbb{1}_{f(X)=0}\eta(X) + \mathbb{1}_{f(X)=1}(1 - \eta(X))\right] \\ &\geq E_X(\eta(X) \wedge (1 - \eta(X)), \end{aligned}$$

avec égalité si et seulement si $\eta(x) > (1 - \eta(x)) \Rightarrow f(x) = 1$ et $\eta(x) < (1 - \eta(x)) \Rightarrow f(x) = 0$ P_X p.s., ce qui prouve les 3 premiers points.

Pour le dernier point, on peut écrire

$$R_P(f) = E_X(1 - \eta(X)) + E_X((2\eta(X) - 1)\mathbb{1}_{f(X)=0}),$$

et donc

$$\ell_P(f, f^*) = E_{(X,Y)}((2\eta(X) - 1)(\mathbb{1}_{f(X)=0} - \mathbb{1}_{f^*(X)=0}).$$

On conclut en regardant les signes : si $\eta(x) < 1/2$, alors $(\mathbb{1}_{f(x)=0} - \mathbb{1}_{f^*(x)=0}) \leq 0$ et $(2\eta(x) - 1)(\mathbb{1}_{f(x)=0} - \mathbb{1}_{f^*(x)=0}) = |2\eta(x) - 1|\mathbb{1}_{f(x) \neq f^*(x)}$. Si $\eta(x) > 1/2$, $(\mathbb{1}_{f(x)=0} - \mathbb{1}_{f^*(x)=0}) \geq 0$, et $(2\eta(x) - 1)(\mathbb{1}_{f(x)=0} - \mathbb{1}_{f^*(x)=0}) = |2\eta(x) - 1|\mathbb{1}_{f(x) \neq f^*(x)}$ là encore. □

En classification 0/1, tout se ramène donc à la fonction de régression $\eta(X)$. On peut alors formellement relier le problème de classification et celui de régression

comme suit : considérons un classifieur de la forme $f(x) = \mathbb{1}_{g(x)\geq 1/2}$, où moralement g va être un estimateur de la fonction de régression. On peut relier le risque en régression au risque en classification par

$$\begin{aligned}\ell_P(f, f^*) &= 2E_{(X,Y)}\left[|\eta(X) - 1/2|\, \mathbb{1}_{\mathbb{1}_{g(X)\geq 1/2} \neq \mathbb{1}_{\eta(X)\geq 1/2}}\right] \\ &\leq 2E_{(X,Y)}|g(X) - \eta(X)| \leq 2\sqrt{E_{(X,Y)}((g(X) - \eta(X))^2}.\end{aligned}$$

On verra par la suite que si ce genre de bornes permettent de prouver la consistance de classifieurs, elles mènent rarement à des vitesses optimales.

Autres pertes en classification

Un autre type de perte utilisée en classification binaire est donné par les pertes asymétriques, c'est à dire

$$c(y', y) = \omega_y \mathbb{1}_{y\neq y'},$$

où ω_0 et ω_1 sont des poids (positifs) correspondants respectivement à ce que l'on va payer en se trompant si la vraie classe est 0 ou 1. Ce choix de perte peut être approprié dans des situations où un type d'erreur est plus importante que l'autre (pour les problèmes de diagnostic médicaux par exemple, où il est plus grave de dire à un patient malade qu'il est sain que l'inverse), ou dans des cas où la proportion d'une classe est fortement plus petite que l'autre. Dans ce dernier cas c'est une manière de forcer notre règle de décision à regarder finement cette petite classe (enfin surtout en pratique, lorsque l'on regardera des règles basées sur échantillon).

Les risques et classifieurs de Bayes pour les pertes asymétriques s'expriment encore sans trop de difficulté.

Proposition 1.8 : Perte 0/1 asymétrique - Classifieur de Bayes

Pour la perte $0/1$ *asymétrique de poids* (ω_0, ω_1)*, en notant* $\eta(X) = \mathbb{P}(Y = 1 \mid X)$*, on a*

1. *Le classifieur* $f^*(X) = \mathbb{1}_{\eta(X)\geq \omega_0/(\omega_0+\omega_1)}$ *est de Bayes.*
2. *Un classifieur* f *est de Bayes si et seulement si,* P_X *p.s.,*

$$\begin{cases}\eta(x) > \omega_0/(\omega_0 + \omega_1) & \Rightarrow \quad f(x) = 1, \\ \eta(x) < \omega_0/(\omega_0 + \omega_1) & \Rightarrow \quad f(x) = 0.\end{cases}$$

3. *Le risque de Bayes vaut*

$$R_P^* = E_X(\omega_1\eta(X) \wedge \omega_0(1 - \eta(X))).$$

4. *Si* f *est un classifieur,*

$$\ell_P(f, f^*) = (\omega_0 + \omega_1)E_X\left[\left|\eta(X) - \frac{\omega_0}{\omega_0 + \omega_1}\right| \mathbb{1}_{f\neq f^*}\right].$$

On remarque qu'on retrouve les résultats pour la perte standard dès lors que $\omega_0 = \omega_1$.

Démonstration. Même preuve qu'avant : si f est un classifieur,

$$\begin{aligned} R_P(f) &= E_X\left[\omega_1 \mathbb{1}_{f(X)=0}\eta(X) + \omega_0 \mathbb{1}_{f(X)=1}(1-\eta(X))\right] \\ &\geq E_X\left[\omega_1\eta(X) \wedge \omega_0(1-\eta(X))\right], \end{aligned}$$

avec égalité si et seulement si $\omega_1\eta(x) > \omega_0(1-\eta(x)) \Rightarrow f(x) = 1$ et $\omega_1\eta(x) < \omega_0(1-\eta(x)) \Rightarrow f(x) = 0$, P_X p.s.. On en déduit encore les 3 premiers points.

Le dernier point part encore de

$$R_P(f) = E_X\left[((\omega_1+\omega_0)\eta(X) - \omega_0)\mathbb{1}_{f(X)=0}\right] + \omega_0 E_X(1-\eta(X)),$$

ce dont on déduit

$$\begin{aligned} \ell_P(f, f^*) &= E_X\left[((\omega_1+\omega_0)\eta(X) - \omega_0)(\mathbb{1}_{f(X)=0} - \mathbb{1}_{f^*(X)=0})\right] \\ &= (\omega_0+\omega_1)E_X\left[\left|\eta(X) - \frac{\omega_0}{\omega_0+\omega_1}\right| \mathbb{1}_{f(X)\neq f^*(X)}\right], \end{aligned}$$

en regardant le signe de $\left(\eta(X) - \frac{\omega_0}{\omega_0+\omega_1}\right)(\mathbb{1}_{f(X)=0} - \mathbb{1}_{f^*(X)=0})$. □

Remarque 1.9. Si on se restreint aux fonctions de pertes sur $\{0,1\}$ vérifiant $y = y' \Rightarrow c(y,y') = 0$, alors une telle fonction de perte sera totalement déterminée par $\omega_0 = c(0,1)$ et $\omega_1 = c(1,0)$, et on est forcément dans le cadre d'une perte 0/1 asymétrique.

Remarque 1.10 *: Lien avec les tests.* Si on note P_0 la loi de X sachant $Y = 0$, et P_1 celle de $X \mid Y = 1$, attribuer un label à un X qui arrive peut être vu comme un problème de test d'hypothèses

$$\begin{cases} H_0 & X \sim P_0 \\ H_1 & X \sim P_1, \end{cases}$$

où moralement $X \sim P_j$ signifie que l'on a tiré X suivant P_j. En prenant μ une mesure dominante (par exemple $P_0 + P_1$), le test du rapport de vraisemblance s'écrit

$$T(X) = \mathbb{1}_{\frac{g_1}{g_0}(X) \geq t_\alpha},$$

où t_α est un seuil calibré pour atteindre une erreur de première espèce souhaitée. Avec la formule de Bayes, on peut exprimer $\eta(x)$ comme

$$\begin{aligned} \eta(x) &= \frac{pg_1(x)}{pg_1(x) + (1-p)g_0(x)} \\ &= \psi_p\left(\frac{g_1}{g_0}(x)\right), \end{aligned}$$

où $p = \mathbb{P}(Y=1)$ et $\psi_p(u) = pu/(pu + (1-p))$. ψ_p étant croissante sur $[0, +\infty[$,

$$T(x) = 1 \Leftrightarrow \eta(x) \geq \psi_p(t_\alpha),$$

et donc, pour $\omega_{0,\alpha}, \omega_{1,\alpha}$ tels que $\psi_p(t_\alpha) = \frac{\omega_{0,\alpha}}{\omega_{0,\alpha}+\omega_{1,\alpha}}$, T est un classifieur de Bayes pour la perte asymétrique associée. On retrouve l'heuristique de 'coût asymétrique d'une erreur' : pour des α petits, on veut le moins possible rejeter à tort, ce qui conduit à un t_α plus grand et donc un $\omega_{0,\alpha}$ plus grand (le prix à payer lorsque qu'on dit 1 alors que la réalité est 0, c'est à dire rejeter à tort).

On pourrait aussi montrer l'équivalence entre classification avec perte asymétrique et test bayésien.

1.3 Apprentissage proprement dit

Dans les deux sections précédentes, on a vu comment bâtir des prédicteurs optimaux (après avoir défini ce pouvait être l'optimalité en prédiction) **à partir de la connaissance de la loi** $P_{(X,Y)}$. Dans la "vraie vie", on a pas accès à $P_{(X,Y)}$, mais éventuellement à un échantillon

$$D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$$

tiré suivant cette loi inconnue (que l'on supposera la plupart du temps i.i.d.). On appelle usuellement D_n l'échantillon d'entraînement (ou train). Le but est alors de construire des prédicteurs basés sur D_n, qui ressemblent le plus possible à des prédicteurs optimaux.

Definition 1.11 : Règle de prédiction - Prédicteur

Une **règle de prédiction** est une fonction

$$f : \begin{cases} \left(\bigcup_{n=1}^{+\infty} (\mathcal{X} \times \mathcal{Y})^n\right) \times \mathcal{X} & \to & \mathcal{Y} \\ (D_n, x) & \mapsto & f(D_n, x). \end{cases}$$

Une règle de prédiction est donc un prédicteur aléatoire en l'échantillon D_n. Par convention un peu abusive, on confond souvent prédicteur et règle de prédiction (donc à partir de maintenant un prédicteur sera entendu comme "prédicteur aléatoire dépendant d'un échantillon"), et pour rappeler cette dépendance en l'échantillon on utilisera la notation $\hat{f}$ ou $\hat{f}_n$ (le n rappelle la taille de l'échantillon), et

$$\hat{f}_n(x) = f(D_n, x).$$

Definition 1.12 : Risque d'un prédicteur

Le risque d'un prédicteur $\hat{f}$ est la quantité **aléatoire**

$$R_P(\hat{f}) = \mathbb{E}\left[c(f(D_n, X), Y) \mid D_n\right] = E_{(X,Y)} c(f(D_n, X), Y) = E_{(X,Y)} c(\hat{f}(X), Y),$$

où on a résumé toutes les conventions : $E_{(X,Y)}$ correspond à l'intégration par rapport à la loi de (X, Y) sachant tout le reste, donc une espérance conditionnelle sachant D_n.

Le risque (tout court) d'un prédicteur est aussi appelé *erreur de généralisation* : cela correspond à ce que vous payez en moyenne sur de nouvelles données (X, Y), l'échantillon d'entraînement D_n étant fixé.

1.3.1 Consistance, consistance uniforme et vitesses d'apprentissage

Pour une loi $P_{(X,Y)}$ donnée mais cachée, le but va être de construire des prédicteurs $\hat{f}$ basés sur D_n qui approchent de mieux en mieux f_P^* lorsque n grandit (moralement lorsqu'on a de plus en plus d'information sur P via D_n). La propriété

minimale attendue d'un prédicteur va être la convergence de $R_P(\hat{f})$ vers R_P^*, pour tout P, en proba.

Definition 1.13 : Consistance

Le prédicteur $\hat{f}$ est consistant si, pour toute loi P sur $\mathcal{X} \times \mathcal{Y}$,

$$R_P(\hat{f}) \xrightarrow[n\to+\infty]{\mathbb{P}} R_P^*.$$

Il s'agit d'une convergence à P fixé, pour tout P, une convergence simple en somme.

De manière générale, pour avoir la consistance, il faut a minima que le prédicteur prenne en compte D_n.

Exemple 1.14. Supposons que $n \geq 1$ $\hat{f}_n = f$, où f est un classifieur ne dépendant pas des données. Soit $x \in \mathcal{X}$, et supposons que $f(x) = 1$ (ou 0, peu importe). On prend alors pour loi P la loi $\delta_x \times \delta_0$, pour laquelle $R_P(f) = 1$ et $R_P(f_P^*) = 0$.

Regardons cette-fois ci un exemple qui marche à peu près : dans le cas où $\mathcal{X}$ est discret un classifieur sensé consiste à suivre la majorité.

Proposition 1.15

Si $\mathcal{X} = \mathbb{N}$ et $\mathcal{Y} = \{0,1\}$, on définit $\hat{f}_{maj}(D_n, x)$ par

$$\hat{f}_{maj}(D_n, x) = \begin{cases} 1 & si \quad N_1(x) \geq N_0(x) \\ 0 & si \quad N_1(x) < N_0(x) \end{cases},$$

avec $N_j(x) = \sum_{i=1}^n \mathbb{1}_{X_i=x}\mathbb{1}_{Y_i=j} = |\{i \mid (X_i, Y_i) = (x, j)\}|$, pour $j \in \{0,1\}$. Alors $\hat{f}_{maj}$ est consistant.

Démonstration. Soit P une loi sur $\mathcal{X} \times \{0,1\}$, décrite par $p(k) = \mathbb{P}(X = k)$ et $\eta(k) = \mathbb{P}(Y = 1|X = k)$. On note $S = \{k|p(k) \neq 0\} \cap \{k|\eta(k) \neq \frac{1}{2}\}$. D'après le théorème sur l'excès de risque, on a

$$\begin{aligned} R(\hat{f}_{maj}) - R(f^*) &= \sum_{k\in S} p(k)|2\eta(k) - 1|\mathbb{1}_{\hat{f}_{maj}(k)\neq f^*(k)} \\ &\leq \sum_{k\in S} p(k)\mathbb{1}_{\hat{f}_{maj}(k)\neq f^*(k)}. \end{aligned}$$

Soit maintenant $\varepsilon > 0$, et S^ε un ensemble fini tel que $\sum_{k \notin S^\varepsilon} p(k) < \varepsilon/2$. On a

$$\begin{aligned}
&\mathbb{P}(R(\hat{f}_{maj}) - R(f^*) \geq \varepsilon) \\
&\leq \mathbb{P}\left(\sum_{k \in S} p(k) \mathbb{1}_{\hat{f}_{maj}(k) \neq f^*(k)} \geq \varepsilon\right) \\
&\leq \mathbb{P}\left(\max\left\{\sum_{k \in S \cap S^\varepsilon} p(k) \mathbb{1}_{\hat{f}_{maj}(k) \neq f^*(k)}, \sum_{k \in S \setminus S^\varepsilon} p(k) \mathbb{1}_{\hat{f}_{maj}(k) \neq f^*(k)}\right\} \geq \varepsilon/2\right) \\
&= \mathbb{P}\left(\sum_{k \in S \cap S^\varepsilon} p(k) \mathbb{1}_{\hat{f}_{maj}(k) \neq f^*(k)} \geq \varepsilon/2\right) \\
&\leq \mathbb{P}\left(\sum_{k \in S \cap S^\varepsilon} \mathbb{1}_{\hat{f}_{maj}(k) \neq f^*(k)} \neq 0\right) \\
&\leq \sum_{k \in S \cap S^\varepsilon} \mathbb{P}(\hat{f}_{maj}(k) \neq f^*(k)).
\end{aligned}$$

De plus, pour tout $k \in S \cap S^\varepsilon$, la loi des grands nombres donne

$$\frac{N_1(k) - N_0(k)}{n} \xrightarrow{\mathbb{P}} p(k)(2\eta(k) - 1) \neq 0,$$

donc $\mathbb{P}(\hat{f}_{maj}(k) \neq f^*(k)) \to 0$. On a donc, par sommation (finie) sur $S \cap S^\varepsilon$,

$$\mathbb{P}(R(\hat{f}_{maj}) - R(f^*) \geq \varepsilon) \leq \sum_{k \in S \cap S^\varepsilon} \mathbb{P}(\hat{f}_{maj}(k) \neq f^*(k)) \to 0,$$

□

On peut donc décemment espérer pouvoir construire des classifieurs consistants. On remarque ici que $\hat{f}_{maj}$ est une règle plug-in, c'est à dire de la forme

$$\hat{f}_{maj}(x) = \mathbb{1}_{\hat{\eta}(x) \geq (1/2)},$$

où ici $\hat{\eta}(x) = \frac{N_1(x)}{N_0(x)+N_1(x)}$. La preuve de la consistance est en fait basée sur la convergence de $\hat{\eta}$ vers η en probabilité en utilisant la loi des grands nombres. De manière générale, la consistance des classifieurs de type plug-in se prouve de cette manière (on vérifiera cela pour les méthodes basées sur du moyennage local), mais pour ce qui est des vitesses de convergence c'est rarement la manière optimale de faire.

La définition de consistance peut être vue comme une convergence ponctuelle sur l'ensemble des mesures de probabilités sur $\mathcal{X} \times \mathcal{Y}$. On peut demander plus que cette convergence ponctuelle en demandant une convergence uniforme sur l'ensemble des lois.

Definition 1.16 : Consistance uniforme

Un prédicteur $(\hat{f}_n)_{n \geq 1}$ est dit **uniformément** consistant si

$$\lim_{n \to \infty} \sup_P E_{D_n \sim P^{\otimes n}} \left(R_P(\hat{f}_n) - R_P(f_P^*)\right) = 0.$$

Un prédicteur uniformément consistant est donc ponctuellement consistant, et on demande en plus que la convergence de l'excès de risque vers 0 soit uniforme pour

l'ensemble des mesures de probabilités possibles sur $\mathcal{X} \times \mathcal{Y}$. On va voir que les cas où l'uniforme consistance est possible sont relativement peu nombreux.

PROPOSITION 1.17

Dans le cas où $\mathcal{X}$ est un ensemble fini, le classifieur par majorité $\hat{f}_{maj}$ est uniformément consistant.

Démonstration. Le classifieur par majorité dans ce cas étant du type ERM, on prouvera au prochain chapitre que

$$\sup_P E_{D_n \sim P^{\otimes n}} \left(R_P(\hat{f}_n) - R_P(f_P^*) \right) \leq 2\sqrt{\frac{2(|\mathcal{X}|+1)\log(2)}{n}}.$$

□

Malgré ce premier résultat encourageant, on peut montrer que demander l'uniforme consistance pour le problème de classification dans d'autres cas que l'exemple précédent est voué à l'échec. C'est l'objet du fameux *No Free Lunch Theorem*.

THÉORÈME 1.18 : NO FREE LUNCH

Si $|\mathcal{X}| = \infty$, alors aucun classifieur ne peut être uniformément consistant.

Démonstration. On se donne f un classifieur (en gardant en tête qu'il dépend de D_n), et le but va être de trouver un ensemble de lois sur $\mathcal{X} \times \{0,1\}$ bien choisi sur lequel l'excès de risque va être minoré par quelque chose ne tendant pas vers 0. De manière générale, minorer un excès de risque se fait avec des techniques bayésiennes ou assimilées (cf le cours de statistiques non asymptotiques).

Sans perdre en généralité, prenons $\mathcal{X} = \mathbb{N}$, et $K \in \mathbb{N}$ (grand). Pour $r \in \{0,1\}^K$, on définit la loi P_r par $P_r(j \times r_j) = \frac{1}{K}$, pour $j \in \{1, \dots, K\}$. En d'autres termes, P_r charge uniquement les entiers entre 1 et K, et à valeur de X connue le label est certain (0 ou 1). On a alors évidemment $f^*_{P_r}(j) = r_j$ pour $j \leq K$ et $R_{P_r}(f^*_{P_r}) = 0$ (classification parfaite). Le défaut d'apprentissage va venir du fait que, à n fixé, on ne peut pas explorer suffisamment $\{1, \dots, K\}$ pour trouver la règle optimale de classification.

De manière immédiate on a

$$\sup_P E_{D_n \sim P^{\otimes n}} \left(R_P(f) - R_P(f_P^*) \right) \geq \sup_{r \in \{0,1\}^K} E_{D_n \sim P_r^{\otimes n}} R_{P_r}(f).$$

Le bayésien intervient ici : on se donne R de loi uniforme sur $\{0,1\}^K$, et on suppose que l'on observe $\tilde{D}_n, (\tilde{X}, \tilde{Y})$ de loi $P_r^{\otimes(n+1)}$ conditionnellement à $R = r$. On peut alors écrire

$$\sup_{r \in \{0,1\}^K} E_{D_n \sim P_r^{\otimes n}} R_{P_r}(f) \geq E_{R \sim \mathcal{U}(\{0,1\}^K} E_{D_n \sim P_R^{\otimes n}} R_{P_R}(f) = \mathbb{P}(f(\tilde{D}_n, \tilde{X}) \neq \tilde{Y}).$$

On remarque que les R_j sont indépendants. On peut alors brutalement minorer le risque "intégré" (c'est en fait un risque Bayésien au sens de la statistique Bayésienne)

via

$$\begin{aligned}\mathbb{P}(f(\tilde{D}_n,\tilde{X})\neq\tilde{Y}) &= \mathbb{P}(f(\tilde{D}_n,\tilde{X})\neq R_{\tilde{X}})\\ &= \mathbb{E}\left[\mathbb{E}\left(\mathbb{1}_{f((\tilde{X}_i,R_{\tilde{X}_i})_i,\tilde{X})\neq R_{\tilde{X}}} \mid \tilde{X},(\tilde{X}_i,R_{\tilde{X}_i})_i\right)\right]\\ &\geq \mathbb{E}\left[\mathbb{E}\left(\mathbb{1}_{\tilde{X}\notin\{\tilde{X}_1,\dots,\tilde{X}_n\}}\mathbb{1}_{f((\tilde{X}_i,R_{\tilde{X}_i})_i,\tilde{X})\neq R_{\tilde{X}}} \mid \tilde{X},(\tilde{X}_i,R_{\tilde{X}_i})_i\right)\right].\end{aligned}$$

Or, on a

$$R_{\tilde{X}}|\left(\tilde{X},(\tilde{X}_i,R_{\tilde{X}_i})_i\right)=\sum_{i=1}^{n}\mathbb{1}_{\tilde{X}_i=\tilde{X}}\delta_{R_{\tilde{X}_i}}+\mathbb{1}_{\tilde{X}\notin\{\tilde{X}_1,\dots,\tilde{X}_n\}}\mathcal{B}(\frac{1}{2}),$$

ce qui signifie que si $\tilde{X}$ a été observé dans l'échantillon $\tilde{X}_1,\dots,\tilde{X}_n$, alors on connait $R_{\tilde{X}}$ (déterministe si on connaît les $(\tilde{X}_i,R_{\tilde{X}_i})_i$), et que dans le cas contraire on n' a aucune information sur $R_{\tilde{X}}$ via les données et $\tilde{X}$, c'est à dire $R_{\tilde{X}} \mid [(\tilde{X}_i,R_{\tilde{X}_i})_i),\tilde{X}] \sim R_{\tilde{X}} \mid \tilde{X}\sim\mathcal{B}(\frac{1}{2})$.

Donc

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left(\mathbb{1}_{\tilde{X}\notin\{\tilde{X}_1,\dots,\tilde{X}_n\}}\mathbb{1}_{f((\tilde{X}_i,R_{\tilde{X}_i})_i,\tilde{X})\neq R_{\tilde{X}}} \mid \tilde{X},(\tilde{X}_i,R_{\tilde{X}_i})_i\right)\right] &= \frac{1}{2}\mathbb{P}\left(\tilde{X}\notin\{\tilde{X}_1,\dots,\tilde{X}_n\}\right)\\ &= \frac{1}{2}\mathbb{E}\left[\mathbb{E}\left(\mathbb{1}_{\tilde{X}\notin\{\tilde{X}_1,\dots,\tilde{X}_n\}}|(R,\tilde{D}_n)\right)\right]\\ &= \frac{1}{2}\left(1-\frac{1}{K}\right)^n\\ &\xrightarrow[K\to\infty]{}\frac{1}{2}.\end{aligned}$$

□

On a le pendant immédiat en régression, en utilisant l'inégalité

$$\ell_P(f,f^*)\leq 2\sqrt{E_{X,Y)}((g(X)-\eta(X))^2}$$

qui moralement stipule que le problème de régression (au sens quadratique) est plus difficile que le problème de classification.

Il y a deux enseignements à tirer de ce résultat. Premièrement, dans la majorité des cas pratiques, il est déraisonnable de demander à ce qu'un prédicteur soit uniformément consistant. Deuxièmement, pour minorer des supremum d'excès de risque sur un ensemble de loi, on passe généralement par une astuce de type "supremum sur un ensemble de lois" superieur à "moyenne sur un ensemble de lois", qui est dans l'esprit de la statistique Bayésienne et en pratique l'objet de techniques de type "bornes inférieures" (cf le cours de Statistique non asymptotique).

Ce résultat négatif ne doit pas être considéré comme décourageant pour autant : on ne peut certes pas espérer avoir un prédicteur uniformément bon sur toutes les lois (d'ailleurs si c'était le cas de nombreux statisticiens se retrouveraient au chômage), mais on peut tout du moins espérer trouver, pour une certaine classe de problèmes (c'est-à-dire de lois sur $\mathcal{X}\times\mathcal{Y}$) des prédicteurs performants uniformément sur cette sous-classe. C'est d'ailleurs un des enjeux cruciaux de la statistique moderne : trouver, pour une classe de problèmes, des prédicteurs uniformément bons, voire optimaux. On peut essayer de formaliser ceci par la notion de vitesse d'apprentissage (sur une classe).

Soit $\mathcal{P}$ une classe de problèmes (c'est-à-dire un sous-ensemble des mesures de probabilité sur $\mathcal{X} \times \mathcal{Y}$. On définit la vitesse d'apprentissage de la classe $\mathcal{P}$ comme suit.

DEFINITION 1.19

Soit $\mathcal{P} \subset \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, et $(a_n)_n$ une suite positive. (a_n) est appelée vitesse d'apprentissage de la classe $\mathcal{P}$ si

$$\inf_{\hat{f}_n} \sup_{P \in \mathcal{P}} E_{D_n \sim P^{\otimes n}} \left(R_P(\hat{f}_n) - R_P(f_P^*) \right) \underset{n \to \infty}{=} O(a_n),$$

et

$$a_n \underset{n \to \infty}{=} O\left(\inf_{\hat{f}_n} \sup_{P \in \mathcal{P}} E_{D_n \sim P^{\otimes n}} \left(R_P(\hat{f}_n) - R_P(f_P^*) \right) \right).$$

Cette notion de vitesse d'apprentissage revient à étudier la dépendance en n du *risque minimax*

$$\sup_{P \in \mathcal{P}} E_{D_n \sim P^{\otimes n}} \left(R_P(\hat{f}_n) - R_P(f_P^*) \right)$$

sur la classe $\mathcal{P}$. D'autres paramètres de cette vitesse minimax peuvent être intéressant, notamment la dimension d de l'espace $\mathcal{X}$ par exemple. De manière générale, on essaye d'étudier la dépendance en des paramètres $p_1, \dots, p_k$ de ces vitesses minimax en trouvant des encadrements de type

$$f^-(n, p_1, \dots, p_k) \leq \sup_{P \in \mathcal{P}} E_{D_n \sim P^{\otimes n}} \left(R_P(\hat{f}_n) - R_P(f_P^*) \right) \leq f^+(n, p_1, \dots, p_k).$$

Vous verrez cela plus en détail dans le cours de statistique non-asymptotique, mais dans les grandes lignes, les minorations du risque minimax se font à renfort de techniques Bayésiennes au sens large (comme dans le No-free Lunch Theorem), et les majorations s'obtiennent en trouvant des bornes sur l'excès de risque d'un classifieur approprié (via des inégalités de type oracle dont on reparlera dans la section suivante).

On peut traduire le No free lunch theorem en termes de vitesse d'apprentissage comme suit :

PROPOSITION 1.20

Si $\mathcal{P} = \mathcal{M}_1((\mathcal{X} \times \{0,1\})$ (l'ensemble des mesures de probabilités sur $\mathcal{X} \times \{0,1\}$), et $\mathcal{X}$ est infini, alors $a_n = 1$ est une vitesse d'apprentissage de la classe $\mathcal{P}$.

Démonstration. Le No free lunch théorème donne

$$1 \underset{n \to \infty}{=} O\left(\inf_{\hat{f}_n} \sup_{P \in \mathcal{P}} E_{D_n \sim P^{\otimes n}} \left(R_P(\hat{f}_n) - R_P(f_P^*) \right) \right).$$

Il ne reste qu'à vérifier l'autre sens. De manière générale, trouver une vitesse dans ce sens revient à choisir un bon classifieur et étudier son excès de risque uniformément sur la classe choisie. Ici un classifieur stupide suffit : prenons $f = \mathcal{B}(\frac{1}{2})$, c'est-à-dire un classifieur qui tire à pile ou face. On à, pour tout $n \geq 1$, pour tout $P \in \mathcal{P}$, $R_P(f) = \frac{1}{2}$ (même pas aléatoire donc), et de manière évidente

$$\sup_{P \in \mathcal{P}} E_{D_n \sim P^{\otimes n}} \left(R_P(f) - R_P(f_P^*) \right) \leq \sup_{P \in \mathcal{P}} E_{D_n \sim P^{\otimes n}} R_P(f) \leq \frac{1}{2} \underset{n \to \infty}{=} O(1).$$

□

En d'autres termes, si $\mathcal{X}$ n'est pas fini, la vitesse d'apprentissage pour le problème de classification sur l'ensemble des lois sur $\mathcal{X} \times \{0,1\}$ est de l'ordre de $1/2$, qui est la vitesse d'apprentissage de l'estimateur qui tire à pile ou face. Donc, si vous espérez trouver un estimateur universellement bon dans ce cas, autant tirer à pile ou face cela ira plus vite.

Dans le cas où $\mathcal{X}$ est fini, on peut aussi donner une vitesse d'apprentissage sur $\mathcal{M}_1((\mathcal{X} \times \{0,1\})$.

Proposition 1.21

Si $|\mathcal{X}| < \infty$, et $\mathcal{P} = \mathcal{M}_1((\mathcal{X} \times \{0,1\})$, alors

$$a_n = \frac{1}{\sqrt{n}}$$

est une vitesse d'apprentissage pour la classe $\mathcal{P}$.

Démonstration. Pour la borne sup, attendre le chapitre ERM. Pour la borne inf, cf le cours de non asymptotique. □

La vitesse en $1/\sqrt{n}$ est typique des vitesses paramétriques, on la retrouvera pour les minimiseurs de risque empirique sur certaines classes de lois (moralement celles dépendant d'un nombre fini de paramètres réels, d'où le nom).

Établir des vitesses pour les problèmes de classification et régression sur certaines classes de loi n'est pas un objectif de ce cours. On donnera quelques exemples de vitesses pour des problèmes classiques, qui nous serviront à attester de la qualité de prédicteurs.

1.3.2 Attester de la qualité d'un prédicteur

Il s'agit de répondre à la situation pratique suivante : vous avez inventé un prédicteur $\hat{f}$ pour répondre à un certain type de problème, comment pouvez-vous attester que votre idée est bonne ?

Point de vue théorique

Attester théoriquement de la qualité de votre prédicteur préféré se fait généralement en 2/3 temps :

1. Avoir une idée de la classe de lois $\mathcal{P} \subset \mathcal{M}_1((\mathcal{X} \times \{\mathcal{Y}\})$ sur laquelle votre idée a des chances de marcher.
2. Etablir une inégalité de type **oracle**, c'est à dire de type

$$\forall P \in \mathcal{P} \quad \mathbb{P}\left(R_P(\hat{f}) - R_P^* \geq C(P,n) + \varepsilon\right) \leq g(\varepsilon),$$

avec idéalement $g(\varepsilon) \to 0$ lorsque ε grandit pour une borne en déviation, ou de type

$$\forall P \in \mathcal{P} \quad E_{D_n}(R_P(\hat{f}) - R_P^*) \leq v(P,n)$$

pour une borne en espérance.

3. Comparer votre inégalité oracle avec les vitesses d'apprentissage pour votre problème (si elles sont établies, sinon il faut établir aussi la borne inférieure). Il existe aussi des vitesses pour les déviations (cruciales dans un cadre de prédiction robuste par exemple), dans ce cours on regardera les vitesses en espérance introduites plus haut. Il s'agit maintenant de comparer $\sup_{P\in\mathcal{P}} v(P,n)$ avec a_n la vitesse d'apprentissage de votre classe. Si elles sont du même ordre de grandeur, c'est gagné (vous pouvez comparer les dépendances en n, mais aussi en d'autres paramètres).

Bien sûr vous pouvez aussi prendre en compte le temps de calcul de votre prédicteur, d'autres avantages comme de la robustesse, etc. . Le canevas proposé n'est qu'une des manières de regarder théoriquement les performances d'un prédicteur, qui permet de faire le lien avec votre cours de non-asymptotique.

Exemple 1.22 *: Classification à seuil.*

On pose $\mathcal{X} = [0,1[$, $\mathcal{Y} = \{0,1\}$, et on considère $\mathcal{P}$ l'ensemble des lois de variables (X,Y) sur $\mathcal{X}\times\mathcal{Y}$ vérifiant

1. X a une densité g comprise entre 10^{-1} et 10,
2. $\eta(x)$ est de la forme $h_{t_0}(x) = (3/4)\mathbb{1}_{x\geq t_0} + (1/4)\mathbb{1}_{x<t_0}$, pour un $t_0 \in [0,1]$.

Cet ensemble de lois est paramétré par deux choses : la densité g de X et le paramètre t_0 qui détermine entièrement la loi de $Y \mid X$. Bien que cet espace soit de dimension infinie (du fait de la densité g de X), on peut montrer que la vitesse d'apprentissage pour cette classe est $1/\sqrt{n}$ (en fait seul l'apprentissage du paramètre t_0 compte pour le problème de classification, d'où cette vitesse paramétrique).

Pour ce problème, $f_{t_0} = \mathbb{1}_{t\geq t_0}$ est évidemment un classifieur de Bayes, et il semble pertinent d'aller chercher un classifieur de type $f_{\hat{t}} = \mathbb{1}_{t\geq\hat{t}}$, où $\hat{t}$ va être un seuil construit à partir des données $D_n = \{(X_1,Y_1),\dots,(X_n,Y_n)\}$ i.i.d de loi $P\in\mathcal{P}$. On va donc chercher empiriquement un endroit de saut dans le comportement des Y_i. Pour ce faire, on va découper $[0,1[$ en cases $(I_k)_{k=1,\dots,K}$ définies par

$$I_k = \left[\frac{k-1}{K}, \frac{k}{K}\right[,$$

où K est un entier que l'on va calibrer, regarder le vote majoritaire sur les cases, défini par

$$\hat{u}_k = \mathbb{1}_{\sum_{i=1}^n \mathbb{1}_{X_i\in I_k} Y_i \geq \sum_{i=1}^n \mathbb{1}_{X_i\in I_k}(1-Y_i)},$$

et déterminer un estimateur du temps de saut via

$$\hat{j} = \min\left\{j \in [\![1, K-1]\!] \mid \hat{u}_j = 0 \text{ et } \hat{u}_{j+1} = 1\right\}.$$

On pose alors $\hat{t} = \hat{j}/K$, et

$$\hat{f} = f_{\hat{t}}.$$

On commence par montrer que $\hat{t}$ est proche de t_0. Pour cela, on note $j^* = \lfloor Kt_0 + 1\rfloor$, et on se donne $j \leq j^* - 2$. On a alors $I_{j+1} \subset [0, t_0[$, et donc

$$\begin{aligned} E_{(X,Y)}(2Y-1)\mathbb{1}_{X\in I_j} &= E_{(X,Y)}\mathbb{1}_{X\in I_j}(2\eta(X)-1) \\ &\leq -\frac{1}{20K}. \end{aligned}$$

On a alors, en utilisant l'inégalité de Hoeffding

$$\begin{aligned}\mathbb{P}(\hat{u}_{j+1}=1) &= \mathbb{P}\left(\sum_{i=1}^{n}\mathbb{1}_{X_i\in I_k}(2Y_i-1)\geq 0\right)\\ &\leq \mathbb{P}\left(\sum_{i=1}^{n}[\mathbb{1}_{X_i\in I_k}(2Y_i-1)-E_{(X,Y)}(2Y-1)\mathbb{1}_{X\in I_j}\geq \frac{1}{20K}\right)\\ &\leq e^{-\frac{2n}{(20K)^2\times 2^2}}=e^{-\frac{n}{800K^2}},\end{aligned}$$

ce dont on déduit, en utilisant une borne d'union

$$\begin{aligned}\mathbb{P}(\hat{j}\leq j^*-2) &\leq \sum_{j=1}^{j^*-1}\mathbb{P}(\hat{u}_j=1)\\ &\leq (j^*-1)e^{-\frac{n}{800K^2}}.\end{aligned}$$

Dans l'autre sens, pour $j\geq j^*+1$, on peut montrer que

$$E_{(X,Y)}(2Y-1)\mathbb{1}_{X\in I_j}\geq \frac{1}{20K},$$

et le raisonnement est plus simple

$$\begin{aligned}\mathbb{P}(\hat{j}\geq j^*+1) &\leq \mathbb{P}(\hat{u}_{j^*+1}=0)\\ &\leq e^{-\frac{n}{800K^2}}.\end{aligned}$$

On en déduit

$$\mathbb{P}\left(\left|\hat{t}-t_0\right|>\frac{1}{K}\right)\leq Ke^{-\frac{n}{800K^2}}.$$

Relions maintenant proximité en t et excès de risque : pour un t quelconque, on a

$$\begin{aligned}R(f_t)-R_(f_{t_0}) &= E_{(X,Y)}(|2\eta(X)-1|\mathbb{1}_{f_t(X)\neq f_{t_0}(X)}\\ &= \frac{1}{2}P_X(f_t(X)\neq f_{t_0}(X)\\ &= \frac{1}{2}\mathbb{P}\left(X\in[t\wedge t_0, t\vee t_0[\right)\leq 5|t-t_0|.\end{aligned}$$

On en déduit directement

$$\mathbb{P}\left(R_P(\hat{f})-R_P^*>\frac{5}{K}\right)\leq Ke^{-\frac{n}{800K^2}},$$

et en choissisant K en $C\frac{\sqrt{n}}{\sqrt{\log(n)}}$ on obtient

$$\begin{aligned}E_{D_n}(R_P(\hat{f})-R_P^*) &= E_{D_n}((R_P(\hat{f})-R_P^*)\mathbb{1}_{|\hat{t}-t|\leq K^{-1}})+E_{D_n}((R_P(\hat{f})-R_P^*)\mathbb{1}_{|\hat{t}-t|>K^{-1}})\\ &\leq C\frac{\sqrt{\log(n)}}{\sqrt{n}},\end{aligned}$$

où C est une constante absolue. Au facteur $\log(n)$ près on récupère la vitesse d'apprentissage sur cette classe (on peut enlever ce facteur en utilisant des techniques plus évoluées).

Point de vue pratique 1 : risque empirique et surapprentissage

Le point de vue théorique décrit plus haut permet de faire face aux situations simples où d'une part on connait à peu près le type de problème auquel on fait face (classe P) et pour lesquelles on peut établir des bornes supérieurs et inférieures relativement précises (même au niveau des constantes). En pratique, si on ne peut pas décemment s'imaginer être dans une telle situation ou si les bornes sont peu informatives (à n fixé vous pouvez avoir des constantes énormes qui donnent une borne peu intéressante pour l'excès de risque), il existe d'autres manières de se faire une idée de la performance de votre prédicteur.

L'objet de base en pratique est l'erreur moyenne observée **sur les données**, c'est à dire le **risque empirique**

$$R_n(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} c(f(X_i), Y_i)) = P_n(dx, dy)c(f(x), y),$$

où P_n désigne la **mesure empirique** $\frac{1}{n}\sum_{i=1}^{n}\delta_{(X_i,Y_i)}$.

La question centrale devient : à quel point ce risque empirique est un bon témoin de la performance du prédicteur ? Regardons sur un exemple simple.

Exemple 1.23 *: Classification linéaire.* Un suppose $\mathcal{X} = \mathbb{R}^D$, $\mathcal{Y} = \{0, 1\}$, et que la décision optimale est de la forme

$$f^*(x) = \mathbb{1}_{\langle\theta^*, x\rangle \geq 0},$$

pour un risque optimal $R^* = R_P(f^*) = 1/4$. Il semble alors sensé de chercher un prédicteur de la forme

$$\hat{f}(x) = f_{\hat{\theta}}(x) = \mathbb{1}_{\langle\hat{\theta}, x\rangle \geq 0}.$$

Pour un θ fixé, le risque empirique est une approximation pas trop mauvaise du vrai risque : l'inégalité de Hoeffding fourniti

$$\mathbb{P}\left(R_n(f_\theta) \leq R(f_\theta) - \sqrt{\frac{x}{2n}}\right) \leq e^{-x},$$

ce qui atteste que $R_n(f_\theta)$ ne sous-estime pas trop le vrai risque $R(f_\theta)$.

En revanche, pour un θ construit à partir des données il peut en aller tout autrement : si on prend $\hat{\theta}$ qui minimise le rique empirique, et que l'on suppose $D \geq n$, pour peu que $X_1, \ldots, X_n$ forme une famille libre de $\mathbb{R}^D$ on aura $R_n(\hat{f}) = 0$ (FAIRE DESSIN).

Si X a une densité par rapport à $\mathcal{L}_D$, $X_1, \ldots, X_n$ sera libre avec proba 1, ce dont on peut déduire

$$\mathbb{P}\left(R_n(f_{\hat{\theta}}) \leq R_P(f_{\hat{\theta}}) - 1/4\right) \geq \mathbb{P}\left(R_n(f_{\hat{\theta}}) = 0\right) = 1.$$

Le risque empirique s'avère alors être un très mauvais témoin des performances du prédicteur $f_{\hat{\theta}}$. Cela vient du fait que l'on a utilisé le même échantillon en *train* et en *test*, c'est à dire qu'on a utilisé les même données pour construire $\hat{\theta}$ et pour le "valider", ce qui conduit forcément à des biais négatifs. Par exemple ici,

$$E_{D_n} R_n(f_{\hat{\theta}}) = 0 \leq E_{D_n} R(f_{\hat{\theta}}) - 1/4,$$

autrement dit le risque empirique est un estimateur fortement biaisé du vrai risque, pour un θ construit à partir des données via minimisation du risque empirique.

L'exemple ci-dessus devrait suffire à convaincre que le risque empirique seul ne doit pas être utilisé comme critère de performance sans réfléchir. On peut néanmoins l'utiliser conjointement avec un peu de théorie dans certains cas : il s'agit d'aller regarder théoriquement à quel point R_n sous-estime R_P, pour un prédicteur $\hat{f}$ donné. Typiquement, il s'agit d'obtenir des bornes du genre

$$\mathbb{P}\left(R_P(\hat{f}) \geq R_n(\hat{f}) + C(P,n) + x\right) \leq g(x),$$

où g est une fonction décroissante. Une manière brutale mais assez révélatrice en première approche est de regarder le supremum des déviations sur la classe de prédicteur considérés, c'est à dire

$$\Delta_n(\mathcal{F}) = \sup_{f \in \mathcal{F}} R_P(f) - R_n(f),$$

et de majorer brutalement $R_P(\hat{f})$ par

$$R_n(\hat{f}) + \Delta_n(\mathcal{F}).$$

Dès lors, à la pénalité $\Delta_n(\mathcal{F})$ près, le risque empirique de votre prédicteur peut être utilisé comme garantie.

L'ordre de grandeur à avoir en tête (on y reviendra dans le chapitre suivant) est

$$E_{D_n}\Delta_n(\mathcal{F}) \lesssim \frac{\sqrt{\dim(\mathcal{F})}}{\sqrt{n}},$$

où la dimension de $\mathcal{F}$ peut s'exprimer de plusieurs manières (une VC dimension en classification par exemple). Dans le cas des classifieurs linéaires, on a $\dim\left(\mathbb{1}_{\langle\theta,.\rangle\geq 0} \mid \theta \in \mathbb{R}^D\right) = D$, et la borne sur le risque du minimiseur de R_n devient

$$E_{D_n}\left(R(f_{\hat{\theta}})\right) \lesssim \sqrt{\frac{D}{n}},$$

et est donc non informative la plupart du temps. On a deux enseignements généraux à retenir de tout ça :

1. Lorsque l'on choisit $\hat{f}$ comme solution d'un problème d'optimisation de type risque empirique sur un espace $\mathcal{F}$ de dimension D, si D est grand par rapport à n, cela mène la plupart du temps au phénomène de **surapprentissage** : votre prédicteur $\hat{f}$ va être bon sur les données D_n d'entraînement, mais son risque en généralisation peut être très mauvais.
2. Attester de la qualité d'un prédicteur en se basant sur R_n simplement peut se faire en ajoutant un terme issu de la théorie. Si votre prédicteur est de type du point 1, le terme à rajouter est de l'ordre de $\sqrt{D/n}$, et donc potentiellement très grand.

On verra par la suite que la stratégie de minimisation du risque empirique est valable dans le cas $D << n$, et que dans le cas contraire il faut concevoir des prédicteurs plus subtils, adaptés à la grande dimension.

Point de vue pratique 2 : validation croisée

Tout le problème vient donc, lorsque l'on veut utiliser le risque empirique comme juge de paix, que l'on entraîne nos prédicteurs sur le même jeu de données que celui qui sert à le valider. Une manière de garantir quelque chose à partir d'un risque empirique consiste alors à découper son jeu de données en deux parties : une partie train/entraînement notée $I_e \subset \{1, \dots, n\}$ et une partie test/validation notée $I_v \subset \{1, \dots, n\}$, avec $I_e \cap I_v = \emptyset$, et $n_e + n_v = n$. Dès lors, le risque empirique **hold-out**

$$R_{I_v}(\hat{f}_{I_e}) = \frac{1}{n_v} \sum_{i \in I_v} c(\hat{f}(X_i, D_{I_e}), Y_i)$$

est une approximation non-biaisée de $R_P(\hat{f}(., D_{I_e})$. En effet, conditionnellement à I_e, D_{I_e}, on a

$$\begin{aligned} \mathbb{E}\left(R_{I_v}(\hat{f}_{n_e}) \mid D_{I_e}, I_e\right) &= \frac{1}{n_v} \sum_{i \in I_v} \mathbb{E}\left(c(\hat{f}(X_i, I_e), Y_i) \mid I_e, D_{n_e}\right) \\ &= \mathbb{E}(R_P(\hat{f}(., D_{I_e}) \mid D_{I_e}, I_e). \end{aligned}$$

Maintenant, si le découpage en entraînement/validation est **indépendant des données**, vous avez $D_{I_e} \sim D_{I_e} \mid I_e \sim D_{n_e}$, où $D_{n_e} = \{(X_1, Y_1), \dots, (X_{n_e}, Y_{n_e})\}$, et donc

$$\mathbb{E}(R_P(\hat{f}(., D_{I_e}) \mid D_{I_e}, I_e) \sim \mathbb{E}\left(R_P, \hat{f}(., D_{n_e}) \mid D_{n_e}\right),$$

où $D_{n_e} = \{(X_1, Y_1), \dots, (X_{n_e}, Y_{n_e})\}$. On conclut alors

$$E_{D_n} R_{I_v}(\hat{f}_{I_e}) = E_{D_{n_e}} R_P(\hat{f}_{n_e}),$$

où $\hat{f}_{n_e} = \hat{f}(., D_{n_e})$, c'est à dire votre prédicteur entraîné sur n_e points d'échantillon. Si votre prédicteur est sensé, normalement $E_{D_n} R_P(\hat{f}_n)$ est une fonction décroissante de n, de sorte que le risque **hold-out**, en moyenne, est un estimateur biaisé de $E_{D_n} R_P(\hat{f}_n)$, avec une tendance à surestimer ce risque.

Dès lors, on peut être tenté de prendre n_e maximal, c'est à dire $n_e = n - 1$. Le prix à payer est en terme de variance : en effet, par le même raisonnement on a (toujours dans le cadre d'un découpage indépendant des données)

$$\begin{aligned} \mathrm{Var}(R_{I_v}(\hat{f}_{I_e})) &= \mathbb{E}\left[\mathrm{Var}(R_{I_v}(\hat{f}_{I_e}) \mid I_e, D_{I_e})\right] + \mathrm{Var}\left(\mathbb{E}\left[R_{I_v}(\hat{f}_{I_e}) \mid I_e, D_{I_e}\right]\right) && (1.1) \\ &= \mathbb{E}\left[\frac{1}{n_v}\mathrm{Var}(c(\hat{f}(., D_{I_e}), .) \mid I_e, D_{I_e})\right] + \mathrm{Var}\left(R_P(\hat{f}_{I_e}) \mid I_e, D_{I_e}\right) && (1.2) \\ &= \mathrm{Var}(R_P(\hat{f}_{n_e})) + \frac{1}{n_v}\mathbb{E}\left[\mathrm{Var}(c(\hat{f}_{n_e}(X), Y) \mid D_{n_e})\right]. && (1.3) \end{aligned}$$

On remarque alors que la variance du risque hold-out est la variance standard du risque du prédicteur bâti sur n_e points d'entraînement, plus un terme de variance *à train fixé* : celui-ci décroît en $1/n_v$, d'où l'intérêt de choisir n_v grand.

Plutôt que de calibrer n_e, n_v pour équilibrer biais et variance du risque hold-out, une stratégie commune pour réduire la variance est de moyenner sur plusieurs découpage, c'est ce qu'on appelle la **validation croisée**. Le principe est le suivant :

vous construisez B découpages en I_e, I_v : $(I_{e,j}, I_{v,j})_{j=1,\dots,B}$ de $\{1,\dots,n\}$, toutes de cardinalités n_e, n_v, et vous regardez

$$R^{CV}(\hat{f}) = \frac{1}{B}\sum_{j=1}^{B} R_{I_v}(\hat{f}_{I_e}).$$

Il y a plusieurs manières de choisir des découpes :

1. **Exploration exhaustive/leave-p-out** : prendre pour $I_{e,j}$ tous les sous-ensembles possibles de $\{1,\dots,n\}$ à n_e éléments. Nombre de découpages : $\binom{n}{n_v}$, ce qui fait long à calculer pour $n_v \geq 5$ par exemple. Aussi appelé risque "leave p out" (où $p = n_v$), abrégé en R^{lpo}.
2. **Exploration au hasard/VC Monte Carlo** : on choisit $I_{e,1},\dots,I_{e,B}$ i.i.d. de loi uniforme $\mathcal{U}(\mathcal{C}_n^{n_e})$. Aussi appelé risque "Monte Carlo CV" (R^{MCCV}). Potentiel défaut : ne fait pas forcément un usage "équilibré" des données (certaines peuvent être sous-employées au détriment d'autres).
3. **Validation croisée V-fold**. Le plus utilisé en pratique. Consiste à découper $\{1,\dots,n\}$ en V tranches $E_1,\dots,E_V$, puis, pour $j = 1,\dots,V$, $I_{j,v} = E_j$, $I_{j,e} = (I_{j,v})^c$. Noté R^{VFCV}. Avantage potentiel : usage équilibré des données. Potentiel défaut : associe toujours les mêmes en train/test. Dans certains cas on peut éviter les deux écueils de MCCV et VFCV, par de la validation croisée incomplète équilibrée notamment (cf Arlot et Celisse 2010). C'est un champ de recherches toujours actif.

De manière informelle, l'estimateur de $R(\hat{f}_{n_e})$ de plus petite variance basé sur validation croisée est $R^{lpo}(\hat{f})$. On peut donner des bornes pour la variance de R^{MCCV} :

$$\mathrm{Var}(R^{MCCV}) = \mathbb{E}\left(\mathrm{Var}(R^{MCCV} \mid D_n\right) + \mathrm{Var}(\mathbb{E}(R^{MCCV} \mid D_n).$$

D'une part on a $\mathbb{E}(R^{MCCV} \mid D_n) = R^{lpo}$. D'autre part, on a

$$\mathrm{Var}(R^{MCCV} \mid D_n) = \frac{1}{V}\mathrm{Var}(R_{I_{1,v}}(\hat{f}_{I_{1,e}}) \mid D_n),$$

et donc

$$\begin{aligned}
\mathbb{E}\left(\mathrm{Var}(R^{MCCV} \mid D_n)\right) &= \frac{1}{V}\mathbb{E}\left[\mathrm{Var}(R_{I_{1,v}}(\hat{f}_{I_{1,e}}) \mid D_n)\right] \\
&= \frac{1}{V}\left(\mathrm{Var}(R_{I_{1,v}}(\hat{f}_{I_{1,e}})) - \mathrm{Var}(\mathbb{E}\left(R_{I_{1,v}}(\hat{f}_{I_{1,e}}) \mid D_n\right))\right) \\
&= \frac{1}{V}\left(\mathrm{Var}(R_{I_v}(\hat{f}_{I_e}))\right) - \mathrm{Var}(R^{lpo}).
\end{aligned}$$

On en déduit

$$\mathrm{Var}(R^{MCCV}) = \frac{1}{V}\mathrm{Var}(R_{I_v}(\hat{f}_{I_e})) + \left(1 - \frac{1}{V}\right)\mathrm{Var}(R^{lpo}).$$

Le premier terme correspond à la variance d'une procédure hold-out simple, le deuxième à la variance du risque leave p-out. La variance du risque Monte Carlo CV va donc interpoler entre ces deux bornes. De manière générale on aura toujours

$$\mathrm{Var}(R^{hold-out}) \geq \mathrm{Var}(R^{CV}) \geq \mathrm{Var}(R^{lpo}),$$

et $\text{Var}(R^{CV})$ va converger vers R^{lpo} qui est la plus petite valeur atteignable lorsque V grandit (ce n'est prouvé que dans le cas de R^{MCCV}).

La plupart du temps, les procédures standard sous `R` ou `Python` calculent un estimateur de la variance

$$\hat{V} = \frac{1}{V}\sum_{j=1}^{V}(R_{I_{j,v}}(\hat{f}_{I_{j,e}}) - \bar{R})^2,$$

où $\bar{R} = \frac{1}{V}\sum_{j=1}^{V} R_{I_{j,v}}(\hat{f}_{I_{j,e}}) = R^{CV}$. Cet estimateur approche en fait la variance du hold-out :

$$\begin{aligned}
\mathbb{E}(\hat{V}) &= \mathbb{E}\left(\frac{1}{V}\sum_{j=1}^{V}(R_{I_{j,v}}(\hat{f}_{I_{j,e}}) - \mathbb{E}_{D_n}R_P(\hat{f}_{n_e}))^2\right) - \mathbb{E}\left(\bar{R} - \mathbb{E}_{D_n}R_P(\hat{f}_{n_e})\right)^2 \\
&= \text{Var}(R_{n_v}(\hat{f}_{n_e})) - \mathbb{E}\left(\bar{R} - \mathbb{E}_{D_n}R_P(\hat{f}_{n_e})\right)^2 \\
&= \text{Var}(R_{n_v}(\hat{f}_{n_e})) - \text{Var}(R^{CV}).
\end{aligned}$$

Pour peu que V soit assez grand pour que $\text{Var}(R^{CV}) \simeq \text{Var}(R^{lpo}) << \text{Var}(R^{hold-out})$, on aura $\mathbb{E}(\hat{V}) \simeq \text{Var}(R^{hold-out})$. On peut utiliser $\hat{V}$ dans un second temps pour borner $\text{Var}(R^{CV})$, étant acquis que $\text{Var}(R^{CV}) << \text{Var}(R^{hold-out})$ (c'est la plupart du temps une borne assez pessimiste).

En conclusion, en pratique il s'agit surtout de choisir V assez grand pour que

$$\text{Var}(R^{CV}) \simeq \text{Var}(R^{lpo}),$$

la contrainte venant la plupart du temps du temps de calcul (le plus coûteux en V-fold CV étant le *leave one out*).

Point de vue théorique : oracle, minimax (se mettre d'accord avec Adrien), no-free lunch

Chapitre 2

Méthodes standard 1 : Minimiseurs de risque empirique (ERM's)

Ce chapitre traite de l'aspect théorique des procédures basées sur la minimisation d'un risque empirique, ce qui regroupe beaucoup de méthodes utilisées en pratique.

2.1 Principe général

On a vu au chapitre précédent que le risque empirique, pour un prédicteur fixé, pouvait être une approximation raisonnable du vrai risque (en généralisation). Dès lors, une stratégie raisonnable peut être de choisir comme prédicteur

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} R_n(f) = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} c(f(X_i), Y_i)),$$

c'est à dire le meilleur prédicteur possible au sens du risque empirique, sur une classe $\mathcal{F}$ donnée. Parenthèse : ces stratégies rentrent dans le cadre de la M-estimation (elles en sont un cas particulier).

Deux questions se posent : pourquoi minimiser un risque empirique sur une classe garantirait-il de bonnes prédictions, et comment choisir une classe $\mathcal{F}$. Ces deux questions se rejoignent lorsque l'on considère l'excès de risque global :

$$R_P(\hat{f}) - R_P^* = \underbrace{\min_{f \in \mathcal{F}} (R_P(f) - R_P^*)}_{\text{erreur d'approximation}} + \underbrace{R_P(\hat{f}) - \min_{f \in \mathcal{F}} R_P(f)}_{\text{erreur d'estimation}}.$$

Cette décomposition de l'excès de risque est valable pour toute stratégie qui sélectionne son prédicteur dans une classe $\mathcal{F}$ donnée.

L'erreur d'approximation est un terme déterministe, qui comme son nom l'indique témoigne de la qualité d'approximation des prédicteurs considérés. Ce terme **décroît** avec la taille de $\mathcal{F}$, par exemple en régression les capacités d'approximation des fonctions polynomiales d'ordre k seront toujours plus grandes que celles d'ordre $k-1$.

L'erreur d'estimation est un terme aléatoire, témoignant de la qualité de la procédure de choix de prédicteur par rapport au choix optimal sur la classe $\mathcal{F}$. Cette qualité **croît** au fur et à mesure que $\mathcal{F}$ grandit (tout du moins pour les ERM). Un exemple de cas extrême peut être donné en régression :

Exemple 2.1 *: Régression - overfit.* On suppose $\mathcal{X} = \{x_1, \dots, x_d\}$ (fixés, on est dans un cadre de design fixe), $\mathcal{Y} = \mathbb{R}^d$, $Y_i = f^*(x_i) + \varepsilon_i$, où les ε_i sont centrés i.i.d. de variance σ^2. Pour $f \in \mathbb{R}^d$ (assimilé à ses prédiction), on prend pour perte

$$c(f, y) = \|y - f\|^2.$$

Pour faire simple on suppose $n = 1$. Le risque empirique est donc $\|Y - f\|^2$, où $Y = (Y_1, \dots, Y_d)$.

Plaçons nous dans le cas où $f^* = 0$, et regardons les classes $\mathcal{F}_0 = \{f \equiv c\}$, $\mathcal{F}_d = \{f \in \mathbb{R}^d[X]\}$. On a alors que

$$\begin{aligned}\arg\min_{f \in \mathcal{F}_0} R_n(f) &= \bar{Y}\mathbb{1}_d \\ \arg\min_{f \in \mathcal{F}_d} R_n(f) &= (Y_1, \dots, Y_d),\end{aligned}$$

la dernière égalité étant à prendre au sens "n'importe quel polynôme donnant à x_i la valeur Y_i". FAIRE DESSIN.

Dans les deux cas $\min_{f \in \mathcal{F}} R_P(f)$ vaut $R^* = d\sigma^2$. En notant $\hat{f}_0$, $\hat{f}_d$ les prédicteurs ERM correspondant, on a

$$\begin{aligned}\mathbb{E}(R_P(\hat{f}_0) - R_P^*) &= d\mathbb{E}(\bar{Y}^2) = \sigma^2 \\ \mathbb{E}(R_P(\hat{f}_d) - R_P^*) &= d\mathbb{E}\bar{Y}^2 = d\sigma^2.\end{aligned}$$

2.1.1 Majoration de l'erreur d'estimation

L'erreur d'approximation étant purement déterministe, son contrôle relève la plupart du temps de résultats d'analyse (qualité d'approximation des polynômes, des réseaux de neurones, etc.). Pour l'erreur d'estimation dans le cas des ERM, une majoration directe découle de la majoration du supremum des écarts entre risque et risque empirique sur la classe $\mathcal{F}$. En effet, notons

$$\Delta_n(\mathcal{F}) = \sup_{f \in \mathcal{F}} |(R - R_n)(f)|.$$

On a alors, si $\hat{f}$ est un ERM, et $\bar{f} \in \arg\min_{f \in \mathcal{F}} R(f)$,

$$\begin{aligned}R(\hat{f}) - R(\bar{f}) &= R_n(\hat{f}) - R_n(\bar{f}) + (R - R_n)(\hat{f}) - (R - R_n)(\bar{f}) \\ &\leq 2\Delta_n(\mathcal{F}).\end{aligned} \tag{2.1}$$

On peut remarquer que cette majoration est brutale : on a besoin du sup uniquement pour contrôler $(R - R_n)(\hat{f})$, donc a priori un seul sup suffit, portant sur $(R - R_n)$ et pas sur $|R - R_n|$. Autre remarque : il peut aussi être intéressant de majorer une quantité du type

$$\Delta_n^{\omega}(\mathcal{F}) = \sup_{f \in \mathcal{F}} \frac{\left|(R - R_n)(f) - (R - R_n)(\bar{f})\right|}{\omega(f - \bar{f})},$$

où $\omega(f - \bar{f})$ est typiquement la variance de $c(f(X), Y) - c(\bar{f}(X), Y)$, idéalement comparable à $R_P(f) - R_P(\bar{f})$. Dans ce cas on s'intéresse à des déviation **renormalisées**, cette technique est à la source des méthodes dites de localisation. On

en verra un exemple simple dans le cadre de la régression linéaire. Par exemple, si $\omega(f-\bar{f}) \leq \sqrt{R_P(f) - R_P(\bar{f})}$, l'équation (2.1) devient

$$\begin{aligned} R(\hat{f}) - R(\bar{f}) &\leq \Delta_n^\omega(\mathcal{F})\sqrt{R(\hat{f}) - R(\bar{f})} \\ &\leq \frac{1}{2}(R(\hat{f}) - R(\bar{f})) + \frac{1}{2}\Delta_n^\omega(\mathcal{F})^2, \end{aligned} \tag{2.2}$$

ce dont on déduit $R(\hat{f}) - R(\bar{f}) \lesssim \Delta_n^\omega(\mathcal{F})^2$, qui peut donner lieu à des vitesses plus rapides que la majoration via (2.1).

Majoration d'un supremum de déviations

La plupart des bornes sur les supremum de déviations mélangent trois ingrédients : inégalités de concentration, principe de symétrisation et principe de contraction (éventuellement), dans des ordres qui peuvent différer. Commençons par l'approche classique, qui fait intervenir les deux derniers principes en espérance seulement.

Voie 1 : via concentration sur le sup

Cette approche est basée sur la connaissance a priori d'une inégalité de concentration portant sur $\Delta_n(\mathcal{F})$. De fait, supposons que l'on ait, avec forte probabilité

$$\Delta_n(\mathcal{F}) \leq \mathbb{E}(\Delta_n(\mathcal{F})) + \text{deviation},$$

alors il reste à étudier $\mathbb{E}(\Delta_n(\mathcal{F}))$. En pratique, lorsque l'on a une fonction de perte bornée on peut utiliser des inégalités de type MacDiarmid (différences bornées), ou Talagrand-Bousquet (surtout pour la version localisée). On donnera un exemple dans la section qui vient, pour une vision globale vous êtes renvoyés à votre cours de statistique non asymptotique.

Il reste alors à majorer

$$\mathbb{E}(\Delta_n(\mathcal{F})) = \mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n c(f(X_i),Y_i) - \mathbb{E}(c(f(X_i),Y_i))\right|\right).$$

La première étape consiste souvent en une application du principe de symétrisation.

Théorème 2.2 : Symétrisation - Espérance

Soit $(Z_{1,f},\dots,Z_{n,f})_{f\in\mathcal{F}}$ est une famille de vecteurs aléatoires indexée par $\mathcal{F}$, vérifiant, à f fixé

1. *$Z_{1,f},\dots,Z_{n,f}$ sont indépendants ;*
2. *$\mathbb{E}(Z_{i,f}) = 0$, pour tout $i \in [\![1,n]\!]$.*

Si $(\varepsilon_i)_{i=1,\dots,n}$ est une famille i.i.d. de variables de Rademacher indépendante des $(Z_{i,f})_{i\in[\![1,n]\!],f\in\mathcal{F}}$, alors

$$\frac{1}{2}\mathbb{E}_{\varepsilon,Z}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i Z_{i,f}\right| \leq \mathbb{E}_Z\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n Z_{i,f}\right| \leq 2\mathbb{E}_{\varepsilon,Z}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i Z_{i,f}\right|.$$

Preuve du Théorème 2.2. On se donne $(Z'_{1,f}, \dots, Z'_{n,f})$ copie indépendante de $(Z_{1,f}, \dots, Z_{n,f})$. Commençons par l'inégalité de droite.

$$\begin{aligned}
\mathbb{E}_Z \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n Z_{i,f}\right| &= \mathbb{E}_Z \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n (Z_{i,f} - \mathbb{E}(Z'_{i,f}))\right| \\
&\leq \mathbb{E}_{Z,Z'} \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n (Z_{i,f} - Z'_{i,f})\right| \qquad \text{(Jensen)} \\
&= \mathbb{E}_{Z,Z',\varepsilon} \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \varepsilon_i (Z_{i,f} - Z'_{i,f})\right| \qquad \Big(\varepsilon_i(Z_{i,f} - Z'_{i,f}) \sim (Z_{i,f} - Z'_{i,f})\Big) \\
&\leq \mathbb{E}_{Z,\varepsilon} \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \varepsilon_i Z_{i,f}\right| + \mathbb{E}_{Z',\varepsilon} \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \varepsilon_i Z'_{i,f}\right| \\
&\leq 2\mathbb{E}_{\varepsilon,Z} \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \varepsilon_i Z_{i,f}\right|.
\end{aligned}$$

Même principe pour l'inégalité de gauche :

$$\begin{aligned}
\frac{1}{2}\mathbb{E}_{\varepsilon,Z} \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \varepsilon_i Z_{i,f}\right| &\leq \frac{1}{2}\mathbb{E}_{\varepsilon,Z} \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \varepsilon_i (Z_{i,f} - \mathbb{E}(Z'_{i,f}))\right| \\
&\leq \frac{1}{2}\mathbb{E}_{Z,Z',\varepsilon} \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \varepsilon_i (Z_{i,f} - Z'_{i,f})\right| \qquad \text{(Jensen)} \\
&= \frac{1}{2}\mathbb{E}_{Z,Z',\varepsilon} \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n (Z_{i,f} - Z'_{i,f})\right| \qquad \Big(\varepsilon_i(Z_{i,f} - Z'_{i,f}) \sim (Z_{i,f} - Z'_{i,f})\Big) \\
&\leq \frac{1}{2}\mathbb{E}_Z \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n Z_{i,f}\right| + \frac{1}{2}\mathbb{E}_{Z'} \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n Z'_{i,f}\right| \\
&= \mathbb{E}_Z \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n Z_{i,f}\right|.
\end{aligned}$$

□

L'inégalité de symétrisation montre que, en espérance, le supremum de déviations sur la classe $\mathcal{F}$ est comparable à la *complexité de Rademacher* de la classe $\mathcal{F}$:

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\varepsilon,Z} \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \varepsilon_i Z_{i,f}\right|.$$

L'intérêt pour borner l'espérance du supremum de déviations est immédiat : on peut écrire

$$\mathbb{E}(\Delta_n(\mathcal{F})) \leq \mathbb{E}_{D_n}\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i c(f(X_i), Y_i)\right|.$$

Á échantillon fixé $\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i c(f(X_i), Y_i)\right|$ se majore bien, et fait généralement apparaître la "dimension du modèle".

Exemple 2.3 *: $\mathcal{F}$ finie/fonction de perte bornée.* Dans le cas où $\mathcal{F} = \{f_1, \dots, f_p\}$ et où $\|c\|_\infty \leq M$, à D_n et f_j fixé, on a, pour tout $\lambda > 0$,

$$\mathbb{E}_\varepsilon(\exp(\sum_{i=1}^n \lambda\varepsilon_i c(f_j(X_i), Y_i)) = \prod_{i=1}^n \mathbb{E}_{\varepsilon_i}(\exp(\lambda\varepsilon_i c(f(X_i), Y_i))).$$

Par ailleurs, le Lemme de Hoeffding (voir par exemple [Boucheron et al., 2013, Lemme 2.2]) assure que si Y est centrée et prend ses valeurs dans $[a, b]$, alors

$$\Psi_Y(\lambda) = \mathbb{E}(\exp(\lambda Y)) \leq \exp\left(\lambda^2 \frac{(b-a)^2}{8}\right).$$

On en déduit alors que

$$\mathbb{E}_\varepsilon(\exp(\sum_{i=1}^n \lambda \varepsilon_i c(f_j(X_i), Y_i)) \leq \exp\left(\frac{n\lambda^2 M^2}{2}\right).$$

Maintenant, on peut majorer l'espérance du maximum par

$$\begin{aligned}
\mathbb{E}_\varepsilon\left(\exp\left[\lambda \max_{j=1,\dots,p} \sum_{i=1}^n \varepsilon_i c(f_j(X_i), Y_i))\right]\right) &= \mathbb{E}_\varepsilon\left(\max_{j=1,\dots,p}\left[\exp\left(\lambda \sum_{i=1}^n \varepsilon_i c(f_j(X_i), Y_i))\right)\right]\right) \\
&\leq \sum_{j=1}^p \mathbb{E}_\varepsilon\left(\exp\left[\lambda \sum_{i=1}^n \varepsilon_i c(f_j(X_i), Y_i))\right]\right) \\
&\leq p \exp\left(\frac{n\lambda^2 M^2}{2}\right).
\end{aligned}$$

En utilisant l'inégalité de Jensen, on obtient, pour tout $\lambda > 0$,

$$\begin{aligned}
\lambda \mathbb{E}_\varepsilon \max_{j=1,\dots,p} \sum_{i=1}^n \varepsilon_i c(f_j(X_i), Y_i)) &\leq \log\left(\mathbb{E}_\varepsilon\left(\exp\left[\lambda \max_{j=1,\dots,p} \sum_{i=1}^n \varepsilon_i c(f_j(X_i), Y_i))\right]\right)\right) \\
&\leq \log(p) + n\frac{\lambda^2 M^2}{2}.
\end{aligned}$$

Pour $\lambda = \sqrt{\frac{2\log(p)}{nM^2}}$, on obtient

$$\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \left|\frac{1}{n} \sum_{i=1}^n \varepsilon_i c(f(X_i), Y_i)\right| \leq \sqrt{\frac{2\log(p)}{n}}.$$

Comme cette borne est indépendante de D_n, on obtient immédiatement

$$\mathbb{E}_{D_n,\varepsilon} \sup_{f \in \mathcal{F}} \left|\frac{1}{n} \sum_{i=1}^n \varepsilon_i c(f(X_i), Y_i)\right| \leq \sqrt{\frac{2\log(p)}{n}}.$$

Dans cet exemple ($\mathcal{F}$ fini), la dimension de l'espace $\mathcal{F}$ est présente via le facteur $\log(|\mathcal{F}|)$. Par ailleurs, le passage par la symétrisation permet d'utiliser une sorte d'inégalité maximale en espérance pour les variables bornées. En fait, on peut généraliser cette méthode.

Commençons par remarquer que, dans la preuve bornant l'espérance d'un max, le point crucial est la majoration de $\mathbb{E}(\exp(\lambda Z))$, pour une variable Z bornée. On peut faire légèrement plus général en parlant de variable **sous-Gaussienne** :

Definition 2.4 : Variable sous-Gaussienne

Une variable Z centrée est dit sous-Gaussienne de variance v si, pour tout $\lambda \in \mathbb{R}$,

$$\Psi_Z(\lambda) = \mathbb{E}\left(\exp\left[\lambda Z\right]\right) \leq e^{\frac{\lambda^2 v}{2}}.$$

En d'autres termes, une variable est sous-Gaussienne de variance v si sa transformée de Laplace est bornée par la transformée de Laplace d'une variable Gaussienne de variance v. Le Lemme de Hoeffding présenté plus haut se résume alors à : "si Y est centrée et à valeurs dans $[a, b]$, alors Y est sous-Gaussienne de variance $(b-a)^2/4$". La sous-Gaussiannité permet de contrôler les déviations (cf cours de statistique non asymptotique), mais aussi les espérances de maximum. La généralisation de ce que l'on a fait dans le cadre $\mathcal{F}$ fini s'écrit :

PROPOSITION 2.5 : INÉGALITÉ MAXIMALE - VARIABLES SOUS- GAUSSIENNES

Si $Z_1, \dots, Z_p$ sont des variables centrées sous-Gaussiennes de variance v, alors

$$\mathbb{E} \max_{j=1,\dots,p} Z_j \leq \sqrt{2v \log(p)}.$$

La preuve consiste en la manipulation effectuée dans l'exemple précédent. On peut montrer que cette borne est asymptotiquement optimale dans le cadre de variables Gaussiennes i.i.d. $\mathcal{N}(0, v)$.

Revenons au cas d'une classe $\mathcal{F}$ infinie. On peut exploiter l'exemple dans le cas fini si on peut approcher l'espérance du sup par une espérance de sup, sur un ensemble fini. C'est à dire, si pour un $\delta > 0$, on a $\mathcal{F}_\delta$ sous ensemble fini de $\mathcal{F}$ tel que

$$\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i c(f(X_i), Y_i) \right| \leq \mathbb{E}_\varepsilon \sum_{f \in \mathcal{F}_\delta} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i c(f(X_i), Y_i) \right| + \delta,$$

alors, pour peu que c soit bornée, on aura immédiatement

$$\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i c(f(X_i), Y_i) \right| \leq \delta + \sqrt{\frac{2 \log\left(|\mathcal{F}_\delta|\right)}{n}}.$$

En particulier, pour $\delta = 1/\sqrt{n}$ on paiera un terme en $\sqrt{\log\left|\mathcal{F}_{1/\sqrt{n}}\right| / n}$. La notion de dimension d'un espace de prédicteurs va intervenir ici, via le nombre de prédicteurs requis pour approcher ces supremum à une échelle fixée, pour n'importe quelle échelle.

La première idée est de regarder pour quelle notion de distance sur l'ensemble $\mathcal{F}$ on a une approximation qui se traduit sur le supremum.

LEMME 2.6 : DISTANCE $L_2(P_n)$

On définit

$$\mathrm{d}(f_1, f_2)^2 = \frac{1}{n}\sum_{i=1}^{n}(c(f_1(X_i), Y_i) - c(f_2(X_i), Y_i))^2.$$

Pour tout $\delta > 0$, si $\mathcal{F}_\delta$ est un δ-covering de $\mathcal{F}$ pour la distance d*, et si $f_\delta(f)$ désigne* $\arg\min_{g\in\mathcal{F}_\delta} \mathrm{d}(f, g)$*, alors on a*

$$\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i (c(f(X_i), Y_i) - c(f_\delta(f)(X_i), Y_i)) \right| \leq \delta.$$

En particulier, si c est bornée par M, on a

$$\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i c(f(X_i), Y_i) \right| \leq \delta + M\sqrt{\frac{2\log\left(|\mathcal{F}_\delta|\right)}{n}}$$

Preuve du Lemme 2.6. La première inégalité vient de Cauchy-Schwartz :

$$\begin{aligned}
&\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i (c(f(X_i), Y_i) - c(f_\delta(f)(X_i), Y_i)) \right| \\
&\quad \leq \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \left(\sum_{i=1}^{n} \frac{\varepsilon_i^2}{n} \right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} \frac{(c(f(X_i), Y_i) - c(f_\delta(f)(X_i), Y_i))^2}{n} \right)^{\frac{1}{2}} \\
&\quad \leq \delta \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \left(\sum_{i=1}^{n} \frac{\varepsilon_i^2}{n} \right)^{\frac{1}{2}} \\
&\leq \delta \quad \text{(Jensen)}.
\end{aligned}$$

Pour la deuxième inégalité, il suffit de remarquer que

$$\begin{aligned}
&\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i c(f(X_i), Y_i) \right| \\
&\quad \leq \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i (c(f(X_i), Y_i) - c(f_\delta(f)(X_i), Y_i)) \right| + \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}_\delta} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i c(f(X_i), Y_i) \right|,
\end{aligned}$$

et on conclut en utilisant le cas fini. □

Pour obtenir une inégalité sur la complexité de Rademacher, il suffit alors (toujours dans le cadre d'une perte bornée) d'obtenir une borne sur les covering de $\mathcal{F}$ pour la distance d, et c'est là où la notion de dimension intervient : si on a

$$|\mathcal{F}_\delta| \leq \left(\frac{C}{\delta}\right)^d,$$

ce qui peut recolle avec la notion de dimension au sens Hausdorff, alors, pour δ en $n^{-1/2}$ on obtient des bornes en

$$\mathbb{E}\left(\Delta_n\right) \lesssim M\sqrt{\frac{d\log(n)}{n}}.$$

On peut se débarasser du facteur $\log(n)$ au moyen de techniques de chaînage, dont le principe est le suivant : plutôt que de choisir une échelle δ, on va regarder des échelles δ_j de plus en plus fines et approximer $c(f(x), y)$ par $c(f_0(f)(x), y) + \sum_j c(f_{j+1}(f)(x), y) - c(f_j(f)(x), y)$.

THÉORÈME 2.7 : INÉGALITÉ ENTROPIQUE DE DUDLEY

Pour la distance d *définie plus haut, on note*

$$H(\delta) = \log\left(|\mathcal{F}_\delta|\right),$$

le logarithme du covering number à l'échelle δ. *Pour* $f_0 \in \mathcal{F}$, *en notant* $\delta_{max}(f_0) = \sup_{f\in\mathcal{F}} \mathrm{d}(f, f_0)$ *on a*

$$\mathbb{E}_\varepsilon\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(c(f(X_i), Y_i) - c(f_0(X_i), Y_i))\right|\right) \leq \frac{12}{\sqrt{n}}\int_0^{\delta_{max}(f_0)/2}\sqrt{H(u)}du.$$

Preuve du Théorème 2.7. Posons $\delta_j = 2^{-j}\delta_{max}(f_0)$, et, pour $f \in \mathcal{F}$, $f_j(f) \in \arg\min_{g\in\mathcal{F}_{\delta_j}} \mathrm{d}(f, g)$. D'après le Lemme 2.6, on a $\frac{1}{n}\sum_{i=1}^n \varepsilon_i c(f_j(f)(X_i), Y_i) \to \frac{1}{n}\sum_{i=1}^n \varepsilon_i c(f(f)(X_i), Y_i)$ au sens $L_1(P_\varepsilon)$, de sorte que l'on peut écrire, pour tout f,

$$\sum_{i=1}^n \varepsilon_i c(f(X_i), Y_i) = \sum_{j=0}^{+\infty}\sum_{i=1}^n \varepsilon_i(c(f_{j+1}(f)(X_i), Y_i) - c(f_j(f)(X_i), Y_i)),$$

en gardant en tête que la convergence a lieu dans $L_1(P_\varepsilon)$. On peut alors écrire

$$\begin{aligned}&\mathbb{E}_\varepsilon\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(c(f(X_i), Y_i) - c(f_0(X_i), Y_i))\right|\right)\\ &\quad= \mathbb{E}_\varepsilon\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=0}^{+\infty}\sum_{i=1}^n \varepsilon_i(c(f_{j+1}(f)(X_i), Y_i) - c(f_j(f)(X_i), Y_i))\right|\right).\end{aligned}$$

Par ailleurs, pour tout $f \in \mathcal{F}$, une inégalité triangulaire donne

$$\mathrm{d}(f_{j+1}(f), f_j(f)) \leq \delta_{max}(f)(2^{-(j+1)} + 2^{-j}) = 3\delta_{j+1}.$$

On en déduit

$$\begin{aligned}&\mathbb{E}_\varepsilon\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=0}^{+\infty}\sum_{i=1}^n \varepsilon_i(c(f_{j+1}(f)(X_i), Y_i) - c(f_j(f)(X_i), Y_i))\right|\right)\\ &\quad\leq \sum_{j=0}^{+\infty}\mathbb{E}_\varepsilon \sup_{f_1\in\mathcal{F}_{\delta_{j+1}}, f_2\in\mathcal{F}_{\delta_j}, \mathrm{d}(f_1,f_2)\leq 3\delta_{j+1}}\left|\frac{1}{n}\sum_{j=0}^{+\infty}\sum_{i=1}^n \varepsilon_i(c(f_1(X_i), Y_i) - c(f_2(X_i), Y_i))\right|.\end{aligned}$$

Maintenant, à j fixé, on a d'une part que si $\mathrm{d}(f_1, f_2) \leq 3\delta_{j+1}$, alors

$$\frac{1}{n}\sum_{i=1}^n \varepsilon_i(c(f_1(X_i), Y_i) - c(f_2(X_i), Y_i))$$

est sous-Gaussienne de variance $9\delta_{j+1}^2/n$ (en utilisant le Lemme de Hoeffding comme dans le cas $\mathcal{F}$ finie). Par ailleurs, on a

$$\begin{aligned}\left|\left\{f_1 \in \mathcal{F}_{\delta_{j+1}}, f_2 \in \mathcal{F}_{\delta_j}, \mathrm{d}(f_1, f_2) \leq 3\delta_{j+1}\right\}\right| &\leq \left|\mathcal{F}_{\delta_{j+1}}\right| \times \left|\mathcal{F}_{\delta_j}\right| \\ &\leq \exp(H(\delta_{j+1}) + H(\delta_j)) \\ &\leq \exp(2H(\delta_{j+1})).\end{aligned}$$

En utilisant l'inégalité dans le cas fini, on obtient

$$\mathbb{E}_\varepsilon \sup_{f_1 \in \mathcal{F}_{\delta_{j+1}}, f_2 \in \mathcal{F}_{\delta_j}, \mathrm{d}(f_1,f_2) \leq 3\delta_{j+1}} \left| \frac{1}{n} \sum_{j=0}^{+\infty} \sum_{i=1}^{n} \varepsilon_i (c(f_1(X_i), Y_i) - c(f_2(X_i), Y_i)) \right| \leq 6 \frac{\delta_{j+1}}{\sqrt{n}} \sqrt{H(\delta_{j+1})},$$

et donc

$$\mathbb{E}_\varepsilon \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (c(f(X_i), Y_i) - c(f_0(X_i), Y_i)) \right| \right) \leq \sum_{j=1}^{+\infty} \frac{6\delta_j}{\sqrt{n}} \sqrt{H(\delta_j)}.$$

Comme $u \mapsto \sqrt{H(u)}$ est décroissante, et $\delta_j = 2(\delta_j - \delta_{j+1})$, une comparaison série/intégrale donne

$$\begin{aligned}\sum_{j=1}^{+\infty} \frac{6}{\sqrt{n}} \delta_j \sqrt{H(\delta_j)} &\leq \frac{12}{\sqrt{n}} \sum_{j=1}^{+\infty} (\delta_j - \delta_{j+1}) \sqrt{H(\delta_j)} \\ &\leq \frac{12}{\sqrt{n}} \sum_{j=1}^{+\infty} \int_{\delta_{j+1}}^{\delta_j} \sqrt{H(u)} du \\ &\leq \frac{12}{\sqrt{n}} \int_0^{\delta_{max}(f_0)/2} \sqrt{H(u)} du.\end{aligned}$$

□

Il existe des versions plus générales pour ce résultat, vous pouvez par exemple vous référer à la Section 13.1 de Boucheron et al. [2013]. On a un corollaire immédiat dans le cas d'une fonction de perte bornée pour un ensemble $\mathcal{F}$ de dimension d.

Corollaire 2.8 : Perte bornée, dimension d

Si la fonction de perte c est bornée par M, et si $\mathcal{F}$ est de dimension d (relativement à cette perte), c'est à dire que

$$H(\delta) \leq d \log \left(\frac{CM}{\delta} \vee 1 \right),$$

pour une constante $C > 1$, et , alors

$$\mathbb{E}_\varepsilon \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i c(f(X_i), Y_i) \right| \right) \leq \frac{M}{\sqrt{n}} \left(12C\sqrt{d} + 1 \right).$$

Preuve du Corollaire 2.8. Comme c est bornée par M, le diamètre de $\mathcal{F}$ est borné par $2M$, et on peut trouver f_0 tel que $\delta_{max}(f_0) \leq 2M$. On écrit alors

$$\begin{aligned}&\mathbb{E}_\varepsilon \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i c(f(X_i), Y_i) \right| \right) \\ &\quad \leq \mathbb{E}_\varepsilon \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (c(f(X_i), Y_i) - c(f_0(X_i), Y_i)) \right| \right) + \mathbb{E}_\varepsilon \left(\left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i c(f_0(X_i), Y_i) \right| \right).\end{aligned}$$

Le deuxième terme se traite grossièrement via Jensen :

$$\mathbb{E}_\varepsilon\left(\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i c(f_0(X_i),Y_i)\right|\right) \leq \sqrt{\mathbb{E}_\varepsilon\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i c(f_0(X_i),Y_i)\right)^2} \leq \frac{M}{\sqrt{n}}.$$

Le premier terme fait appel au Théorème 2.7. On a

$$\begin{aligned}
&\mathbb{E}_\varepsilon\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(c(f(X_i),Y_i)-c(f_0(X_i),Y_i))\right|\right)\\
&\leq \frac{12}{\sqrt{n}}\int_0^M \sqrt{d\log\left(\frac{CM}{u}\vee 1\right)}du\\
&\leq \frac{12\sqrt{d}}{\sqrt{n}}\int_0^{CM}\sqrt{\log\left(\frac{CM}{u}\right)}\\
&\leq \frac{12CM\sqrt{d}}{\sqrt{n}}\int_0^1\sqrt{\log\left(\frac{1}{u}\right)}du.
\end{aligned}$$

On conclut en remarquant que $\int_0^1\sqrt{\log\left(\frac{1}{u}\right)}du \leq 1$. □

Plusieurs remarques sur ce corollaire. En premier chef, le type de borne en

$$d\log\left(\frac{CM}{\delta}\right)$$

fait intervenir le "diamètre" M de façon relativement standard. Pour s'en convaincre, si on regarde la boule Euclidienne de dimension d et de diamètre M, on obtient

$$H(u) \leq d\log\left(\frac{3M}{u}\right).$$

On peut aussi remarquer que le choix de f_0 ici est un peut arbitraire, un choix plus malin est de prendre pour f_0 un $f^* \in \arg\min_{f\in\mathcal{F}} R_P(f)$, notamment pour les stratégies où on borne des supremum renormalisés. Enfin, remarque très importante : cette borne est **à D_n fixé**. Si la borne sur $H(\delta)$ vaut **uniformément en D_n**, on peut en déduire, en intégrant sur l'échantillon

$$\mathbb{E}(\Delta_n(\mathcal{F})) \lesssim CM\frac{\sqrt{d}}{\sqrt{n}}.$$

On pourrait raffiner en intégrant une borne éventuelle à D_n fixé (cela existe). Concluons sur le fait que ce n'est qu'une méthode classique et un peu générale. Dans des cas particuliers on peut s'en affranchir (cf l'exemple en régression linéaire). En première approche, c'est toutefois l'ordre de grandeur à avoir en tête : pour un problème de minimisation de risque empirique en dimension d, les "vitesses lentes" sont en $\sqrt{d/n}$, ce qui pose problème lorsque d devient grand.

On termine cette partie en énonçant un dernier outil utilisable pour la majoration d'une espérance de supremum : le principe de contraction.

Théorème 2.9 : Principe de contraction

Soit $(x_{1;f}, \dots, x_{n,f})_{i=1,\dots,n,f\in\mathcal{F}}$ un vecteur de $\mathbb{R}^n$ indexé par $\mathcal{F}$, et $\phi_1, \dots, \phi_n$ des fonctions L-Lipschitz satisfaisant $\phi_j(0) = 0$. On a alors

$$\mathbb{E}_\varepsilon \left| \sup_{f\in\mathcal{F}} \sum_{i=1}^n \varepsilon_i \phi_i(x_{i,f}) \right| \leq 2L\mathbb{E}_\varepsilon \left| \sup_{f\in\mathcal{F}} \sum_{i=1}^n \varepsilon_i x_{i,f} \right|.$$

Ce théorème est utile pour se ramener à une expérance maximale plus facilement contrôlable. Par exemple, si on peut écrire, pout tout $f \in \mathcal{F}$,

$$c(f(x_i), y_i) = \phi(g_f(x_i, y_i)),$$

où ϕ est L-Lipschitz, alors le principe de contraction montre que

$$\begin{aligned}\mathbb{E}(\Delta_n(\mathcal{F})) &\leq 2L\mathbb{E} \left| \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i g_f(x_i, y_i) \right| \\ &\leq 2L\mathbb{E}\Delta_n(\mathcal{G}),\end{aligned}$$

où $\mathcal{G} \supset \{g_f \mid f \in \mathcal{F}\}$. Si on sait contrôler $\Delta_n(\mathcal{G})$, on pourra alors contrôler $\Delta_n(\mathcal{F})$ (c'est un théorème de comparaison).

Voie 2 : via concentration à l'intérieur du sup

Cette approche consiste à travailler directement sur la déviation de Δ_n, sans passer par une décomposition en $\mathbb{E}(\Delta_n)$ et concentration. Pour se fixer les idées, dans le cas $\mathcal{F}$ finie et c bornée par M, on peut directement écrire, pour tout $t > 0$,

$$\begin{aligned}\mathbb{P}\left(\Delta_n(\mathcal{F}) \geq t\right) &= \mathbb{P}\left(\max_{j=1,\dots,p} \left| \frac{1}{n} \sum_{i=1}^n c(f_j(X_i), Y_i) - \mathbb{E}(c(f_j(X_i), Y_i)) \right| \geq t \right) \\ &\leq \sum_{j=1}^p \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n c(f_j(X_i), Y_i) - \mathbb{E}(c(f_j(X_i), Y_i)) \right| \geq t \right) \\ &\leq p\exp(-2nt^2/M^2),\end{aligned}$$

en utilisant directement Hoeffding. Dans le cadre général on peut aussi utiliser des techniques d'approximation par une grille, voire de chaînage, avec les mêmes outils que précédemment : symétrisation et contraction.

Théorème 2.10 : Symétrisation - En déviation

Avec les mêmes notations, si c est bornée par M, alors, pour tout $t \geq 4$,

$$\mathbb{P}\left(\Delta_n(\mathcal{F}) \geq 4M\sqrt{\frac{2t}{n}} \right) \leq 4\mathbb{P}\left(\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i c(f(X_i), Y_i) \right| \geq M\sqrt{\frac{2t}{n}} \right).$$

C'est une version allégée de [van de Geer, 2016, Lemme 16.1], dont on peut trouver une preuve dans van de Geer [2000]. On a le pendant pour le principe de contraction :

Théorème 2.11 : Principe de contraction - En déviation

Avec les mêmes notations que pour le Théorème 2.9, on a pour tout $t > 0$,

$$\mathbb{P}\left(\left|\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\varepsilon_i\phi_i(x_{i,f})\right| \geq t\right) \leq 2\mathbb{P}\left(\left|\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\varepsilon_i x_{i,f}\right| \geq \frac{t}{L}\right).$$

On trouvera la preuve de ce résultat (ainsi que du Théorème 2.9) dans Ledoux and Talagrand [2011]. De manière informelle, l'approche via concentration à l'intérieur du sup est un poil plus technique, et un poil plus générale que la première (au sens qu'elle permet de traiter un peu plus de cas, notamment celui des fonctions de coût non bornées en introduisant des conditionnements). Signalons enfin que la plupart des inégalités présentées ont des adaptations pour les processus renormalisés de type $\Delta_n^{\omega}(\mathcal{F})$ (cf votre cours de stat non asymptotique), dont l'intérêt est clairement exposé dans Boucheron, Stéphane et al. [2005].

2.1.2 Sélection de modèles

Dans la partie précédente on a regardé les propriétés des ERM $\hat{f}_n$ sur une classe $\mathcal{F}$, cette classe étant fixée a priori. La question qui se pose maintenant est : étant donnée une collection de modèle $(\mathcal{F}_m)_{m\in\mathbb{N}}$, comment choisir un bon modèle ?

Heuristiquement parlant, la réponse est "ni trop petit, ni trop gros". Dans le cas de modèles emboîtés $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \subset \mathcal{F}_m \subset \ldots$, si on regarde la performance des minimiseurs de risques empiriques $\hat{f}_{n,m}$, on a

$$R_P(\hat{f}_{n,m}) - R_P^* = \text{Biais}(m) + \text{Var}(m),$$

avec

$$\begin{aligned}\text{Biais}(m) &= \inf_{f\in\mathcal{F}_m} R_P(f) - R_P^* \\ \text{Var}(m) &= R_P(\hat{f}_{n,m}) - \inf_{f\in\mathcal{F}_m} R_P(f).\end{aligned}$$

Dans la suite on notera $R_m^* = \inf_{f\in\mathcal{F}_m} R_P(f)$ le plus petit risque atteignable sur le modèle $\mathcal{F}_m$. Typiquement on a $\text{Biais}(m)$ décroissante en m et $\text{Var}(m)$ croissante (un ordre de grandeur est $\mathbb{E}(\text{Var}(m)) \simeq \sqrt{d_m/n}$, où $d_m = \dim(\mathcal{F}_m)$ dans un cadre de vitesses lentes), ce qui mène à ce genre de graphique : FAIRE DESSIN.

On n'a donc pas généralement intérêt à choisir le plus gros modèle en général. Le problème est que si l'on sélectionne m sur la base du risque empirique, c'est à dire en choisissant

$$\hat{m} \in \arg\min_{m\in\mathbb{N}} R_n(\hat{f}_{n,m}),$$

on va systématiquement se retrouver à sélectionner le plus gros modèle (et le prédicteur correspondant $\hat{f}_{n,\hat{m}}$. On se place donc dans le cadre plus général où on cherche à minimiser un critère $\text{Crit}(m, D_n) = \text{Crit}_n(m)$, c'est à dire

$$\hat{m} \in \arg\min_{m} \text{Crit}_n(m),$$

et à regarder la performance de $\hat{f}_{n,\hat{m}}$, c'est à dire

$$R_P(\hat{f}_{n,\hat{m}}).$$

Idéalement, on voudrait faire à peu près aussi bien que le meilleur prédicteur pour la meilleure classe possible, c'est à dire obtenir quelque chose en

$$R_P(\hat{f}_{n,\hat{m}}) \lesssim \inf_{m\in\mathbb{N}} R_P(\hat{f}_{n,m}).$$

On en vient alors à la notion de *pénalité idéale*. En effet, si on pose

$$\begin{aligned}\mathrm{pen}_{id}(m) &= R_P(\hat{f}_{n,m}) - R_n(\hat{f}_{n,m})\\ \mathrm{Crit}_n(m) &= R_n(\hat{f}_{n,m}) + \mathrm{pen}_{id}(m),\end{aligned}$$

on a immédiatement (pour $\hat{m} \in \arg\min_m \mathrm{Crit}_n(m)$),

$$\begin{aligned}R_P(\hat{f}_{n,\hat{m}}) &= R_n(\hat{f}_{n,\hat{m}}) + \mathrm{pen}_{id}(\hat{m})\\ &\leq \inf_{m\in\mathbb{N}} R_n(\hat{f}_{n,m}) + \mathrm{pen}_{id}(m)\\ &\leq \inf_{m\in\mathbb{N}} R_P(\hat{f}_{n,m}),\end{aligned}$$

ce qui serait l'idéal. Le problème est qu'on ne connaît pas $\mathrm{pen}_{id}(m)$, tout au plus peut-on essayer de la majorer. D'après la section précédente, on a

$$\mathrm{pen}_{id}(m) \leq \Delta_n(\mathcal{F}_m),$$

et on peut construire dessus. Commençons par une fausse bonne idée : comme on connaît $\mathbb{E}(\Delta_n(\mathcal{F}_m))$, pourquoi ne pas l'utiliser comme pénalité ? La raison est la suivante : certes, on peut garantir, modèle par modèle que

$$\mathbb{E}(R_P(\hat{f}_{n,m})) = \mathbb{E}(R_n(\hat{f}_{n,m}) + (R_P - R_n)(\hat{f}_{n,m}) \leq \mathbb{E}(R_n(\hat{f}_{n,m}) + \mathrm{pen}(m)) = \mathbb{E}(\mathrm{Crit}_n(m)),$$

mais on ne peut rien dire sur

$$\mathbb{E}(R_P(\hat{f}_{n,\hat{m}})).$$

En effet, $\hat{m}$ étant lui aussi aléatoire, rien ne garantit que

$$\mathbb{E}((R_P - R_n)(\hat{f}_{n,\hat{m}})) \leq \mathbb{E}(\Delta_n(\mathcal{F}_{\hat{m}})),$$

qui est ce qu'il faudrait pour poursuivre, vu qu'on "intègre aussi en $\hat{m}$". On peut même aller plus loin : dans les cas où on peut calculer $\mathbb{E}(\mathrm{pen}_{id}(m)) = \mathbb{E}(R_P(\hat{f}_{n,m}) - R_n(\hat{f}_{n,m})$, comme en régression Gaussienne par moindre carrés, choisir $\mathrm{pen}(m) = \mathbb{E}(\mathrm{pen}_{id}(m))$ dans ce cas peut mener à des performances arbitrairement mauvaises (et à de très mauvais choix de modèles), voir par exemple la Section 2 de Giraud [2022].

L'idée centrale est que, pour pouvoir dire des choses sur $\mathbb{E}(R_P(\hat{f}_{n,\hat{m}}))$, il faut que le contrôle sur

$$|R_P - R_n|(\hat{f}_{n,m})$$

soit **uniforme** en m. Plus précisément, si on a une pénalité $\text{pen}(m)$ satisfaisant

$$\mathbb{P}\left(\forall m \in \mathbb{N} \quad |(R_P - R_n)|(\hat{f}_{n,m}) \leq \text{pen}(m)\right) \geq 1 - \varepsilon,$$

alors, sur ce même évènement de probabilité plus grande que $1-\varepsilon$ on aura

$$\begin{aligned} R_P(\hat{f}_{n,\hat{m}}) &= R_n(\hat{f}_{n,\hat{m}}) + R_P(\hat{f}_{n,\hat{m}}) - R_n(\hat{f}_{n,\hat{m}}) \\ &\leq R_n(\hat{f}_{n,\hat{m}}) + \text{pen}(\hat{m}) \\ &\leq \inf_{m\in\mathbb{N}} R_n(\hat{f}_{n,m}) + \text{pen}(m) \\ &\leq \inf_{m\in\mathbb{N}} R_P(\hat{f}_{n,m}) + 2\text{pen}(m). \end{aligned}$$

On peut raffiner légèrement si on veut comparer à R^*_m, auquel cas on a juste besoin d'une borne de déviation uniforme sur $(R_P - R_n)(\hat{f}_{n,m})$, et d'une autre sur $(R_n - R_P)(f^*_m)$. Enfin, d'autres stratégies prenant en compte des inégalités de type $(R_P - R_n)(\hat{f}_{n,m} - f^*_m) \lesssim R_P(\hat{f}_{n,m}) - R_P(f^*_m)$ sont possibles, le lecteur intéressé pourra se référer à Massart [2007].

Exemple 2.12 *: Pénalisation d'ERM - perte bornée.* On se place dans le cas où la fonction de perte est bornée par M, et où le modèle $\mathcal{F}_m$ est de dimension d_m, de sorte que

$$\mathbb{E}(\Delta_n(\mathcal{F}_m)) \leq CM\sqrt{\frac{d_m}{n}}.$$

Une borne en espérance ne va pas suffire, introduisons ici l'inégalité des différences bornées (suffisante dans ce cas).

Théorème 2.13 : Inégalité de MacDiarmid

Si $g : \mathcal{Z}^n \to \mathbb{R}$ est à différences bornées, c'est à dire

$$\sup_{z_1,\ldots,z_n,z_i'} |g(z_1,\ldots,z_i,\ldots,z_n) - g(z_1,\ldots,z_i',\ldots,z_n)| \leq c_i,$$

pour tout $i \in [\![1,n]\!]$, et $U = g(Z_1,\ldots,Z_n)$, où $Z_1,\ldots,Z_n$ sont indépendantes, alors U est sous-Gaussienne de variance

$$v = \frac{1}{4}\sum_{i=1}^n c_i^2.$$

En d'autres termes

$$\begin{aligned} \mathbb{P}\left(U \geq \mathbb{E}(U) + \sqrt{2vx}\right) &\leq e^{-x} \\ \mathbb{P}\left(U \leq \mathbb{E}(U) - \sqrt{2vx}\right) &\leq e^{-x}. \end{aligned}$$

La preuve de ce résultat fait appel à la méthode entropique, comme décrit dans [Boucheron et al., 2013, Section 6]. Dans notre cas, posons $z_i = (x_i, y_i)$, et

$$g(z_1,\ldots,z_n) = \sup_{f\in\mathcal{F}_m} \left|\frac{1}{n}\sum_{i=1}^n (c(f(x_i),y_i) - \mathbb{E}(c(f(X_i),Y_i))\right|.$$

On vérifie facilement que g est à différences bornées, avec $c_i \leq M/n$. On obtient alors

$$\mathbb{P}\left(\Delta_n(\mathcal{F}_m) \geq CM\sqrt{\frac{d_m}{n}} + M\sqrt{\frac{2x}{n}}\right) \leq e^{-x}.$$

Pour avoir une borne **uniforme en** m, il faut rajouter un terme. Si on se donne α_m tel que $\sum_{m\in\mathbb{N}} e^{-\alpha_m} = 1$, on a

$$\mathbb{P}\left(\bigcup_{m\in\mathbb{N}}\left\{\Delta_n(\mathcal{F}_m) \geq CM\sqrt{\frac{d_m}{n}} + M\sqrt{\frac{2(x+\alpha_m)}{n}}\right\}\right) \leq \left(\sum_{m\in\mathbb{N}} e^{-\alpha_m}\right) = e^{-x} \quad (2.3)$$

Et donc, en posant $\text{pen}(m) = CM\sqrt{\frac{d_m}{n}} + M\sqrt{\frac{2(x+\alpha_m)}{n}}$, la procédure de sélection de modèle donne

$$R_P(\hat{f}_{n,\hat{m}}) \leq \inf_{m\in\mathbb{N}}\left[R_P(\hat{f}_{n,m}) + 2CM\sqrt{\frac{d_m}{n}} + M\sqrt{\frac{2(x+\alpha_m)}{n}}\right],$$

avec probabilité plus grande que e^{-x}. Si on veut une inégalité en espérance, on peut en déduire

$$\mathbb{E}(R_P(\hat{f}_{n,\hat{m}})) \leq \inf_{m\in\mathbb{N}}\left[\mathbb{E}(R_P(\hat{f}_{n,m})) + C_1 M\left(\sqrt{\frac{d_m}{n}} + \sqrt{\frac{\alpha_m}{n}}\right)\right].$$

Concernant le choix des α_m, cela dépend du nombre de modèles ayant même dimension d_m. On a moralement deux extrémités possibles : la suite croissante de modèles, avec $d_m = m$, dans ce cas $\alpha_m = \frac{\pi^2}{6} \times 2\log(m)$ convient, et, dans le cas où $\mathcal{X} = \mathbb{R}^D$, on regarde les modèles basés sur m variables uniquement, qui sont au nombre de $\binom{d}{m} \leq d^m$. Dans ce dernier cas on prendra plutôt α_m de l'ordre de $m\log(d)$.

2.1.3 Aggrégation

L'idée générale de l'aggrégation est de combiner différents prédicteurs (plutôt que d'en sélectionner un comme en sélection de modèles). Le cadre le plus simple est celui de la régression, où, à partir d'une collection de prédicteurs de base $(\hat{f}_{n,m})_{m\in\mathbb{N}}$, on peut construire un prédicteur combiné par aggrégation, c'est à dire de type

$$\hat{f}_n = \sum_{m\in\mathbb{N}} \omega_m \hat{f}_{n,m},$$

où les poids ω_m satisfont $\omega_m \geq 0, \sum_{m\in\mathbb{N}} \omega_m = 1$ et dépendent usuellement des observations : en effet, il semble sensé de donner plus de poids aux prédicteurs les plus performants qu'aux autres. De ce point de vue, la sélection de modèle est un cas particulier de l'aggrégation, avec pour poids

$$\omega_m = \mathbb{1}_{m=\arg\min_k \text{Crit}(k)},$$

avec décision arbitraire en cas d'égalité. Pour mesurer la performance d'un prédicteur particulier, on se base sur un estimateur r_m de $R(\hat{f}_{n,m})$, et on prend pour ω_m une fonction décroissante de r_m. Un choix couramment utilisé en pratique (et souvent

optimal) est **l'aggrégation à poids exponentiels**, ou mélange de Gibbs, où les poids sont donnés par

$$\omega_m = \frac{\pi_m \exp\left(-\frac{1}{\beta} r_m\right)}{\sum_{m\in\mathbb{N}} \pi_m \exp\left(-\frac{1}{\beta} r_m\right)}, \tag{2.4}$$

où π_m est une pondération a priori sur les prédicteurs satisfaisant $\pi_m \geq 0, \sum_{m\in\mathbb{N}} \pi_m = 1$ (ce qui permet l'interprétation de l'aggrégation exponentielle comme une méthode bayésienne), et β est un paramètre de température. On remarque que lorsque $\beta \to 0$, ω_m converge vers

$$\omega_m = \frac{1}{\sum_{\ell\in\arg\min_k r_k} \pi_\ell} \pi_m \mathbb{1}_{m\in\arg\min_k r_k},$$

et on retrouve asymptotiquement le cadre de la sélection de modèles. Lorsque $\beta \to +\infty$, on a $\omega_m \to \pi_m$, et le mélange ne tient plus compte des performances des prédicteurs de base. Le régime intéressant se trouve entre les deux, pour plus de détails on peut se référer à Dalalyan and Tsybakov [2007], Catoni [2004]. La propriété cruciale de ces poids est qu'ils sont minimiseurs de la fonctionnelle suivante.

Lemme 2.14

Soient $(\pi_m)_{m\in\mathbb{N}}$ des poids a priori et $(r_m)_{m\in\mathbb{N}}$ une suite positive. Soient $(\omega_m)_{m\in\mathbb{N}}$ les poids exponentiels définis par (2.4). Alors

$$\sum_{m\in\mathbb{N}} \omega_m r_m + \beta \mathrm{d}_{KL}(\omega, \pi) = \inf_{\{q | \sum_{m\in\mathbb{N}} q_m = 1, q\geq 0\}} \left[\sum_{m\in\mathbb{N}} q_m r_m + \beta \mathrm{d}_{KL}(q, \pi)\right].$$

Ce lemme est généralisable : on peut enlever la condition de positivité de r, et remplacer la somme sur $\mathbb{N}$ par une intégrale. Un lecteur intéressé pourra trouver les détails dans [Catoni, 2004, p.160].

Preuve du Lemme 2.14. On rappelle que

$$\mathrm{d}_{KL}(q, \pi) = \begin{cases} \sum_{m\in\mathbb{N}} q_m \log\left(\frac{q_m}{\pi_m}\right) & \text{si} \quad q \ll \pi \\ +\infty & \text{sinon} \end{cases}$$

En remarquant que $\omega \ll \pi$, l'inégalité est triviale si π ne domine pas q. Supposons $q \ll \pi$, et notons $Z = \sum_{m\in\mathbb{N}} \pi_m \exp(-r_m/\beta)$. On remarque que $r_m = \beta \log\left(\frac{\pi_m}{Z\omega_m}\right)$. On calcule alors

$$\begin{aligned}
&\sum_{m\in\mathbb{N}} (q_m - \omega_m) r_m + \beta(\mathrm{d}_{KL}(q, \pi) - \mathrm{d}_{KL}(\omega, \pi)) \\
&\quad = \sum_{m\in\mathbb{N}} (q_m - \omega_m) \beta \log\left(\frac{\pi_m}{Z\omega_m}\right) + \beta\left(q_m \log\left(\frac{q_m}{\pi_m}\right) - \omega_m \log\left(\frac{\omega_m}{\pi_m}\right)\right) \\
&\quad = \beta \sum_{m\in\mathbb{N}} q_m \log\left(\frac{q_m}{\omega_m}\right) + (\omega_m - q_m) \log(Z) \\
&\quad = \mathrm{d}_{KL}(q, \omega) \geq 0.
\end{aligned}$$

□

Pour exploiter le Lemme 2.14 en vue de fournir une inégalité oracle, des inégalités de type convexité $R_P(\hat{f}_n) \leq \sum_{m\in\mathbb{N}} R_P(\hat{f}_{m,n})$ ou des pendants avec des estimateurs du risque sont souvent requis. Dans le cadre précédent d'une fonction de coût bornée par M et d'une fonction de risque convexe, on peut énoncer le corollaire suivant :

Corollaire 2.15

Dans le cadre de la section précédente, où la fonction de coût c est bornée par $M > 0$, si l'on suppose de plus que la fonction de risque R_P est convexe, alors, pour le choix de poids

$$\omega_m = \frac{\pi_m \exp\left(-\frac{1}{\beta}\left(R_n(\hat{f}_{n,m}) + CM\sqrt{\frac{d_m+\log(1/\pi_m)}{n}}\right)\right)}{Z},$$

le prédicteur aggrégé $\hat{f}$ vérifie, avec probabilité plus grande que $1-e^{-x}$,

$$R_P(\hat{f}) \leq \inf_q \left[\sum_{m\in\mathbb{N}} q_m(R_p(\hat{f}_{n,m}) + 2\mathrm{pen}(m)) + \beta \mathrm{d}_{KL}(q,\pi)\right] + 2\delta(x).$$

En particulier, pour $\beta \leq \frac{CM}{\sqrt{n}}$, on a, avec probabilité plus grande que $1-e^{-x}$,

$$R_P(\hat{f}) \leq \inf_{m\in\mathbb{N}} \left[R_P(\hat{f}_{n,m}) + 3CM\left(\sqrt{\frac{d_m}{n}} + \frac{\log(1/\pi_m)}{\sqrt{n}}\right)\right] + CM\sqrt{\frac{2x}{n}}.$$

Ce résultat s'applique en particulier dans le cas où la fonction de coût est convexe et bornée. On remarque qu'on peut obtenir pour le prédicteur agrégé à peu près les mêmes garanties que pour le prédicteur obtenu par sélection de modèle. C'est un phénomène général : dans la plupart des cas les prédicteurs par aggrégation font au moins aussi bien que les prédicteurs par sélection de modèle en termes de prédiction, parfois strictement mieux (on donnera un exemple en Section 3.2.2).

Ces prédicteurs aggrégés ont d'autres désavantages, notamment dans un contexte de grande dimension (voir Section 3.2.2 encore) ou en termes d'interprétabilité. On remarque aussi que le Corollaire 2.15 permet de comparer la performance du prédicteur agrégé de cette manière aux autres prédicteurs obtenus par aggrégation. Pour un point de vue plus général sur ces questions vous pouvez consulter Tsybakov [2013].

Preuve du Corollaire 2.15. Notons $\mathrm{pen}(m) = CM\sqrt{\frac{d_m+\log(1/\pi_m)}{n}}$, $r_{n,m} = R_n(\hat{f}_{n,m}) + \mathrm{pen}(m)$, $\beta = CM/\sqrt{n}$, et $\delta(x) = M\sqrt{2x/n}$. L'inégalité (2.3) se traduit alors en

$$\mathbb{P}\left(\bigcup_{m\in\mathbb{N}} \{\Delta_n(\mathcal{F}_m) \geq \mathrm{pen}(m) + \delta(x)\}\right) \leq e^{-x}.$$

On se place sur l'évènement complémentaire. On a alors, pour tous poids q,

$$\begin{aligned} R_P(\hat{f}) &\leq \sum_{m\in\mathbb{N}} \omega_m R_P(\hat{f}_{n,m}) \qquad \text{(Convexité de } R_P\text{)} \\ &\leq \sum_{m\in\mathbb{N}} \omega_m r_{n,m} + \delta(x) + \beta \mathrm{d}_{KL}(\omega,\pi) \qquad \text{(Inégalité (2.3))} \\ &\leq \sum_{m\in\mathbb{N}} q_m r_{n,m} + \delta(x) + \beta \mathrm{d}_{KL}(q,\pi) \qquad \text{(Lemme 2.14)} \\ &\leq \sum_{m\in\mathbb{N}} q_m (R_p(\hat{f}_{n,m}) + 2\mathrm{pen}(m)) + 2\delta(x) + \beta \mathrm{d}_{KL}(q,\pi) \end{aligned}$$

Pour $\beta \leq \frac{CM}{\sqrt{n}}$ et $q_m = \mathbb{1}_m$, on déduit le deuxième résultat. □

2.2 Exemples fondamentaux : dimension de Vapnik en classification et régression linéaire

Nous allons regarder les deux exemples standard en classification et régression, et notamment quelle notion de dimension intervient dans ces deux cas.

2.2.1 Classification binaire/ VC dimension

On se place dans le cadre où $\mathcal{F}$ est une classe de *classifieurs* à valeurs dans $\{0,1\}$, où la perte est la perte 0/1 standard, et où on choisit comme classifieur un minimiseur du risque empirique, c'est à dire

$$\hat{f}_n \in \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{f(X_i)\neq Y_i} = \arg\min_{f\in\mathcal{F}} R_n(f).$$

Pour majorer son excès de risque, on est amené à regarder

$$\Delta_n(\mathcal{F}) = \sup_{f\in\mathcal{F}} |(R - R_n)(f)|.$$

La fonction de perte étant bornée, on a immédiatement (en utilisant l'inégalité des différences bornées, Théorème 2.13),

$$\Delta_n(\mathcal{F}) \leq \mathbb{E}(\Delta_n(\mathcal{F})) + \sqrt{\frac{2x}{n}},$$

avec probabilité au moins $1 - e^{-x}$. Il s'agit alors de contrôler l'espérance (et comme expliqué auparavant c'est là où va intervenir la notion de dimension). En utilisant le principe de symétrisation, on a

$$\mathbb{E}(\Delta_n(\mathcal{F})) \leq 2\mathcal{R}_n(\mathcal{F}) \leq 2\mathbb{E}_{D_n}\mathbb{E}_\varepsilon\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{1}_{f(X_i)\neq Y_i}\right).$$

Raisonnons maintenant à D_n fixé : pour un f donné, $\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{1}_{f(X_i)\neq Y_i}$ est sous-Gaussienne de variance $1/n$. Par ailleurs, à D_n fixé, le supremum est en fait un supremum sur un ensemble fini : si on note

$$\mathcal{S}_{\mathcal{F}}(D_n) = \left\{(\mathbb{1}_{f(X_i)\neq Y_i})_{i=1,\dots,n} \mid f \in \mathcal{F}\right\} \subset \{0,1\}^n,$$

alors

$$\mathbb{E}_\varepsilon\left(\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\mathbb{1}_{f(X_i)\neq Y_i}\right)=\mathbb{E}_\varepsilon\left(\sup_{a\in\mathcal{S}_\mathcal{F}(D_n)}\frac{1}{n}\sum_{i=1}^n\varepsilon_i a_i\varepsilon_i\right)$$
$$\leq\sqrt{\frac{2\log\left(|\mathcal{S}_\mathcal{F}(D_n)|\right)}{n}},$$

en utilisant l'inégalité maximale sur les variables sous-Gaussiennes (Proposition 2.5). On peut remarquer que le cardinal de $S_\mathcal{F}(D_n)$ ne dépend pas des Y_i : de fait seules les n-uplets de valeurs possibles de $(f(X_i))_{i=1,\dots,n}$ peuvent varier, et donc, avec un léger abus on note

$$S_\mathcal{F}(X_1,\dots,X_n)=|\{(f(X_i))_{i=1,\dots,n}\}|=|\mathcal{S}_\mathcal{F}(D_n)|\leq 2^n.$$

Une autre manière de voir (qui est la manière originelle), est d'assimiler un classifieur f à un ensemble $C=\{f(x)=1\}$, et de regarder l'ensemble des $C\cap\{X_1,\dots,X_n\}$ possibles, c'est à dire

$$S_\mathcal{F}(X_1,\dots,X_n)=\left|\{\{X_1,\dots,X_n\}\cap C\mid C\in\mathcal{C}\}\right|.$$

A $x_{1:n}$ fixé, $S_\mathcal{F}(x_{1:n})$ est parfois appelé "VC-shatter coefficient", où VC est pour "Vapnik Cervonenkis" à qui l'on doit ces concepts. Ces coefficients sont étroitement reliés une notion de dimension combinatoire : la VC dimension (dimension de Vapnik Cervonenkis encore).

Definition 2.16 : VC dimension

— Soit $x_{1:n}\in\mathcal{X}^n$. $x_{1:n}$ est dit explosé par $\mathcal{F}$ (shattered) si $S_\mathcal{F}(x_{1:n})=2^n$.
— La dimension de Vapnik de $\mathcal{F}$, $d_{VC}(\mathcal{F})$, est défini comme le plus grand n tel qu'il existe un n-uplet explosé par $\mathcal{F}$. En d'autres termes

$$d_{VC}=\sup\left\{n\geq 1\mid\exists x_{1:n}\in\mathcal{X}^n\quad S_\mathcal{F}(x_{1:n})=2^n\right\}.$$

L'idée est la suivante : la complexité/dimension d'une classe $\mathcal{F}$ est attestée par sa capacité à produire toutes les 2^n étiquettes possibles, pour n allant jusqu'à d_VC. Les cas où n (taille d'échantillon) est plus petit que d_VC correspondent à des situations où le risque empirique peut être nul (surapprentissage le plus souvent). Évidemment, si $n\leq d_{VC}$, alors $S_\mathcal{F}(x_{1:n})\leq 2^n$ semble la meilleure borne atteignable. Le lemme de Sauer permet de borner $S_\mathcal{F}(x_{1:n})$ lorsque $n\geq d_{VC}$.

Lemme 2.17 : Lemme de Sauer

Si $\mathcal{C}$ est de dimension de Vapnik d_{VC}, alors, pour tout $n\geq d_{VC}$,

$$|\{x_1,\dots,x_n\}\cap C\mid C\in\mathcal{C}|\leq\sum_{i=0}^{d_{VC}}\binom{n}{i}\leq(n+1)^{d_{VC}}\wedge\left(\frac{en}{d_{VC}}\right)^{d_{VC}}.$$

La première partie de l'inégalité est prouvée dans [Wellner et al., 2013, Section 2], la deuxième l'est dans d'autres bon bouquins. Toujours est-il qu'une application

directe du Lemme de Sauer donne, si $\mathcal{F}$ est de dimension de Vapnik d,

$$\mathbb{E}(\Delta_n(\mathcal{F})) \leq 2\sqrt{\frac{2d\log(n+1)}{n}}.$$

Si on veut se débarasser du $\log(n)$, des techniques de chaînage comme dans la section précédente peuvent être employée. On aura toutefois besoin d'un résultat connectant la notion de dimension de Vapnik avec les covering numbers de $\mathcal{F}$ pour la distance $L_2(P_n)$.

Théorème 2.18 : Théorème 14.12 dans Ledoux and Talagrand [2011]

Pour Q une mesure sur $\mathcal{X}$, on note d_Q la distance L_2 induite sur $\mathcal{F}$. Si $\mathcal{F}$ est de dimension de Vapnik d, on a alors, pour tout $0 < \varepsilon < 1$,

$$\log\left(\mathcal{N}(\mathcal{F}, \varepsilon, \mathrm{d}_Q)\right) \leq Kd\left(1 + \log\left(\frac{1}{\varepsilon}\right)\right),$$

où K est une constante universelle (ne dépend ni de $\mathcal{F}$ ni de $\mathcal{Q}$).

La preuve n'est pas très compliquée, vous êtes encouragés à aller la voir. En utilisant le Théorème 2.7 avec les normes $L_2(P_n)$, où P_n désigne la mesure empirique associée à l'échantillon, on en déduit

$$\mathbb{E}(\Delta_n(\mathcal{F})) \leq C\sqrt{\frac{d_{VC}(\mathcal{F})}{n}},$$

et donc

$$\mathbb{E}(R(\hat{f}_n) - R^*_{\mathcal{F}}) \leq C\sqrt{\frac{d_{VC}(\mathcal{F})}{n}}.$$

Cette borne est uniforme en P (loi de (X, Y)), ce qui n'est pas contradictoire avec le no-free Lunch Theorem : on se contente de comparer au meilleur classifieur dans la classe $\mathcal{F}$. On peut enfin montrer que c'est la vitesse "optimale" pour cette classe :

Théorème 2.19 : Classes de Vapnik - Bornes inférieures

Pour une classe $\mathcal{F}$ de classifieurs de dimension de Vapnik $d \geq 2$, si $\mathcal{P}(\mathcal{F})$ désigne l'ensemble des lois P telles que $f^* \in \mathcal{F}$, on a

$$\inf_{\hat{f}} \sup_{P \in \mathcal{P}(\mathcal{F})} \mathbb{E}_{D_n}(R_P(\hat{f}) - R^*_P) \geq c\sqrt{\frac{d}{n}},$$

où c est une constante absolue.

Ce résultat est prouvé dans Devroye and Lugosi [1995]. Cela montre que les stratégies de minimisation du risque empiriques en classification sont optimales et ont vitesse grosso modo $\sqrt{d/n}$ **sans restrictions supplémentaires sur le problème**. En fait, on peut faire mieux si le problème de classification est plus simple (cas 0 erreur, ou conditions de marge, et dans ce cas une vitesse en $(d/n)^\alpha$, avec $\alpha \in [1/2, 1]$

est optimale et atteignable avec un ERM (voir par exemple Massart and Nédélec [2006]), ou sous des contraintes de structure type parcimonie, on reviendra dessus dans le cadre de la régression.

On termine avec un peu de manipulation de classes de Vapnik.

Classes de Vapnik standard

Proposition 2.20 : Dimension des classifieurs linéaires

Si $\mathcal{X} = \mathbb{R}^d$, *et* $\mathcal{F} = \{\mathrm{sg}(\langle\theta, .\rangle + b) \mid (\theta, b) \in \mathbb{R}^{d+1}\}$, *alors*

$$d_{VC}(\mathcal{F}) = d + 1.$$

Par ailleurs, si $\mathcal{F}_{lin} = \{\mathrm{sg}(\langle\theta, .\rangle) \mid \theta \in \mathbb{R}^d\}$,

$$d_{VC}(\mathcal{F}_{lin}) = d.$$

Preuve de la Proposition 2.20. Commençons par la minoration. Sans perte de généralité, on se ramène à $\mathcal{X}' = \{(x, 1) \mid x \in \mathbb{R}^d\} \subset \mathbb{R}^{d+1}$, et $\mathcal{F} = \{\mathrm{sg}(\langle\theta, .\rangle \mid \theta \in \mathbb{R}^{d+1}\}$. En notant e_j le j-ème vecteur de la base canonique de $\mathbb{R}^{d+1}$, on a, en posant $f_j = e_j + e_{d+1}$, pour $j \leq d$ et $f_{d+1} = e_{d+1}$, que les $f_j \in \mathcal{X}'$ pour tout j et sont libres. On peut alors définir f_j^* forme linéaire sur $\mathbb{R}^{d+1}$, valant 1 en f_j et 0 sur f_p, $p \neq j$.

Soit $\sigma_i \in \{0, 1\}^{d+1}$. On définit alors

$$f_\sigma^* = \sum_{i=1}^{d+1} \sigma_i f_i^*,$$

et θ_σ le vecteur tel que $f_\sigma^* = \langle\theta_\sigma, .\rangle$. On a alors bien $\mathrm{sg}(\langle\theta_\sigma, f_i\rangle) = \sigma_j$, pour tout j, et $f_1, \ldots, f_{d+1}$ est bien éclaté par $\mathcal{F}$. D'où $d_{VC}(\mathcal{F}) \geq d + 1$.

Minorons maintenant $d_{VC}(\mathcal{F})$. On aura besoin pour cela du Lemme de Radon :

Lemme 2.21 : Lemme de Radon

Si $x_1, \ldots, x_{d+2} \in \mathbb{R}^d$, *alors il existe une partition* X_1, X_2 *de* $\{x_1, \ldots, x_{d+2}\}$ *telle que*

$$\mathrm{Conv}(X_1) \cap \mathrm{Conv}(X_2) \neq \emptyset.$$

Preuve du Lemme 2.21. Regardons le système d'équations

$$\begin{cases} \sum_{i=1}^{d+2} \alpha_i x_i = 0, \\ \sum_{i=1}^{d+2} \alpha_i = 0, \end{cases}$$

qui est équivalent à $\sum_{i=1}^{d+2} \alpha_i \tilde{x}_i = 0$, avec $\tilde{x}_i = (x_i, 1) \in \mathbb{R}^{d+1}$. Les $\tilde{x}_i$ étant liés dans $\mathbb{R}^{d+2}$, une solution non-nulle α^* à ce système existe. Groupons d'un côté les coefficients positifs et de l'autre les coefficients négatifs, cela donne

$$\sum_{i \in I_+} \alpha_i^* x_i = \sum_{i \in I_-} (-\alpha_i^*) x_i,$$

et $\sum_{i\in I_+} \alpha_i^* = -\sum_{i\in I_-} \alpha_i^* = |\alpha|/2$. Cela donne

$$\frac{2}{|\alpha|}\sum_{i\in I_+} \alpha_i^* x_i = \frac{2}{|\alpha|}\sum_{i\in I_-} (-\alpha_i^*) x_i,$$

ce point est donc dans l'enveloppe convexe de $\{x_i \mid i \in I_+\}$ et $\{x_i \mid i \in I_-\}$. □

On se donne maintenant $x_1, \dots, x_{d+2}$ dans $\mathbb{R}^{d+1}$, la partition X_1, X_2 correspondante par le Lemme de Radon, et x_0 dans l'intersection des deux enveloppes convexes. Soit alors $\sigma \equiv 0$ sur X_1 et $\equiv 1$ sur X_2, si $x_{1:d+2}$ était explosé on aurait f_σ tel que $f_\sigma(x_i) = \sigma_i$, or un tel f_σ donnerait $f_\sigma(x_0) = 0$ et $f_\sigma(x_0) = 1$, d'où la contradiction.

Pour les classifieurs linéaires : avec les mêmes arguments que précédemment $e_1, \dots, e_d$ est explosé, et si $x_1, \dots, x_{d+1} \in \mathbb{R}^d$, alors il existe α non-nulle telle que $\sum_i \alpha_i x_i = 0$. En décomposant en I_+ et I_- comme précédemment, et en prenant $\sigma \equiv 0$ sur I_+, $\sigma \equiv 1$ sur I_-, on a, en supposant $I_+ \neq \emptyset$,

$$\begin{aligned} 0 &= \langle \theta_\sigma, 0\rangle \\ &= \sum_{i\in I_-} \alpha_i \langle \theta_\sigma, x_i\rangle - \sum_{i\in I_+} \alpha_i \langle \theta_\sigma, x_i\rangle \\ &< 0. \end{aligned}$$

Si $I_+ = \emptyset$, 0 est dans l'enveloppe convexe des x_i et on conclut comme auparavant. □

On a un corollaire quasiment immédiat.

Corollaire 2.22 : Dimension des boules Euclidiennes

Si $\mathcal{X} = \mathbb{R}^d$ *et* $\mathcal{F} = \left\{\mathbb{1}_{x\in \mathrm{B}(c,r)} \mid c \in \mathbb{R}^d, r > 0\right\}$, *alors*

$$d_{VC}(\mathcal{F}) = d + 1.$$

Preuve du Corollaire 2.22. On peut remarquer que pour c, r donnés, l'équation $x \in \mathrm{B}(c, r)$ s'écrit

$$\|x\|^2 - 2\langle x, c\rangle + \|c\|^2 \leq r^2,$$

soit encore

$$\langle 2c, x\rangle - \|x\|^2 + (r^2 - \|c\|^2) \geq 0.$$

En posant $\tilde{x} = (x, \|x\|^2)$, on a immédiatement que $d_{VC}(\mathcal{F}) \leq d + 2$ (séparation par hyperplans dans $\mathbb{R}^{d+1}$). Cette borne immédiate est toutefois inutile.

En fait, il suffit de se rendre compte que si $x_1, \dots, x_{d+2}$ est explosé par les boules, il l'est aussi par des hyperplans. En effet, supposons que $x_{1:d+2}$ soit explosé par des boules. Alors, pour tout I_1, I_2 partition de $x_{1:d+2}$ il existe c_1, c_2, r_1, r_2 tels que

$$\begin{aligned} &\forall x \in I_1 \quad \|x - c_1\|^2 \leq r_1^2 \quad \text{et} \quad \|x - c_2\|^2 > r_2^2, \\ &forall x \in I_2 \quad \|x - c_1\|^2 > r_1^2 \quad \text{et} \quad \|x - c_2\|^2 \leq r_2^2. \end{aligned}$$

On en déduit alors

$$\begin{aligned}\forall x \in I_1 \quad & 2\langle c_2 - c_1, x\rangle + \|c_1\|^2 - \|c_2\|^2 + r_2^2 - r_1^2 < 0\\ \forall x \in I_2 \quad & 2\langle c_2 - c_1, x\rangle + \|c_1\|^2 - \|c_2\|^2 + r_2^2 - r_1^2 > 0.\end{aligned}$$

On en déduit alors que $d_{VC}(\mathcal{F}) \leq d+1$. Pour l'existence d'un $d+1$-uplet explosé, il suffit de se convaincre qu'un $d+1$-uplet explosé par les hyperplans l'est aussi par les boules : en effet un demi-plan n'est rien d'autre que la limite d'une boule dont le centre s'en va dans une direction orthogonale. □

Un autre exemple avec les rectangles.

PROPOSITION 2.23 : DIMENSION DES RECTANGLES

Si $\mathcal{X} = \mathbb{R}^d$, et $\mathcal{F} = \{\mathbb{1}_R \mid R$ rectangle de $\mathbb{R}^d\}$, alors

$$d_{VC}(\mathcal{F}) = 2d.$$

Preuve de la Proposition 2.23. Si $x_j = e_j$, pour $j \in [\![1, d]\!]$, et $x_j = -e_{j-d}$ pour $j \in [\![d+1, 2d]\!]$, et $\sigma \in \{-1, 1\}^{2d}$, le rectangle défini par

$$R_\sigma = \bigcap_{j=1}^{d} \{e_j^* \leq 1 + \sigma_j/2\} \cap \bigcap_{j=1}^{d} \{e_j^* \geq -1 - \sigma_{j+d}/2\}$$

vérifie bien $\sigma_j = 1 \Leftrightarrow x_j \in R_\sigma$. Maintenant, si $x_1, \ldots, x_{2d+1} \in \mathbb{R}^d$, $x_j^{+,-} \in \arg\min, \arg\max e_j^*(x_i)$, et $c_j^{+,-}$ sont les valeurs correspondantes, on a forcément $x_{i_0} \notin \{x_j^{+,-} \mid j = 1, \ldots, d\}$ et donc

$$x_{i_0} \in \bigcap_{j=1}^{d} \{c_j^- \leq e_j^*(x) \leq c_j^+\}.$$

Dès lors la configuration $\sigma_{i_0} = -1$ et $\sigma_i = 1$ si $i \neq i_0$ est impossible à obtenir. □

Arrivé ici, on pourrait être tenté de croire que la dimension de Vapnik correspond aux "degrés de libertés" de la classe considérée, c'est à dire la dimension du paramétrage. Il n'en est rien.

PROPOSITION 2.24 : CONTRE-EXEMPLE

Si $\mathcal{X} = \mathbb{R}$, et $\mathcal{F} = \{\mathrm{sg}(\sin(\theta x)) \mid \theta > 0\}$, alors

$$d_{VC}(\mathcal{F}) = +\infty.$$

Preuve de la Proposition 2.24. Soit $p \in \mathbb{N}^*$, $x_i = 2^{-i}$, pour $i = 1, \ldots, p$, et $\sigma \in \{0, 1\}^p$. En posant

$$\theta = \pi(1 + \sum_{i=1}^{p} 2^i(1 - y_i)),$$

on va montrer que $\mathrm{sg}(\sin(\theta x_j)) = y_j$ pour tout j. On commence par écrire

$$\theta x_j = 2^{-j}\pi + \pi \sum_{i=1}^{p} 2^{i-j}(1 - y_i).$$

Pour $i > j$ les termes dans la somme sont congrus à 2π, donc

$$\sin(\theta x_j) = \sin\left(2^{-j}\pi + (1 - y_j)\pi + \pi \sum_{i=1}^{j-1} 2^{i-j}(1 - y_i)\right).$$

Or, on a

$$\begin{aligned} 2^{-j}\pi + \pi \sum_{i=1}^{j-1} 2^{i-j}(1 - y_i) &\leq \pi \sum_{i=1}^{j} 2^{-i} < \pi, \\ 2^{-j}\pi + \pi \sum_{i=1}^{j-1} 2^{i-j}(1 - y_i) &\geq (1 - y_j)\pi + 2^{-j}\pi > 0. \end{aligned}$$

On en déduit $\mathrm{sg}(\sin(\theta x_j)) = y_j$, et donc que $x_{1:p}$ est explosé. On en déduit $d_{VC}(\mathcal{F}) \geq p$. □

Donc en général il ne faut pas confondre dimension de Vapnik et degrés de liberté, sauf dans des cas bien précis (ensemble de 0 de polynômes de degré k par exemple). Moralité : calculer une dimension de Vapnik de manière exacte est souvent compliqué. En revanche, en obtenir une majoration à partir de dimensions de base connues est souvent possible.

Arithmétique de classes de Vapnik

On peut raisonner simplement sur les shatter coefficients d'union, intersection, produit et composition de classes de Vapnik. Pour une classe $\mathcal{A}$ de sous-ensembles de $\mathcal{X}$ (ou la classe $\mathcal{F}$ de classifieurs correspondante), on note

$$\begin{aligned} S_{\mathcal{A}}(n) &= \sup_{x_{1:n} \in \mathcal{X}^n} S_{\mathcal{A}}(x_{1:n}) \\ &= \sup_{x_{1:n}} \left|\mathcal{F}(x_{1:n})\right|. \end{aligned}$$

Proposition 2.25 : Arithmétique - shatter coefficients

— *Soient $\mathcal{A}$ et $\mathcal{B}$ deux classes de sous-ensembles de $\mathcal{X}$. Alors, si $\mathcal{A} \cap \mathcal{B} = \{A \cap B \mid A \in \mathcal{A}, B \in \mathcal{B}\}$, et $\mathcal{A} \cap \mathcal{B} = \{A \cap B \mid A \in \mathcal{A}, B \in \mathcal{B}\}$, alors*

$$\begin{aligned} S_{\mathcal{A}\cup\mathcal{B}}(n) &\leq S_{\mathcal{A}}(n) \times S_{\mathcal{B}}(n), \\ S_{\mathcal{A}\cap\mathcal{B}}(n) &\leq S_{\mathcal{A}}(n) \times S_{\mathcal{B}}(n). \end{aligned}$$

— *Soient $\mathcal{A}$ et $\mathcal{B}$ deux classes de sous-ensembles de $\mathcal{X}_1$ et $\mathcal{X}_2$, et $\mathcal{A} \times \mathcal{B} = \{A \times B \mid A \in \mathcal{A}, B \in \mathcal{B}\}$, alors*

$$S_{\mathcal{A}\times\mathcal{B}}(n) \leq S_{\mathcal{A}}(n) \times S_{\mathcal{B}}(n).$$

Preuve de la Proposition 2.25. . Soit $x_{1:n} \in \mathcal{X}^n$, et $A \in \mathcal{A}$, en notant $x_A = \{x_i \mid x_i \in A\}$, comme $|x_A| \leq n$, on a

$$|\{B \cap x_A \mid B \in \mathcal{B}\}| \leq S_{\mathcal{A}}(n).$$

On en déduit

$$S_{\mathcal{A}\cap\mathcal{B}}(n) \leq S_{\mathcal{A}}(n) \times S_{\mathcal{B}}(n).$$

Comme $S_{\mathcal{A}^c}(n) = S_{\mathcal{A}}(n)$, on en déduit le résultat sur $S_{\mathcal{A}\cup\mathcal{B}}(n)$.

Soient maintenant $z_1 = (x_1, y_1), \ldots, z_n = (x_n, y_n)$ n points de $\mathcal{X}_1 \times \mathcal{X}_2$. Avec les mêmes notations qu'au dessus, pour tout $A \in \mathcal{A}$,

$$\begin{aligned} |(A \times \mathcal{B} \cap z_{1:n})| &= |\{\mathcal{B}(y_A)\}|, \\ &\leq S_{\mathcal{B}}(n), \end{aligned}$$

où $y_A = \{y_i \mid x_i \in A\}$. On en déduit

$$S_{\mathcal{A}\times\mathcal{B}}(n) \leq S_{\mathcal{A}}(n) \times S_{\mathcal{B}}(n).$$

□

En conjonction avec le Lemme de Sauer, on peut en déduire des bornes sur les dimensions de Vapnik des classes correspondantes.

Corollaire 2.26 : Arithmétique - VC dimensions

> *Avec les notations de la Proposition précédente, il existe une constante absolue* $c_0 > 0$ *telle que*
>
> $$d_{\mathcal{A}\cup\mathcal{B}}, d_{\mathcal{A}\cap\mathcal{B}}, d_{\mathcal{A}\times\mathcal{B}} \leq c_0(d_{\mathcal{A}} + d_{\mathcal{B}}).$$

Preuve du Corollaire 2.26. On le prouve pour $\mathcal{A}\times\mathcal{B}$, c'est exactement la même chose pour les deux autres cas. Commençons par remarquer que de manière évidente

$$d_{\mathcal{A}\times\mathcal{B}} \geq d_{\mathcal{A}} \vee d_{\mathcal{B}}.$$

En effet, si on se donne un n-uplet (x_i) explosé par $\mathcal{A}$, il suffit de prendre $z_i = (x_i, y_B)$, où $y_B \in B$ pour un B quelconque pour avoir (z_i) explosé par $\mathcal{A}\times\mathcal{B}$. Posons $n = d_{\mathcal{A}\times\mathcal{B}}$. La Proposition 2.25 et le Lemme de Sauer donnent alors (car $n \geq d_{\mathcal{A}} \vee d_{\mathcal{B}}$)

$$\begin{aligned} 2^n &= S_{\mathcal{A}\times\mathcal{B}}(n) \\ &\leq S_{\mathcal{A}}(n) \times S_{\mathcal{B}}(n) \\ &\leq \left(\frac{en}{d_{\mathcal{A}}}\right)^{d_{\mathcal{A}}} \left(\frac{en}{d_{\mathcal{B}}}\right)^{d_{\mathcal{B}}}. \end{aligned}$$

En passant au logarithme on obtient

$$\begin{aligned} n\log(2) &\leq (d_{\mathcal{A}} + d_{\mathcal{B}})\log(en) - d_{\mathcal{A}}\log(d_{\mathcal{A}}) - d_{\mathcal{B}}\log(d_{\mathcal{B}}) \\ &\leq (d_{\mathcal{A}} + d_{\mathcal{B}})\left[1 + \log(n) - p_{\mathcal{A}}\log(p_{\mathcal{A}}) - p_{\mathcal{B}}\log(p_{\mathcal{B}}) - \log(d_{\mathcal{A}} + d_{\mathcal{B}})\right]. \end{aligned}$$

Or $-p_{\mathcal{A}}\log(p_{\mathcal{A}}) - p_{\mathcal{B}}\log(p_{\mathcal{B}}) \leq \log(2)$, on en déduit

$$n\log(2) \leq (d_{\mathcal{A}} + d_{\mathcal{B}})\left[1 + \log(2) + \log(n) - \log(d_{\mathcal{A}} + d_{\mathcal{B}})\right].$$

Si on pose $n = d_{\mathcal{A}\times\mathcal{B}} = C(d_{\mathcal{A}} + d_{\mathcal{B}})$, cette équation devient

$$C\log(2) \leq \left[1 + \log(2) + \log(C)\right],$$

ce qui n'est possible que pour $C \leq c_0$, pour une constante absolue c_0. □

Une constante optimale est donnée dans Van Der Vaart and Wellner [2009], au prix de calculs un peu plus techniques. On a vu que les techniques à l'oeuvre pour majorer des dimensions de Vapnik passent par la majoration (souvent plus facile) des shatter coefficients. Une autre tactique peut être de majorer les covering number de ces classes (au sens $L_2(P_n)$), et d'utiliser le Théorème 2.18. C'est d'ailleurs une stratégie adaptable à des extensions de la VC dimension, voir par exemple Men [2003].

Un exemple concret : réseau de neurones (feedforward)

Commençons par définir ce qu'est un neurone (ou Perceptron). Pour des entrées x de dimension d va réaliser les opérations suivantes (qui vont le définir) : (DESSIN) !

1. un produit scalaire avec un vecteur de poids w : $\langle w, x\rangle + w_0$, où $w \in \mathbb{R}^d$, $w \in \mathbb{R}$,
2. un passage par une fonction d'activation σ, ce qui donne la sortie

$$\sigma(\langle w, x\rangle + w_0).$$

Un neurone est donc défini par ses poids et sa fonction d'activation. Comme fonctions d'activations standard on a

— Fonction *binaire* ou *seuil* : $\sigma(t) = \mathrm{sg}(t)$.
— Fonction *sigmoïde*, ou *logistique* : $\sigma(t) = \frac{1}{1+e^{-t}}$.
— Fonction *tangente hyperbolique* : $\sigma(t) = \mathrm{th}(t)$.
— Fonction *Relu* (Rectified linear unit) : $\sigma(t) = t \vee 0$.

Un réseau de neurones **feedworward** (à k couches cachées) est composé de couches de neurones, dont les sorties vont **alimenter la couche suivante**. (DESSIN). On peut le paramétrer par

1. couche d'entrée : c'est juste l'input $x \in \mathbb{R}^d$.
2. couche j : formée par les neurones $N_{j,1}, \ldots, N_{j,d_j}$. Chaque neurone $N_{j,i}$ prend en entrée $x^{(j-1)}$ (sortie de la couche d'avant), et ressort $\sigma(\left\langle w_{j,i}, x^{(j-1)}\right\rangle + w^0_{j,i})$. **Remarque** : le fait que le neurone $N_{j,i}$ ne prenne en compte que certaines sorties de neurones parmi la couche $j-1$ se traduit par la nullité **figée** des poids correspondants. On note $d_{j,i}$ le nombre de poids non nuls, de sorte que le neurone $N_{j,i}$ a moralement $\omega_{j,i} = d_{j,i} + 1$ degrés de libertés.
3. couche de sortie : un seul neurone $N_{k+1,1}$.

Usuellement, on choisit une fonction d'activation commune à tout le monde, et à la fin on classifie suivant le signe de la sortie. On peut se faire une idée de la dimension de Vapnik d'un tel objet assez facilement dans le cas d'une activation seuil.

Théorème 2.27 : VC dim - activation binaire

Si $\mathcal{N}$ est un réseau de neurones feedworward, avec activations binaires, et

$$N = \text{nombre de neurones de } \mathcal{N},$$
$$D = \sum_{p\in\mathcal{N}} \omega_p \quad \text{nombre total de poids non nuls, ou degré de liberté total,}$$

alors

$$d_{VC}(\mathcal{N}) \leq c_0 D \log(N),$$

pour une constante absolue c_0.

Si la constante vous intéresse, vous pouvez la trouver dans [Anthony and Bartlett, 1999, Section 6], qui prouve au passage (avec une minoration) que l'ordre de grandeur en $D\log(N)$ est le bon. Cet ordre de grandeur en $D\log(n)$ est le meilleur des cas : pour des activations sigmoïdes, toujours dans [Anthony and Bartlett, 1999, Section 6] on a une borne en $(ND)^2$ (qui semble sous-optimale, un travail sur les covering numbers en $L_2(P_n)$ serait peut-être plus adapté), et pour des activations RELU on a plutôt du $kD\log(N)$, où k est le nombre de couches Bartlett et al. [2019].

Preuve du Théorème 2.27. On aura besoin d'un petit lemme pour le shatter coefficient d'une composition de fonctions :

Lemme 2.28

Si $\mathcal{F}_1 \subset \mathcal{Y}_1^{\mathcal{Y}_0}$, et $\mathcal{F}_2 \subset \mathcal{Y}_2^{\mathcal{Y}_1}$, et, pour $j \in \{1,2\}$,

$$S_{\mathcal{F}_j}(n) := \max_{z_1,\dots,z_n\in\mathcal{Y}^{j-1}} |\mathcal{F}_j(z_{1:n})|,$$

alors, en notant $\mathcal{F} = \mathcal{F}_2 \circ \mathcal{F}_1$, on a

$$S_{\mathcal{F}}(n) \leq S_{\mathcal{F}_1}(n) S_{\mathcal{F}_2}(n).$$

Preuve du Lemme 2.28. Pour $y_{1:n} \in \mathcal{Y}_1^n$ fixé, on a

$$|\mathcal{F}_2(y_{1:n})| \leq S_{\mathcal{F}_2}(n).$$

Donc, pour $x_{1:n} \in \mathcal{Y}_0^n$

$$\begin{aligned} |\mathcal{F}_2 \circ \mathcal{F}_1(x_{1:n})| &\leq \sum_{y_{1:n}\in\mathcal{F}_1(x_{1:n})} |\mathcal{F}_2(y_{1:n})| \\ &\leq \sum_{y_{1:n}\in\mathcal{F}_1(x_{1:n})} S_{\mathcal{F}_2}(n) \\ &\leq S_{\mathcal{F}_1}(n) S_{\mathcal{F}_2}(n). \end{aligned}$$

On procède maintenant par récurrence : si $\mathcal{N}^{\leq j}$ désigne le réseau de neurones jusqu'à la couche j (à sorties dans $\{0,1\}^{d_j}$ donc), on a $\mathcal{N}^{\leq j} = \mathcal{C}_j \circ \mathcal{N}^{\leq j-1}$, avec

$$\mathcal{C}_j : \begin{cases} \{0,1\}^{d_{j-1}} & \to & \{0,1\}^{d_j} \\ x^{(j-1)} & \mapsto & (N_{j,1}(x^{(j-1))}, \dots, N_{j,d_j}(x^{(j-1)})), \end{cases}$$

ce dont on déduit

$$S_{\mathcal{N}^{\leq j}}(n) \leq S_{\mathcal{N}^{\leq j-1}}(n) S_{\mathcal{C}_j}(n),$$

et

$$S_{\mathcal{N}}(n) \leq \prod_{j=1}^{k+1} S_{\mathcal{C}_j}(n).$$

Maintenant, on peut aussi facilement montrer que, si $\mathcal{F} = (\mathcal{F}_1, \mathcal{F}_2)$, alors

$$S_{\mathcal{F}}(n) \leq S_{\mathcal{F}_1}(n) S_{\mathcal{F}_2}(n),$$

de sorte que, à j fixé,

$$S_{\mathcal{C}_j}(n) \leq \prod_{i=1}^{d_j} S_{N_{j,i}}(n).$$

Enfin, le neurone $N_{i,j}$ étant de dimension de Vapnik $\omega_{j,i}$, une application du lemme 2.17 donne, pour $n \geq d_{VC}(\mathcal{N})$,

$$S_{\mathcal{N}}(n) \leq \prod_{p \in \mathcal{N}} \left(\frac{en}{\omega_p} \right)^{\omega_p}.$$

On conclut comme dans la preuve du corollaire 2.26. Pour $n = d_{VC}(\mathcal{N})$, on a

$$2^n \leq \prod_{p \in \mathcal{N}} \left(\frac{en}{\omega_p} \right)^{\omega_p},$$

ce qui, en passant au log donne

$$\begin{aligned} n \log(2) &\leq \left(\sum_{p \in \mathcal{N}} \omega_p \right) (1 + \log(n)) - \sum_{p \in \mathcal{N}} \omega_p \log(\omega_p) \\ &\leq D \left[1 + \log(n) - \sum_{p \in \mathcal{N}} q_p \log(q_p) - \log(D) \right], \end{aligned}$$

avec $q_p = \frac{\omega_p}{D}$. En reconnaissant une entropie, on a

$$- \sum_{p \in \mathcal{N}} q_p \log(q_p) \leq \log(N),$$

et donc

$$n \log(2) \leq D \left[1 + \log(n) + \log(N) - \log(D) \right].$$

Posons $n = CD \log(N)$ (on se place pour $N \geq 3$, $N = 1, 2$ se traitent à part), en utilisant $\log(\log(N)) \leq \log(N) - 1$ et $\log(N) \geq \log(3)$, on a

$$\begin{aligned} C \log(N) &\leq [1 + \log(C) + \log(\log(N)) + \log(N)] \\ &\leq 2 \log(N) + \log(C), \end{aligned}$$

où encore

$$C \leq 2 + \frac{\log(C)}{\log(3)},$$

ce qui n'est possible que pour $C \leq c_0$ (c_0 constante absolue). □

□

reglin standard (dimension), "généralisée" (ex histogrammes, splines).

2.2.2 Régression linéaire moindres carrés

On se place ici dans le problème de régression, avec $\mathcal{Y} = \mathbb{R}$, avec pour fonction de perte

$$c(y', y) = (y' - y)^2.$$

Pour des variables prédictives dans $\mathbb{R}^d$ on va chercher des prédicteurs linéaires, c'est à dire de la forme

$$f_\theta : \begin{cases} \mathbb{R}^d & \to & \mathbb{R}, \\ x & \mapsto & \langle \theta, x \rangle . \end{cases}$$

Un bref rappel sur le fait que les prédicteurs linéaires ne sont pas si frustes que cela : pour des variables originales $z_1, \dots, z_{d_1}$, si on note $\psi_1, \dots, \psi_d$ des fonctions de $\mathbb{R}^{d_1}$ à valeurs dans $\mathbb{R}$, les classifieurs de type

$$z_{1:d_1} \mapsto \sum_{j=1}^{d} \theta_j \psi_j(z_{1:d_1})$$

sont bien des régresseurs linéaires, de la forme

$$x \mapsto \langle \theta, x \rangle ,$$

avec $x = (\psi_1(z_{1:d_1}), \dots, \psi_d(z_{1:d_1}))^T$.

Exemple 2.29 *: Histogrammes.*

Pour $\mathcal{Z} = [0, 1]$, $A_1, \dots, A_d$ une partition de $[0, 1]$, en notant $x = (\mathbb{1}_{A_1}(z), \dots, \mathbb{1}_{A_d}(z))^T$, un prédicteur par histogrammes, de la forme

$$\begin{aligned} f_\theta(z) &= \sum_{j=1}^{d} \theta_j \mathbb{1}_{A_j}(z) \\ &= \langle \theta, x \rangle , \end{aligned}$$

est linéaire en x. On peut complexifier la chose en remplaçant une prédiction constante sur les éléments de la partition (histogrammes) par des prédictions polynomiales. Pour $s_1, \dots, s_d \in \mathbb{N}^d$ (degrés des polynômes sur la partition), on note, pour $1 \leq j \leq d$ et $0 \leq k \leq s_j$,

$$\psi_{j,k}(z) = z^k \mathbb{1}_{A_j}(z),$$

de sorte que un estimateur polynômial par morceaux se met sous la forme

$$x \mapsto \langle \theta, x \rangle ,$$

avec $x = (\psi_{j,k}(z))_{1 \leq j \leq d, 0 \leq k \leq s_j} \in \mathbb{R}^D$, avec $D = \sum_{j=1}^{d} (s_j + 1)$.

Si on veut assurer une certaine régularité à un prédicteur polynomial par morceaux, plutôt que les $\psi_{j,k}$ on peut considérer une base de splines adaptée à la partition. Par exemple, si on veut un polynôme par morceaux de degré (uniforme) 3 et continu au global, $1, z, z^2, z^3, ((z - z_j)_+, (z - z_j)_+^2, (z - z_j)_+^3)_{j=1,\dots,d}$ est une base adaptée (avec $A_j = [z_j, z_{j+1}[$).

Les régresseurs linéaires peuvent donc être des méthodes d'approximation précise de $\eta(x) = \mathbb{E}(Y \mid x)$, pour peu que l'on intuite une forme a priori de η et choisisse une base adaptée. Une fois une telle base choisie, on se ramène au cas $x \mapsto \langle \theta, x \rangle$ (on aura compris que c'est le choix du x qui est crucial dans cette affaire de modélisation), et on trouve un $\hat{\theta}$ en minimisant un risque empirique associé à la perte quadratique : d'où l'appellation régression linéaire par moindre carrés. On va décrire trois situations, par ordre croissant de difficulté technique.

Design fixe - Modèle bien spécifié

Dans le cadre d'un design fixe, les variables prédictives $x_1, \dots, x_n$ sont considérées comme fixes (ça peut être le cas par exemple lorsque l'on peut les choisir à l'avance dans le cadre d'une expérience). On suppose alors que l'on observe $Y \in \mathbb{R}^n$, vecteur de coordonnées indépendantes, de moyenne μ, et de variances par coordonnées $(\sigma_i^2)_{i=1,\dots,n}$.

Le régresseur linéaire associé à θ s'écrit $(\langle \theta, x_i \rangle)_{i=1,\dots,n}$, ou encore $X\theta$, où X est la matrice $n \times d$, de lignes les x_i.

Le modèle est **bien spécifié** s'il existe θ^* tel que $\mu = X\theta^*$, c'est à dire si on peut écrire

$$Y = X\theta^* + \varepsilon,$$

où $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$. On parle aussi de **modèle linéaire**. Ce modèle est dit **homoscédastique** lorsque, pour tout $1 \leq i \leq n$, $\sigma_i^2 = \sigma^2$ (dans le cas contraire il est dit hétéroscédastique). Lorsque de plus ε est supposé Gaussien (vecteur), on parle de modèle linéaire **Gaussien**. Ce dernier cas est le plus favorable pour l'estimation de θ^* (cas X^TX inversible, nécessairement $d \leq n$).

Dans cette partie on supposera le modèle bien spécifié, et $\sigma_i^2 \leq \sigma^2$, pour tout i. On note $\mathcal{P}_{V(X)}(\sigma^2)$ l'ensemble des lois (de Y) satisfaisant ces hypothèses.

A un prédicteur f_θ on a la fonction de coût $c(f_\theta(X), Y) = \|Y - X\theta\|^2$, la fonction de risque associée

$$\begin{aligned} R(\theta) &= \mathbb{E}\|Y_{new} - X\theta\|^2 = \mathbb{E}\|X(\theta^* - \theta) + \varepsilon_{new}\|^2 \\ &= \|X(\theta - \theta^*)\|^2 + \mathbb{E}(\|\varepsilon_{new}\|^2), \end{aligned}$$

pour laquelle on peut vérifier que le θ^* du modèle est bien un minimiseur. La perte en θ vaut alors

$$\ell(\theta, \theta^*) = \|X(\theta - \theta^*)\|^2.$$

La régression moindres carrés consiste à minimiser le risque empirique

$$R_n(\theta) = \|Y - X\theta\|^2,$$

ce qui, dans le cas X^TX inversible (ce que l'on supposera ici) donne

$$\hat{\theta} = (X^TX)^{-1}X^TY,$$

d'excès de risque

$$\begin{aligned} \ell(\hat{\theta}, \theta^*) &= \|X(X^TX)^{-1}X^TY - \theta^*\|^2 \\ &= \|X(X^TX)^{-1}X^T\varepsilon\|^2, \end{aligned}$$

ce dont on déduit

$$\begin{aligned} \mathbb{E}\ell(\hat{\theta}, \theta^*) &= \mathbb{E}\left(\varepsilon^TX(X^TX)^{-1}X^T\varepsilon\right) \\ &= \sum_{i=1}^n \sigma_i^2 (X(X^TX)^{-1}X^T)_{i,i} \\ &= \sum_{i=1}^n \sigma_i^2 \|X(X^TX)^{-1/2}_{i,\cdot}\|^2 \\ &\leq \sigma^2 \sum_{i=1}^n \|X(X^TX)^{-1/2}_{i,\cdot}\|^2 \\ &= \sigma^2 \mathrm{Tr}(X(X^TX)^{-1}X^T) = d\sigma^2, \end{aligned}$$

en reconnaissant que $X(X^TX)^{-1}X^T$ est la matrice de projection sur $V(X)$. On en déduit

$$\inf_{\hat{f}} \sup_{P \in \mathcal{P}_{V(X)}(\sigma^2)} \mathbb{E}\ell(\hat{f}, f_{\theta^*}) \leq d\sigma^2.$$

On peut montrer avec des arguments bayésiens que

$$\inf_{\hat{f}} \sup_{P \in \mathcal{P}_{V(X)}(\sigma^2)} \mathbb{E}\ell(\hat{f}, f_{\theta^*}) \geq d\sigma^2,$$

de sorte que la régression moindre carré est optimale du point de vue minimax dans ce cas. La dimension naturelle $\dim(V(X) = d$ (rappelons encore qu'ici X^TX est supposée inversible) est donc bien la dimension au sens statistique du terme. Le fait qu'elle n'intervienne pas en $\sqrt{d}$ provient du fait qu'ici on a plutôt une vitesse rapide, ce point sera plus clair dans ce qui suit.

Design aléatoire - Modèle bien spécifié

On se place dans le même modèle que précédemment, à savoir que

$$y = \langle x, \theta^* \rangle + \varepsilon,$$

mais cette fois-ci x est supposé aléatoire, avec $\mathbb{E}(\|x\|^2) < +\infty$, $m(x) = \mathbb{E}(\varepsilon \mid x) = 0$, et $\sigma^2(x) = \mathbb{E}(\varepsilon^2 \mid x) \leq \sigma^2$ (on ne demandera pas nécessairement l'indépendance de x et ε). Pour des raisons que l'on va tout de suite comprendre on va demander à ce que P_x soit non-dégénérée, c'est à dire $\Sigma = E(xx^T) \succ 0$, ou intuitivement que P_x est bien de dimension d. On notera $\mathcal{P}_{well}(\sigma^2)$ l'ensemble de telles lois sur $\mathbb{R}^d \times \mathbb{R}$.

Le risque d'un classifieur f quelconque redevient alors le classique

$$E_{x,y}(f(x) - y)^2 = E_x((f(x) - \langle x, \theta^* \rangle)^2) + E_x(\sigma^2(x)).$$

On constate alors que f^*_θ est bien un minimiseur, et, que pour un θ quelconque,

$$\begin{aligned} \ell(\theta, \theta^*) &= E_x(\langle x, (\theta - \theta^*) \rangle)^2 \\ &= \|\Sigma^{1/2}(\theta - \theta^*)\|^2, \end{aligned}$$

où on rappelle que $\Sigma = E_x(xx^t)$.

On suppose que l'on observe (x_i, y_i) i.i.d. de loi $P \in \mathcal{P}_{well}(\sigma^2)$, et, avec les mêmes notations qu'à la section précédente, le risque empirique pour un prédicteur f_θ s'écrit

$$\begin{aligned} R_n(\theta) &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle x_i, \theta \rangle)^2 \\ &= \frac{1}{n} \|Y - X\theta\|^2. \end{aligned}$$

Ce qui fait que l'estimateur par moindres carrés s'écrit encore

$$\hat{\theta} = (X^TX)^{-1}X^TY,$$

où là encore X^TX est supposée inversible (ce qui arrive ici presque sûrement si $d \leq n$ et P_x est non dégénérée). L'espérance de la perte en $\hat{\theta}$ a cette fois-ci une expression

un peu plus compliquée :

$$\begin{aligned}\mathbb{E}\ell(\hat{\theta},\theta^*) &= \mathbb{E}\|\Sigma^{1/2}(\hat{\theta}-\theta^*)\|^2 \\ &= \mathbb{E}\|\Sigma^{1/2}(X^TX)^{-1}X^T\varepsilon\|^2 \\ &= \mathbb{E}\sum_{i=1}^n \sigma_i^2\left[X(X^TX)^{-1}\Sigma(X^TX)^{-1}X^T\right]_{i,i} \\ &= \mathbb{E}\sum_{i=1}^n \sigma_i^2\left\|(X(X^TX)^{-1}\Sigma^{1/2})_{i,\cdot}\right\|^2 \\ &\leq \frac{\sigma^2}{n}\mathbb{E}\mathrm{Tr}\left(\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2}\right),\end{aligned}$$

avec $\hat{\Sigma} = \frac{1}{n}X^TX$. Par ailleurs, on peut prouver que, à P_x fixée,

$$\inf_{\hat{f}}\sup_{P\in\mathcal{P}_{well}(\sigma^2),P_x}\mathbb{E}\ell(\hat{f},f_{\theta^*}) \geq \sigma^2\mathbb{E}\mathrm{Tr}\left(\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2}\right),$$

ce qui montre que les moindres carrés restent optimaux dans cette situation (P_x non dégénérée et $d \leq n$). Reste à essayer de comprendre le terme $\mathbb{E}\mathrm{Tr}\left(\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2}\right)$. Moralement, pour n grand (et sous conditions sur P_x), $\hat{\Sigma} \to \Sigma$, et donc ce terme devrait tendre vers d.

Commençons par remarquer que, comme $A \mapsto \mathrm{Tr}(A^{-1})$ est convexe,

$$\mathbb{E}\mathrm{Tr}\left(\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2}\right) \geq \mathrm{Tr}\left(\Sigma(\mathbb{E}(\hat{\Sigma}))^{-1}\right) = d.$$

On a donc $\mathbb{E}\ell(\hat{\theta},\theta^*) \geq d/n$. Pour obtenir une majoration, c'est plus compliqué, et cela dépend de la vitesse à laquelle $\hat{\Sigma}^{-1}$ estime Σ^{-1}. Des conditions ad-hoc font intervenir d'une part la proximité de Σ à 0 (via $\lambda_{\min}(\Sigma)$ par exemple), et d'autre part des conditions de concentration pour xx^T (par exemple x sous-Gaussien, voire borné). Voici un exemple de tels hypothèses et résultats.

Théorème 2.30

Soit $P \in \mathcal{P}_{well}(\sigma^2)$ satisfaisant

1. *pour tout $\theta \in \mathbb{R}^d$, $P_x(|\left\langle\Sigma^{1/2}X,\theta\right\rangle| \leq t) \leq (Ct)^\alpha$, pour C et $\alpha \leq 1$ fixés (condition d'étalement minimum de x),*
2. *$\|x\| \leq M$ (variables prédictives bornées).*

Alors, pour n assez grand, on a

$$\mathbb{E}\ell(\hat{\theta},\theta^*) \leq \frac{\sigma^2 d}{n}\left(1+C'\frac{d}{n}\right).$$

On peut trouver une preuve de ce résultat dans Mourtada [2022], avec des conditions plus faibles que x bornée. Ce qu'on peut retenir dès à présent est que les moindres carrés sont optimaux quand $d \leq n$ et que $\hat{\Sigma}$ estime à peu près bien Σ, ce qui ne sera pas forcément le cas en grande dimension. Enfin, ces résultats ont été obtenus en regardant explicitement la forme de $\hat{\theta}$, et non en suivant l'autoroute de preuve pour les minimiseurs de risque empirique. On peut arriver à des résultats similaires par cette voie, on en donne un exemple dans ce qui suit.

Design aléatoire - Modèle mal spécifié

On suppose maintenant que le modèle est mal spécifié, c'est à dire que

$$\eta(x) = \mathbb{E}(y \mid x) \notin V(x).$$

En notant $\theta^* = \arg\min_\theta R(\theta)$, on a

$$f_{\theta^*} = \langle \theta^*, x \rangle = \pi_{\mathcal{L}(V(x))}(\eta),$$

où la projection est au sens $L_2(P_x)$, de sorte que

$$y = \langle \theta^*, x \rangle + m(x) + \varepsilon,$$

avec $\mathbb{E}(\varepsilon \mid x) = 0$, $\mathbb{E}(\varepsilon^2 \mid x) = \sigma^2(x) \leq \sigma^2$, $\mathbb{E}(xm(x)) = 0$. Pour un θ candidat, on a encore

$$\begin{aligned}
\ell(\theta, \theta^*) &= E_{x,y}\left[(y - \langle x, \theta \rangle)^2 - (y - \langle x, \theta^* \rangle)^2\right] \\
&= E_{x,\varepsilon}\left[(\varepsilon + m(x) + \langle x, \theta^* - \theta \rangle)^2 - (\varepsilon + m(x))^2\right] \\
&= E_x \langle x, \theta - \theta^* \rangle^2 + 2E_{x,\varepsilon} \langle (\varepsilon + m(x))x, \theta^* - \theta \rangle \\
&= E_x \langle x, \theta - \theta^* \rangle^2 + 2E_{x,\varepsilon} \langle (\mathbb{E}(\varepsilon + m(x))x \mid x), \theta^* - \theta \rangle \\
&= E_x \langle x, \theta - \theta^* \rangle^2 + 2E_{x,\varepsilon} \langle m(x)x, \theta^* - \theta \rangle \\
&= E_x \langle x, \theta - \theta^* \rangle^2 = \|\Sigma^{1/2}(\theta - \theta^*)\|^2.
\end{aligned}$$

Le prédicteur moindre carrés $\hat{\theta}$ a la même expression qu'avant. Son excès de risque peut se contrôler de la façon suivante, sans faire intervenir son expression spécifique (seulement en utilisant le fait qu'il minimise R_n).

Théorème 2.31

Supposons que $\|\Sigma^{-1/2}x\| \leq M$ P_x *p.s.. Alors*

$$\mathbb{E}\ell(\hat{\theta}, \theta^*) \leq \left(1 - \frac{8M^2}{\sqrt{n}}\right)^{-1} \frac{16M^2}{n} E_{x,y}\left(y - \langle x, \theta^* \rangle\right)^2.$$

On remarque qu'on retombe sur les ordres de grandeurs du Théorème 2.30 dans le cas bien spécifié : dans ce cas le terme $E_{x,y}\left(y - \langle x, \theta^* \rangle\right)^2$ vaut σ^2, et, si les coordonnées de x sont bornées par M_∞ (version fortes de coordonées sous Gaussiennes, ou avec hypothèses de queues faibles), on a $M^2 \leq \frac{dM_\infty^2}{\lambda_{min}(\Sigma)}$. On peut aussi remarquer que les constantes devant le $d\sigma^2/n$ ne permettent pas de retrouver le facteur 1, ce qui est assez caractéristique des méthodes de preuve par déviation entre mesure et mesure empirique, cela dit on n'a pas eu besoin d'hypothèse d'étalement minimum de x (là aussi c'est courant avec ces méthodes). En résumé, l'autoroute de preuve donne souvent le bon ordre de grandeur à moindres frais, avec des constantes sous-optimales la plupart du temps comparés à des méthodes où l'expression du prédicteur est prise en compte.

Preuve du Théorème 2.31. On travaille dans un premier temps à D_n fixé, et on note $\eta = m(x) + \varepsilon$. On commence le même raisonnement qu'en Section 2.1.1.

$$\begin{aligned}
\ell(\hat{\theta}, \theta^*) &= P_{(x,y)} \left[\left(y - \left\langle x, \hat{\theta} \right\rangle \right)^2 - (y - \langle x, \theta^* \rangle)^2 \right] \\
&\leq (P - P_n) \left[\left(y - \left\langle x, \hat{\theta} \right\rangle \right)^2 - (y - \langle x, \theta^* \rangle)^2 \right] \\
&= (P - P_n) \left[\left(\langle x, \theta^* \rangle + \eta - \left\langle x, \hat{\theta} \right\rangle \right)^2 - \eta^2 \right] \\
&= (P - P_n) \left\langle x, \hat{\theta} - \theta^* \right\rangle^2 + (P - P_n) 2\eta \left\langle x, \theta^* - \hat{\theta} \right\rangle . \qquad (2.5)
\end{aligned}$$

On voit alors que contrôler $\sup_\theta (P - P_n) \langle x, \theta^* - \theta \rangle$ risque de poser problème. L'idée générale est de *renormaliser* ces déviations, en les comparant à $\ell(\hat{\theta}, \theta^*) = \|\Sigma^{1/2}(\hat{\theta} - \theta^*)\|^2$. Pour le second terme, on écrit

$$(P - P_n) 2\eta \left\langle x, \theta^* - \hat{\theta} \right\rangle \leq \|\Sigma^{1/2}(\hat{\theta} - \theta^*)\| \sup_{\|u\| \leq 1} \left| (P - P_n) 2\eta \left\langle \Sigma^{-1/2} x, u \right\rangle \right| .$$

Et, en utilisant le principe de symétrisation,

$$\begin{aligned}
\sup_{\|u\| \leq 1} \left| (P - P_n) 2\eta \left\langle \Sigma^{-1/2} x, u \right\rangle \right| &\leq 2E_\delta \sup_{\|u\| \leq 1} \left| \frac{1}{n} \sum_{i=1}^n 2\delta_i \eta_i \left\langle \Sigma^{-1/2} X_i, u \right\rangle \right| \\
&\leq \frac{4}{n} E_\delta \sup_{\|u\| \leq 1} \left| \left\langle \sum_{i=1}^n \delta_i \eta_i \Sigma^{-1/2} X_i, u \right\rangle \right| \\
&\leq \frac{4}{n} E_\delta \left\| \sum_{i=1}^n \delta_i \eta_i \Sigma^{-1/2} X_i \right\| \\
&\leq \frac{4}{n} \left(E_\delta \left\| \sum_{i=1}^n \delta_i \eta_i \Sigma^{-1/2} X_i \right\|^2 \right)^{1/2} \\
&\leq \frac{4M}{n} \left(\sum_{i=1}^n \eta_i^2 \right)^{1/2} .
\end{aligned}$$

On peut maintenant en déduire

$$\begin{aligned}
\mathbb{E}(P - P_n) 2\eta \left\langle x, \theta^* - \hat{\theta} \right\rangle &\leq \frac{1}{2} \mathbb{E} \ell(\hat{\theta}, \theta^*) + \frac{1}{2} \frac{16M^2}{n^2} \mathbb{E} \left(\sum_{i=1}^n \eta_i^2 \right) \\
&\leq \frac{1}{2} \mathbb{E} \ell(\hat{\theta}, \theta^*) + \frac{8M^2}{n} E_{x,y} (y - \langle x, \theta^* \rangle)^2 . \qquad (2.6)
\end{aligned}$$

On a utilisé $ab \leq \frac{1}{2}(a^2 + b^2)$, mais on aurait aussi pu réarranger différemment avec du $\frac{1}{2}\varepsilon a^2 + \frac{1}{2}\varepsilon^{-1} b^2$, au prix d'un peu plus de lourdeur calculatoire.

Passons au premier terme. On a

$$(P - P_n) \left\langle x, \hat{\theta} - \theta^* \right\rangle^2 \leq \|\Sigma^{1/2}(\hat{\theta} - \theta^*)\|^2 \sup_{\|u\| \leq 1} (P - P_n) \left\langle \Sigma^{-1/2} x, u \right\rangle^2 .$$

En utilisant symétrisation et contraction ($x \mapsto x^2$ est $2M$-Lipschitz sur $[-M, M]$),

on obtient

$$\begin{aligned}
\sup_{\|u\|\leq 1} (P - P_n) \left\langle \Sigma^{-1/2} x, u \right\rangle^2 &\leq 2E_\delta \sup_{\|u\|\leq 1} \frac{1}{n} \sum_{i=1}^n \delta_i \left\langle \Sigma^{-1/2} X_i, u \right\rangle^2 \\
&\leq 4ME_\delta \sup_{\|u\|\leq 1} \left| \frac{1}{n} \sum_{i=1}^n \delta_i \left\langle \Sigma^{-1/2} X_i, u \right\rangle \right| \\
&\leq 4ME_\delta \left\| \frac{1}{n} \sum_{i=1}^n \delta_i \Sigma^{-1/2} X_i \right\| \\
&\leq \frac{4M}{n} \left(E_\delta \left\| \sum_{i=1}^n \delta_i \Sigma^{-1/2} X_i \right\|^2 \right)^{1/2} \\
&\leq \frac{4M^2}{\sqrt{n}},
\end{aligned}$$

ce dont on déduit

$$\mathbb{E}(P - P_n) \left\langle x, \hat{\theta} - \theta^* \right\rangle^2 \leq \frac{4M^2}{\sqrt{n}} \mathbb{E}\ell(\hat{\theta}, \theta^*). \tag{2.7}$$

En utilisant (2.6) et (2.7) dans (2.5), on obtient

$$\frac{1}{2}\mathbb{E}\ell(\hat{\theta}, \theta^*) \leq \frac{4M^2}{\sqrt{n}} \mathbb{E}\ell(\hat{\theta}, \theta^*) + \frac{8M^2}{n} E_{x,y}(y - \langle x, \theta^* \rangle)^2,$$

ce dont on déduit le résultat. □

2.3 ERM en pratique

Dans cette section on s'intéresse aux méthodes pour trouver des solutions effectives au problème de minimisation du risque empirique. Dans le cadre de la régression moindre carrés c'est plutôt facile, aussi on se concentrera sur le problème de classification binaire. Pour des raisons techniques on supposera que le label $Y \in \{-1, 1\}$ plutôt que $\{0, 1\}$.

2.3.1 SVM

Pour cette partie on aura besoin de quelques résultats d'optimisation, que l'on ne prouvera pas (si cela vous intéresse, vous êtes renvoyés à Boyd and Vandenberghe [2004]). Mettons que l'on cherche à résoudre le problème suivant

$$\min_{x \in \mathcal{C}} f(x), \qquad \mathcal{C} = \{h_i(x) = 0, g_j(x) \leq 0, i = 1, \ldots, m, j = 1, \ldots, p\}, \tag{P}$$

où f, h_i, g_j sont des fonctions, et $\mathcal{C}$ un ensemble de contraintes sur $x \in \mathbb{R}^d$. Ce problème est appelé *problème primal*. En introduisant des paramètres λ_i, $i = 1, \ldots, m$ dans $\mathbb{R}$ et $\mu_j \geq 0$, $j = 1, \ldots, p$, ce problème est équivalent à

$$\min_{x \in \mathbb{R}^d} \max_{\lambda_i, \mu_j \geq 0} f(x) + \sum_{i=1}^m \lambda_i h_i(x) + \sum_{j=1}^p \mu_j g_j(x) := \min_{x \in \mathbb{R}^d} \max_{\lambda_i, \mu_j \geq 0} \mathcal{L}(x, \lambda, \mu), \tag{P'}$$

qui lui est un problème avec contraintes linéaires. Tout le jeu consiste maintenant à intervertir min et max. Pour λ, μ tel que $\mu_j \geq 0$, on définit

$$g(\lambda, \mu) = \min_{x \in \mathbb{R}^d} \mathcal{L}(x, \lambda, \mu).$$

Le problème dual consiste alors à trouver (λ^*, μ^*) avec $\mu_j^* \geq 0$ résolvant

$$\max_{\lambda_i, \mu_j \geq 0} g(\lambda, \mu) = \max_{\lambda_i, \mu_j \geq 0} \min_{x \in \mathbb{R}^d} \mathcal{L}(x, \lambda, \mu), \tag{D}$$

qui est appelé *problème dual*, souvent plus facile à résoudre. Si x^* est solution du problème primal, et (λ^*, μ^*) solution du problème dual, on a

$$g(\lambda^*, \mu^*) \leq \mathcal{L}(x^*, \lambda^*, \mu^*) \leq \max_{\lambda_i, \mu_j \geq 0} \mathcal{L}(x^*, \lambda, \mu) = f(x^*),$$

de sorte que $(D) \leq (P)$ usuellement. On dit qu'il y a *dualité forte* lorsque cette inégalité devient une égalité. Il y a une palanquée de conditions suffisantes pour la dualité forte. Citons-en au moins une.

Théorème 2.32 : Conditions de Slater

On appelle conditions de Slater les hypothèses suivantes :

1. *f est convexe,*
2. *g_j est convexe pour tout j,*
3. *h_i est affine pour tout i,*
4. *il existe $x \in \mathbb{R}^d$ satisfaisant*

$$\begin{aligned} &\forall 1 \leq i \leq m \quad h_i(x) = 0, \\ &\forall 1 \leq j \leq p \quad g_j(x) < 0. \end{aligned}$$

Si ces conditions sont vérifiées, alors il y a dualité forte.

Une fois la dualité forte acquise, il reste à résoudre (D). On peut faire cela au moyen des conditions de Karush-Kuhn-Tucker (extension de la CNS $\nabla_{x^*} f = 0$ au cadre avec contraintes).

Théorème 2.33 : Karush-Kuhn-Tucker

Sous les conditions de Slater, (x^*, λ^*, μ^*) sont solutions de (P) (ou (D)) ssi les conditions suivantes sont vérifiées :

— ***Stationnarité*** *:*

$$0 \in \partial_{x^*} f + \sum_{i=1}^{m} \lambda_i^* \nabla_{x^*} h_i + \sum_{j=1}^{p} \mu_j^* \partial_{x^*} g_j,$$

où $\partial_y F$ désigne le sous-gradient d'une fonction F convexe.

— ***Faisabilité primale*** *:*

$$\begin{aligned} &\forall 1 \leq i \leq m \quad h_i(x^*) = 0, \\ &\forall 1 \leq j \leq p \quad g_j(x^*) \leq 0. \end{aligned}$$

— ***Faisabilité duale*** *:*

$$\forall 1 \leq j \leq p \quad \mu_j^* \geq 0$$

— ***Complementary slackness*** *:*

$$\forall 1 \leq j \leq p \quad g_j(x^*)\mu_j^* = 0.$$

Rappelons ici que le sous-gradient en x d'une fonction f convexe, $\partial_x f$, est défini par l'ensemble des directions h telles que pour tout y

$$f(y) \geq f(x) + \langle h, (y - x) \rangle .$$

La dernière condition (complementary slackness) peut se traduire en "les coefficients μ_j non nuls correspondent aux contraintes saturées". En fait, un énoncé plus précis serait : "si un x^* et λ^*, μ^* satisfont les conditions KKT alors il y a dualité forte et le triplet est solution", et "sous conditions de dualité forte (comme Slater), un triplet optimal vérifie les conditions KKT". Pour une approche plus subtile, voyez Boyd and Vandenberghe [2004].

Cas linéairement séparable

On considère les classifieurs linéaires, du type $x \mapsto \text{sg}(f_{\beta,\beta_0}(x))$, avec $f_{\beta,\beta_0}(x) = \langle \beta, x \rangle + \beta_0$, et on suppose que les classes sont linéairement séparable, c'est à dire $Y_i f_{\beta,\beta_0}(X_i) > 0$ pour tout i, ou encore $R_n(f_{\beta,\beta_0}) = 0$, pour au moins un (β, β_0). Le but va être alors de trouver un hyperplan qui sépare au mieux les données, au sens de la *marge*, définie, pour un (β, β_0) séparateur, par

$$M = \min_{i=1,\dots,n} \frac{Y_i(\langle \beta, X_i \rangle + \beta_0)}{\|\beta\|}.$$

FAIRE DESSIN. Notre problème peut donc s'écrire

$$\begin{aligned} &\max_{\beta,\beta_0} M \\ &\text{Contraintes} : \forall 1 \leq i \leq n \quad \frac{Y_i(\langle \beta, X_i \rangle + \beta_0)}{\|\beta\|} \geq M. \end{aligned}$$

Comme β peut être transformé en $\lambda\beta$ sans changer l'hyperplan (quitte à changer la paramétrisation du β_0), on peut choisir, pour un hyperplan donné, de poser $\|\beta\| = 1/M$, de sorte que le problème se réécrit

$$\begin{aligned}&\max_{\beta,\beta_0} \frac{1}{\|\beta\|}\\ &\text{Contraintes : } \forall 1 \leq i \leq n \quad Y_i(\langle \beta, X_i\rangle + \beta_0) \geq 1,\end{aligned}$$

ou encore

$$\begin{aligned}&\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2\\ &\text{Contraintes : } \forall 1 \leq i \leq n \quad Y_i(\langle \beta, X_i\rangle + \beta_0) \geq 1, \qquad \text{(P)}\end{aligned}$$

ce qui va constituer notre problème primal. On remarque qu'un (β_0^*, β^*) solution de (P) correspond à un choix particulier de minimiseur de risque empirique. Pour être un peu plus précis, on peut montrer qu'elle minimise

$$C\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i f_{\beta,\beta_0}(X_i)-1<0} + \frac{1}{2}\|\beta\|^2,$$

pour $C \geq n\|\beta^*\|^2/2$, ou encore

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i f_{\beta,\beta_0}(X_i)-1<0} + \lambda\|\beta\|^2,$$

pour $\lambda \leq 2/(n\|\beta^*\|^2)$. Cette formulation exprime le classifieur SVM comme solution d'une minimisation de risque empirique *pénalisé*.

Revenons au problème (P), et introduisons le problème dual correspondant

$$\max_{\alpha_i \geq 0} \min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + \sum_{i=1}^{n} \alpha_i(1 - Y_i f_{\beta,\beta_0}(X_i)) := \max_{\alpha_i \geq 0} \min_{\beta,\beta_0} \mathcal{L}(\beta, \beta_0, \alpha). \qquad \text{(D)}$$

Comme $\beta \mapsto \|\beta\|^2$ est convexe, $(\beta, \beta_0) \mapsto (1 - Y_i f_{\beta,\beta_0}(X_i))$ aussi, et qu'il existe β, β_0 tel que, pour tout i, $Y_i f_{\beta,\beta_0}(X_i) > 1$ (prendre un β, β_0 séparateur et le multiplier assez fort), les conditions de Slater sont vérifiées, et on peut donc obtenir les solutions de (P) via les solutions de (D).

Soit $\alpha^*, \beta^*, \beta_0^*$ une solution de (D). On a de manière évidente

$$\begin{aligned}&\forall 1 \leq i \leq n \quad Y_i f_{\beta^*,\beta_0^*}(X_i) \geq 1,\\ &Y_i f_{\beta^*,\beta_0^*}(X_i) > 1 \Rightarrow \alpha_i^* = 0,\end{aligned}$$

et donc on aura toujours

$$(Y_i f_{\beta^*,\beta_0^*}(X_i) - 1)\alpha_i^* = 0.$$

On a redémontré la propriété de "Complementary slackness" des conditions KKT (Théorème 2.33). Les indices tels que $\alpha_i^* > 0$ correspondent aux **vecteurs supports**, et sont ceux placés pile sur la marge. Essayons maintenant de résoudre (D). Pour $\alpha_i \geq 0$, $\mathcal{L}(., \alpha)$ est minimisé pour β satisfaisant

$$\beta = \sum_{i=1}^{n} \alpha_i Y_i X_i. \qquad (2.8)$$

Par ailleurs, si $\sum_{i=1}^n \alpha_i Y_i \neq 0$, $\min_{\beta,\beta_0} \mathcal{L}(\beta, \beta_0, \alpha) = -\infty$. Un optimal vérifie donc

$$\sum_{i=1}^{n} \alpha_i Y_i = 0. \tag{2.9}$$

Une solution α^* de (D) est donc solution de

$$\max_{\sum_{i=1}^n \alpha_i Y_i = 0, \alpha_i \geq 0} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{1 \leq i,j \leq n} \alpha_i \alpha_j Y_i Y_j \langle X_i, X_j \rangle , \tag{(D')}$$

qui est un problème quadratique et peut se résoudre explicitement. Une fois (D') résolu par α^* :

1. les α_i^* donnent les points supports,
2. à partir des points supports, on construit

$$\beta^* = \sum_{\alpha_i^* > 0} \alpha_i^* Y_i X_i,$$

3. on obtient β_0^* à partir d'un point support,

$$\beta_0^* = Y_i - \langle X_i, \beta^* \rangle .$$

Cas non linéairement séparable-relaxation convexe

Si on définit le risque empirique

$$R_n(f_{\beta,\beta_0}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i f_{\beta,\beta_0}(X_i) - 1 < 0},$$

le problème initial que l'on cherchait à résoudre était

$$\begin{aligned} &\min_{\beta} \frac{1}{2} \|\beta\|^2 \\ &\text{Contrainte} : R_n(f_{\beta,\beta_0}) = 0, \end{aligned}$$

qui admet des solutions uniquement dans le cas de données linéairement séparables. Dans le cas de données non séparables, on pourrait être tenté de résoudre les problèmes

$$\begin{aligned} &\min_{\beta} \frac{1}{2} \|\beta\|^2 \\ &\text{Contrainte} : \quad R_n(f_{\beta,\beta_0}) \leq \frac{K}{n}, \end{aligned}$$

puis de choisir le plus petit K possible pour lequel ce problème admet des solutions. Le problème est que la contrainte $R_n(f_{\beta,\beta_0}) \leq \frac{K}{n}$ n'est pas convexe, ce qui rend cette optimisation difficile. Une idée consiste alors à non pas imposer une contrainte sur le nombre de points du mauvais côté de la marge, mais d'imposer une contrainte sur les *écarts à la marge*. On se donne alors $\xi_i \geq 0$ famille d'écarts possibles, et on va essayer de résoudre

$$\begin{aligned} &\min_{\beta} \frac{1}{2} \|\beta\|^2 \\ &\text{Contraintes} : \quad \xi_i \geq 0, Y_i f_{\beta,\beta_0}(X_i) - 1 \geq -\xi_i, \sum_{i=1}^{n} \xi_i \leq K, \end{aligned}$$

où K est un paramètre. FAIRE DESSIN. Ce problème est équivalent au problème primal suivant

$$\begin{aligned} &\min_{\beta,\xi} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i, \\ &\text{Contraintes :} \quad \xi_i \geq 0, Y_i f_{\beta,\beta_0}(X_i) \geq 1-\xi_i, \end{aligned} \tag{P}$$

où la fonction objectif ainsi que les contraintes sont convexes. Le Lagrangien de ce problème est

$$\mathcal{L}(\beta,\beta_0,\xi,\alpha,\mu) = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1-\xi_i - Y_i f_{\beta,\beta_0}(X_i)) - \sum_{i=1}^{n}\mu_i\xi_i.$$

Les conditions de Slater étant vérifiées (prendre $\beta,\beta_0 = 0$ et $\xi_i = 2$), résoudre (P) revient à résoudre

$$\max_{\alpha,\mu\geq 0}\min_{\beta,\beta_0,\xi} \mathcal{L}(\beta,\beta_0,\xi,\alpha,\mu). \tag{D}$$

Les conditions KKT donnent

— **Stationnarité** :

$$\begin{aligned} \beta^* &= \sum_{i=1}^{n}\alpha_i^* Y_i X_i \\ \sum_{i=1}^{n}\alpha_i^* Y_i &= 0 \\ \mu_i^* &= C - \alpha_i^*. \end{aligned}$$

— **Faisabilités duales et primales** :

$$\begin{aligned} &Y_i f_{\beta^*,\beta_0^*}(X_i) \geq 1-\xi_i^*, \qquad \xi_i^* \geq 0 \\ &\alpha_i^* \geq 0 \qquad \mu_i^* \geq 0. \end{aligned}$$

— **Complementary slackness** :

$$\begin{aligned} &\alpha_i^*(Y_i f_{\beta^*,\beta_0^*}(X_i) - (1-\xi_i^*)) = 0 \\ &\xi_i^*\mu_i^* = 0. \end{aligned}$$

En intégrant les conditions de stationnarité dans (D) et en regardant les différentes contraintes, on montre que α^* est solution de

$$\begin{aligned} &\max_{\alpha_i}\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{1\leq i,j\leq n}\alpha_i\alpha_j Y_i Y_j \langle X_i, X_j\rangle \\ &\text{Contraintes :} \quad 0 \leq \alpha_i \leq C, \sum_{i=1}^{n}\alpha_i Y_i = 0, \end{aligned}$$

qui lui peut se résoudre plus facilement. Une fois les α^* obtenus, on revient au classifieur obtenu via

$$\begin{aligned} \beta^* &= \sum_{\alpha_i^*>0}\alpha_i^* Y_i X_i \\ \xi_i^* &= (1 - Y_i f_{\beta^*,\beta_0^*}(X_i)) \quad \text{si } \alpha_i^* > 0. \end{aligned}$$

Les points tels que $\alpha_i^* > 0$ sont appelés *points supports*. Parmi ces points, ceux pour lesquel $\xi_i^* = 0$ sont situés exactement sur la marge (et sont caractérisés par $0 < \alpha_i^* < C$), et peuvent être utilisés pour calculer β_0^*. Remarquons enfin que le problème (P) peut se réécrire

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\frac{1}{n}\sum_{i=1}^{n} \tilde{c}(f_{\beta,\beta_0}(X_i), Y_i),$$

où $\tilde{c}(y', y) = (1 - y'y)_+$ est une fonction de coût convexe en $y' \in \mathbb{R}$. Cette fonction de coût est appelée *Hinge Loss*, et est une relaxation convexe du coût binaire parmi beaucoup d'autres (on verra quelques exemples dans la partie suivante). Le critère à optimiser étant devenu convexe, des méthodes de type descente de gradient sont applicables en général (en dehors du cas particulier des SVM pour lesquels on a une solution exacte). Concluons sur le paramètre C : il traduit la force de "l'attache aux données". Plus il est grand, moins votre classifieurs s'autorise d'écart à la marge, plus vous aurez de points supports (et usuellement plus le temps de calcul sera long). A l'inverse, un C qui tend vers 0 fait tendre β vers 0 (le terme de pénalité en $\|\beta\|^2$ devient dominant).

Kernel SVM

Comme en régression linéaire, pour un espace de variables x de départ dans $\mathbb{R}^d$, on peut construire un espace en envoyant $\mathbb{R}^d$ dans $\mathbb{R}^M$ via une transformation Φ (par exemple en prenant une base de polynômes en les variables originelles). Une séparation linéaire dans $\Phi(\mathbb{R}^d)$ correspond alors à une séparation non-linéaire dans l'espace de départ $\mathbb{R}^d$. On peut donc donner à manger n'importe quelle transformation d'un espace de départ en entrée d'un SVM, pour peu qu'elle ait valeurs dans un espace de Hilbert (requis pour calculer des produits scalaires).

Le point de vue "Kernel SVM" généralise cette idée, en se basant sur la remarque que pour calculer un optimal linéaire dans l'espace transformé on résoud

$$\max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\alpha_i\alpha_j Y_i Y_j \left\langle \Phi(X_i), \Phi(X_j) \right\rangle,$$

où le produit scalaire a lieu dans l'espace de Hilbert $\mathcal{H} = \Phi(\mathbb{R}^d)$. On se rend alors compte que, pour trouver un α^* optimal (et les β_0^*, β^* correspondant), on n'a pas besoin de spécifier Φ, il est juste nécessaire de spécifier

$$\mathbb{K} = (\langle \Phi(X_i), \Phi(X_j) \rangle)_{i,j},$$

soit les produits scalaires entre points de l'échantillon. Pareillement, pour une nouvelle donnée x, pour calculer la prédiction associée on aura uniquement besoin des $\langle \Phi(X_i), \Phi(x) \rangle$. Un *noyau* est alors juste un choix de produit scalaire, c'est à dire une fonction k

$$k : \begin{cases} \mathcal{X} \times \mathcal{X} & \to & \mathbb{R}, \\ (x, x') & \mapsto & k(x, x'). \end{cases}$$

Pour que tout marche bien, k doit être symétrique positive semi-définie (au sens que $(k(x_i, x_j))_{i,j}$ doit l'être pour tout $(x_i)_i$). Quelques noyaux usuels sont :

— Linéaire : $k(x_i, x_j) = \langle x_i, x_j \rangle$, correspondant à un SVM linéaire comme avant.

— Polynôme de degré d : : $k(x_i, x_j) = (1 + \langle x_i, x_j \rangle)^d$.

— Radial : $k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$.

— Réseau de neurones : $k(x_i, x_j) = \tanh(\kappa_1 \langle x_i, x_j \rangle + \kappa_2)$.

Pour $\mathcal{X} = \mathbb{R}^2$ et le noyau polynomial d'ordre 2, on a

$$\begin{aligned} k(x, x') &= (1 + \langle x, x' \rangle)^2 \\ &= 1 + (x_1 x'_1)^2 + (x_2 x'_2)^2 + 2x_1 x'_1 + 2x_2 x'_2 + 2x_1 x_2 x'_1 x'_2. \end{aligned}$$

Si on pose

$$\Phi : \begin{cases} \mathbb{R}^2 & \to & \mathbb{R}^6 \\ (x_1, x_2) & \mapsto & (1, x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2), \end{cases}$$

on a bien $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$. De manière générale on peut quasiment toujours associer à un noyau k (symétrique, positif, semi-défini) une représentation Φ à valeur dans un *Reproductive Kernel Hilbert Space*, le lecteur intéressé est renvoyé au théorème de Moore Aronszajn. Il est quand même plus rapide de directement spécifier le noyau. Comme en régression linéaire, un bon espace de paramètre (i.e. un bon noyau) dépend du problème, il est souvent choisi graphiquement. On peut conclure sur le fait que les SVM (et KSVM) existent aussi en régression, cela revient grosso modo à trouver une solution de

$$\min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2 + \frac{1}{n} \sum_{i=1}^{n} \left(|Y_i - f_{\beta, \beta_0}(X_i)| - \varepsilon \right)_+ ,$$

ou d'autres variantes de fonctions de coût.

2.3.2 Convexification du risque

On se place ici dans le cadre de la classification encore (en régression les critères à optimiser sont plus souvent convexes), où par commodité technique on prend $\mathcal{Y} = \{0, 1\}$. A quelques exceptions près (comme celle des SVM dans le cas séparable), minimiser un risque empirique pour le coût 0/1 est un problème difficile en pratique, ce qui justifie des méthodes de contournement (comme la convexification du risque en SVM).

Une première idée est la suivante : comme tout classifieur peut se mettre sous la forme $\mathrm{sg}(f(x))$, où f est à valeurs réelles, pourquoi ne pas ajuster f avec une autre fonction de coût et prendre son signe à la fin ? Par exemple, si on prend le coût quadratique, cela revient à estimer la fonction de régression $\eta(x) = 2\mathbb{P}(Y = 1 \mid x) - 1$ ($\mathbb{E}(Y \mid x)$ si $Y \in \{-1, 1\}$) via f, puis considérer le classifieur plug-in $\mathrm{sg}(f)$. Dans la suite on prend la convention que la fonction signe vaut 1 si $f(x) \geq 0$, -1 sinon (convention 1 en 0).

Si l'intention paraît louable, ce genre de méthode souffre d'un défaut conceptuel : elle tend à vouloir approcher η dans sa globalité, là ou pour la classification on a surtout besoin de l'estimer finement lorsqu'elle est proche de 0, moins lorsque qu'elle s'en éloigne (en ce sens le problème de classification est moralement plus "facile" que le problème de régression). Par ailleurs, si $\eta > 0$, on devrait payer moins cher en surestimant η qu'en le sous-estimant. On aimerait donc remplace la fonction de coût quadratique par une autre fonction de coût convexe, plus adaptée au problème de classification.

Φ-risque

Le cadre est celui décrit plus haut : on prend $\mathcal{Y} = \{-1, 1\}$, et on regarde les classifieurs de la forme $\mathrm{sg}(f)$, où $f : \mathcal{X} \to \bar{\mathbb{R}}$ (on autorise a priori les valeurs infinies). Rappelons ici la convention que $\mathrm{sg}(0) = 1$. Un tel f est aussi appelé *pseudo-classifieur*.

Soit maintenant $\Phi : \bar{\mathbb{R}} \to [0, +\infty]$ une fonction mesurable, qui peut ne prendre pour valeur $+\infty$ qu'en $\pm\infty$, et idéalement convexe. La fonction de coût associée à Φ est définie par

$$c_\Phi : \begin{cases} \bar{\mathbb{R}} \times \{-1, 1\} & \to & [0, +\infty] \\ (y', y) & \mapsto & \Phi(y'y). \end{cases}$$

Le risque associé à Φ, ou Φ-risque est alors

$$R_P^\Phi(f) = \mathbb{E}(c(f(X)Y)),$$

où (X, Y) est de loi P. Lorsque Φ est décroissante (ce qui est presque toujours le cas), $|f|$ donne une information sur la "certitude" que l'on a sur le label, en plus du label donné par $\mathrm{sg}(f)$, le signe du pseudo-classifieur. Les exemples classiques sont

— **Risque** $0/1$: correspond à $\Phi_{0/1}(u) = \mathbb{1}_{u \leq 0}$, et donc

$$R_P^{\Phi_{0/1}}(f) = P(f(X)Y \leq 0).$$

Le Φ-risque $0/1$ du pseudo-classifieur correspond quasiment au risque binaire en classification du classifieur associé. De fait, si $f : \mathcal{X} \to \bar{\mathbb{R}}$, on a

$$\begin{aligned} R_P^{\Phi_{0/1}}(f) &= P(Yf(X) \leq 0) \\ &= P\left(\{\mathrm{sg}(f) \neq Y\} \cup \{f(X) = 0\}\right), \\ &\geq R_P(\mathrm{sg}(f)), \\ &\leq R_P(\mathrm{sg}(f)) + P(f(X) = 0). \end{aligned}$$

On a donc égalité si $P(f(X) = 0) = 0$.

— **Risque quadratique** : correspond à $\Phi(u) = (u-1)^2$, et

$$R_P^\Phi(f) = E((f(X)Y - 1)^2) = E((f(X) - Y)^2).$$

Cela revient à faire de la classification du point de vue de la régression (pas forcément adapté). Par ailleurs, la fonction Φ n'étant pas décroissante, $|f|$ ne peut pas être interprétée comme une certitude sur le label.

— **Risque charnière/Hinge Loss** : correspond à $\Phi(u) = (1-u)_+ = (1-u) \vee 0$, et au risque

$$R_P^\Phi(f) = E((1 - Yf(X))_+),$$

qui est celui qui intervient dans le risque SVM dans le cas non linéairement séparable. Un intérêt majeur de la hinge loss est qu'un classifieur de Bayes pour le risque $0/1$ est un pseudo-classifieur de Bayes. Ce n'est pas le cas pour les autres pertes classiques.

— **Risque logistique/Logistic loss** : correspond à $\Phi(u) = \log(1 + e^{-u})$, et au risque

$$R_P^\Phi(f) = E\left(\log(1 + \exp(-f(X)Y))\right).$$

C'est le risque associé à la régression logistique (qui en pratique revient à minimiser ce Φ-risque).

— **Risque exponentiel/Exponential loss** : correspond à $\Phi(u) = \exp(-u)$, et au risque

$$R_P^\Phi(f) = E\left(\exp(-f(X)Y)\right).$$

C'est la fonction de coût associée à l'algorithme AdaBoost (algorithme de boosting standard).

A l'exception du Φ-risque 0/1, tous les Φ-risque usuels sont convexes et sont des majorants du Φ-risque 0/1 FAIRE DESSIN. Usuellement, pour un pseudo classifieur f et un tel Φ-risque, on a

$$R_P(\mathrm{sg}(f)) \leq R_P^{\Phi_{0/1}}(f) \leq R_P^\Phi(f),$$

de sorte qu'une garantie sur le Φ-risque du pseudo-classifieur donne une garantie sur le risque du classifieur (et on peut appliquer les méthodes précédemment vues), et par ailleurs minimiser un Φ-risque empirique convexe est un problème plus facile que le précédent risque empirique binaire.

Dorénavant on suppose que Φ est convexe On peut maintenant se poser la question du Φ-risque de Bayes, ou meilleur Φ-risque atteignable par un pseudo-classifieur. On introduit $p(X) = \mathbb{P}(Y = 1 \mid X)$. Pour un pseudo-classifieur f on a

$$\begin{aligned}\mathbb{E}(c_\Phi(f(X), Y) \mid X) &= p(X)\Phi(f(X)) + (1 - p(X))\Phi(-f(X)) \\ &:= C_{p(X)}^\Phi(f(X)).\end{aligned}$$

Un pseudo-classifieur optimal vérifie donc, P_X-p.s.,

$$C_{p(X)}^\Phi(f(X)) = \inf_{\alpha \in \bar{\mathbb{R}}} C_{p(X)}^\Phi(\alpha) := H^\Phi(p(X)).$$

On en déduit le résultat suivant.

Proposition 2.34 : Pseudo-classifieurs de Bayes

Pour une fonction Φ convexe et le Φ-risque correspondant, le Φ-risque de Bayes vaut

$$\inf_{f \in \mathcal{F}} R^\Phi(f) = \mathbb{E}H^\Phi(p(X)).$$

Par ailleurs, f^* est un pseudo-classifieur de Bayes (i.e. $R^\Phi(f^*) = \inf_{f \in \mathcal{F}} R^\Phi(f)$) si et seulement si

$$f^*(X) \in \arg\min_{\alpha \in \bar{\mathbb{R}}} C_{p(X)}^\Phi(\alpha),$$

P_X presque sûrement.

Preuve de la Proposition 2.34. D'après ce qui précède, pour un pseudo-classifieur f, on a

$$\mathbb{E}\left(\Phi(Yf(X)) \mid X\right) = C^{\Phi}_{p(X)}(f(X)) \geq H^{\Phi}(p(X)),$$

et donc

$$\mathbb{E}(\Phi(Yf(X)) \geq \mathbb{E}(H^{\Phi}(p(X)),$$

ce dont on déduit

$$\inf_{f\in\mathcal{F}} R^{\Phi}(f) \geq \mathbb{E}H^{\Phi}(p(X)).$$

Or Φ est convexe, donc C^{Φ}_p l'est aussi pour tout $p \in [0,1]$, on en déduit l'existence de $f^*(X) \in \arg\min_{\alpha\in\bar{\mathbb{R}}} C^{\Phi}_{p(X)}(\alpha)$ (ici ressort l'utilité de travailler sur $\bar{\mathbb{R}}$). Il existe donc f^* tel que $\mathbb{E}(\Phi(Yf^*(X)) = \mathbb{E}(H^{\Phi}(p(X))$, d'où le premier point.

Pour le second point, remarquons que pour un pseudo-classifieur f, on a

$$R^{\Phi}(f) - \inf_{f\in\mathcal{F}} R^{\Phi}(f) = \mathbb{E}\left[C^{\Phi}_{p(X)}(f(X)) - H^{\Phi}(p(X))\right],$$

et donc f^* est un pseudo-classifieur de Bayes si et seulement si $f^*(X) \in \arg\min_{\alpha\in\bar{\mathbb{R}}} C^{\Phi}_{p(X)}(\alpha)$. □

On peut remarquer que lorsque Φ est strictement convexe, $C^{\Phi}_{p(X)}(\alpha)$ admet un unique minimiseur sur $\bar{\mathbb{R}}$, ce qui rend le pseudo-classifieur de Bayes unique lui aussi (P_X presque sûrement).

Exemple 2.35 *: Pseudos-classifieurs optimaux usuels.*

— **Perte quadratique** : Pour $\Phi(u) = (u-1)^2$, on a

$$C^{\Phi}_p(\alpha) = p(\alpha-1)^2 + (1-p)(\alpha+1)^2,$$

qui est minimal en $\alpha^* = 2p-1$, et donc $f^*(X) = \eta(X)$ est le pseudo-classifieur de Bayes.

— **Hinge loss** : Pour $\Phi(u) = (1-u)_+$, on a

$$C^{\Phi}_p(\alpha) = p(1-\alpha)_+ + (1-p)(1+\alpha)_+.$$

On a alors $C^{\Phi}_p(\alpha) = p(1-\alpha)$ pour $\alpha \leq -1$, $C^{\Phi}_p(\alpha) = 1 + (1-2p)\alpha$ pour $\alpha \in [-1,1]$, et $C^{\Phi}_p(\alpha) = (1-p)(1+\alpha)$ pour $\alpha \geq 1$. FAIRE DESSIN. On en déduit que $f^*(X) = \mathrm{sg}(2p(X)-1) = \mathrm{sg}(\eta(X))$ est optimal. Par ailleurs il est unique si $p(X) \notin \{0, (1/2), 1\}$ P_X-p.s. (sinon, lorsque $p(X) = 1/2$, $f^*(X) \in [-1,1]$ suffit à minimiser $C^{\Phi}_{p(X)}$, et pour $p(X) = 0/1$, $f^*(X) \leq -1/\geq 1$ suffit aussi).

— **Risque logistique** : Pour $\Phi(u) = \log(1+e^{-u})$, on a

$$C^{\Phi}_p(\alpha) = p\log(1+e^{-\alpha}) + (1-p)\log(1+e^{\alpha}).$$

On a alors $(C^{\Phi}_p)'(\alpha) = 0 \Leftrightarrow (-p + (1-p)e^{\alpha}) / (e^{\alpha}+1) = 0$, ce dont on déduit

$$f^*(X) = \log\left(\frac{p(X)}{1-p(X)}\right),$$

avec les conventions $\pm\infty$ lorsque $p(X) \in \{0,1\}$.

— **Risque exponentiel** : Pour $\Phi(u) = \exp(-u)$, on a

$$C_p^\Phi(\alpha) = pe^{-\alpha} + (1-p)e^{\alpha}.$$

On a alors $(C_p^\Phi)'(\alpha) = 0 \Leftrightarrow -p + (1-p)e^{2\alpha} = 0$, ce dont on déduit

$$f^*(X) = \frac{1}{2}\log\left(\frac{p(X)}{1-p(X)}\right).$$

Dans tous ces exemples classiques, les valeurs $\pm\infty$ des pseudos-classifieurs de Bayes correspondents aux situations où $p(X) \in \{0, 1\}$, ce qui semble assez moral.

Calibration des Φ-risques et excès de Φ-risques

Dans tous les exemples usuels, on a bien $f^*(X) \geq 0 \Leftrightarrow p(X) \geq 1/2$, et donc, si f^* est un pseudo-classifieur de Bayes, alors $\mathrm{sg}(f^*)$ est un classifieur de Bayes (pour le problème de classification binaire). On peut donner des conditions générales sur Φ pour que cette propriété soit générique : il s'agit de la *calibration* d'un Φ-risque.

Definition 2.36 : Calibration des Φ-risques

Une fonction $\Phi : \bar{\mathbb{R}} \to [0, +\infty]$ est calibrée pour la classification 0/1 si

$$\begin{cases} \forall p > \frac{1}{2}, \alpha < 0 & \inf_{\alpha\in\bar{\mathbb{R}}} C_p^\Phi(\alpha) < C_p^\Phi(\alpha), \\ \forall p < \frac{1}{2}, \alpha \geq 0 & \inf_{\alpha\in\bar{\mathbb{R}}} C_p^\Phi(\alpha) < C_p^\Phi(\alpha). \end{cases}$$

Cette définition est relativement intuitive : pour un p donné, on demande à ce que le pseudo-classifieur de Bayes ait même signe que $p(X) - 1/2$. Cette condition de calibration suffit à généraliser la bonne propriété des Φ-risques usuels.

Proposition 2.37

Si Φ est convexe, alors Φ est calibrée si et seulement si, pour toute loi P et f^* pseudo-classifieur de Bayes, $\mathrm{sg}(f^*)$ est un classifieur de Bayes (pour le problème de classification binaire).

Preuve de la Proposition 2.37. Commençons par supposer que Φ est calibrée, soit P une loi sur $\mathcal{X} \times \{-1, 1\}$, et f^* un pseudo-classifieur de Bayes. Rappelons que, C_p^Φ étant convexe, son minimum sur $\bar{\mathbb{R}}$ est atteint. Si $p(X) < 1/2$, alors la propriété de calibration assure que $f^*(X) < 0$, et donc que $\mathrm{sg}(f^*(X)) = -1$. Si $p(X) > 1/2$, cette même propriété donne $f^*(X) \geq 0$, et donc $\mathrm{sg}(f^*(X)) = 1$. D'après la Proposition 1.7, cela suffit à montrer que $\mathrm{sg}(f^*)$ est un classifieur de Bayes.

Regardons maintenant l'autre sens, et essayons de montrer que Φ est nécessairement calibrée. Soit $x \in \mathcal{X}$ et $p > (1/2)$, et posons $P = \delta_x \times \mathrm{Rad}(p)$ (variable de Rademacher de paramètre p). Soit $\alpha^* \in \arg\min_{\alpha\in\bar{\mathbb{R}}} C_p^\Phi(\alpha)$ **un** minimum de C_p^Φ. Alors $f^* \equiv \alpha^*$ est un pseudo-classifieur de Bayes. Comme, par hypothèse, $\mathrm{sg}(f^*)$ est un classifieur de Bayes, on a $\mathrm{sg}(f^*) = 1$, soit $\alpha^* \geq 0$. On en déduit que tous les minimiseurs de C_p^Φ sont nécessairement positifs ou nuls (ce qui prouve le premier cas). Un raisonnement analogue pour $p < 1/2$ permet de conclure. □

La définition de la calibration est assez intuitive, mais peut être pénible à vérifier en pratique. On peut en donner une condition nécessaire et suffisante (toujours dans le cas Φ convexe).

Proposition 2.38

Soit $\Phi : \bar{\mathbb{R}} \to [0, +\infty]$ convexe. Φ est calibrée pour la classification 0/1 si et seulement si Φ est dérivable en 0, avec $\Phi'(0) < 0$.

On a donc un moyen simple et rapide de vérifier qu'un Φ-risque est calibré. On peut vérifier aisément cette propriété sur les Φ-risques usuels convexes.

Preuve de la Proposition 2.38. Commençons par supposer que Φ est calibrée. Soit $p < 1/2$. Φ étant convexe, elle admet des dérivées à gauche et à droite en 0. La dérivée à gauche de C_p^Φ en 0 s'écrit alors

$$(C_p^\Phi)'_g(0) = p\Phi'_g(0) - (1-p)\Phi'_d(0).$$

Φ étant calibrée, les minimums de C_p^Φ sont nécessairement dans $[-\infty, 0[$, ce qui implique $(C_p^\Phi)'_g(0) > 0$, ou encore

$$p\Phi'_g(0) > (1-p)\Phi'_d(0).$$

En faisant tendre p vers $1/2$ par valeurs négatives on obtient $\Phi'_d(0) \leq \Phi'_g(0)$. Or Φ est convexe, donc $\Phi'_g(0) \leq \Phi'_d(0)$, ce dont on déduit que Φ est dérivable en 0. L'inégalité précédente se récrit alors, pour $p < 1/2$,

$$(2p-1)\Phi'(0) > 0,$$

ce qui donne $\Phi'(0) < 0$.

Pour la réciproque, si Φ est dérivable en 0, alors C_p^Φ l'est aussi, avec $(C_p^\Phi)'(0) = (2p-1)\Phi'(0)$. Si $\Phi'(0) < 0$, et $p < 1/2$, alors $(C_p^\Phi)'(0) > 0$, donc C_p^Φ admet ses minimums dans $[-\infty, 0[$. De même, si $p > 1/2$, C_p^Φ admet ses minimums dans $]0, +\infty]$, et elle est donc calibrée. □

Pour la plupart des Φ-risques usuels, on a toujours $R_P^\Phi \geq R_P^{\Phi_{0/1}}$, du fait que la fonction de coût usuellement majore $\mathbb{1}_{u \leq 0}$. Cela suffit à donner une majoration du Φ-risque 0/1 (le risque en classification binaire, à peu de choses près). En revanche, si on a une majoration de *l'excès* de Φ risque, $R_P^\Phi(f) - R_P^{\Phi,*}$, peut-on en déduire une majoration de l'excès de risque en classification binaire ? La réponse est oui, pour les fonctions Φ convexes et calibrées.

THÉORÈME 2.39

Soit $\Phi : \bar{\mathbb{R}} \to [0, +\infty]$ convexe et calibrée. On définit

$$\Psi : \begin{cases} [-1,1] & \to & \bar{\mathbb{R}} \\ u & \mapsto & \Phi(0) - H^\Phi\left(\frac{1+u}{2}\right), \end{cases}$$

où on rappelle que $H^\Phi(p) = \min_{\alpha \in \bar{\mathbb{R}}} C_p^\Phi(\alpha)$. Alors Ψ est positive, paire, convexe, et $\Psi(u) = 0$ si et seulement si $u = 0$. De plus, si f est un pseudo-classifieur, alors

$$\Psi\left(R_P(\mathrm{sg}(f)) - R_P^*\right) \leq R_P^\Phi(f) - R_P^{\Phi,*}.$$

Démonstration. Soit $u \in [-1,1]$. On a

$$\begin{aligned} \Psi(u) &= \Phi(0) - H^\Phi\left(\frac{1+u}{2}\right) \\ &= C^\Phi_{\frac{1+u}{2}}(0) - H^\Phi\left(\frac{1+u}{2}\right) \\ &\geq 0. \end{aligned}$$

D'autre part, pour tout α,

$$\begin{aligned} C^\Phi_{\frac{1-u}{2}}(\alpha) &= \frac{1-u}{2}\Phi(\alpha) + \frac{1+u}{2}\Phi(-\alpha) \\ &= C^\Phi_{\frac{1+u}{2}}(-\alpha). \end{aligned}$$

On en déduit $H^\Phi\left(\frac{1+u}{2}\right) = H^\Phi\left(\frac{1-u}{2}\right)$, et donc $\Psi(-u) = \Psi(u)$.

Pour la convexité, on a $H^\Phi(p) = \min_{\alpha \in \bar{\mathbb{R}}} p\Phi(\alpha) + (1-p)\Phi(-\alpha)$, H^Ψ est donc un infimum de fonctions linéaires et est donc concave. On en déduit que Ψ est convexe.

Soit maintenant $\theta \in [-1, 0[$. Comme Φ est calibrée,

$$C^\Phi_{\frac{1+u}{2}}(0) > H^\Phi\left(\frac{1+u}{2}\right),$$

ce qui revient à $\Psi(u) > 0$. Comme Ψ est paire, on conclut à $u \neq 0 \Rightarrow \Psi(u) \neq 0$. Enfin, pour $u = 0$ et $\alpha \in \mathbb{R}$,

$$C^\Phi_{\frac{1}{2}}(\alpha) = \frac{1}{2}\left(\Phi(\alpha) + \Phi(-\alpha)\right) \geq \Phi(0),$$

avec égalité en $\alpha = 0$. On en déduit $H^\Phi\left(\frac{1}{2}\right) = \Phi(0)$, et $\Psi(0) = 0$.

Pour la dernière partie, soit f un pseudo-classifieur et P une loi sur $\mathcal{X} \times \{-1, 1\}$. En rappelant que $g^* = \mathbb{1}_{p(x) \geq 1/2} - \mathbb{1}_{p(x) < 1/2}$ est de Bayes (pour la classification binaire), on a

$$\begin{aligned} \Psi\left(R_P(\mathrm{sg}(f)) - R_P^*\right) &= \Psi\left(E_X\left[|2p(X) - 1| \mathbb{1}_{\mathrm{sg}(f) \neq f^*}\right]\right) \\ &\leq E_X \Psi\left(|2p(X) - 1| \mathbb{1}_{\mathrm{sg}(f) \neq f^*}\right) \qquad \text{(convexité)} \\ &= E_X\left(\mathbb{1}_{\mathrm{sg}(f) \neq f^*} \Psi\left(|2p(X) - 1|\right)\right) \qquad (\Psi(0) = 0) \\ &= E_X\left(\mathbb{1}_{\mathrm{sg}(f) \neq f^*} \Psi((2p(X) - 1)\right) \qquad \text{(parité)}, \\ &= E_X\left(\mathbb{1}_{\mathrm{sg}(f) \neq f^*}\left[\Phi(0) - H^\Phi(p(X))\right]\right). \end{aligned}$$

Or, si $\mathrm{sg}(f(x)) \neq f^*(x)$, c'est à dire "$f$ se trompe de signe", alors son Φ-risque est minoré par

$$p(x)\Phi(f(x)) + (1-p(x))\Phi(-f(x)) \geq \Phi(0),$$

comme Φ est calibrée. Donc

$$C^{\Phi}_{p(X)}(f(X))\mathbb{1}_{\mathrm{sg}(f(X))\neq f^*(X)} \geq \Phi(0)\mathbb{1}_{\mathrm{sg}(f(X))\neq f^*(X)}.$$

On en déduit

$$\begin{aligned}\Psi\left(R_P(\mathrm{sg}(f)) - R_P^*\right) &\leq E_X\left(\mathbb{1}_{\mathrm{sg}(f)\neq f^*}\left[C^{\Phi}_{p(X)}(f(X)) - H^{\Phi}(p(X))\right]\right)\\ &\leq R_P^{\Phi}(f) - R_P^{\Phi,*}.\end{aligned}$$

□

Cette borne est "optimale", au sens qu'on peut, pour tout excès de risque possible $\theta \in [0,1]$, trouver des distributions P pour lesquelles la borne devient une égalité, dès lors que $|\mathcal{X}| \geq 2$.

Ce résultat permet de transformer des résultats de Φ-consistance en consistance pour la classification binaire ($\Psi(0) = 0$), et peut aussi permettre de déduire des vitesses de convergence pour le classifieur obtenu à partir d'un pseudo-classifieur. Là encore, des stratégies de minimisation de Φ-risque empirique sur des classes peuvent s'analyser avec les méthodes vues précédemment, avec cependant des erreurs d'estimation plutôt de l'ordre de $\dim(\mathcal{F})/n$ lorsque la Φ est fortement convexe. La notion de dimension pour des pseudo-classifieurs est à peu près la même que pour les classifieurs standards, on parle plutôt de pseudo-dimension dans ce cas.

Reste à donner les fonctions Ψ des Φ-risques usuels.

— **Perte quadratique** : On a $H^{\Psi}(p) = 4p(1-p)$, et donc

$$\Psi(u) = 1 - (1-u^2) = u^2.$$

— **Hinge Loss** : On a $H^{\Psi}(p) = 2(p \wedge (1-p))$, et donc

$$\begin{aligned}\Psi(u) &= 1 - ((1+u) \wedge (1-u))\\ &= |u|.\end{aligned}$$

— **Risque logistique** : On a $H^{\Psi}(p) = -p\log(p) - (1-p)\log(1-p)$, et donc

$$\begin{aligned}\Psi(u) &= \log(2) + \left[\frac{1+u}{2}\log\left(\frac{1+u}{2}\right) + \frac{1-u}{2}\log\left(\frac{1-u}{2}\right)\right]\\ &= \frac{1}{2}\left((1+u)\log(1+u) + (1-u)\log(1-u)\right)\\ &\geq \frac{1}{2}u^2. \qquad\qquad (\Psi''(u) = (1-u^2)^{-1} \geq 1).\end{aligned}$$

— **Risque exponentiel** : On a $H^{\Psi}(p) = 2\sqrt{p(1-p)}$, et

$$\Psi(u) = 1 - \sqrt{1-u^2} \geq \frac{u^2}{2}.$$

Pour un problème donné, le choix d'un Φ-risque adapté se fait essentiellement sur la forme du pseudo-classifieur optimal f^* et la classe $\mathcal{F}$ de pseudo-classifieurs que vous pouvez ajuster. Par exemple, si vous êtes dans un problème ou vous pouvez décemment penser que $\log(p(X)/(1-p(X))$ est à peu près linéaire en X, alors la perte logistique, associée à $\mathcal{F}$ l'ensemble des classifieurs linéaires semble un choix pertinent. Pour la perte Hinge Loss, approcher $\mathrm{sg}(2p(X)-1)$ avec des fonctions linéaires n'est pas forcément approprié, on devra plutôt se tourner vers des pseudos-classifieurs moins réguliers (par exemple NN avec activation non régulière).

2.3.3 Méthodes de descente de gradient

Toujours dans le cas où la méthode de prédiction choisie est celle de la minimisation d'un risque empirique, on a vu précédemment que le calcul effectif de ces prédicteurs ERM peut s'avérer compliqué, voire inextricable en toute généralité.

Une première remarque importante est que usuellement on n'a pas vraiment besoin d'un minimiseur exact du risque empirique : si on trouve $\hat{f}$ tel que $R_n(\hat{f}) \leq \inf_f R_n(f) + a_n$, où a_n est de l'ordre de l'erreur d'estimation sur la classe $\mathcal{F}$ (usuellement en d/n pour des fonctions coût convexes), alors on aura les mêmes garanties théoriques pour $\hat{f}$ que pour un ERM, à un facteur 2 près. On peut donc se contenter d'un minimiseur approché de R_n.

Une deuxième remarque est que, lorsque le risque empirique est une fonction convexe en le prédicteur courant f, les méthodes de type descente de gradient fournissent naturellement des minimiseurs approchés, pour un coût algorithmique qui devient intéressant lorsque la taille d'échantillon devient grande.

Cela justifie donc l'étude des prédicteurs fournis pas des algorithmes de type descente de gradient, qui sont de plus couramment utilisés en pratique.

Descente de gradient simple

On se place dans la situation générique suivante : on considère un espace de prédicteurs paramétré par $\theta \in \mathbb{R}^d$ (et avec abus on confondra θ et f_θ le prédicteur associé dans ce qui suit), et on suppose que le risque empirique à minimiser

$$\theta \mapsto R_n(\theta)$$

est différentiable, μ-**fortement convexe** et L-**smooth**, on rappelle ces définitions en dessous.

Definition 2.40

Une fonction F différentiable est μ-fortement convexe, pour $\mu > 0$, si

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d \quad F(\theta_2) \geq F(\theta_1) + \langle \nabla_{\theta_1} F, (\theta_2 - \theta_1) \rangle + \frac{\mu}{2} \|\theta_2 - \theta_1\|^2.$$

Une fonction F différentiable est L-smooth, pour $L > 0$, si

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d \quad |F(\theta_2) - [F(\theta_1) + \langle \nabla_{\theta_1} F, (\theta_2 - \theta_1) \rangle]| \leq \frac{L}{2} \|\theta_2 - \theta_1\|^2.$$

Pour une fonction μ-convexe et L-smooth, le conditionnement κ est défini par

$$\kappa = \frac{L}{\mu} \geq 1.$$

En d'autres termes, une fonction L smooth est approximable par une fonction linéaire à une fonction quadratique de coefficient L près, et une fonction est μ convexe si elle est minorée localement, à un terme linéaire près, par une fonction quadratique de coefficient μ. FAIRE DESSIN (en sandwich entre 2 quadratiques, supérieure à une quadratique).

La condition L-smooth est équivalente à la condition de gradients L-Lipschitz, c'est à dire $\|\nabla_{\theta_1} F - \nabla_{\theta_2} F\| \leq L\|\theta_1 - \theta_2\|$. La condition de μ convexité est quant à elle équivalente à une conditions "d'inverse gradient" Lipschitz, c'est à dire $\|\nabla_{\theta_1} F - \nabla_{\theta_2} F\| \geq \mu\|\theta_1 - \theta_2\|$.

Enfin, lorsque la fonction F est deux fois différentiable, la μ-convexité est équivalente à $H_\theta F \succcurlyeq \mu I_d$, tandis que la condition L-smooth est équivalente à $-LI_d \preccurlyeq H_\theta F \preccurlyeq LI_d$.

On aura besoin par la suite d'un petit résultat technique concernant les fonctions μ-convexes.

Lemme 2.41 : Inégalité de Lojasiewicz

Si F est différentiable et μ-convexe, et θ^* est le minimiseur de F, alors, pour tout $\theta \in \mathbb{R}^d$,

$$\|\nabla_\theta F\|^2 \geq 2\mu \left(F(\theta) - F(\theta^*)\right).$$

Preuve du Lemme 2.41. On part de la définition de la μ-convexité. Á θ_1 fixé, $\theta_2 \mapsto F(\theta_1) + \langle \nabla_{\theta_1} F, (\theta_2 - \theta_1) \rangle + \frac{\mu}{2}\|\theta_2 - \theta_1\|^2$ est fortement convexe, et admet pour unique minimiseur $\tilde{\theta}_2 = \theta_1 - \frac{1}{\mu}\nabla_{\theta_1} F$. On en déduit alors que, pour tout θ_1, θ_2,

$$\begin{aligned} F(\theta_2) &\geq F(\theta_1) + \langle \nabla_{\theta_1} F, (\theta_2 - \theta_1) \rangle + \frac{\mu}{2}\|\theta_2 - \theta_1\|^2 \quad (\mu\text{-convexité}) \\ &\geq F(\theta_1) + \left\langle \nabla_{\theta_1} F, (\tilde{\theta}_2 - \theta_1) \right\rangle + \frac{\mu}{2}\|\tilde{\theta}_2 - \theta_1\|^2 \\ &\geq F(\theta_1) - \frac{1}{2\mu} \|\nabla_{\theta_1} F\|^2. \end{aligned}$$

En prenant $\theta_2 = \theta^*$ on obtient le résultat. □

Pour ce qui est du risque empirique

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} c(f_\theta(X_i), Y_i),$$

ces deux propriétés découlent des caractéristiques de $\theta \mapsto c(f_\theta(x), y)$. Lorsque cette dernière est différentiable, L-smooth et μ convexe, le risque empirique l'est. Par exemple, dans un contexte de Φ-risque où $c(f_\theta(x), y) = \Phi(yf_\theta(x))$, et prédictions linéaires $f_\theta = \langle \theta, . \rangle$, ces propriétés résultent de celles de la fonction Φ combinées généralement avec la matrice de Gram des observations. Pour faire simple, supposons Φ deux fois dérivable. On a alors

$$H_\theta R_n = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 \Phi''(\langle \theta, y_i X_i \rangle) X_i X_i^T.$$

Dans ce cas, si on suppose $Y \in \{-1, 1\}$, que $\mu_\Phi \leq \Phi'' \leq L_\Phi$, et $\hat{\lambda}_{\min} I_d \preccurlyeq \hat{\Sigma} \preccurlyeq \hat{\lambda}_{max} I_d$, alors R_n sera μ-convexe et L-smooth, avec $\mu = \mu_\Phi \hat{\lambda}_{min}$ et $L = L_\Phi \hat{\lambda}_{max}$. C'est donc un cas (relativement courant) où on a accès à μ et L. Il arrive que $\hat{\Sigma}$ ne soit pas inversible, notamment en grande dimension, c'est pourquoi usuellement on remplace le Φ risque par un Φ-risque pénalisé du type

$$R_n(\theta) + \frac{\mu}{2}\|\theta\|^2,$$

qui lui est μ-fortement convexe (comme en SVM).

Dans ce cas, on connaît structurellement les paramètres μ et L, et on peut appliquer des méthodes de descente de gradient : on part d'un état initial θ_0, que l'on actualise via

$$\theta_t = \theta_{t-1} - \gamma_t \nabla_{\theta_{t-1}} R_n.$$

Le résultat classique suivant fournit un ordre de grandeur sur le nombre d'itérations nécessaires.

Théorème 2.42 : Convergence : cas smooth et fortement convexe

Si R_n est μ-convexe et L-smooth, alors, en choissisant $\gamma_t \equiv \gamma = \frac{1}{L}$, on a

$$R_n(\theta_t) - R_n(\hat{\theta}) \leq \left(1 - \frac{1}{\kappa}\right)^t (R_n(\theta_0) - R_n(\hat{\theta})),$$

où $\hat{\theta}$ est le minimiseur de R_n.

Preuve du théorème 2.42. On part de

$$\begin{aligned} R_n(\theta_t) &= R_n\left(\theta - t - 1 - \frac{1}{L}\nabla_{\theta_{t-1}} R_n\right) \\ &\leq R_n(\theta_{t-1}) - \frac{1}{L}\left\langle \nabla_{\theta_{t-1}} R_n, \nabla_{\theta_{t-1}} R_n \right\rangle + \frac{L}{2}\left\| \frac{1}{L}\nabla_{\theta_{t-1}} R_n \right\|^2 \qquad (L\text{-smooth}) \\ &= R_n(\theta_{t-1}) - \frac{1}{2L}\left\| \nabla_{\theta_{t-1}} R_n \right\|^2, \end{aligned}$$

ou encore

$$R_n(\theta_t) - R_n(\hat{\theta}) \leq R_n(\theta_{t-1}) - R_n(\hat{\theta}) - \frac{1}{2L} \left\| \nabla_{\theta_{t-1}} R_n \right\|^2 .$$

On utilise maintenant le Lemme 2.41 pour arriver à

$$\begin{aligned} R_n(\theta_t) - R_n(\hat{\theta}) &\leq R_n(\theta_{t-1}) - R_n(\hat{\theta}) - \frac{1}{2L} 2\mu (R_n(\theta_{t-1}) - R_n(\hat{\theta})) \\ &\leq \left(1 - \frac{1}{\kappa}\right) (R_n(\theta_{t-1}) - R_n(\hat{\theta})), \end{aligned}$$

et on conclut par récurrence. □

L'utilité du Théorème 2.42 est la suivante : supposons que l'erreur stochastique soit de l'ordre de d/n (arrive souvent si la fonction de coût est convexe). Alors, pour obtenir un d/n-minimiseur de R_n, une descente de gradient partant de θ_0 aura besoin de

$$T_n = \lceil \kappa \log \left(\frac{n(R_n(\theta_0) - R_n(\hat{\theta}))}{d} \right) \rceil$$

itérations pour sortir un prédicteur quasiment optimal. En termes de complexité algorithmique, cela ressort une dépendance en $O(n \log(n))$ (si on suppose par exemple que la fonction coût est bornée).

On remarque plusieurs choses :

— La calibration du pas et le temps d'arrêt dépendent fortement du conditionnement κ, qui dépend lui-même des données d'entraînement. Pour sortir une borne en généralisation "globale", il faut souvent faire un peu de concentration sur $\hat{\lambda}_{min}$ et $\hat{\lambda}_{max}$.

— Dans le cas où on ne connaît pas κ, on peut prendre des pas tendant vers 0 mais dont la somme diverge (en $t^{-\alpha}$, $\alpha \leq 1$, au prix d'une moins bonne borne théorique. En pratique on peut aussi calibrer ce pas par line search.

— Même si R_n n'est pas fortement convexe (juste convexe), on peut avoir des garanties de convergence sur $R_n(\theta_t)$, moins rapides certes. Le lecteur intéressé trouvera des résultats là-dessus dans l'oeuvre de F. Bach et les références associées.

— Cette méthode nécessite de stocker en mémoire tout les points de l'échantillon, ce qui est une limitation pour n vraiment très grand.

En guise de remarque finale : pour donner des garanties théoriques sur une sortie de descente de gradient simple, on est obligé de passer par des garanties sur des minimiseurs de risque empirique approché. On peut se départir de cette étape pour des algorithmes type gradient stochastique.

Descente de gradient stochastique

A gros traits, le principe de la descente de gradient stochastique est toujours itératif, de la forme

$$\theta_t = \theta_{t-1} - \gamma_t g_{t-1},$$

mais ici, plutôt que de prendre $g_{t-1} = \nabla_{\theta_{t-1}R_n}$, ce qui nécessite de regarder tout l'échantillon à chaque étape, on prend plutôt

$$g_{t-1} = \nabla_{\theta_{t-1}} c(f_{\theta_{t-1}}(X_t), Y_t),$$

c'est à dire une approximation du gradient basée sur la t-ième observation. On voit alors que ce type de procédure est *single-pass*, c'est à dire que chaque point d'échantillon sera visité une seule fois, et en termes de stockage seul le point courant et son estimée de gradient sont nécessaires. On peut utiliser les méthodes de gradient stochastique pour 2 types de tâches :

1. Donner un prédicteur $\hat{\theta}$, avec des garanties sur $\mathbb{E}(R(\hat{\theta}))$. Dans ce cas on regarde séquentiellement les points de l'échantillon.
2. Minimiser R_n (essayer), dans ce cas on re-tire au hasard dans les points de D_n et c'est ce qui donne l'ordre des points à visiter. Cette stratégie peut être utile lorsque n est vraiment trop grand pour que même la descente de gradient simple ne soit pas d'usage commode.

On présentera ici quelques résultats sur la première tâche, dans un cadre de forte convexité et smooth. Les résultats pour le second cas sont similaires. On se place dans un contexte de minimisation de **vrai risque**

$$R(\theta) = \int c(f_\theta(x), y) P_{(X,Y)}(dx, dy),$$

et les hypothèses de forte convexité et smooth vont porter sur cette fonction. Comme auparavant, dans un cadre de classification avec Φ-risque, il suffit que Φ soit μ_Φ-convexe et L-smooth, et que $\lambda_{\min} I_d \preccurlyeq \mathbb{E}(XX^T) = \Sigma \preccurlyeq \lambda_{max} I_d$. Les hypothèses portent donc ici sur la matrice de Gram théorique (et les constantes de μ convexité sont souvent inconnues). Comme auparavant, on peut forcer la convexité en rajoutant un terme de pénalisation convexe au risque.

Comme les points de l'échantillon sont visités successivement (et ces points sont considérés i.i.d.), la suite θ_t est formellement une chaîne de Markov. Pour travailler dessus, on aura naturellement besoin de conditionner par rapport au passé jusqu'à l'étape t : $\mathcal{F}_t = \sigma((X,Y)_{1:t})$. On a alors que θ_t est $\mathcal{F}_t$-mesurable. Pour simplifier les notations, on notera E_t l'espérance conditionnelle sachant $\mathcal{F}_{t-1}$ (on intègre par rapport à $(X,Y)_t$ quoi. Les conditions de succès des méthodes de gradient stochastiques reposent sur deux hypothèses, portant essentiellement sur les gradients et leurs approximations.

Hypothèse 2.43 : Conditions SGD

1. *(Gradients non biaisés) :* $\forall t \geq 1 \ E_t(g_t) = \nabla_{\theta_{t-1}} R$.
2. *(Gradients quasi-bornés) :* $\exists G > 0 \quad \forall t \geq 1 \ E_t \|g_t\|^2 \leq G^2$, *presque sûrement.*

Regardons ces hypothèses dans le cadre le plus simple possible : régression linéaire moindre carrés. Dans ce cas, $c(f_\theta(x), y) = (y - \langle \theta, x \rangle)^2$, et

$$g_t = 2(Y_t - \langle X_t, \theta_{t-1} \rangle) X_t,$$

On a bien

$$E_t(g_t) = 2(\mathbb{E}(YX) - \mathbb{E}(XX^T)\theta_{t-1}) = \nabla_{\theta_{t-1}} R.$$

Par ailleurs,

$$\begin{aligned} E_t(\|g_t\|^2) &= 4\mathbb{E}\left(\mathbb{E}\left[(Y - \langle X, \theta_{t-1}\rangle)\|X\|^2 \mid X\right]\right) \\ &= 4\mathbb{E}\left(\|X\|^2\sigma^2(X)\right) + 4\mathbb{E}\left(\|X\|^2(\eta(X) - \langle X, \theta_{t-1}\rangle)^2\right), \end{aligned}$$

avec $\eta(X) = \mathbb{E}(Y \mid X)$, et $\sigma^2(X) = \mathbb{E}(Y - \eta(X))^2 \mid X)$. La condition de gradients quasi-bornés sera donc vérifiée sous des conditions usuelles, de type $\sigma(X) \leq \sigma$, X à support borné, et θ dans une boule. On peut enlever cette dernière hypothèse en modifiant légèrement la preuve qui va suivre.

THÉORÈME 2.44

Si R est μ-convexe, L-smooth, et si les gradients vérifient les hypothèses 2.43, alors, en posant $\gamma_t = \frac{1}{\mu t}$, on obtient

$$\mathbb{E}\left(R(\theta_t) - R(\theta^*)\right) \leq \frac{LG^2}{2\mu^2 t}.$$

Preuve du Théorème 2.44. On commence comme pour la descente de gradient standard :

$$R(\theta_t) \leq R(\theta_{t-1}) + \left\langle \nabla_{\theta_{t-1}} R, \gamma_t g_t \right\rangle + \frac{L}{2}\left\|\gamma_t g_t\right\|^2.$$

En prenant l'espérance conditionnelle par rapport à $\mathcal{F}_{t-1}$, on obtient

$$\begin{aligned} E_t(R(\theta_t) - R(\theta^*)) &\leq R(\theta_{t-1}) - R(\theta^*) + \gamma_t \left\langle \nabla_{\theta_{t-1}} R, E_t(g_t) \right\rangle + \frac{L\gamma_t^2}{2} E_t(\|g_t\|^2) \\ &\leq R(\theta_{t-1}) - R(\theta^*) - \gamma_t \|\nabla_{\theta_{t-1}} R\|^2 + \frac{LG^2\gamma_t^2}{2} \qquad \text{(Hypothèses 2.43)}. \end{aligned}$$

En utilisant le Lemme 2.41, on obtient

$$E_t(R(\theta_t) - R(\theta^*)) \leq (1 - 2\mu\gamma_t)(R(\theta_{t-1}) - R(\theta^*)) + \frac{LG^2\gamma_t^2}{2}.$$

En utilisant la forme de $\gamma_t = (\mu t)^{-1}$ et en prenant une espérance,

$$\mathbb{E}(R(\theta_t) - R(\theta^*)) \leq \left(1 - \frac{2}{t}\right)\mathbb{E}(R(\theta_{t-1}) - R(\theta^*)) + \frac{LG^2}{2\mu^2 t^2}. \tag{2.10}$$

Montrons maintenant que $\mathbb{E}(R(\theta_t) - R(\theta^*)) \leq \frac{LG^2}{2\mu^2 t}$ par récurrence.
Cas $t = 1$: Pour $t = 1$, (2.10) donne

$$\mathbb{E}(R(\theta_1) - R(\theta^*)) \leq -\mathbb{E}(R(\theta_0) - R(\theta^*)) + \frac{LG^2}{2\mu^2} \leq \frac{LG^2}{2\mu^2}.$$

Passage $t \to t+1$: Supposons la propriété vraie au rang t. La même équation (2.10) donne alors

$$\begin{aligned}
\mathbb{E}(R(\theta_{t+1}) - R(\theta^*)) &\leq \left(1 - \frac{2}{t+1}\right) \mathbb{E}(R(\theta_t) - R(\theta^*)) + \frac{LG^2}{2\mu^2(t+1)^2} \\
&\leq \left(1 - \frac{2}{t+1}\right) \frac{LG^2}{2\mu^2 t} + \frac{LG^2}{2\mu^2(t+1)^2} \qquad \text{(Hypothèse de récurrence)} \\
&\leq \frac{LG^2}{2\mu^2 t(t+1)^2} \left(t^2 - 1 + t\right) \leq \frac{LG^2}{2\mu^2(t+1)},
\end{aligned}$$

ce qui conclut la preuve. □

Plusieurs remarques sur ce résultat.

— Avec une autre méthode de preuve, on peut faire ressortir la dépendance en la condition initiale θ_0. Avec la méthode exposée, on peut remarquer que n'importe quel choix de pas initial plus grand que $(2\mu)^{-1}$ aurait fait l'affaire.

— Pour un n-échantillon, et $\|g_t\|_\infty \leq M$ p.s., on obtient une borne en $LdM^2/(\mu^2 n)$, qui est l'ordre grandeur de la vitesse optimale usuellement. Dans le cas fortement convexe et smooth, la dernière étape d'un gradient stochastique fournit donc un prédicteur optimal au sens minimax.

— Cette performance en $1/n$ est uniquement en espérance, en $\mathbb{E}(R(\theta_n) - R(\theta^*))$. Pour un résultat en déviation, il est nécessaire de combiner avec un résultat de concentration global (voir θ_n comme une fonction globale de $(X, Y)_{1:n}$) ou à chaque étape (et travailler conditionnellement à des évènements à chaque étape).

— Usuellement, le prédicteur SGD est d'assez forte variance. Des modifications de l'algorithme (visant à prendre en compte un peu des étapes précédentes) permettent de pallier un peu ce défaut, comme l'algorithme SAGA ou d'autres variantes. Vous pouvez vous référer à l'oeuvre complète de F. Bach pour plus de détails. Une autre méthode couramment employée est d'utiliser des mini-batchs de données plutôt que des données individuelles à chaque itération, ce qui évite les comportements trop aberrants.

— Une manière simple de réduire la variance peut consister à agréger les prédicteurs de chaque étape, c'est à dire de définir $\bar{\theta}_t$ via $\sum_{j=1}^t \omega_{j,t}\theta_j$ comme prédicteur final. On peut montrer que les choix $\omega_{j,t} = 1/t$ et $\omega_{j,t} = (2j)/[t(t+1)]$ mènent à des prédicteurs avec les mêmes performances en espérance.

— Si on a à minimiser un risque empirique, et qu'on se demande quelle méthode choisir entre GD et SGD, la réponse dépendra de la précision requise et de la taille d'échantillon : pour une précision requise donnée, GD convergera en moins d'itérations, mais chacune d'elle nécessitera de regarder l'ensemble des données. Pour une précision requise trop importante, SGD prendra trop d'itérations à arriver à quelque chose d'intéressant, GD sera donc plus intéressant, mais peut s'avérer coûteux lorsque n est grand. Pour des n vraiment trop grand, SGD reste l'une des rare méthodes encore utilisable.

Chapitre 3

Introduction à la grande dimension

3.1 Méthodes locales et grande dimension

Annonçons la couleur d'emblée : méthodes locales et grande dimension font rarement bon ménage, on explique brièvement pourquoi ici. Si on se base sur les vitesses en régression pour une régularité s : $n^{-\frac{s}{2s+d}}$, on voit que la dépendance en la dimension d est exponentielle, elle se déteriore donc assez vite. Intuitivement parlant, les méthodes locales vont marcher pour des tailles de voisinage garantissant au moins $\log(n)$ points. En dimension d, une telle taille de voisinage est de l'ordre de

$$h = \left(\frac{\log(n)}{n}\right)^{\frac{1}{d}},$$

qui grandit lui aussi exponentiellement en d : si pour une dimension d donnée cette échelle est de l'ordre de 1/100, en doublant la dimension on récupère une échelle de l'ordre de 1/10 et la précision qui y est relative.

On peut faire un peu plus précis. Donnons-nous une précision 1, et cherchons le nombre n de points minimal pour recouvrir $[0,1]^d$ avec n boules de rayon 1 (le covering number à l'échelle 1 quoi). Ce covering number donnera à son tour une borne inférieure sur la taille d'échantillon requise pour atteindre cette précision. Par un argument de volume, on a

$$n\omega_d \geq 1,$$

où ω_d est le volume de la boule unité d-dimensionelle, et vaut

$$\omega_d = \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)}.$$

On en déduit alors $n \geq \Gamma\left(\frac{d}{2}+1\right)/(\pi^{d/2}) \sim_{d\to+\infty} \left(\frac{d}{2\pi e}\right)^{\frac{d}{2}} \sqrt{d\pi}$, ce qui devient vite très gros. Pour donner une idée, en dimension 20, 39 points suffisent, en dimension 30 autour de 45000, mais en dimension 50 on arrive à un ordre de grandeur en 10^12 (et une dimension 50 n'est pas quelque chose de déraisonnable en soi). Bref, même pour des dimensions modérées, les méthodes non paramétriques risquent fort d'être sous optimales, de ce fait elles sont le plus souvent réservées aux problèmes de faible dimension.

Au delà du nombre de points nécessaire pour arriver à une échelle donnée, la géométrie même de la répartition des points en grande dimension est suffisamment particulière pour susciter la méfiance vis à vis des méthodes locales. Grosso modo, en très grande dimension, la configuration générique de n points tirés au hasard est en "ballon de foot" : tous sur la même sphère, et tous à peu près à la même distance les uns des autres. FAIRE DESSIN (ou inclure simus).

Par exemple, supposons $X_1, \dots, X_n$ i.i.d. $\mathcal{U}(]0,1[^d)$. Comme $\|X_i\|^2 = \sum_{j=1}^d (X_i^{(j)})^2$, on a, pour tout $i \in [\![1,n]\!]$,

$$\mathbb{P}\left(\left|\|X_i\|^2 - \frac{d}{2}\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2}{d}\right),$$

en utilisant l'inégalité de Hoeffding. On en déduit que, pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\forall i \in [\![1,n]\!] \quad (1-\varepsilon)\frac{d}{2} \leq \|X_i\|^2 \leq (1+\varepsilon)\frac{d}{2}\right) \geq 1 - 2n\exp\left(-\frac{d\varepsilon^2}{4}\right) \to_{d\to+\infty} 1. \tag{3.1}$$

La première assertion se vérifie donc : au fur et à mesure que la dimension augmente oun n-échantillon aura tendance à se concentrer autour d'une $d-1$-sphère.

Pour la deuxième assertion, notons $Z_{i,j} = \|X_i - X_j\|^2$. On commence par calculer

$$\mathbb{E}(Z_{i,j}) = \sum_{k=1}^d \mathbb{E}((X_i^{(k)} - X_j^{(k)})^2) = d\mathbb{E}((U-V)^2) = 2d\mathrm{Var}(U) = \frac{d}{6},$$

où U, V sont des variables $\mathcal{U}(]0,1[)$ indépendantes. Comme auparavant, pour $\varepsilon > 0$, on a

$$\mathbb{P}\left(|Z_{i,j} - \mathbb{E}(Z_{i,j})| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2}{4d}\right),$$

toujours en utilisant Hoeffding. On en déduit alors

$$\mathbb{P}\left(\forall 1 \leq i < j \leq n \quad (1-\varepsilon)\frac{d}{6} \leq Z_{i,j} \leq (1+\varepsilon)\frac{d}{6}\right) \geq 1 - n^2\exp\left(-\frac{d\varepsilon^2}{72}\right) \to_{d\to+\infty} 1. \tag{3.2}$$

La deuxième assertion est alors montrée : au fur et à mesure que la dimension augmente les points d'un n-échantillon se comportent comme des points équidistants. La notion de voisinage devient alors assez peu pertinente.

3.1.1 Quelle dimension pour ces méthodes ?

Plaçons-nous pour simplifier dans le cadre de la régression moindre carrés via méthodes locales/ On vient de voir que si la variable prédictive X à valeurs dans $\mathbb{R}^D$ était de dimension "pleine" D, par exemple à coordonnées indépendantes, pour des D modérément grands la géométrie de la grande dimension ne plaidait pas en faveur des méthodes locales. La condition de coordonnées indépendantes n'est pas nécessaire pour "être de dimension D" : avoir une densité par rapport à $\mathcal{L}_D$ suffit. En fait, la bonne notion de dimension dans ce cas est intimement liée à la **distribution** P_X plutôt qu'à l'espace d'observation.

Definition 3.1 : (a, d)-standard

Soit $\mathcal{X}$ un espace métrique, et $a, d, r_0 > 0$. Une distribution P sur $\mathcal{X}$ est (a, d)-standard à l'échelle r_0 si et seulement si

$$\forall r \leq r_0, x \in \text{Supp}(P) \quad P(\text{B}(x, r)) \geq ar^d.$$

On dira qu'une distribution est de dimension d si elle est (a, d)-standard, pour un certain $a > 0$ et une certaine échelle. Cette notion de dimension prend en compte la dimension du support de P et la densité éventuelle.

Proposition 3.2

Si $\text{Supp}(P) \subset \mathcal{M}$, *où* $\mathcal{M}$ *est une sous-variété de dimension* d *de* $\mathbb{R}^D$, *alors* P *ne peut être de dimension* d', *pour* $d' < d$.

De plus, si $\text{Supp}(P) = \mathcal{M}$, *et si* P *admet une densité par rapport à* $\mathcal{H}_d$ *(mesure de Hausdorff* d*-dimensionnelle) minorée par* $f_{\min} > 0$, *alors* P *est* $(c_{\mathcal{M}} f_{\min}, d)$*-standard.*

Ce genre de résultat peut se trouver dans la littérature d'inférence de support (comme dans Aamari and Levrard [2019]). En d'autre termes, pour une mesure "relativement uniforme" sur un support $\mathcal{M}$, la dimension de P et celle de son support coïncident. Si on autorise P à avoir une densité qui tend vers 0 quelque part sur son support, on peut éventuellement gagner un peu en dimension (par exemple P de densité $6x(1-x)$ sur $[0, 1]$ est de dimension 2).

Les deux résultats qui suivent témoignent du fait que la bonne notion de dimension pour témoigner des phénomènes géométriques décrits précédemment est cette notion de distribution (a, d)-standard. Commençons par un lemme sur les covering number.

Lemme 3.3 : (a, d)-standard-Covering numbers

Si P *est* (a, d)*-standard à l'échelle* r_0, *alors, pour tout* $r \leq r_0$,

$$\mathcal{N}(\text{Supp}(P), r) \leq \frac{2^d}{ar^d} \vee 1.$$

Preuve du Lemme 3.3. Soit $\{c_1, \ldots, c_p\}$ un packing maximal de $\text{Supp}(P)$ à l'échelle $r \leq r_0$. On a alors

$$\begin{aligned} 1 &\geq P\left(\bigcup_{j=1}^{p} \text{B}(c_j, r/2)\right) \\ &\geq \left(p\frac{a}{2^d}r^d \wedge 1\right). \end{aligned}$$

Un packing maximal étant nécessairement un covering, on en déduit le résultat. □

On peut maintenant en déduire "l'échelle" à laquelle les points tirés suivant une distribution (a, d)-standard se dispersent.

Théorème 3.4

Si P est (a,d)-standard à l'échelle r_0, et $r_n := \left(C\frac{\log(n)}{n}\right)^{\frac{1}{d}}$, alors, pour $C > a^{-1}$ et n assez grand pour que $r_n \leq r_0$, on a

$$\mathbb{P}\left(\mathrm{d_H}(\mathbb{X}_n, \mathrm{Supp}(P)) \geq 2r_n\right) \underset{n\to+\infty}{\longrightarrow} 0.$$

Preuve du Théorème 3.4. En notant $c_1, \ldots, c_p$ un r_n covering de $\mathrm{Supp}(P)$, on a

$$\begin{aligned}\mathbb{P}\left(\mathrm{d_H}(\mathbb{X}_n, \mathrm{Supp}(P)) \geq 2r_n\right) &\leq \mathbb{P}\left(\exists j \in [\![1,p]\!] \quad \mathbb{X}_n \cap \mathrm{B}(c_j, r_n) = \emptyset\right)\\ &\leq p\left(1 - ar_n^d\right)^n\\ &\leq p\exp(-anr_n^d) = pn^{-aC}.\end{aligned}$$

En utilisant le Lemme 3.3, on a $p \leq \frac{2^d n}{aC\log(n)} \vee 1$, ce qui donne le résultat. □

Ce résultat montre que l'échelle à laquelle les points d'une distribution (a,d)-standard se répartissent est de l'ordre de $n^{-1/d}$, peu importe la dimension ambiante D. On peut étendre ce genre de résultat pour caractériser la distance au k-ième plus proche voisin, où sur le nombre de points dans un r_n voisinage, tout cela démontrant que les méthodes basées sur des voisinages locaux s'adaptent à la dimension **intrinsèque** d plutôt qu'à la dimension ambiante. De fait, on peut montrer que les vitesses en régression et estimation de densité sont de l'ordre de $n^{-\frac{s}{2s+d}}$, où d est la dimension intrinsèque Bickel and Li [2007], Pelletier [2006]. A noter que pour ces méthodes il n'est pas nécessaire de connaître le support, l'adaptation se fait automatiquement.

Concluons sur le fait que pour un problème de prédiction, une hypothèse sur la dimension des variables prédictives X est par essence un peu trop forte. On a plutôt besoin de caractériser la dimension des variables utiles pour prédire Y, ce qui n'est pas forcément X tout entier. Une hypothèse adaptée en régression serait alors

$$\eta(X) = \mathbb{E}(Y \mid X) = h \circ g(X),$$

où $g : \mathbb{R}^D \to \mathbb{R}^d$ et $h : \mathbb{R}^d \to \mathbb{R}$, caractérisant le fait que le problème de prédiction de Y est un problème d-dimensionnel. Dans ce cas de figure, il faut au préalable identifier g, ou de manière équivalente une transformation des variables d'entrée qui ne déteriore pas trop le problème de prédiction.

Trouver de telles transformations "générales" à valeurs dans des variétés est un problème actif (par exemple Raducanu and Dornaika [2012]). Dans le cadre de transformations linéaires, l'hypothèse de réduction du problème est connue sous le nom de "Sufficient dimension reduction" (SDR), et plusieurs méthodes comme la Sliced Inverse Regression fournissent une approximation de la transformation ad-hoc Li [1991]. Dans la Section suivante, on montrera des résultats pour des transformations linéaires bien particulières : celles qui consistent à évacuer des variables.

3.2 Régression parcimonieuse en grande dimension

On se place ici dans le contexte de la régression linéaire par moindres carrés à design fixe, pour un modèle Gaussien bien spécifié, c'est à dire

$$Y = X\theta^* + \varepsilon,$$

où les ε_i sont i.i.d. $\mathcal{N}(0,\sigma^2)$, et $X \in \mathbb{R}^{n\times D}$. Lorsque $D > n$, le modèle est non identifiable (on a aussi $Y = X(\theta^* + \theta^\perp) + \varepsilon$, où $X\theta^\perp = 0$). Par convention, on choisit θ^* comme l'élément de plus petite norme satisfaisant cette équation, c'est à dire

$$\theta^* = \arg\min_\theta \{\|\theta\| \mid Y = X\theta + \varepsilon\}.$$

Dans un contexte de grande dimension, on voit deux problèmes arriver.

Bornes générales en la dimension : les fluctuations s'additionnent.

On a déjà montré que pour ce problème, la vitesse minimax est

$$\inf_{\hat\theta}\sup_{\theta^*} \mathbb{E}\ell(\theta,\theta^*) = D\sigma^2,$$

où $\ell(\theta,\theta^*) = \|X(\theta-\theta^*)\|^2$ est l'excès de risque en prédiction. On voit alors que la dépendance de l'excès de risque en la dimension est linéaire : les sources d'erreurs s'additionnent. En grande dimension, cela pose problème : on ne peut espérer de méthode uniformément performante sur toutes les configurations possibles de θ^*. Il va donc falloir rajouter des hypothèses (de parcimonie).

La covariance empirique devient peu fiable.

Ce défaut concerne essentiellement le design aléatoire. On rappelle que l'estimateur par moindre carrés est

$$\hat\theta = \hat\Sigma^{-1}\frac{1}{n}X^TY,$$

où $\hat\Sigma = \frac{1}{n}X^TX = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$ est la matrice de covariance empirique. Lorsque $D > n$, cette matrice n'est plus inversible. Cela n'est pas très grave, on peut utiliser un pseudo-inverse de Moore-Penrose, $\hat\Sigma^\dagger$, défini par

$$\hat\Sigma^\dagger = Q\mathrm{Diag}((\mu_i^{-1}\mathbb{1}_{\mu_i>0} + 0\mathbb{1}_{\mu_i=0})_{i=1,\dots,D})Q^T,$$

si $\hat\Sigma = Q\mathrm{Diag}((\mu_i)_{i=1,\dots,D})Q^T$, ce qui correspond à prendre le θ de plus petite norme satisfaisant $X\theta = Y$ (on est donc clairement dans le régime du surapprentissage, la solution proposée étant interpolante). Ce qui est plus préoccupant est que les bornes usuelles pour la convergence des moindres carrés (ou des méthodes de descentes de gradient) se basent sur les propriétés spectrales de $\Sigma = \mathbb{E}(xx^T)$ et sur la qualité d'approximation de Σ par $\hat\Sigma$ (au sens spectral, par exemple pour pouvoir garantir $\hat\lambda_{\min}$ assez grand).

De fait, si les x_i sont i.i.d. et de variance bornée, on a bien, à D fixé,

$$\hat\Sigma \xrightarrow[n\to+\infty]{p.s.} \Sigma,$$

par la loi des grands nombres. Sous conditions de moments d'ordre 4, on peut même donner un théorème central limite. Le problème est que l'asymptotique D fixe, $n \to +\infty$ ne correspond pas au régime de la grande dimension, où le cas $D >> n$ est fréquent. Il convient donc de changer de régime : on autorise D à croître avec n. Posons par exemple $\gamma = \frac{D}{n}$, supposons γ constant, et regardons la mesure empirique associée aux valeurs propres de $\hat\Sigma$:

$$\Lambda_{\hat\Sigma} = \frac{1}{D}\sum_{j=1}^D \delta_{\hat\lambda_j},$$

où les $\hat{\lambda}_j$ sont les valeurs propres de $\hat{\Sigma}$ (on remarque qu'on a au moins $D-n$ valeurs propres nulles). Cette mesure empirique est parfois appelée "empirical spectral distribution". Dans le régime D fixe, et pour faire simple $\Sigma = I_D$, $n \to +\infty$, on aurait $\Lambda_{\hat{\Sigma}} \to \delta_1$, par la loi des grands nombres. On pourrait énoncer des théorèmes centraux limites sous des conditions de moment d'ordre 4. Dans le régime γ constant, le résultat suivant montre que cette distribution est loin de ce concentrer autour de sa limite.

THÉORÈME 3.5 : MARCENKO-PASTUR

Si $x_i = (x_{i,1}, \dots, x_{i,D})^T$, où les $(x_{i,j})$ sont i.i.d., centrés et de variance v, alors

$$\Lambda_{\hat{\Sigma}} \underset{n\to+\infty}{\rightsquigarrow} \Lambda^{MP} p.s.,$$

où Λ^{MP} est la loi définie par

$$\Lambda^{MP}(dx) = \frac{1}{2\pi\gamma v x}\sqrt{(b-x)(x-a)}dx + \mathbb{1}_{\gamma>1}\left(1-\frac{1}{\gamma}\right)\delta_0(dx),$$

avec $a = v(1-\sqrt{\gamma})^2$, $b = a = v(1+\sqrt{\gamma})^2$.

FAIRE DESSIN. A noter que la convergence a lieu dans un sens un peu plus fort que la convergence en loi : on peut montrer que la norme sup entre fonctions de répartitions tend vers 0. Le lecteur intéressé par la preuve de ce résultat est revoyé vers Bai and Silverstein [2010]. Dans le régime où D/n ne tend pas vers 0 (ce qui est le cas en grande dimension), il est donc illusoire d'espérer que $\hat{\Sigma}$ se concentre bien autour de Σ.

Tout les résultats qui suivent sont de le cas de la régression linéaire Gaussienne à designe fixe pour un modèle Gaussien bien spécifié, qui est le cas le plus facile. Les adaptations au cadre sous-Gaussien et/ou design aléatoire s'en déduisent modulo un travail sur les matrices de covariance (pas toujours évident, voir la Section 3.3). Pour des pertes différentes, on a des résultats et des heuristiques de preuve similaires, modulo des résultats intermédiaires de concentration plus techniques.

3.2.1 Pénalisation $L2$, prédicteur Ridge

On reste dans le cadre de la régression linéaire à design fixe pour un modèle bien spécifié. En termes de prédiction, on a vu que le risque optimal pour le problème de régression était

$$\inf_{\hat{\theta}} \sup_{\theta} E_\theta \|X(\hat{\theta}-\theta)\|^2 = D\sigma^2,$$

dans le cas où X^TX est inversible. Si X^TX n'est plus supposée inversible (par exemple pour $D > n$), un prédicteur minimisant $\|Y - X\theta\|^2$ vérifie

$$X^TX\theta = X^TY. \tag{3.3}$$

L'ensemble des solutions de cette équation est de la forme $\theta_0 + \mathrm{Ker}(X^TX)$, où θ_0 est une solution particulière. Si on pose

$$\hat{\theta} = \hat{\Sigma}^\dagger X^TY,$$

où $\hat{\Sigma}$ est l'inverse de Moore-Penrose précédemment défini, on a

$$\begin{aligned} X^T X\hat{\theta} &= (Q\mathrm{Diag}(\mu_i)Q^T)(Q\mathrm{Diag}(\mu_i^{-1}\mathbb{1}_{\mu_i>0})Q^T)X^TY \\ &= \pi_{V(X^T)}X^TY \\ &= X^TY, \end{aligned}$$

où $V(X^T) = \mathrm{Im}(X^TX)$ est le sev de $\mathbb{R}^D$ engendré par les colonnes de X^T. Par ailleurs, si θ est solution de (3.3), on a $\hat{\theta} - \theta \in \mathrm{Ker}(X^TX)$. $\mathrm{Ker}(X^TX)$ et $\mathrm{Im}(X^TX)$ étant orthogonaux, on a alors

$$\|\theta\|^2 = \|\hat{\theta}\|^2 + \|\theta - \hat{\theta}\|^2 \geq \|\hat{\theta}\|^2.$$

On en déduit que $\hat{\theta} = \hat{\Sigma}^\dagger X^TY$ est la solution de (3.3) de norme minimale. Par ailleurs il vérifie

$$\begin{aligned} \mathbb{E}\ell(\hat{\theta}, \theta^*) &= \mathbb{E}\|X(\hat{\theta} - \theta^*)\|^2 \\ &= \mathbb{E}\|X\left[\hat{\Sigma}^\dagger X^T(X\theta^* + \varepsilon) - \theta^*\right]\|^2 \\ &= \|\pi_{V(X)}(X\theta^*) - X\theta^*\|^2 + \mathbb{E}\|\pi_{V(X)}(\varepsilon)\|^2 \\ &\leq r\sigma^2, \end{aligned}$$

où $\sigma^2 = \mathbb{E}(\varepsilon_i^2)$ et $r = \mathrm{rang}(X) \leq D \wedge n$. On peut aussi montrer que

$$\inf_{\hat{\theta}} \sup_{\theta^*} E_{\theta^*}\|X(\hat{\theta} - \theta^*)\|^2 \geq r\sigma^2.$$

L'estimateur $\hat{\theta}$ obtenu par inversion de Moore-Penrose est donc optimal.

Si on ne veut pas passer par l'inversion de Moore-Penrose, qui passe par un calcul de SVD assez coûteux lorsque D est grand (en $O(Dn(D \wedge n))$), une solution (très utilisée en pratique) est de biaiser légèrement le problème, en introduisant un *terme de régularisation* au problème à minimiser. On va maintenant rechercher

$$\hat{\theta}_{ridge} \in \arg\min_{\theta} \|Y - X\theta\|^2 + \lambda\|\theta\|^2,$$

c'est à dire un estimateur des moindres carrés *pénalisés* (ou régularisés). Si le terme de régularisation est en norme 2, de type $\lambda\|u\|^2$, on parle d'estimateur *ridge*.

Le risque pénalisé étant maintenant strictement convexe, le paramètre ridge est totalement déterminé par l'équation

$$\nabla_u(\|Y - Xu\|^2 + \lambda\|u\|^2) = 2(H + \lambda I_k)u - 2X^TY = 0,$$

avec $H(\lambda) = \hat{\Sigma} + \lambda I_D$, et donc

$$\hat{\theta}_{ridge} = H(\lambda)^{-1}X^TY.$$

Remarque : Lien avec le shrinkage. Dans le cas où $X^TX = I_D$ (ce qui implique $D \leq n$), on a $\hat{\theta}_{ridge} = \frac{1}{1+\lambda}\hat{\theta}$. Dans ce cas le ridge revient à comprimer les coefficients de l'estimateur par moindres carrés. La cible devient alors biaisée : en espérance on a $\mathbb{E}(\hat{\theta}_{ridge}) = \frac{1}{1+\lambda}\theta^*$. De l'introduction de ce biais on peut espérer un gain en prédiction. En toute généralité, cette compression des coefficients a lieu pour ceux de $Q^T\hat{\theta}$ (voir plus loin).

Remarque : En pratique, on peut éviter l'inversion de $H(\lambda)$ (coûteux si $D >> n$) et se ramener à une inversion de matrice du type $(I_n + \lambda^{-1} X X^T)^{-1})$ (via l'identité de Woodbury) : en effet

$$(X^T X + \lambda I_D) X^T = X^T (X X^T + \lambda I_n),$$

ce dont on peut déduire

$$X^T((X X^T + \lambda I_n)^{-1}) = X^T H(\lambda)^{-1},$$

en multipliant à droite et à gauche par les inverses. Cela justifie l'intérêt des méthodes ridge en grande dimension (pour peu que n soit de taille raisonnable).

On peut aussi calculer explicitement son risque en prédiction. On rappelle la décomposition de $\hat{\Sigma}$:

$$\hat{\Sigma} = Q \mathrm{Diag}((\mu_i)_{i=1,\dots,D}) Q^T. \tag{3.4}$$

PROPOSITION 3.6

En notant $\beta_i = (Q^T \theta^*)_i$, pour $i \in \{1, \dots, D\}$, on a

$$E_\theta(\|X(\hat{\theta}_{ridge} - \theta^*)\|^2) = \lambda^2 \sum_{j=1}^{D} \frac{\mu_j \beta_j^2}{(\mu_j + \lambda)^2} + \sigma^2 \sum_{j=1}^{D} \frac{\mu_j^2}{(\mu_j + \lambda)^2}.$$

Preuve de la Proposition 3.6. On commence par écrire

$$\begin{aligned} X(\hat{\theta}_{ridge} - \theta^*) &= X H(\lambda)^{-1} X^T (X\theta + \varepsilon) - X\theta \\ &= (X H(\lambda)^{-1} \hat{\Sigma} - X)\theta + X H(\lambda)^{-1} X^T \varepsilon, \end{aligned}$$

de sorte qu'on puisse décomposer

$$\mathbb{E}\|X(\hat{\theta}_{ridge} - \theta^*)\|^2 = \left\|(X H(\lambda)^{-1} \hat{\Sigma} - X)\theta^*\right\|^2 + \mathbb{E}\left\|X H(\lambda)^{-1} X^T \varepsilon\right\|^2,$$

une décomposition biais/variance habituelle. Commençons par le terme de variance. On a

$$\begin{aligned} \mathbb{E}\left\|X H(\lambda)^{-1} X^T \varepsilon\right\|^2 &= \mathbb{E}\left(\varepsilon^T X H(\lambda)^{-1} \hat{\Sigma} H(\lambda)^{-1} X^T \varepsilon\right) \\ &= \sigma^2 \left(\mathrm{Tr}(X H(\lambda)^{-1} \hat{\Sigma} H(\lambda)^{-1} X^T\right) \\ &= \sigma^2 \mathrm{Tr}((\hat{\Sigma} H(\lambda)^{-1})^2). \end{aligned}$$

Or, si $\hat{\Sigma} = Q D Q^T$, $H(\lambda)^{-1} = Q D(\lambda)^{-1} Q^T$ (en notant $D = \mathrm{Diag}((\mu_i)_{i=1,\dots D})$ et $D(\lambda) = D + \lambda I_D$). On en déduit

$$\begin{aligned} \mathrm{Tr}((\hat{\Sigma} H(\lambda)^{-1})^2) &= \mathrm{Tr}(Q D^2 D(\lambda)^{-2} Q^T) \\ &= \sum_{j=1}^{D} \frac{\mu_j^2}{(\mu_j + \lambda)^2}. \end{aligned}$$

Passons au terme de biais. On a

$$(X H(\lambda)^{-1} H - X)\theta^* = X(I_k - \lambda H(\lambda)^{-1} - I_k)\theta^* = \lambda X H(\lambda)^{-1} \theta^*,$$

et donc

$$\left\|(XH(\lambda)^{-1}\hat{\Sigma} - X)\theta^*\right\|^2 = \lambda^2\theta^T H(\lambda)^{-1}\hat{\Sigma}H(\lambda)^{-1}\theta^*$$

$$= \lambda^2\theta^T Q D(\lambda)^{-1} D D(\lambda)^{-1} Q^T\theta^*$$
$$= \lambda^2 \sum_{j=1}^{D} \frac{\mu_j}{(\mu_j+\lambda)^2}\beta_j^2.$$

□

On peut remarquer que le terme de variance est plus petit que celui associé à celui des moindres carrés. En effet

$$\sigma^2 \sum_{j=1}^{D} \frac{\mu_j^2}{(\mu_j+\lambda)^2} = \sigma^2 \sum_{j=1}^{r} \frac{\mu_j^2}{(\mu_j+\lambda)^2} < r\sigma^2,$$

si $\lambda > 0$. La quantité $\sum_{j=1}^{r} \frac{\mu_j}{(\mu_j+\lambda)}$ est parfois appelé *degré de liberté effectif* de l'estimateur ridge (plus petit donc que r, celui de l'estimateur moindre carrés). Le prix à payer pour une réduction de variance est un terme de biais. On peut toutefois montrer qu'il existe un λ pour lequel on a un gain strict.

Théorème 3.7

Il existe $\lambda > 0$ tel que

$$\mathbb{E}\|X(\hat{\theta}_{ridge} - \theta^*)\|^2 < r\sigma^2.$$

Démonstration. Ce λ est à chercher parmi les λ petits. On remarque que le terme de biais est $O(\lambda^2)$. Le terme de variance lui peut s'écrire

$$\sigma^2 \sum_{j=1}^{D} \frac{\mu_j^2}{(\mu_j+\lambda)^2} = \sigma^2 \sum_{j=1}^{r} \frac{1}{(1+\lambda/\mu_j)^2}$$
$$= \sigma^2 \sum_{j=1}^{r} (1 - \frac{2\lambda}{\mu_j} + O(\lambda^2))$$
$$= \sigma^2 \left[r - 2\lambda \left(\sum_{j=1}^{r} \frac{1}{\mu_j} \right) \right] + O(\lambda^2),$$

On a alors

$$\mathbb{E}\|X(\hat{\theta}_{ridge} - \theta^*)\|^2 - r\sigma^2 = -2\lambda\sigma^2 \left(\sum_{j=1}^{r} \frac{1}{\mu_j} \right) + O(\lambda^2) < 0$$

pour λ assez petit. □

Cela ne contredit en rien le fait que $\hat{\theta}_{LS}$ soit minimax : le λ du Théorème 3.7 dépend de σ^2, $\hat{\Sigma}$ et surtout θ^*. On ne sert pas de ce Théorème en pratique pour calibrer un λ, on utilise plutôt des techniques de cross-validation.

Moralement un bon λ est donc à chercher assez petit (on donnera un peu plus de précision là-dessus plus loin). Toutefois, lorsque $\lambda \to 0$, on peut montrer que

$\hat{\theta}_{ridge}$ tend vers l'estimateur par moindre carrés de norme minimale. Si on note $X^T = Q\sqrt{D}R$ (SVD de X^T), et $D_{>0}^{-1} = \mathrm{Diag}((\mu_i^{-1}\mathbb{1}_{\mu_i>0})_{i=1,\dots,D})$, on a

$$\begin{aligned}\hat{\theta}_{ridge} &= H(\lambda)^{-1}X^TY \\ &= Q(D+\lambda I_D)^{-1}\sqrt{D}RY \\ &\xrightarrow[\lambda\to 0]{} Q\sqrt{D_{>0}^{-1}}RY \\ &= QD_{>0}^{-1}Q^TQ\sqrt{D}RY = \hat{\theta}.\end{aligned}$$

Le fait qu'un minimiseur de $\|Y - X\theta\|^2 + \lambda\|\theta\|^2$ tende vers un minimiseur de $\|Y - X\theta\|^2$ n'a rien de surprenant. La contrainte sur $\|\theta\|$ force en plus la convergence vers un minimiseur de $\|Y - X\theta\|^2$ de plus petite norme.

Cette notion de contrainte en norme peut s'interpréter ce phénomène de deux autres manières.

Point de vue bayésien

Revenons au modèle $Y = X\theta^* + \varepsilon$, où cette fois-ci $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Prenons maintenant la loi a priori $\pi(d\theta) \sim \mathcal{N}(0, \frac{1}{\lambda})$. Le modèle étant dominé, la formule de Bayes s'applique et on retrouve

$$Q_y(d\theta) \propto e^{-\frac{1}{2}\left[\|y - X\theta\|^2 + \lambda\|\theta\|^2\right]}d\theta.$$

On reconnaît une loi normale a posteriori. Plutôt que de calculer ses paramètres, remarquons que

$$\hat{\theta}_{ridge} \in \arg\max p(\theta \mid \tilde{Y}),$$

$P_{\tilde{Y}}$ presque sûrement, où $p(\theta \mid \tilde{Y})$ représente la densité de la loi a posteriori par rapport à $\mathcal{L}_k$. Cette densité a posteriori étant celle d'une Gaussienne, on en déduit que

$$\hat{\theta}_{ridge} = \mathbb{E}\left(\tilde{\theta} \mid \tilde{Y}\right),$$

et donc que l'estimateur ridge correspond à un estimateur bayésien pour la loi a priori $\mathcal{N}(0, \frac{1}{\lambda})$. Cette loi a priori charge d'autant plus les θ de faible norme que λ est grand, ce qui est cohérent avec la version en risque pénalisé. Le paradigme bayésien s'applique alors : le prédicteur ridge (vu comme un estimateur bayésien) sera d'autant plus performant que $\|\theta^*\|$ sera petit (on aura intégré une bonne information a priori).

Point de vue M-estimation sous contraintes

On peut remarquer que, si $\hat{c}_\lambda = \|\hat{\theta}_{ridge}\|$, alors

$$\hat{\theta}_{ridge} = \arg\min_{\|\theta\| \le \hat{c}_\lambda} \|Y - X\theta\|^2.$$

De fait, résoudre un problème ridge est équivalent à résoudre un problème de moindre carrés sous contrainte en norme 2, avec un rayon $\hat{c}_\lambda$ décroissant en λ. Pour se convaincre que ce point de vue justifie la possible meilleure performance du ridge en prédiction, plaçons nous dans le cas (équivalent) où on veut estimer la

moyenne θ^* (correspond à anciennement $QD^{\frac{1}{2}}\theta^*$) d'un vecteur $Y \in \mathbb{R}^r$ (anciennement $D_{>0}^{-\frac{1}{2}}Q^TX^TY$, cf section sur la minimaxité de l'estimateur par moindre carrés), dans le modèle $Y = \theta^* + \varepsilon$, les ε_i étant i.i.d. centrés de variance σ^2.

Si on regarde

$$\hat{\theta}_M \in \arg\min_{u \in \mathrm{B}(0,M)} \|Y - u\|^2 = \pi_{\mathrm{B}(0,M)}(Y),$$

on a, dès lors que $M \geq \|\theta^*\|$,

$$\mathbb{E}\|\hat{\theta}_M - \theta^*\|^2 = \mathbb{E}\|\pi_{\mathrm{B}(0,M)}Y - \theta^*\|^2 \ < \mathbb{E}\|Y - \theta^*\|^2 \mathbb{1}_{Y \notin \mathrm{B}(0,M)} + \mathbb{E}\|Y - \theta^*\|^2 \mathbb{1}_{Y \in \mathrm{B}(0,M)},$$

car $\mathrm{B}(0, M)$ est strictement convexe et $\theta^* \in \mathrm{B}(0, M)$. On a alors que $\hat{\theta}_M$ a un meilleur risque en θ^* que Y si $\mathbb{P}(Y \notin \mathrm{B}(0, M)) > 0$. Si on prend M très grand (correspond à λ tend vers 0), on retombe sur l'estimateur par moindre carrés de norme minimale classique. L'idéal est de connaître $\|\theta^*\|$ à l'avance pour pouvoir gagner à coup sûr (réduit la variance, pas de biais).

On va essayer d'illustrer quantitativement le fait qu'un petit $\|\theta^*\|$ (information connue a priori) améliore significativement la prédiction. On regarde un problème équivalent au ridge

$$\hat{\theta}_M \in \arg\min_{\theta \in \mathrm{B}(0,M)} \|Y - X\theta\|^2,$$

pour un M quelconque. On a déjà vu qu'on pouvait concevoir les moindres carrés comme un problème de minimisation de risque empirique, la fonction de risque empirique étant donnée par $R_n(\theta) = \|Y - X\theta\|^2$ visant à approcher le risque idéal $\|X(\theta - \theta^*)\|^2 + n\sigma^2$. Pour appliquer les recettes ERM, il faut contrôler

$$\begin{aligned}
\mathbb{E}\Delta_n &= \mathbb{E} \sup_{\theta \in \mathrm{B}(0,M)} R(\theta) - R_n(\theta) \\
&= \mathbb{E} \sup_{\theta \in \mathrm{B}(0,M)} \|X(\theta - \theta^*)\|^2 + n\sigma^2 - \|X(\theta - \theta^*)\|^2 - \|\varepsilon\|^2 + 2\langle \varepsilon, X(\theta - \theta^*)\rangle \\
&\leq \mathbb{E} \sup_{\theta \in \mathrm{B}(0,M)} \langle 2\varepsilon, X(\theta - \theta^*)\rangle = \mathbb{E}\left[\langle -2\varepsilon, X\theta^*\rangle + 2 \sup_{\theta \in \mathrm{B}(0,M)} \langle \varepsilon, X\theta\rangle\right] \\
&\leq 2\mathbb{E}(\|\pi_{V(X)}(\varepsilon)\|) \sup_{\theta \in \mathrm{B}(0,M)} \|X\theta\|,
\end{aligned}$$

en utilisant l'inégalité de Cauchy-Schwartz. En utilisant l'inégalité de Jensen,

$$\mathbb{E}(\|\pi_{V(X)}(\varepsilon)\|) \leq \sqrt{\mathbb{E}\|\pi_{V(X)}(\varepsilon)\|^2} \leq \sigma\sqrt{r}.$$

Par ailleurs, pour $\theta \in \mathrm{B}(0, M)$, on a $\|X\theta\| \leq \sqrt{\mu_1}M$ (en rappelant que μ_1 est la plus grande valeur propre de $\hat{\Sigma} = X^TX$). Mis bout à bout, on obtient

$$\mathbb{E}\Delta_n \leq 2M\sigma\sqrt{r\mu_1}.$$

Il reste à dérouler le fil pour borner l'excès de risque en M-estimation

$$\theta_M^* \in \arg\min_{\mathrm{B}(0,M)} R(\theta) = \arg\min_{\theta \in \mathrm{B}(0,M)} \|X(\theta - \theta^*)\|^2,$$

c'est à dire un des meilleur θ atteignable au sens du risque en prédiction sur $\mathrm{B}(0, M)$, avec pour convention que si il y a plusieurs possibilités on prend l'élément de plus petite norme (comme pour θ^*).

On peut écrire

$$\begin{aligned}\mathbb{E}(\|X(\hat{\theta}_M - \theta^*)\|^2) &= \mathbb{E}R(\hat{\theta}_M) - R(\theta^*) \qquad (3.5)\\ &= \mathbb{E}(R_n(\hat{\theta}_M + (R - R_n)(\hat{\theta}_M)) - R(\theta^*)\\ &\leq \mathbb{E}(R_n(\theta^*_M)) - R(\theta^*) + \mathbb{E}\Delta_n\\ &\leq \inf_{\theta \in \mathrm{B}(0,M)} \|X(\theta - \theta^*)\|^2 + 2M\sigma\sqrt{r\mu_1}. \qquad (3.6)\end{aligned}$$

On retrouve une *inégalité oracle* : elle compare le risque d'un estimateur au meilleur risque atteignable sur une classe (inconnu, sauf si on a un oracle sous la main). On retrouve une formulation en erreur d'estimation/erreur d'approximation, de type

$$\mathbb{E}(\|X(\hat{\theta}_M - \theta^*)\|^2) = B^2(M) + V(M),$$

avec $B^2(M) = \inf_{\theta \in \mathrm{B}(0,M)} \|X(\theta - \theta^*)\|^2$ l'erreur d'approximation et $V(M) = 2M\sigma\sqrt{r\mu_1}$ l'erreur d'estimation.

Cette erreur peut être meilleure que l'erreur pour $\hat{\theta}$ donnée par $r\sigma^2$. Par exemple, pour le choix optimal $M = \|\theta^*\|$ (inatteignable en pratique), on a

$$\mathbb{E}(\|X(\hat{\theta}_M - \theta^*)\|^2) \leq 2\|\theta^*\|\sigma\sqrt{r\mu_1},$$

ce dont on peut déduire que pour $\|\theta^*\|$ assez petit (et pour un choix de M approprié),

$$\mathbb{E}(\|X(\hat{\theta}_M - \theta^*)\|^2) < r\sigma^2 = \mathbb{E}(\|X(\hat{\theta} - \theta^*)\|^2).$$

Grosso modo, plus $\|\theta^*\|$ est petit, plus performante sera l'estimation sous contrainte (et donc le ridge) sera.

Remarque : Cette meilleure performance dans certains cas du prédicteur sous contraintes ne contredit en rien l'optimalité de $\hat{\theta}$ au sens minimax : tout dépend de la classe de problème que l'on considère.

On peut déduire de ce qui précède une majoration de la vitesse minimax sur $\{\theta^* \in \mathrm{B}(0, M)\}$:

$$\sup_{\theta^* \in \mathrm{B}(0,M)} \mathbb{E}(\|X(\hat{\theta}_M - \theta^*)\|^2) \leq 2M\sigma\sqrt{r\mu_1} \wedge r\sigma^2,$$

passant en dessous de $r\sigma^2$ pour M assez petit. Il est évident que le problème restreint est plus "facile" statistiquement parlant que le problème non restreint (supremum sur un ensemble plus petit). Cette borne donne une idée du niveau de contrainte à partir duquel le problème restreint est sensiblement plus facile.

Remarque 2 : Le choix du M en pratique se fait par validation croisée. Théoriquement parlant, on pourrait utiliser une stratégie par pénalisation, comme ce qu'on va voir juste après.

3.2.2 Parcimonie et sélection de modèle

A partir d'ici on supposera que le modèle est Gaussien (c'est à dire $\varepsilon \sim \mathcal{N}(O, \sigma^2 I_n)$).

On a vu précédemment que si on a une information a priori sur $\|\theta\|^*$, alors on pouvait espérer aller plus vite que le risque minimax $r\sigma^2$ correspondant à la classe $\theta^* \in \mathbb{R}^D$, du fait que l'on pouvait se restreindre à la classe $\theta^* \in \mathrm{B}(0, M)$, pour un certain M.

L'hypothèse de parcimonie est une contrainte a priori sur θ^* d'une nature un peu particulière : elle suppose que parmi les θ_j^* seul un petit nombre (d^*) sont réellement pertinents pour le problème de régression, c'est à dire

$$\|\theta^*\|_0 = |S(\theta^*)| = \left|\left\{j \mid \theta_j^* \neq 0\right\}\right| \leq d^*.$$

De la même manière que pour la régression ridge, si le modèle

$$Y = X\theta^* + \varepsilon$$

est non identifiable (plusieurs θ^* marchent), on prendra pour convention tacite un choix de θ^* de support minimal. Bien sur, si $S(\theta^*)$ était connu au préalable, il suffirait de considérer le sous-modèle

$$Y = X_{S^*}\theta^* + \varepsilon,$$

où X_{S^*} correspond aux colonnes de X dont l'indice est dans S^*, et d'effectuer un moindre carré classique dessus (pour arriver à une borne en $r^*\sigma^2$, où $r^* = \text{rang}(X_{S^*}) = d^*$, X_{S^*} étant obligatoirement une famille libre de vecteurs par hypothèse de minimalité de θ^*).

Toute la difficulté provient du fait que l'on ne connaît pas ce support a priori. D'un point de vue applicatif, la connaissance de ce support peut être une information intéressante en soi (identifier les variables qui sont pertinentes pour un problème). Une manière directe de sélectionner les variables pertinentes peut être de calculer les coefficients d'un estimateur par moindre carré $(\hat{\theta}_j)_{j=1,\dots,D}$ (dont on connaît la loi dans un modèle Gaussien), puis de tester la nullité de ces coefficients. Une fois le support identifié, on peut réajuster des moindres carrés standard. C'est une approche couramment utilisée dans les situations "classiques" où D n'est pas trop grand. Plusieurs objections peuvent toutefois être soulevées :

1. Gare aux tests multiples lorsque D devient grand.
2. Pour le problème de prédiction, on se fiche un peu de retrouver le support.

Pour expliciter un peu le dernier point : retrouver le support de θ^* donne une information de type "statistique inférentielle", où on veut dire quelque chose sur les paramètres d'un modèle supposé vrai. D'un point de vue prédictif, on se fiche un peu que le modèle soit vrai, tant que les prédictions sont bonnes. Par exemple, si on trouve $\bar{\theta}$ de support différent de θ^* mais dont les performances en prédictions sont optimales, peu importe que le modèle associé ne soit pas "vrai". En d'autres termes, en prédiction on peut et doit s'autoriser du biais. Dans la même veine, même si θ^* ne satisfait pas une hypothèse de parcimonie intéressante, il peut toutefois être pertinent de considérer des θ de support biaisé (encore cette histoire de compromis approximation/estimation).

On va donc essayer de trouver des prédicteurs de support faible. Dans un cadre idéal, si on connaissait pour chaque candidat support S, le risque $R(\hat{\theta}_S)$ des moindres carrés ajustés sur ce support, on aurait juste à sélectionner

$$\hat{S} \in \arg\min_{S \in \mathcal{M}} R(\hat{\theta}_S),$$

puis regarder $\hat{\theta}_{\hat{S}}$. Ici $\mathcal{M}$ désigne une collection de modèle. Comme on ne connaît pas $R(\hat{\theta}_S)$, on va utiliser des techniques de sélection de modèle par risque empirique

pénalisé, comme dans la Section 2.1.2 (c'est à dire choisir $\hat{S}$ minimisant $R_n(\hat{\theta}_S) + \text{pen}(S)$). Rappelons l'idée générale : une bonne pénalité doit être telle que

$$\forall S \in \mathcal{M} \quad |R - R_n|(\hat{\theta}_S) \lesssim \text{pen}(S),$$

uniformément en S. Dans le cas de la régression, ce genre de borne est innateignable, on devra faire légèrement plus compliqué : on devra faire appel à des déviations renormalisées, comme dans la preuve du Théorème 2.31.

LEMME 3.8

Si on a

$$\mathbb{E}\left[\sup_{S\in\mathcal{M}} \Big((R - R_n)(\hat{\theta}_S - \theta^*) - (1-\delta)\ell(\hat{\theta}_S, \theta^*) - \text{pen}(S)\Big)_+\right] \leq \eta,$$

pour $\delta \in]0, 1]$ *et* $\eta > 0$*, alors*

$$\mathbb{E}\ell(\hat{\theta}_{\hat{S}}, \theta^*) \leq \frac{1}{\delta} \inf_S \left[\ell(\theta^*_S, \theta^*) + \text{pen}(S) + \eta\right],$$

où $X\theta^*_S = \pi_{V(X_S)}(X\theta^*)$.

Si on voulait comparer aux $\ell(\hat{\theta}_S, \theta^*)$, il faudrait aussi une borne sur $(R_n - R)$, et on aurait un facteur 2 devant $\text{pen}(S) + \eta$.

Preuve du Lemme 3.8. Soient donc de telles pénalités et δ, η. On a, pour un S quelconque,

$$\begin{aligned}
\mathbb{E}\ell(\hat{\theta}_{\hat{S}}, \theta^*) &= \mathbb{E}\left(R_n(\hat{\theta}_{\hat{S}}) - R_n(\theta^*)\right) + \mathbb{E}\left[(R - R_n)(\hat{\theta}_{\hat{S}} - \theta^*)\right] \\
&\leq \mathbb{E}\left(R_n(\hat{\theta}_{\hat{S}}) - R_n(\theta^*)\right) + (1-\delta)\mathbb{E}\ell(\hat{\theta}_{\hat{S}}, \theta^*) + \mathbb{E}(\text{pen}(\hat{S})) + \eta \\
&\leq \mathbb{E}\left[R_n(\hat{\theta}_S) - R_n(\theta^*) + \text{pen}(S)\right] + \eta + (1-\delta)\mathbb{E}\ell(\hat{\theta}_{\hat{S}}, \theta^*).
\end{aligned}$$

En réarrangeant, cela donne

$$\begin{aligned}
\delta\mathbb{E}\ell(\hat{\theta}_{\hat{S}}, \theta^*) &\leq \mathbb{E}\left[R_n(\hat{\theta}_S) - R_n(\theta^*) + \text{pen}(S)\right] + \eta \\
&\leq \mathbb{E}\left[R_n(\theta^*_S) - R_n(\theta^*) + \text{pen}(S)\right] + \eta \\
&\leq \ell(\theta^*_S, \theta^*) + \text{pen}(S) + \eta,
\end{aligned}$$

ce dont on déduit le résultat. □

De ce Lemme général on peut déduire le résultat suivant en sélection de modèle ℓ_0.

Théorème 3.9

Pour une collection de modèles $\mathcal{M}$, si, pour tout $S \in \mathcal{M}$, on choisit

$$\mathrm{pen}(S) = \sigma^2(1+K)^2\left(\sqrt{r(S)} + \sqrt{2\log\left(\frac{1}{\pi_S}\right)}\right)^2,$$

pour $r(S) = \mathrm{rang}(X_S)$, $K > 0$ et $\sum_{S\in\mathcal{M}} \pi_S \leq 1$. Alors

$$\mathbb{E}\ell(\hat{\theta}_{\hat{S}}, \theta^*) \leq \frac{1+K}{K}\left[\inf_{S\in\mathcal{M}} \ell(\theta_S^*, \theta^*) + \mathrm{pen}(S)\right] + 3\frac{(1+K)^3}{K^2}\sigma^2.$$

Preuve du Théorème 3.9. Il s'agit de vérifier les hypothèses du Lemme 3.8 pour ce choix de pénalité. Pour un $S \in \mathcal{M}$, on repart de

$$\begin{aligned}
(R - R_n)(\hat{\theta}_S - \theta^*) &= 2\left\langle \varepsilon, X(\hat{\theta}_S - \theta^*)\right\rangle \\
&\leq 2\|\pi_{V(X_S)\oplus\mathbb{R}(X\theta^*)}\varepsilon\|\|X(\hat{\theta}_S - \theta^*)\| \\
&\leq \frac{1}{1+K}\|X(\hat{\theta}_S - \theta^*)\|^2 + (1+K)\|\pi_{V(X_S)\oplus\mathbb{R}(X\theta^*)}\varepsilon\|^2 \\
&\leq \frac{1}{1+K}\ell(\hat{\theta}_S, \theta^*) + (1+K)\|\pi_{\mathbb{R}(X\theta^*))}\varepsilon\|^2 + (1+K)\|\pi_{V(X_S)}\varepsilon\|^2.
\end{aligned} \tag{3.7}$$

Intéressons nous au dernier terme. Comme $x \mapsto \|\pi_{V(X_S)}x\|$ est 1-Lipschitz, un résultat de concentration Gaussienne ([Boucheron et al., 2013, Théorème 5.6] par exemple) garantit que

$$\mathbb{P}\left(\|\pi_{V(X_S)}\varepsilon\| \geq \mathbb{E}(\|\pi_{V(X_S)}\varepsilon\|) + \sigma\sqrt{2t}\right) \leq e^{-t}.$$

Comme $\mathbb{E}(\|\pi_{V(X_S)}\varepsilon\|) \leq \sqrt{E(\|\pi_{V(X_S)}\varepsilon\|^2)} = \sqrt{\sigma^2 r(S)}$, on en déduit

$$\mathbb{P}\left(\|\pi_{V(X_S)}\varepsilon\| \geq \sigma\sqrt{r(S)} + \sigma\sqrt{2\log\left(\frac{1}{\pi_S}\right)} + \sigma\sqrt{2x}\right) \leq \pi_S e^{-x},$$

ou encore

$$\mathbb{P}\left(\|\pi_{V(X_S)}\varepsilon\|^2 \geq (1+K)\sigma^2\left(\sqrt{r(S)} + \sqrt{2\log\left(\frac{1}{\pi_S}\right)}\right)^2 + 2\sigma^2(1+K^{-1})x\right) \leq \pi_S e^{-x}.$$

Cela mène à

$$\begin{aligned}
&\mathbb{E}\left(\sup_{S\in\mathcal{M}}\left(\|\pi_{V(X_S)}\varepsilon\|^2 - (1+K)\sigma^2\left(\sqrt{r(S)} + \sqrt{2\log\left(\frac{1}{\pi_S}\right)}\right)^2\right)_+\right) \\
&\leq \sum_{S\in\mathcal{M}} \mathbb{E}\left(\|\pi_{V(X_S)}\varepsilon\|^2 - (1+K)\sigma^2\left(\sqrt{r(S)} + \sqrt{2\log\left(\frac{1}{\pi_S}\right)}\right)^2\right)_+ \\
&\leq \sum_{S\in\mathcal{M}} \pi_S \int_0^{+\infty} \exp\left(\frac{t}{2(1+K^{-1})\sigma^2}\right) dt \\
&\leq 2\sigma^2(1+K^{-1}),
\end{aligned}$$

en utilisant, pour une variable Z positive, $\mathbb{E}(Z) = \int_0^{+\infty} \mathbb{P}(Z \geq t)$.

On en déduit maintenant, en utilisant (3.7), que

$$\begin{aligned}
&\mathbb{E}\left(\sup_{S\in\mathcal{M}}\left[(R-R_n)(\hat{\theta}_S-\theta^*)-\frac{1}{K+1}\ell(\hat{\theta}_S,\theta^*)-\mathrm{pen}(S)\right]\right)\\
&\leq (1+K)\mathbb{E}\left(\|\pi_{\mathbb{R}(X\theta^*))}\varepsilon\|^2\right)+2\sigma^2(1+K^{-1})(1+K)\\
&\leq 3\sigma^2\frac{(K+1)^2}{K},
\end{aligned}$$

et donc que les conditions du Lemme 3.8 sont vérifiées, avec $\delta = \frac{K}{K+1}$ et $\eta = 3\sigma^2\frac{(K+1)^2}{K}$. Le résultat s'en déduit. □

On peut remarquer plusieurs choses :

1. Avec à peu près les mêmes outils (et un Lemme 3.8 en déviation), on peut obtenir un résultat similaire en déviation. Par ailleurs, le résultat est encore valide si on remplace Gaussien par sous-Gaussien de même variance (pour les ε).
2. Le coefficient devant $\ell(\theta_S^*, \theta^*)$ est toujours plus grand que 1, et le faire tendre vers 1 fait exploser le terme de pénalité. C'est un défaut courant des méthodes par pénalisation. On peut arriver à un facteur 1 et une pénalité "stable" en utilisant plutôt des méthodes **d'aggrégation** (toujours meilleures en prédiction).
3. Si on prend $K > 0$, on peut montrer que la pénalité n'est pas suffisante, au sens que le modèle sélectionné sera avec proba minorée toujours le plus gros (overfitting).

Le Théorème 3.9 traite le cadre général, pour n'importe quelle famille de modèles et famille de poids. Dans un cas simple où on fait rentrer les variables une par une dans le support, c'est à dire $\mathcal{M} = \{[\![1, s]\!] \mid s = 1, \ldots, D\}$, une pondération uniforme $\pi_S \equiv \frac{1}{D}$ donne un facteur $\log(D)$ dans la borne, et si $\bar{d} = \min\{s \mid S(\theta^*) \subset [\![1, s]\!]\}$, le Théorème 3.9 donne alors

$$\begin{aligned}
\mathbb{E}\ell(\hat{\theta}_{\hat{S}}, \theta^*) &\leq \frac{(1+K)^3}{K}\sigma^2\left(\sqrt{\bar{d}}+\sqrt{2\log(D)}\right)^2+3\frac{(1+K)^3}{K^2}\sigma^2\\
&\lesssim C\sigma^2(\bar{d}+\log(D)),
\end{aligned}$$

la dimension ambiante n'intervient plus qu'en log. On obtient se résultat en prenant $S = [\![1, \bar{d}]\!]$ dans la borne (c'est à dire en comparant au "bon" modèle), mais on peut aussi considérer la borne totale

$$\mathbb{E}\ell(\hat{\theta}_{\hat{S}}, \theta^*) \leq C \inf_{s=1,\ldots,D} \ell(\theta_s^*, \theta^*) + \sigma^2(s + \log(D)),$$

attestant que $\hat{s}$ sera toujours plus petit que $\bar{d}$, parfois strictement, ce qui correspond à une sélection de modèle "faux" mais avec une meilleure performance en prédiction (encore un compromis approximation/estimation). Cet exemple est pour la forme : $\bar{d}$ ne reflète pas vraiment l'hypothèse de parcimonie de début de chapitre $\|\theta^*\|_0 = d$ (de fait on peut avoir $\|\theta^*\|_0 = d$ et $\bar{d} = D$).

Pour traiter le contexte de parcimonie dans le cas général, il va falloir prendre comme modèle

$$\mathcal{M} = \bigcup_{j=1}^{D} \mathcal{S}_j,$$

où $\mathcal{S}_j = \{S \subset [\![1, D]\!] \mid |S| = j\}$, c'est à dire tous les supports possibles. Cela nous amène à réfléchir à la pondération π_S. Idéalement, on devrait avoir $\log(1/\pi_S)$ du même ordre de grandeur (ou plus petit) que $r(S)$, pour ne pas trop déteriorer la borne, avec toutefois la contrainte de somme à 1. On peut se tourner vers des pénalités qui ne dépendent que de la taille du modèle S, c'est à dire en $\pi_S = f(|S|)$. La contrainte de sommation à 1 se réécrit alors

$$\begin{aligned} 1 &\geq \sum_{S \in \mathcal{M}} \pi_S \\ &= \sum_{j=1}^{D} \binom{D}{j} f(j). \end{aligned}$$

On va devoir prendre en compte la massivité de la classe $\mathcal{S}_j$, c'est à dire le nombre de modèles ayant pour sparsité j. On peut dès maintenant s'apercevoir que demander $\log(1/\pi_S)$ de la forme $\alpha|S|$ va nécessiter un coefficient α satisfaisant

$$1 \geq \sum_{j=1}^{D} \binom{D}{j} e^{-\alpha j} = (1 + e^{-\alpha})^D - 1.$$

On en déduit

$$(1 + e^{-\alpha})^D \leq 2,$$

ce qui mène à

$$\frac{D}{2} e^{-\alpha} \leq D \log(1 + e^{-\alpha}) \leq \log(2),$$

soit

$$\alpha \geq \log(D) - \log(2 \log(2)).$$

Bref, les $\log(1/\pi_S)$ vont devoir faire intervenir la dimension ambiante D, moralement au moins en $\log(D)$. On peut montrer que cette dépendance minimale en $\log(D)$ peut être atteinte.

Corollaire 3.10

Sous les conditions du Théorème 3.9, pour

$$\pi_S = \frac{e^{-|S|}(e-1)}{\binom{D}{|S|}(1 - e^{-D})},$$

on a

$$\mathbb{E}\ell(\hat{\theta}_{\hat{S}}, \theta^*) \leq C \inf_{\theta} \left[\ell(\theta, \theta^*) + \sigma^2 \|\theta\|_0 \left(1 + \log\left(\frac{D}{\|\theta\|_0} \right) \right) \right].$$

Preuve du Corollaire 3.10. On commence par remarquer que si $S(\theta) = S$, alors

$$\begin{aligned} &\inf_{S(\theta)=S} \left[\ell(\theta, \theta^*) + \sigma^2 \|\theta\|_0 \left(1 + \log\left(\frac{D}{\|\theta\|_0} \right) \right) \right] \\ &= \ell(\theta_S^*, \theta^*) + \sigma^2 |S| \left(1 + \log\left(\frac{D}{|S|} \right) \right), \end{aligned}$$

de sorte que le membre de droite dans le résultat du corollaire vaut en fait

$$C\left[\inf_{S\in\mathcal{M}} \ell(\theta_S^*,\theta^*) + \sigma^2|S|\left(1+\log\left(\frac{D}{|S|}\right)\right)\right],$$

où $\mathcal{M} = \bigcup_{j=1}^{D} \mathcal{S}_j$ (précédemment défini). Reste à travailler sur les poids π_S. Pour cela on aura besoin d'un lemme combinatoire.

Lemme 3.11

Pour $0 \le s \le D$, on a

$$\log\left(\binom{D}{s}\right) \le s\left(1+\log\left(\frac{D}{s}\right)\right),$$

avec la convention $0\log(0) = 0$.

Preuve du Lemme 3.11. On prouve ça par récurrence. Pour $s = 0$, c'est évident. Supposons la propriété vraie au rang s, on a alors

$$\begin{aligned}
\binom{D}{s+1} &= \binom{D}{s} \times \frac{D-s}{s+1} \\
&\le \left(\frac{eD}{s}\right)^s \times \frac{D}{s+1} \\
&\le \left(\frac{eD}{s+1}\right)^s \times \left(1+\frac{1}{s}\right)^s \frac{D}{s+1} \\
&\le \left(\frac{eD}{s+1}\right)^s \frac{eD}{s+1},
\end{aligned}$$

où on a utilisé $\left(1+\frac{1}{n}\right)^n \le e$. □

Du Lemme 3.11, on déduit immédiatement que le choix de poids du Corollaire 3.10, on a

$$\log(\pi_S^{-1}) = \log\left(\binom{D}{s}(1-e^{-D})\right) + s - \log(e-1) \le 2s\left(1+\log\left(\frac{D}{s}\right)\right).$$

Par ailleurs, comme $r_S \le s$, le terme de droite dans Théorème 3.9 est bien de la forme

$$C\left[\inf_{S\in\mathcal{M}} \ell(\theta_S^*,\theta^*) + \sigma^2|S|\left(1+\log\left(\frac{D}{|S|}\right)\right)\right].$$

Il reste donc à vérifier que $\sum_S \pi_S \le 1$. On redécompose via les $\mathcal{S}_j$:

$$\begin{aligned}
\sum_{S\in\mathcal{M}} \pi_S &= \sum_{j=1}^{D} \binom{D}{j} \pi_j \\
&= \sum_{j=1}^{D} \frac{e^{-j}(e-1)}{1-e^{-D}} \\
&= \frac{e-1}{1-e^{-D}} e^{-1} \frac{e^{-D}-1}{e^{-1}-1} = 1.
\end{aligned}$$

□

Deux remarques :

1. Un choix en $\pi_S = \frac{1}{D}\binom{D}{s}^{-1}$ donnerait plus facilement un terme en $s\log(D)$ dans l'inégalité finale. Pour récupérer un peu mieux, on est obligés de pénaliser un peu plus fortement les modèles de grandes cardinalités, en prendant quelque chose de proportionnel à $\binom{D}{s}^{-1}e^{-s}$ (le $\binom{D}{s}^{-1}$ n'est là que pour tenir compte de la cardinalité de $\mathcal{S}_s$, à ce facteur près cela revient à comparer une uniforme et une exponentielle tronquée).
2. De ce corollaire on déduit immédiatement

$$\mathbb{E}\ell(\hat{\theta}_{\hat{S}}, \theta^*) \leq C\sigma^2\|\theta^*\|_0\left(1 + \log\left(\frac{\|\theta^*\|_0}{D}\right)\right).$$

Là encore on peut éventuellement faire mieux pour un $\hat{\theta}_S$ de support plus petit que $S(\theta^*)$. Cela montre en tout cas que la dimension ambiente n'intervient qu'en logarithme. Ce $\log(D/s)$ est le prix à payer en termes de prédictions à ne pas connaître le bon support (si on le connaissait on pourrait l'enlever). On voit alors qu'on peut a priori faire beaucoup mieux que la borne des moindres carrés standard $r\sigma^2$, si $\|\theta^*\|_0$ est assez petit. Là encore pas de contradiction avec l'optimalité des moindres carrés en général. Si il est prouvé que

$$\inf_{\hat{\theta}} \sup_{\theta\in\mathbb{R}^D} \mathbb{E}\ell(\hat{\theta}, \theta^*) = r\sigma^2,$$

borne atteinte par les moindres carrés ordinaires, nous venons juste de prouver que

$$\inf_{\hat{\theta}} \sup_{\|\theta^*\|_0=d} \mathbb{E}\ell(\hat{\theta}, \theta^*) \leq Cd\sigma^2\left(1 + \log\left(\frac{D}{d}\right)\right),$$

c'est à dire mieux sur une classe de θ plus restreinte. Si on prend le sup en d de cette borne (pour retrouver le cadre général $\theta \in \mathbb{R}^D$), on retrouve le $CD\sigma^2$ classique. Signalons enfin que cette vitesse est au facteur près optimale : on peut prouver que

$$\inf_{\hat{\theta}} \sup_{\|\theta^*\|_0=d} \mathbb{E}\ell(\hat{\theta}, \theta^*) \geq cd\sigma^2\left(1 + \log\left(\frac{D}{d}\right)\right),$$

pour une constante c. On peut trouver une preuve de ce résultat dans Verzelen [2012] ou Giraud [2022]. La pénalisation L_0 est donc optimale (en terme de dépendance en la parcimonie, dimension ambiante, et taille d'échantillon).

Aggrégation à poids exponentiels en régression linéaire

Dans ce contexte de régression linéaire Gaussienne à design fixe, on peut donner des bornes précises en espérance pour les techniques d'aggrégation à poids exponentiels (voir Dalalyan and Tsybakov [2007] par exemple), qui sont strictement meilleures que celles obtenues par sélection de modèle. Ces bornes sont basées sur l'exploitation du Lemme 2.14 encore, combinée avec des estimations sans biais de $\ell(\hat{\theta}_m, \theta^*)$.

On se donne encore la collection de modèle $\mathcal{M}$ (correspondant à des supports différents), et pour $m \in \mathcal{M}$, $\hat{\theta}_m$ un prédicteur par moindres carrés. Un estimateur sans biais de l'excès de risque de $\hat{\theta}_m$ est alors donné par

$$r_m = \|Y - X\hat{\theta}_m\|^2 + 2d_m\sigma^2 - n\sigma^2,$$

où $d_m = \text{rang}(X_m)$. On regarde alors le régresseur obtenu par aggrégation à poids exponentiels,

$$\hat{\theta} = \sum_{m\in\mathcal{M}} \omega_m \hat{\theta}_m,$$

avec pour poids

$$\omega_m = \frac{\pi_m \exp\left(-\frac{r_m}{\beta}\right)}{Z},$$

avec $Z = \sum_{m\in\mathcal{M}} \pi_m \exp\left(-\frac{r_m}{\beta}\right)$, et on s'intéresse à l'excès de risque de $\hat{\theta}$ en espérance.

Théorème 3.12

Pour $\beta \geq 4\sigma^2$, on a

$$\begin{aligned}\mathbb{E}\ell(\hat{\theta}, \theta^*) &\leq \inf_{m\in\mathcal{M}} \mathbb{E}\ell(\hat{\theta}_m, \theta^*) + \beta \log\left(\frac{1}{\pi_m}\right) \\ &= \inf_{m\in\mathcal{M}} \ell(\theta^*_m, \theta^*) + d_m\sigma^2 + \beta \log\left(\frac{1}{\pi_m}\right).\end{aligned}$$

La preuve de ce Théorème est basée sur le Lemme 2.14 et la formule de Stein.

Proposition 3.13 : Formule de Stein

Soit $Y \sim \mathcal{N}(\mu, \sigma^2 I_n)$, et $g : \mathbb{R}^n \to \mathbb{R}^n$ d'applications coordonnées $(g_i)_{i=1,\dots,n}$. Si g est C^1, et satisfait, pour tout $i \in [\![1, n]\!]$:

a) Pour tout $(y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n) \in \mathbb{R}^{n-1}$, $\lim_{|y_i|\to+\infty} g_i(y_{1:n}) \exp(-(\mu_i - y_i)^2/(2\sigma^2)) = 0$,

b) $\mathbb{E}|\partial_i g_i(y_{1:n})| < +\infty$,

alors on a

$$\mathbb{E}\|g(Y) - \mu\|^2 = \mathbb{E}\left(\|g(Y) - Y\|^2 - n\sigma^2 + 2\sigma^2 \text{div}(g)(Y)\right),$$

où $\text{div}(g)(y) = \sum_{i=1}^n \partial_i g_i(y)$.

Preuve de la Proposition 3.13. Soit $(y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n) \in \mathbb{R}^{n-1}$ fixé. Une intégration par partie donne (les conditions a) et b) assurent que tout le monde existe)

$$\sigma^2 \int_{\mathbb{R}} \partial_i g_i(y_{1:n}) \exp(-(y_i - \mu_i)^2/(2\sigma^2)) dy_i = \int_{\mathbb{R}} (y_i - \mu_i) g_i(y_{1:n}) \exp(-(y_i - \mu_i)^2/(2\sigma^2)) dy_i.$$

On en déduit alors que (en intégrant sur les $(n-1)$ coordonnées restantes et en sommant),

$$\begin{aligned}\sigma^2\mathbb{E}(\mathrm{div}(g)(Y)) &= \sum_{i=1}^{n}\mathbb{E}((Y_i-\mu_i)g_i(Y)) \\ &= \mathbb{E}\left\langle (Y-\mu), g(Y)\right\rangle .\end{aligned}$$

On conclut en décomposant

$$\begin{aligned}\mathbb{E}(\|g(Y)-Y\|^2) &= \mathbb{E}(\|g(Y)-\mu\|^2) + \mathbb{E}(\|Y-\mu\|^2) + 2\mathbb{E}\left\langle g(Y)-\mu, \mu - Y\right\rangle \\ &= \mathbb{E}(\|g(Y)-\mu\|^2) + n\sigma^2 - 2\sigma^2\mathbb{E}(\mathrm{div}(g)(Y)).\end{aligned}$$

□

On peut maintenant passer à la preuve du Théorème 3.12.

Preuve du Théorème 3.12. . Pour alléger un peu les notations, on note $\hat{f}_m = X_m\hat{\theta}_m$, $f^* = X\theta^*$, et $\hat{f} = \sum_{m\in\mathcal{M}}\omega_m\hat{f}_m = X\hat{\theta}$. La formule de Stein donne alors

$$\begin{aligned}\mathbb{E}(\ell(\hat{\theta},\theta^*)) &= \mathbb{E}\|\hat{f}-f^*\|^2 \\ &= \mathbb{E}(\|\hat{f}-Y\|^2) - n\sigma^2 + 2\sigma^2\mathbb{E}\mathrm{div}(\hat{f})(Y),\end{aligned}$$

où $\hat{f}$ est vu comme une fonction de Y. On peut remarquer que la formule de Stein s'applique, car $\hat{f}_m(y) = X_m(X_m^TX_m)^\dagger X_m^Ty$ est linéaire en y, et r_m quadratique et donc $\omega_m\hat{f}_m$ satisfait bien les conditions de la Proposition 3.13, ce dont on déduit que c'est le cas pour $\hat{f}$. Il s'agit alors de contrôler

$$\begin{aligned}\hat{r} &= \|\hat{f}-y\|^2 - n\sigma^2 + 2\sigma^2\mathrm{div}(\hat{f})(y) \\ &= \|\hat{f}-y\|^2 - n\sigma^2 + 2\sigma^2\sum_{m\in\mathcal{M}}\mathrm{div}(\omega_m(y)\hat{f}_m(y)).\end{aligned}$$

Commençons par le terme de divergence. À m fixé, on a

$$\mathrm{div}(\omega_m(y)\hat{f}_m(y)) = \omega_m(y)\mathrm{div}(\hat{f}_m(y)) + \left\langle \nabla_y\omega_m, \hat{f}_m\right\rangle .$$

Or $\hat{f}_m(y) = A_my$, pour $A_m = X_m(X_m^TX_m)^\dagger X_m^T$ (matrice de projection orthogonale sur $V(X_m)$). Donc $\partial_i\hat{f}_{m,i}(y) = (A_m)_{i,i}$, et $\mathrm{div}(\hat{f}_m(y)) = \mathrm{Tr}(A_m) = d_m$. Par ailleurs, on a

$$\begin{aligned}\log(\omega_m) &= \log(\pi_m) - (1/\beta)r_m(y) - \log(Z) \\ &= \log(\pi_m) - (1/\beta)\left(\|y-\hat{f}_m(y)\|^2 + 2d_m\sigma^2 - n\sigma^2\right) - \log(Z),\end{aligned}$$

ce dont on déduit

$$\begin{aligned}\frac{\nabla_y\omega_m}{\omega_m} &= -\frac{2}{\beta}\left(D_y(y-\hat{f}_m(y))\right)^T(y-\hat{f}_m(y)) - \frac{\nabla_y Z}{Z} \\ &= -\frac{2}{\beta}(y-\hat{f}_m(y)) - \frac{\nabla_y Z}{Z} \qquad ((y-\hat{f}_m(y)) \text{ est la projection orthgonale sur } V(X_m)^\perp) \\ &= -\frac{2}{\beta}(y-\hat{f}_m(y)) - \frac{1}{Z}\sum_{\ell\in\mathcal{M}}\nabla_y(\pi_\ell e^{-\frac{1}{\beta}r_\ell(y)}) \\ &= -\frac{2}{\beta}(y-\hat{f}_m(y)) + \sum_{\ell\in\mathcal{M}}\frac{2}{\beta}(y-\hat{f}_\ell(y)\omega_\ell(y) \\ &= \frac{2}{\beta}(\hat{f}_m(y) - \sum_{\ell\in\mathcal{M}}\omega_\ell(y)\hat{f}_\ell(y)) = \frac{2}{\beta}(\hat{f}_m - \hat{f}),\end{aligned}$$

et enfin

$$\operatorname{div}(\omega_m(y)\hat{f}_m(y)) = d_m\omega_m + \frac{2}{\beta}\omega_m \left\langle \hat{f}_m - \hat{f}, \hat{f}_m \right\rangle. \tag{3.8}$$

Le premier terme $\|\hat{f} - y\|^2$ peut lui se réécrire

$$\|\hat{f} - y\|^2 = \sum_{m\in\mathcal{M}} \omega_m \|\hat{f}_m - y\|^2 - \sum_{m\in\mathcal{M}} \omega_m \|\hat{f} - \hat{f}_m\|^2 \qquad \text{(décomposition biais/variance).} \tag{3.9}$$

En combinant (3.8) et (3.9), on obtient

$$\begin{aligned}
\hat{r} &= \sum_{m\in\mathcal{M}} \omega_m \|\hat{f}_m - y\|^2 - \sum_{m\in\mathcal{M}} \omega_m \|\hat{f} - \hat{f}_m\|^2 - n\sigma^2 + 2\sigma^2 \sum_{m\in\mathcal{M}} d_m\omega_m + \frac{2}{\beta}\omega_m \left\langle \hat{f}_m - \hat{f}, \hat{f}_m \right\rangle \\
&= \sum_{m\in\mathcal{M}} \omega_m r_m - \sum_{m\in\mathcal{M}} \omega_m \|\hat{f} - \hat{f}_m\|^2 + \frac{4\sigma^2}{\beta} \sum_{m\in\mathcal{M}} \left\langle \hat{f}_m - \hat{f}, \hat{f}_m - \hat{f} \right\rangle \\
&\qquad \left(\left\langle \hat{f}, \textstyle\sum_{m\in\mathcal{M}} \omega_m(\hat{f}_m - \hat{f}) \right\rangle = 0\right) \\
&= \sum_{m\in\mathcal{M}} \omega_m r_m + \left(\frac{4\sigma^2}{\beta} - 1\right) \sum_{m\in\mathcal{M}} \omega_m \|\hat{f} - \hat{f}_m\|^2 \\
&\leq \sum_{m\in\mathcal{M}} \omega_m r_m \qquad (\beta \geq 4\sigma^2).
\end{aligned}$$

On peut maintenant dérouler le Lemme 2.14. Si q est une famille de poids donnée non aléatoire, on peut écrire

$$\begin{aligned}
\mathbb{E}\ell(\hat{\theta}, \theta^*) &= \mathbb{E}\hat{r} \\
&\leq \mathbb{E}\left(\sum_{m\in\mathcal{M}} \omega_m r_m \right) \\
&\leq \mathbb{E}\left(\sum_{m\in\mathcal{M}} q_m r_m \right) + \beta \mathrm{d}_{KL}(q, \pi) \\
&\leq \sum_{m\in\mathcal{M}} q_m \mathbb{E}\ell(\hat{\theta}_m, \theta^*) + \beta \mathrm{d}_{KL}(q, \pi).
\end{aligned}$$

Pour $q_m = \mathbb{1}_m$, on obtient bien

$$\begin{aligned}
\mathbb{E}\ell(\hat{\theta}, \theta^*) &\leq \mathbb{E}\ell(\hat{\theta}_m, \theta^*) + \beta \log\left(\frac{1}{\pi_m}\right) \\
&= \ell(\theta^*_m, \theta^*) + d_m\sigma^2 + \beta \log\left(\frac{1}{\pi_m}\right).
\end{aligned}$$

□

On remarque que là encore, plutôt que de se comparer aux $\hat{\theta}_m$, on pourrait énoncer un résultat plus général comparant les performances de $\hat{\theta}$ à toute combinaison convexe des $\hat{\theta}_m$. Comparativement au Théorème 3.9, c'est à dire à la sélection de modèle, bien que les termes de variances soient du même ordre de grandeur (en $d_m\sigma^2 + \log(1/\pi_m)$), le coefficient devant $\ell(\hat{\theta^*_m}, \theta^*)$ vaut cette fois-ci exactement 1, comme annoncé. Dans ce cadre de la régression linéaire, les méthodes par aggrégation sont donc optimales, comparativement aux méthodes par sélection de modèle.

À noter que dans les deux méthodes la connaissance de σ^2 est requise au préalable pour atteindre les bornes optimales.

Concernant le choix des poids π_m, ce sont globalement les mêmes que ceux utilisés en sélection de modèle, ils vont dépendre du nombre de modèles et de leur dimension. A titre d'exemple on peut énoncer un corollaire du Théorème 3.12 dans le cas où

$$\mathcal{M} = \bigcup_{j=1}^{D} \mathcal{S}_j.$$

Corollaire 3.14

Sous les conditions du Théorème 3.12, pour

$$\pi_m = \frac{e^{-|m|}(e-1)}{\binom{D}{|m|}(1-e^{-D})},$$
$$\beta = 4\sigma^2,$$

on a

$$\mathbb{E}\ell(\hat{\theta}, \theta^*) \leq \inf_{\theta} \left[\ell(\theta, \theta^*) + \sigma^2 \|\theta\|_0 \left(9 + 8 \log \left(\frac{D}{\|\theta\|_0} \right) \right) \right]$$

Preuve du Corollaire 3.14. Cela découle de l'inégalité

$$\log(\pi_m^{-1}) =\leq 2|m| \left(1 + \log \left(\frac{D}{|m|} \right) \right)$$

déjà utilisée, combinée avec le Théorème 3.12, où on remarque que $d_m \leq |m|$ pour arriver au résultat final. □

On a là encore une procédure optimale sous contrainte de parcimonie (en utilisant les bornes inférieures mentionnées dans Verzelen [2012], Giraud [2022]). La question naturelle qui se pose est alors : si les prédicteurs par aggrégation sont toujours meilleurs que ceux obtenus par sélection de modèle, pourquoi ne pas les utiliser systématiquement ? Deux raisons parmi d'autres :

1. Avoir une information sur les variables pertinentes pour le problème de régression est intéressant en soi (interprétabilité), ce que ne permet pas l'aggrégation. On peut toutefois nuancer en remarquant que les poids les plus élevés dans le prédicteur aggrégé correspondent à peu près aux modèles qu'aurait sélectionné une procédure de sélection de modèles.
2. Pour une nouvelle donnée x, calculer la prédiction $\hat{f}(x)$ pour le prédicteur aggrégé nécessite de calculer $\hat{f}_m(x)$ pour tous les modèles dans la collection, ce qui peut prendre beaucoup de temps notamment dans un contexte de grande dimension. À contrario, si la procédure de sélection de modèles a sélectionné un $\hat{m}$ de faible cardinalité, calculer $\hat{f}_{\hat{m}}(x)$ peut se faire très rapidement.

Pénalisation L_0 en pratique

Pour le choix quasiment optimal $\pi_S = D^{-(s+1)}$ on peut réécrire la procédure de sélection de modèle comme

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^D} R_n(\theta) + \lambda\sigma^2 \|\theta\|_0,$$

c'est à dire comme minimiseur d'un risque empirique pénalisé, avec $\lambda = C\log(D)$. La norme L_0 étant non convexe, on ne peut pas vraiment se servir de cette formulation, et en toute généralité on est obligé d'aller regarder les moindres carrés $\hat{\theta}_S$ sur chaque $S \subset \{1, \dots, D\}$ et de choisir le meilleur après pénalisation (du coup choisir plutôt la pénalisation optimale). On est donc amenés à effectuer 2^D ajustements, ce qui n'est faisable, même en parallélisant, que pour des D largement en dessous de la cinquantaine (on est loins de la très grande dimension alors). En pratique, la sélection de modèle s'opère après une préselection grossière de variables (avec du Lasso par exemple) ou de manière approchée (avec des algorithmes itératifs forward ou backward, consistant à rajouter ou enlever des variables à un modèle itérativement, c'est le cas des procédures AIC ou BIC).

Il y a toutefois une situation où cette sélection de modèle est tout le temps faisable : celle où la matrice de design est orthogonale.

Cas orthogonal : lien avec le "seuillage dur"

Regardons de plus près les solutions "pratiques" de

$$\arg\min_S \|Y - X\hat{\theta}_S\|^2 + \lambda\sigma^2|S|,$$

pour un λ donné. Regardons le $\hat{S}$ sélectionné, de cardinal $\hat{s}$. On a alors, pour toute variable $j \in \hat{S}$,

$$\|Y - \pi_{V(X_{\hat{S}})}Y\|^2 + \sigma^2\lambda\hat{s} < \|Y - \pi_{V(X_{\hat{S}-j})}Y\|^2 + \sigma^2\lambda(\hat{s}-1),$$

ce qui équivaut à

$$\|\pi_{V(X_{\hat{S}})}Y - \pi_{V(X_{\hat{S}-j})}Y\|^2 > \sigma^2\lambda.$$

En particulier, on doit avoir $X_{\hat{S}}$ de rang $\hat{s}$, et donc $\hat{\theta}_j \neq 0$ (on rappelle ici que par convention quand X n'est pas inversible on prend le θ de plus petite norme qui donne $\pi_{V(X)}Y = X\theta$). Par ailleurs, on peut remarquer que

$$\pi_{V(X_{\hat{S}})}Y = X_{\hat{S}}\hat{\theta}_{\hat{S}} = X_{\hat{S}-j}\hat{\theta}^{-j}_{\hat{S}} + X^j\hat{\theta}_j,$$

où X^j est la j-ème colonne de X. On en déduit

$$\|\pi_{V(X_{\hat{S}})}Y - \pi_{V(X_{\hat{S}-j})}Y\|^2 \leq \hat{\theta}_j^2\|X^j\|^2,$$

et donc nécessairement

$$|\hat{\theta}_j| > \frac{\sigma\sqrt{\lambda}}{\|X^j\|}.$$

Dans le cas où X est formée de vecteurs orthonormés, cette condition nécessaire est suffisante, et la pénalisation ℓ_0 est équivalente à regarder $\hat{\theta}$ l'estimateur par moindre carrés sur $V(X)$, et à seuiller les coefficients, c'est à dire

$$\hat{\theta}_{\hat{S}} = \left(\hat{\theta}_j \mathbb{1}_{|\hat{\theta}_j| > \sigma\sqrt{\lambda}}\right)^T_{j=1,\dots,D}.$$

Dans le cadre d'une matrice de design X orthonormée, la sélection de modèle ou pénalisation L_0 est donc équivalente au seuillage "dur" des coefficients des moindres

carrés ordinaires. On peut comparer ça avec le prédicteur Ridge, qui dans la même situation effectue une contraction de ces coefficients. On verra juste après un intermédiaire entre les deux : dans ce même cadre le Lasso va effectuer un seuillage "mou" des coefficients.

Lorsque D est vraiment grand et que la matrice de design n'est pas orthogonale, on est obligés de se tourner vers des relaxations convexes de la pénalisation L_0.

3.2.3 Lasso

La sélection de modèle de la section précédente peut se réécrire sous la forme

$$\hat{\theta} \in \arg\min_{\theta} \|Y - X\theta\|^2 + \lambda \|u\|_0,$$

pour un λ assez grand. Comme cette fonction de θ n'est pas convexe, on en est réduit à parcourir les sous-espaces à support fixés ou éventuellement à seuiller dans les cas favorables.

L'idée du Lasso est de prendre l'enveloppe convexe (inférieure) de $\theta \mapsto \|\theta\|_0$, qui est juste $\theta \mapsto \|\theta\|_1$ (faire un dessin en dimension 1 pour s'en convaincre). Pour rappel, l'enveloppe convexe de f est définie par $f_c(x) = \sup\{g(x) \mid g \text{ est convexe et } g \leq f\}$.

On va donc chercher une solution de

$$\arg\min_{\theta} \|Y - X\theta\|^2 + \lambda \|\theta\|_1.$$

Cette fonction est propre et convexe, ses minimums sont donc nécessairement atteints. Si on se place du point de vue équivalent de la minimisation de risque empirique sous contrainte, c'est à dire

$$\arg\min_{\|\theta\|_1 \leq R} \|Y - X\theta\|^2,$$

on peut intuiter que les solutions de ce problème convexe peuvent être parcimonieuses, du fait que les boules L_1 "ont des coins" (FAIRE LES 3 DESSINS : min dans la boule l1, lignes de niveaux ellipses qui tapent un coin, lignes de niveau l1 qui tapent un milieu de segment.).

Regardons cela de plus près.

Point de vue analytique : seuillage doux des coefficients

Soit $\hat{\theta}$ un tel minimum. On va regarder des conditions nécessaires coordonnée par coordonnée. On note e_j le vecteur $(0, \ldots, 0, 1, 0, \ldots, 0)^T$, avec le 1 à la j-ème place, et f la fonction à minimiser.

Supposons $\hat{\theta}_j \neq 0$, et prenons $h \in \mathbb{R}$ petit. On a alors

$$f(\hat{\theta} + he_j) - f(\hat{\theta}) = -2\left\langle X^j, Y \right\rangle h + 2(X^T X\hat{\theta})_j h + \lambda h \mathrm{sg}(\hat{\theta}_j) + O(h^2) \geq 0,$$

où $\mathrm{sg}(x)$ représente le signe de x, à valeurs dans $\{-1, 0, 1\}$. On en déduit alors que nécessairement

$$(X^T X\hat{\theta})_j = \left\langle X^j, Y \right\rangle - \frac{\lambda}{2}\mathrm{sg}(\hat{\theta}_j). \tag{3.10}$$

Supposons maintenant que $\hat{\theta}_j = 0$ et prenons h positif. On a alors

$$f(\hat{\theta} + he_j) - f(\hat{\theta}) - 2\left\langle X^j, Y \right\rangle h + 2(X^T X\hat{\theta})_j h + \lambda h + O(h^2) \geq 0,$$

ce dont on déduit

$$\left\langle X^j, Y \right\rangle - (X^T X\hat{\theta})_j \leq \frac{\lambda}{2}.$$

Pour h négatif, le même raisonnement donne

$$\left\langle X^j, Y \right\rangle - (X^T X\hat{\theta})_j \geq -\frac{\lambda}{2}.$$

Les deux équations combinées donnent

$$\left|\left\langle X^j, Y \right\rangle - (X^T X\hat{\theta})_j\right| \leq \frac{\lambda}{2}. \tag{3.11}$$

En fait, les conditions (3.10) et (3.11) se résument à

$$0 \in \partial_{\hat{\theta}} f,$$

où $f : \theta \mapsto \|Y - X\theta\|^2 + \lambda\|\theta\|_1$. Les conditions de Slater (Théorème 2.32) étant évidemment vérifiées, le Théorème de Karush-Kuhn-Tucker (Théorème 2.33) implique que (3.10) et (3.11) sont des conditions nécessaires et suffisantes d'optimalité.

Cas orthonormé

Dans le cas orthonormé où $X^T X = I_D$, pour un $\hat{\theta}_j$ non nul, la condition (3.11) devient

$$\hat{\theta}_j + \frac{\lambda}{2}\mathrm{sg}(\hat{\theta}_j) = \left\langle X^j, Y \right\rangle.$$

On en déduit que $\mathrm{sg}(\hat{\theta}_j) = \mathrm{sg}(\langle X^j, Y\rangle)$, et

$$\hat{\theta}_j = \mathrm{sg}(\left\langle X^j, Y \right\rangle)(\left|\left\langle X^j, Y \right\rangle\right| - \frac{\lambda}{2}).$$

Les coordonées où $\hat{\theta}_j = 0$ vérifient

$$\left|\left\langle X^j, Y \right\rangle\right| \leq \frac{\lambda}{2}.$$

Ces conditions étant suffisantes, on déduit alors que (toujours dans le cas orthonormé),

$$\hat{\theta}_j = \mathrm{sg}(\left\langle X^j, Y \right\rangle)\left(\left|\left\langle X^j, Y \right\rangle\right| - \frac{\lambda}{2}\right)_+ = \mathrm{sg}(\hat{\theta}_{LS,j})\left(\left|\hat{\theta}_{LS,j}\right| - \frac{\lambda}{2}\right)_+,$$

où $\hat{\theta}_{LS}$ désigne l'estimateur par moindre carrés standard. L'estimateur Lasso dans ce cas effectue un seuillage doux : les coefficients trop petits de $\hat{\theta}_{LS}$ seront mis à 0 (comme en pénalisation L_0), ceux suffisamment grands seront décalés vers 0 d'un facteur $\lambda/2$. Cette contraction vers 0 est différente de celle opérée par un prédicteur Ridge (on retranche λ au coefficient plutôt que de le diviser par $(1 + \lambda)$).

Dans ce cas orthonormé on a évidemment un algorithme simple pour résoudre le problème Lasso, basé sur ce seuillage doux.

Cas non orthonormé

Soit $\hat{S}$ le support de $\hat{\theta}$. La condition (3.10) donne alors

$$0 \leq \hat{\theta}^T X^T X \hat{\theta} = \sum_{j \in \hat{S}} \hat{\theta}_j \left(\left\langle X^j, Y \right\rangle - \frac{\lambda}{2} \mathrm{sg}(\hat{\theta}_j) \right),$$

ce dont on déduit que $\hat{\theta} = 0$ si $\lambda \geq 2\|X^T Y\|_\infty$. Pour $\lambda < 2\|X^T Y\|_\infty$, $\hat{\theta}$ est non nul mais on ne peut plus facilement caractériser son support .

On peut aussi comparer la prédiction Lasso $\hat{f} = X_{\hat{S}} \hat{\theta}_{\hat{S}}$ à la prédiction moindre carrés sur le même support $\hat{S}$, $\hat{f}_{LS} = X_{\hat{S}} (X_{\hat{S}}^T X_{\hat{S}})^\dagger X_{\hat{S}}^T Y = \pi_{V(X_{\hat{S}})}(Y)$. L'équation (3.10) s'écrit, pour tout $j \in \hat{S}$,

$$(X^T X_{\hat{S}} \hat{\theta}_{\hat{S}})_j = (X^T Y)_j - \frac{\lambda}{2} \mathrm{sg}(\hat{\theta}_j),$$

se dont on peut déduire (en regardant les lignes correspondant à $\hat{S}$),

$$X_{\hat{S}}^T X_{\hat{S}} \hat{\theta}_{\hat{S}} = X_{\hat{S}}^T Y - \frac{\lambda}{2} \mathrm{sg}(\hat{\theta}_{\hat{S}}).$$

Ensuite, comme $X_{\hat{S}} (X_{\hat{S}}^T X_{\hat{S}})^\dagger X_{\hat{S}}^T = \pi_{V(X_{\hat{S}})}$, on a

$$\begin{aligned} \hat{f} &= X_{\hat{S}} (X_{\hat{S}}^T X_{\hat{S}})^\dagger \left(X_{\hat{S}}^T X_{\hat{S}} \hat{\theta}_{\hat{S}} \right) \\ &= X_{\hat{S}} (X_{\hat{S}}^T X_{\hat{S}})^\dagger \left(X_{\hat{S}}^T Y - \frac{\lambda}{2} \mathrm{sg}(\hat{\theta}_{\hat{S}}) \right) \\ &= \hat{f}_{LS} - \frac{\lambda}{2} X_{\hat{S}} (X_{\hat{S}}^T X_{\hat{S}})^\dagger \mathrm{sg}(\hat{\theta}_{\hat{S}}). \end{aligned}$$

La prédiction Lasso est donc celle des moindres carrés sur le support du Lasso, à une déviation linéaire en λ près. Dans le cas orthogonal, cette déviation consiste en une diminution de ces coefficients. Dans le cadre général, pour garantir que $\mathrm{sg}(\hat{\theta}_{\hat{S}}) = \mathrm{sg}(\hat{f}_{LS})$ (et donc que la déviation aux moindres carrés ordinaires est une diminution), des conditions d'irrépresentabilité sont suffisantes et quasi-nécessaires (voir Zhao and Yu [2006]).

Performances en prédiction du Lasso

L'idée à retenir est que les performances en prédiction du Lasso sont conditionnées au fait que les colonnes de X sont à peu près orthogonales. La **constante de compatibilité** ci-dessous donne une mesure de cette orthogonalité

$$\kappa(\theta) = \min_{u \in \mathcal{C}(\theta)} \frac{\sqrt{|s|}\|Xu\|}{\|u_s\|_1}, \quad \text{avec } s = \mathrm{Supp}(\theta),\ \mathcal{C}(\theta) = \{u \in \mathbb{R}^D \mid \|u_s\|_1 > 5\|u_{s^c}\|_1\}. \tag{3.12}$$

L'intérêt de cette constante réside dans le théorème suivant.

Théorème 3.15

Pour $\lambda \geq 3\|X^T\varepsilon\|_\infty$, on a

$$\ell(\hat{\theta}_\lambda, \theta^*) \leq \inf_\theta \left[\ell(\theta, \theta^*) + \frac{\lambda^2}{\kappa^2(\theta)}\|\theta\|_0\right].$$

En particulier on a

$$\ell(\hat{\theta}_\lambda, \theta^*) \leq \frac{\lambda^2}{\kappa^2(\theta^*)}\|\theta^*\|_0,$$

c'est à dire un excès de risque dépendant de la parcimonie de θ^*, mais on peut avoir une meilleure borne pour un $\bar{\theta}_\lambda$ plus parcimonieux encore. On voit alors l'importance de la constante de compatibilité $\kappa(\theta)$: dans le terme de droite elle ne donne quelque chose d'intéressant que pour les support s tels que X_s sont à peu près orthogonaux et les X_{s^c} ne sont pas trop corrélés avec les X_s. En effet, si X_s est liée, alors $\kappa(\theta) = 0$. A l'inverse, si $\lambda_{\min}(X^TX) := \lambda_{min} > 0$, alors, pour tout θ,

$$\begin{aligned}
\kappa(\theta) &\geq \sqrt{s\lambda_{\min}} \inf_{u \in \mathcal{C}(\theta)} \frac{\|u\|_2}{\|u_s\|_1} \\
&\geq \sqrt{s\lambda_{\min}} \inf_{u \in \mathbb{R}^D} \frac{\|u_s\|_1}{\sqrt{s}\|u_s\|_1} \qquad \text{(concavité de } \sqrt{}\text{)} \\
&\geq \sqrt{\lambda}_{\min}.
\end{aligned}$$

En somme, tout va bien se passer si X_{s^*} est formé de vecteurs à peu près orthogonaux (rappelons qu'en cas de plusieurs θ^* possible on choisit un de plus petit support), et si les autres variables prédictives sont suffisamment peu corrélées avec celles de X_{s^*}.

Preuve du Théorème 3.15. Cette preuve exploite les conditions d'optimalité. On aura besoin du Lemme suivant sur les sous-gradients des fonctions convexes.

Lemme 3.16

Soit f est une fonction convexe sur $\mathbb{R}^k$. Alors, pour tous $x, y \in \mathbb{R}^k$ et $g_x \in \partial_x f$, $g_y \in \partial_y f$, on a

$$\langle g_y - g_x, y - x\rangle \geq 0.$$

Démonstration. Par définition, $f(y) - f(x) \geq \langle g_x, (y - x)\rangle$, et $f(x) - f(y) \geq \langle g_y, x - y\rangle$. Il suffit d'additionner. □

Revenons à la preuve du Théorème. Soit $\hat{\theta}$ l'estimateur Lasso. Les conditions d'optimalité s'écrivent alors

$$2X^T(Y - X\hat{\theta}) = \lambda\hat{g},$$

où $\hat{g} \in \partial_{\hat{\theta}} = \|.\|_1$. On peut alors écrire, pour un concurrent potentiel $u \in \mathbb{R}^D$,

$$2\left\langle X^TY, \hat{\theta} - u\right\rangle - 2\left\langle X^TX\hat{\theta}, \hat{\theta} - u\right\rangle = \lambda\left\langle \hat{g}, \hat{\theta} - u\right\rangle,$$

soit

$$2\left\langle Y, X(\hat{\theta}-u)\right\rangle - 2\left\langle X\hat{\theta}, X(\hat{\theta}-u)\right\rangle = \lambda\left\langle \hat{g}, \hat{\theta}-u\right\rangle.$$

Comme $Y = X\theta^* + \varepsilon$, on obtient

$$2\left\langle X(\hat{\theta}-\theta^*), X(\hat{\theta}-u)\right\rangle = 2\left\langle \varepsilon, X(\hat{\theta}-u)\right\rangle - \lambda\left\langle \hat{g}, \hat{\theta}-u\right\rangle.$$

En utilisant le Lemme 3.16, pour un $g \in \partial_u\|.\|_1$, on a $\left\langle \hat{g}, \hat{\theta}-u\right\rangle \geq \left\langle g, \hat{\theta}-u\right\rangle$, ce dont on déduit

$$2\left\langle X(\hat{\theta}-\theta^*), X(\hat{\theta}-u)\right\rangle \leq 2\left\langle \varepsilon, X(\hat{\theta}-u)\right\rangle - \lambda\left\langle g, \hat{\theta}-u\right\rangle. \tag{3.13}$$

Le terme de gauche peut s'écrire $\|X(\hat{\theta}-\theta^*)\|^2 + \|X(\hat{\theta}-u)\|^2 - \|X(u-\theta^*)\|^2$. Si $2\left\langle X(\hat{\theta}-\theta^*), X(\hat{\theta}-u)\right\rangle \leq 0$, on a alors directement

$$\|X(\hat{\theta}-\theta^*)\|^2 \leq \|X(u-\theta^*)\|^2,$$

et on a rien de plus à faire.

Supposons maintenant que $2\left\langle X(\hat{\theta}-\theta^*), X(\hat{\theta}-u)\right\rangle > 0$ et explicitons maintenant un choix de $g \in \partial_u\|.\|_1$. En notant J le support de u, n'importe quel $g = \mathrm{sg}(u) + h$, avec $h_J = 0$ et $\|h\|_\infty \leq 1$ convient. On prend $h = \mathrm{sg}(\hat{\theta}-u)_{J^c}$, et alors

$$\left\langle g, \hat{\theta}-u\right\rangle \geq -\|(\hat{\theta}-u)_J\|_1 + \|(\hat{\theta}-u)_{J^c}\|_1.$$

L'équation (3.13) donne alors

$$\begin{aligned} 2\left\langle X(\hat{\theta}-\theta^*), X(\hat{\theta}-u)\right\rangle &\leq 2\left\langle \varepsilon, X(\hat{\theta}-u)\right\rangle + \lambda\|(\hat{\theta}-u)_J\|_1 - \lambda\|(\hat{\theta}-u)_{J^c}\|_1 \\ &\leq 2\|X^T\varepsilon\|_\infty\|\hat{\theta}-u\|_1 + \lambda\|(\hat{\theta}-u)_J\|_1 - \lambda\|(\hat{\theta}-u)_{J^c}\|_1 \\ &\leq 2\lambda\|(\hat{\theta}-u)_J\|_1, \end{aligned} \tag{3.14}$$

en utilisant $\lambda \geq 3\|X^T\varepsilon\|_\infty$. Si $(\hat{\theta}-u) \in \mathcal{C}(u)$, on peut en déduire

$$\begin{aligned} 2\lambda\|(\hat{\theta}-u)_J\|_1 &\leq \frac{\sqrt{\|u\|_0}\|X(\hat{\theta}-u)\|}{\kappa(u)} \\ &\leq \|X(\hat{\theta}-u)\|^2 + \frac{\lambda^2}{\kappa^2(u)}\|u\|_0, \end{aligned}$$

menant à son tour à

$$\|X(\hat{\theta}-\theta^*)\|^2 + \|X(\hat{\theta}-u)\|^2 - \|X(u-\theta^*)\|^2 \leq \|X(\hat{\theta}-u)\|^2 + \frac{\lambda^2}{\kappa^2(u)}\|u\|_0,$$

donc au résultat. Il reste à montrer que $(\hat{\theta}-u) \in \mathcal{C}(u)$. Rappelons que $2\left\langle X(\hat{\theta}-\theta^*), X(\hat{\theta}-u)\right\rangle > 0$, ce qui, avec (3.14) donne

$$\begin{aligned} 0 &< \frac{2\lambda}{3}\|\hat{\theta}-u\|_1 + \lambda\|(\hat{\theta}-u)_J\|_1 - \lambda\|(\hat{\theta}-u)_{J^c}\|_1 \\ &< \frac{5\lambda}{3}\|(\hat{\theta}-u)_J\|_1 - \frac{\lambda}{3}\|(\hat{\theta}-u)_{J^c}\|_1, \end{aligned}$$

ce dont on déduit $(\hat{\theta}-u) \in \mathcal{C}(u)$. □

On peut en déduire le corollaire suivant (dans le cadre du modèle Gaussien).

COROLLAIRE 3.17

Si les colonnes de X sont normées, alors, dans le modèle Gaussien, pour

$$\lambda \geq 3\sigma\sqrt{2(\log(D)+x)},$$

on a, avec probabilité plus grande que e^{-x},

$$\ell(\hat{\theta}_\lambda, \theta^*) \leq \inf_\theta \left[\ell(\theta, \theta^*) + \frac{\lambda^2}{\kappa^2(\theta)}\|\theta\|_0\right].$$

Preuve du Corollaire 3.17. Il suffit de vérifier que

$$\mathbb{P}\left(\|X^T\varepsilon\|_\infty \geq \sigma\sqrt{2(\log(D)+x)}\right) \leq e^{-x}.$$

Pour $j = 1, \ldots, D$, on a

$$Z_j = (X^j)^T\varepsilon \sim \mathcal{N}(0, \sigma^2\|X^j\|^2) = \mathcal{N}(0, \sigma^2),$$

les colonnes de X étant normées. On a alors $\mathbb{P}(|Z_j| \geq \sigma\sqrt{2x}) \leq e^{-x}$, et une borne d'union fait le reste. □

On peut en déduire que, si $\|\theta^*\|_0 = d$, alors le choix optimal $\lambda = 3\sigma\sqrt{2(\log(D)+x)}$ donne

$$\ell(\hat{\theta}_\lambda, \theta^*) \leq Cd\frac{\sigma^2}{\kappa^2(\theta^*)}(\log(D)+x),$$

et on retrouve l'ordre de grandeur des vitesses minimax sur cette classe $\sigma^2 d\log(D)$. Plusieurs remarques toutefois :

1. Le terme en $\kappa^2(\theta^*)$ est quasiment nécessaire (Zhang et al. [2014]). C'est le prix à payer pour une non-orthogonalité des colonnes.
2. Dans cette formulation, le terme λ dépend de la "borne de confiance" que l'on veut sur la précision (paramétré par x). Ce n'est pas nécessaire, on peut prendre un λ fixe et montrer une déviation Bellec et al. [2018].
3. On peut arriver à une borne "optimale" en $s\log(D/s)$ en modifiant un peu le terme de pénalité ($\lambda_j \sim \sigma\sqrt{\log(D/j)}$, en regardant un ordre décroissant sur les coordonnées). C'est l'objet des procédures SLOPE Bellec et al. [2018].

Enfin, cette preuve est vraiment adaptée au cadre de la régression linéaire. Dans un cadre général, on peut s'en sortir avec des processus empiriques renormalisés, sous des conditions légèrements différentes. On donne un exemple ici de tel résultat, un lecteur intéressé se réfèrera aux oeuvres complètes de S. Van de Geer.

THÉORÈME 3.18

Pour $s \subset s^* \subset \{1, \dots d\}$, on note

$$\kappa(s) = \min_{(u-\theta_s^*) \in \mathcal{C}(s)} \frac{\sqrt{|s| \ell(u, \theta_s^*)}}{\|(u-\theta^*)\|_{1,s}},$$

où $\theta_s^* = \arg\min_{\mathrm{Supp}(\theta)=s} R(\theta)$, et $\mathcal{C}(s) = \{v \mid \|v_{s^c}\|_1 < 2\|v_s\|_1$. Si

$$\lambda_0 \geq \sup_{u,v} \frac{(R-R_n)(u-v)}{\|u-v\|_1},$$

alors, pour $\lambda \geq 3\lambda_0$, on a

$$\ell(\hat{\theta}, \theta^*) \leq \inf_{\theta | s(\theta) \subset s^*} \ell(\theta, \theta^*) + 4 \frac{\lambda^2 \|\theta\|_0}{\kappa^2(s(\theta))}.$$

Preuve du Théorème 3.18. Soit $s \subset s^* \subset \{1, \dots, D\}$, et $\theta_s^* = \arg\min_{\mathrm{Supp}(\theta)=s} R(\theta)$. De l'inéquation de base

$$R_n(\hat{\theta}) + \lambda \|\hat{\theta}\|_1 \leq R_n(\theta_s^*) + \lambda \|\theta_s^*\|_1,$$

on déduit

$$\begin{aligned} R(\hat{\theta}) + \lambda \|\hat{\theta}\|_1 &\leq R(\theta_s^*) + \lambda \|\theta_s^*\|_1 + (R-R_n)(\hat{\theta} - \theta_s^*) \\ &\leq R(\theta_s^*) + \lambda \|\theta_s^*\|_1 + \lambda_0 \|\hat{\theta} - \theta_s^*\|_1. \end{aligned}$$

En décomposant les normes, on a alors

$$R(\hat{\theta}) + \lambda \|\hat{\theta}\|_{1,s^c} \leq R(\theta_s^*) + (\lambda + \lambda_0) \|\theta_s^* - \hat{\theta}\|_{1,s} + \lambda_0 \|\hat{\theta}\|_{1,s^c},$$

menant à

$$R(\hat{\theta}) - R(\theta_s^*) + (\lambda - \lambda_0) \|\hat{\theta}\|_{1,s^c} \leq (\lambda + \lambda_0) \|\theta_s^* - \hat{\theta}\|_{1,s}. \tag{3.15}$$

Si $R(\hat{\theta}) \leq R(\theta_s^*)$ on n'a rien à prouver. Sinon, l'inégalité (3.15) donne

$$\|\hat{\theta} - \theta_s^*\|_{1,s^c} < \frac{\lambda + \lambda_0}{\lambda - \lambda_0} \|\theta_s^* - \hat{\theta}\|_{1,s} \leq 2 \|\theta_s^* - \hat{\theta}\|_{1,s},$$

pour $\lambda \geq 3\lambda_0$. On en déduit alors $\|\theta_s^* - \hat{\theta}\|_{1,s} \leq \frac{\sqrt{|s| \ell(\hat{\theta}, \theta_s^*)}}{\kappa(s)}$, et en retravaillant (3.15),

$$\begin{aligned} \ell(\hat{\theta}, \theta_s^*) + (\lambda - \lambda_0) \|\hat{\theta} - \theta_s^*\|_1 &\leq 2\lambda \|\theta_s^* - \hat{\theta}\|_{1,s} \\ &\leq 2\lambda \frac{\sqrt{|s| \ell(\hat{\theta}, \theta_s^*)}}{\kappa(s)} \\ &\leq \frac{1}{2} \ell(\hat{\theta}, \theta_s^*) + 2 \frac{\lambda^2 |s|}{\kappa^2(s)}. \end{aligned}$$

On en déduit

$$\ell(\hat{\theta}, \theta^*) \leq \ell(\theta_s^*, \theta^*) + 4 \frac{\lambda^2 |s|}{\kappa^2(s)}.$$

□

Cette borne est légèrement moins bonne que celle du Corollaire 3.17, mais a le mérite de pouvoir s'étendre pour d'autres fonctions de perte : si on regarde attentivement, le terme crucial dans le Théorème 3.15 dans le cadre moindre carré est

$$2\left\langle \varepsilon, X(u-\hat{\theta})\right\rangle = (R-R_n)(\hat{\theta}-u).$$

On peut aussi déduire de la preuve du Théorème 3.18 un résultat de consistance :

$$\|\hat{\theta}-\theta^*\|_1 \leq 3\frac{\lambda d}{\kappa^2(s^*)}.$$

Pour en déduire un résultat probabiliste, il faut alors contrôler des supremum renormalisés de type $\sup(R-R_n)(u-v)/\|u-v\|_1$, ce qui est un peu technique mais tout à fait faisable (vous êtes renvoyés à votre cours de concentration).

Prix à payer pour la relaxation convexe

Le premier défaut structurel du Lasso se rencontre dans les situations à forte colinéarité entre colonnes. Prenons le cas extrême où $\theta = (\theta_1, 0, \dots, 0)^T$ et $X^1 = X^2$. Pour le critère idéal $\|X(\theta-u)\|^2$ pénalisé en norme 1, n'importe quel u tel que $u_1+u_2=\theta_1$ est convenable, là où la pénalisation ℓ_0 privilégiera $u_1=0$ ou $u_2=0$. En pratique, le Lasso sélectionne non-seulement les variables supports mais aussi les variables fortement corrélées avec, et fournit donc plutôt une "borne sup" sur le support (les conditions de corrélation entre variable garantissant de trouver le bon support peuvent être trouvées dans Zhao and Yu [2006]). Néanmoins, on peut se servir de cette première approximation de support pour refaire tourner une procédure plus coûteuse (comme du ℓ_0) sur les variables sélectionnées pour pallier ce défaut).

Le deuxième défaut tient au rétrécissement des coefficients non-nuls vers 0. Dans le cas orthogonal, les coefficients non nuls sont $\mathrm{sg}(\hat{\theta}_{LS,j})\left(\left|\hat{\theta}_{LS,j}\right| - \frac{\lambda}{2}\right)$, dont les prédictions en performance sont souvent moins bonnes que les moindres carrés standards sur le support sélectionné. Une pratique courante consiste alors à réajuster un moindre carrés standard (ou ridge) sur les variables sélectionnées par le Lasso.

Terminons enfin sur le choix du paramètre λ. En pratique, la "trajectoire" entière pour λ dans un intervalle peut se calculer, on peut alors chercher des "sauts" dans les ensembles sélectionnés et tester plusieurs candidats par cross-validation par exemple. D'un point de vue algorithmique, il existe plusieurs méthodes pour résoudre ce problème d'optimisation convexe, citons par exemple les algorithmes LARS et FISTA. Le lecteur intéressé trouvera les détails dans Giraud [2022].

RQ : compatibilité améliore SMl0 ?

3.3 Overfitting bénin et double descente

Dans cette partie on essaiera d'expliquer un phénomène observé empiriquement : pour une taille d'échantillon donnée et une dimension croissante, si on trace l'excès de risque en fonction de la dimension, une première partie de la courbe illustre bien le compromis biais/variance jusqu'à $D \simeq n$, mais cet excès de risque peut décroître après (on assiste à une "double-descente"), donnant parfois lieu à des meilleurs performances en prédiction lorsque $D >> n$ (FAIRE DESSIN). Les explications théoriques de ce phénomène n'en sont qu'à leurs débuts (par exemple voir Chizat and Bach [2020], Hastie et al. [2022], Tsigler and Bartlett [2023]), on en présentera quelques

uns dans le cadre le plus simple : en régression Gaussienne. On peut dès à présents donner quelques grandes lignes caractérisant l'occurence ou non de ce phénomène :

1. Il ne peut apparaître que pour des dimensions à partir de laquelle on a une interpolation exacte sur l'échantillon d'apprentissage (un vrai overfitting quoi, on parle de modèle surparamétré).
2. Les procédures pour lesquelles il peut apparaître ne doivent pas comporter de régularisation explicite (par exemple OK pour la descente de gradient simple, pas OK pour le ridge).
3. Les prédicteurs X_i doivent être aléatoires (ne marche pas en régression à design fixe par exemple).
4. Il faut que le biais en "petite dimension" ne décroisse pas trop vite, ou en d'autres termes on a suffisamment de signal restant lorsque l'on fait croître la dimension (le point de vue opposé au régime parcimonieux en somme).

3.3.1 Aspect régularisant de la descente de gradient

On se place ici en régression linéaire moindres carrés, avec potentiellement $D > n$. Regardons le résultat d'une descente de gradient **initialisée en** 0 sur le risque empirique

$$R_n(\theta) = \|Y - X\theta\|^2,$$

pour un pas γ_t constant. On a

$$\begin{aligned} \theta_0 &= 0 \\ \theta_{t+1} &= \theta_t - 2\gamma(X^T X\theta_t - X^T Y) \\ &= 2\gamma X^T Y + (I_D - 2\gamma X^T X)\theta_t, \end{aligned}$$

ce dont on déduit par récurrence

$$\theta_t = 2\sum_{k=0}^{t-1}(I_D - 2\gamma X^T X)^k X^T Y.$$

Or, si $\gamma < 1/(2\hat{\lambda}_{max})$,

$$\sum_{k=0}^{t-1}(I_D - 2\gamma X^T X)^k \underset{t\to+\infty}{\longrightarrow} \frac{1}{2}(X^T X)^\dagger,$$

où $(X^T X)^\dagger$ est l'inverse de Moore-Penrose déjà défini en Section 3.2 (pour s'en rendre compte regarder dans une base SVD). On en déduit alors que

$$\theta_t \underset{t\to+\infty}{\longrightarrow} \hat{\theta}_{LS},$$

où $\hat{\theta}_{LS}$ est le régresseur moindre carrés de norme minimale.
Remarque : On peut montrer que cette convergence a lieu exponentiellement vite comme dans le cadre du Théorème 2.42 : les paramètres successifs $\theta^t \in V(X^T)$, seule la forte convexité sur $V(X^T)$ compte (et elle est mesurée par $\hat{\lambda}_{\min}$, plus petite valeur propre non-nulle de $X^T X$).

Cette analyse s'étend pour d'autres fonctions de pertes (Hinge, Exponentielle, Logistique), et parfois pour d'autres méthodes utilisées comme SGD. Le message pour commun est que, pour les méthodes standards utilisées sans régularisation (type GD ou SGD) et initialisées en 0, dans le régime interpolant la sortie sera un interpolant de norme minimale (on parle de biais implicite). Reste à comprendre en quoi chercher un interpolant de norme minimale peut amener à une double descente. Une réponse intuitive serait : l'ensemble des solutions interpolantes étant de dimension $D-n$, lorsque D augmente on peut espérer avoir des solutions de norme minimale de plus en plus faible "naturellement pénalisées" (à la limite si tout se passe bien on convergera même vers 0). On peut peut-être s'attendre alors à un phénomène de réduction de variance (comme pour la projection L_2). Et c'est bien ce qui peut se passer.

3.3.2 Modèle Gaussien isotrope bien spécifié

A partir de maintenant on se place dans un modèle de régression Gaussien bien spécifié, au design aléatoire, c'est à dire qu'on observe

$$y_i = \langle x_i, \theta^* \rangle + \varepsilon_i,$$

où les ε_i sont i.i.d. de loi $\mathcal{N}(0\sigma^2)$, et les x_i sont i.i.d. de loi $\mathcal{N}(0, I_D)$. C'est le modèle le plus simple sur les variables prédictives, qui permet néanmoins d'exhiber ce phénomène de double descente.

Pour un candidat θ, la performance est mesurée par l'excès de risque en prédiction, c'est à dire

$$\mathbb{E}\left\langle x, \theta - \theta^* \right\rangle^2 = \|\theta - \theta^*\|^2.$$

Enfin, on va regarder le comportement de l'estimateur par moindres carrés de plus petite norme, c'est à dire

$$\hat{\theta} = (X^T X)^\dagger X^T Y.$$

En regardant le risque conditionnellement à X on obtient

$$\mathbb{E}(\|\hat{\theta} - \theta^*\|^2 \mid X) = \|(I_D - (X^T X)^\dagger (X^T X))\theta^*\|^2 + \sigma^2 \mathrm{Tr}((X^T X)^\dagger). \tag{3.16}$$

Pour obtenir le risque en espérance on est donc obligé de comprendre le comportement aléatoire de $X^T X$. Dans le cas Gaussien isotrope c'est relativement simple.

Definition 3.19 : Matrice de Wishart

Une matrice $W \in \mathbb{R}^D$ est une matrice de Wishart à n degrés de libertés si

$$W = N^T N,$$

où N est une matrice $n \times D$ dont les lignes sont indépendantes et de loi $\mathcal{N}(0, I_D)$ (vecteurs Gaussien indépendants).

Si $n \geq D+2$, W^{-1} existe presque sûrement, et suit une **loi de Wishart inverse** (de mêmes degrés de libertés).

Dans notre cas de régresseurs Gaussien isotrope, X^TX est bien une matrice de Wishart à n degrés de libertés. Dans toute la suite, on notera $W_D(n)$ une telle matrice. Il reste à comprendre le comportement spectral des matrices de Wishart (inverses). Les lois de telles matrices sont explicites (voir par exemple [Anderson, 1958, Chapitre 7]), on peut se contenter des résultats suivants.

PROPOSITION 3.20 : PROPRIÉTÉS SPECTRALES DES MATRICES DE WISHART (INVERSE)

Si $n \geq D+2$, alors

$$\mathbb{E}(\mathrm{Tr}(W_D(n)^{-1}) = \frac{D}{n-(D+1)} \qquad \textit{[Anderson, 1958, Lemme 7.7.1]} \quad .$$

Si $n \leq D-1$, et $U = (e_1, \dots, e_D)$ est la base de vecteurs propres de $W_D(n)$ vérifiant $e_{j,1} \geq 0$ pour tout j, alors

$$U \sim \mathcal{U}_{\mathcal{S}_{+,D,n}} \qquad \textit{[Anderson, 1958, Chapitre 13]} \quad ,$$

où $\mathcal{S}_{+,D,n}$ est l'ensemble des BON à n éléments de $\mathbb{R}^D$ de premières coordonnées positives (variété de Stiefel), et $\mathcal{U}_{\mathcal{S}_{+,D,n}}$ est la mesure uniforme (de Haar) sur cet espace.

Cette mesure uniforme est entièrement caractérisée par la propriété d'invariance $O \times \mathcal{U}_{\mathcal{S}_{+,D,n}} \sim \mathcal{U}_{\mathcal{S}_{+,D,n}}$, où $O \in \mathcal{O}_D$ qui conserve la positivité des premières coordonnées. La deuxième propriété est alors relativement facile à démontrer : soit O une telle matrice, et U le choix de vecteurs propres (à premiers coefficients positifs, du coup bien défini presque sûrement). On a alors que OU est le choix de vecteurs propres associés pour OWO^T. Or on a

$$OWO^T = (NO^T)^T(NO^T),$$

et les lignes de NO^T sont indépendantes, de loi $\mathcal{N}(OO^T) = \mathcal{N}(0, I_D)$. On a donc bien $OU \sim U$, et $U \sim \mathcal{U}_{\mathcal{S}_{+,D,n}}$. En particulier, si u est un vecteur de norme 1, on a

$$v_j v_j^t u \sim \mathcal{U}_{S(D)},$$

c'est à dire que les projections d'un même vecteur unitaire sur les différentes directions propres sont toutes de même loi uniforme sur la sphère.

Remarque : Si on prend les signes des premières coordonnées au hasard indépendamment comme dans [Bai and Silverstein, 2010, Section 10.1.1], on se retrouve avec la loi uniforme sur $\mathcal{S}_{D,n}$, la variété de Stiefel standard (sans restriction de signe).

Avec ces éléments, on peut calculer explicitement l'excès de risque des moindres carrés (de normes minimale), pour $D \geq n+2$, en reprenant l'équation (3.16).

Terme de Variance :

Comme $\mathrm{Tr}((X^TX)^\dagger) = \mathrm{Tr}((XX^T)^{-1} = \mathrm{Tr}(W_n(D)^{-1})$ (regarder la SVD pour s'en convaincre), une application directe de la Proposition 3.20 donne

$$\sigma^2 \mathbb{E}\mathrm{Tr}((X^TX)^\dagger) = \sigma^2 \frac{n}{D-(n+1)}.$$

C'est là le phénomène intéressant : plus la dimension augmente, plus la **variance** de $\hat{\theta}$ décroît. Moralement parlant, comme $\hat{\theta}$ est censé converger vers 0 (estimateur nul constant), c'est un comportement naturel.

Terme de biais :

On remarque que $(X^TX)^\dagger(X^TX) = W_D(n)^\dagger W_D(n)$ est la matrice de projection orthogonale sur $V(X^TX)$, qui est la matrice de projection orthogonale sur $V(U)$, où U est une base orthonormale propre correspondant aux n plus grandes valeurs propres de X^TX. On a alors

$$\begin{aligned}\|(I_D - (X^TX)^\dagger(X^TX))\theta^*\|^2 =& \|\theta^*\|^2 - \|\pi_{V(U)}\theta^*\|^2 \\ &= \|\theta^*\|^2 - \sum_{j=1}^{n}(\theta^*)^T v_j v_j^T \theta^*\end{aligned}$$

D'après la Proposition 3.20, on en déduit

$$\mathbb{E}\|(I_D - (X^TX)^\dagger(X^TX))\theta^*\|^2 = \|\theta^*\|^2 - n\mathbb{E}\left\langle v_1, \theta^*\right\rangle^2.$$

Or, si on se donne $u_1, \dots, u_D$ base orthonormée de loi uniforme sur $\mathcal{S}_{D,D}$, on a toujours la propriété que pour tout j $u_j \sim u_1 \sim v_1$. On en déduit

$$\begin{aligned}\|\theta^*\|^2 &= \mathbb{E}\sum_{j=1}^{D}\left\langle \theta^*, u_j\right\rangle^2 \\ &= D\mathbb{E}\left\langle \theta^*, u_1\right\rangle^2 \\ &= D\mathbb{E}\left\langle \theta^*, v_1\right\rangle^2.\end{aligned}$$

On conclut alors que le terme de biais vaut

$$\mathbb{E}\|(I_D - (X^TX)^\dagger(X^TX))\theta^*\|^2 = \left(1 - \frac{n}{D}\right)\|\theta^*\|^2.$$

On a donc un biais qui croît avec la dimension, pour atteindre le biais du prédicteur nul. En combinant les deux, on obtient

$$\mathbb{E}\|\hat{\theta} - \theta^*\|^2 = \left(1 - \frac{n}{D}\right)\|\theta^*\|^2 + \sigma^2\frac{n}{D-(n+1)}. \tag{3.17}$$

Lorsque D devient grand, on converge bien vers le risque du prédicteur nul.

Pour donner un sens formel à "pour D croissant" et pouvoir illustrer l'apparition ou non de la double descente, il faut spécifier ce qui bouge lorsque D augmente, notamment ce qui se passe pour θ^* (qui jusque ici était censé vivre dans $\mathbb{R}^D$). Bref, il nous faut un modèle qui autorise une croissance en D. On adopte ici le point de vue utilisé dans Tsigler and Bartlett [2023], à savoir que $\theta^* \in \mathbb{H}$, où $\mathbb{H}$ est un espace de Hilbert séparable, et que l'on observe

$$Y = X\theta^* + \varepsilon,$$

où $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ est indépendante de X, et où les lignes de X, $X_{i,\cdot}$ sont des bruits blancs Gaussiens indépendants (si on se donne une base $e_1, \dots, e_k, \dots$ de $\mathbb{H}$, on peut définir $\langle X_{i,\cdot}, \theta^*\rangle$ dans L^2 via $\sum_{p=1}^{+\infty}\theta_p^*\xi_{i,p}$, où les $\xi_{i,p}$ sont des Gaussiennes standard

i.i.d.). Bref, on conserve l'écriture matricielle (avec des matrices de dimension infinie). Un lecteur contrarié par le caractère informel de cette définition (et ce qui va suivre) peut remplacer $\mathbb{H}$ par $\mathbb{R}^{D_\infty}$, où D_∞ est fini mais très grand, ou transposer l'écriture matricielle en termes d'opérateurs sur $\mathbb{H}$.

Pour une dimension D donnée, on suppose que l'on observe Y et X_D, c'est à dire les D premières colonnes de X, et que l'on construit le prédicteur par moindres carrés (de norme minimale)

$$\hat{\theta}_D = \begin{cases} (X_D^T X_D)^{-1} X_D^T Y & \text{si } D \leq n \\ (X_D^T X_D)^{\dagger} X_D^T Y & \text{si } D > n. \end{cases}$$

Par ailleurs, on notera $X_{>D}$ les colonnes de X après D, et θ^*_D, $\theta^*_{>D}$ les D premières coordonnées (resp. coordonnées après D) de θ^*, et on confondra allègrement $\theta^*_D \in \mathbb{R}^D$ avec $\theta^*_D \in \mathbb{H}$ la projection de θ^* sur l'espace engendré par les D premiers éléments de la base canonique de $\mathbb{H}$. On a la borne suivante sur l'excès de risque du prédicteur par moindres carrés.

Théorème 3.21

Si $D \leq n-2$,

$$\mathbb{E}\|\hat{\theta}_D - \theta^*\|^2 = B^2(D) + \frac{D}{n-(D+1)}(\sigma^2 + B^2(D)).$$

Si $D \geq n+2$,

$$\mathbb{E}\|\hat{\theta}_D - \theta^*\|^2 = \left(1 - \frac{n}{D}\right)\|\theta^*\|^2 + \frac{n}{D}B^2(D) + \frac{n}{D-(n+1)}(\sigma^2 + B^2(D)).$$

Où $B^2(D) = \|\theta^*_{>D}\|^2$.

Preuve du Théorème 3.21. Commençons par le cas $D \leq n-2$. On a

$$\begin{aligned} \|\hat{\theta}_D - \theta^*\|^2 &= \|\hat{\theta}_D - \theta^*_D\|^2 + \|\theta^*_{>D}\|^2 \\ &= B^2(D) + \|\hat{\theta}_D - \theta^*_D\|^2. \end{aligned}$$

Ensuite,

$$\begin{aligned} \hat{\theta}_D - \theta^*_D &= (X_D^T X_D)^{-1} X_D^T (X_D \theta^*_D + X_{>D}\theta^*_{>D} + \varepsilon) - \theta^*_D \\ &= (X_D^T X_D)^{-1} X_D^T X_{>D}\theta^*_{>D} + (X_D^T X_D)^{-1} X_D^T \varepsilon. \end{aligned}$$

En raisonnant conditionnellement à X on obtient

$$\begin{aligned} \mathbb{E}\|\hat{\theta}_D - \theta^*_D\|^2 &= \mathbb{E}\|(X_D^T X_D)^{-1} X_D^T X_{>D}\theta^*_{>D}\|^2 + \sigma^2 \mathbb{E}(\mathrm{Tr}(W_D(n)^{-1})) \\ &= \mathbb{E}\|(X_D^T X_D)^{-1} X_D^T X_{>D}\theta^*_{>D}\|^2 + \sigma^2 \frac{D}{n-(D+1)}, \end{aligned}$$

en utilisant la Proposition 3.20. Par ailleurs, les entrées de X étant i.i.d. $\mathcal{N}(0,1)$, on remarque que $X_{>D}\theta^*_{>D} \sim \mathcal{N}(0, B^2(D) I_n)$ et est indépendante de X_D. En conditionnant par rapport à X_D on a alors

$$\begin{aligned} \mathbb{E}\|(X_D^T X_D)^{-1} X_D^T X_{>D}\theta^*_{>D}\|^2 &= B^2(D)\mathbb{E}(\mathrm{Tr}((X_D^T X_D)^{-1})) \\ &= B^2(D)\mathbb{E}(\mathrm{Tr}(W_D(n)^{-1})) = B^2(D)\frac{D}{n-(D+1)}. \end{aligned}$$

On en déduit

$$\mathbb{E}\|\hat{\theta}_D - \theta^*\|^2 = B^2(D) + \frac{D}{n-(D+1)}(\sigma^2 + B^2(D)).$$

Passons maintenant au cas $D \geq n+2$. Avec la même analyse, on est ramenés à

$$\mathbb{E}\|\hat{\theta}_D - \theta^*_D\|^2 = \mathbb{E}\|(I_D - (X_D^T X_D)^\dagger (X_D^T X_D))\theta^*_D\|^2 + \frac{n}{D-(n+1)}(\sigma^2 + B^2(D)),$$

en utilisant le fait que $\mathrm{Tr}((X_D^T X_D)^\dagger) = \mathrm{Tr}((X_D X_D^T)^{-1}) = \mathrm{Tr}(W_n(D))$ et la Proposition 3.20. En raisonnant comme pour (3.17), on a

$$\begin{aligned}
\mathbb{E}\|(I_D - (X_D^T X_D)^\dagger (X_D^T X_D))\theta^*_D\|^2 &= \|\theta^*_D\|^2 - n\mathbb{E}\,\langle v_1, \theta^*_D\rangle^2 \\
&= \left(1 - \frac{n}{D}\right)\|\theta^*_D\|^2 \\
&= \left(1 - \frac{n}{D}\right)(\|\theta^*\|^2 - B^2(D)).
\end{aligned}$$

En mettant tout bout à bout on obtient

$$\begin{aligned}
\mathbb{E}\|\hat{\theta}_D - \theta^*\|^2 &= B^2(D) + \left(1 - \frac{n}{D}\right)(\|\theta^*\|^2 - B^2(D)) + \frac{n}{D-(n+1)}(\sigma^2 + B^2(D)) \\
&= \left(1 - \frac{n}{D}\right)\|\theta^*\|^2 + \frac{n}{D}B^2(D) + \frac{n}{D-(n+1)}(\sigma^2 + B^2(D)).
\end{aligned}$$

□

On remarque là encore que l'excès de risque du prédicteur moindre carrés tend vers celui du prédicteur nul (qui vaut $\|\theta^*\|^2$). Si, comme dans Hastie et al. [2022], on introduit le ratio bruit/signal $r^2 = \sigma^2/\|\theta^*\|^2$ et biais/signal $b(D) = B^2(D)/\|\theta^*\|^2$, au facteur $\|\theta^*\|^2$ près on a

$$\mathbb{E}\|\hat{\theta}_D - \theta^*\|^2 = \begin{cases} b^2(D) + \frac{D}{n-(D+1)}(r^2 + b^2(D)) & \text{si } D \leq n-2, \\ \left(1 - \frac{n}{D}\right) + \frac{n}{D}b^2(D) + \frac{n}{D-(n+1)}(r^2 + b^2(D)) & \text{si } D \geq n+2. \end{cases}$$

On voit alors que le D optimal (s'il existe), va fortement dépendre du profil du biais $b(D)$. Commençons par comparer avec le prédicteur nul. Pour $D \leq n-2$, on a

$$\mathbb{E}\|\hat{\theta}_D - \theta^*\|^2 \leq \|\theta^*\|^2 \quad \Leftrightarrow \quad b^2(D) \leq 1 - \frac{D(r^2+1)}{n-1}.$$

On peut donc espérer une prédiction meilleure que le prédicteur nul si le biais (au carré) décroit sous-linéairement en la dimension. Comme par ailleurs $\|\theta^*\|^2 = \mathbb{E}\|\hat{\theta}_0 - \theta^*\|^2$ (ou plus rigoureusement est la limite en 0 de la courbe de risque, pour une fonction de biais régulière), on peut déduire que dans ce cas de figure il existera un minimum local entre 0 et n pour la fonction de perte.

Pour $D \geq n-2$, on a

$$\mathbb{E}\|\hat{\theta}_D - \theta^*\|^2 \leq \|\theta^*\|^2 \quad \Leftrightarrow \quad b^2(D) \leq \frac{D(1-r^2) - (n+1)}{2D - (n+1)}.$$

On remarque alors que $r^2 \geq 1$ implique que le prédicteur nul est meilleur que le prédicteur par moindres carrés de norme minimale. En revanche, pour $r^2 < 1$, pour une fonction $b^2(D)$ décroissante on aura toujours une plage du type $[D_+, +\infty[$ sur laquelle le prédicteur par moindres carrés sera meilleur que le prédicteur nul (FAIRE DESSIN). On aura donc, pour $r^2 < 1$, toujours au moins un minimum local en D sur la portion $[n+2, +\infty[$. En revanche il n'est pas garanti que ce minimum local soit un minimum global (à comparer avec la plage $[0, n-2]$).

Enfin, on peut aller plus loin pour le cas $r^2 \geq 1$. Dans ce cas, pour $D \geq n+2$, on a

$$f(D) := \mathbb{E}\|\hat{\theta}_D - \theta^*\|^2 = \left(1 - \frac{n}{D}\right) + \frac{nr^2}{D-(n+1)} + b^2(D)\left(\frac{n}{D} + \frac{n}{D-(n+1)}\right).$$

On remarque que le terme en $b^2(D)$ est décroissant. Posons alors

$$g(D) = \left(1 - \frac{n}{D}\right) + \frac{nr^2}{D-(n+1)}.$$

Un calcul immédiat donne

$$\begin{aligned} g'(D) &= \left[D^2(n(1-r^2)) - 2n(n+1)D + n(n+1)^2\right] \frac{1}{V(D)} \\ &:= \frac{P(D)}{V(D)}, \end{aligned}$$

où $V(D) > 0$. On a alors, pour $D \geq n+2$,

$$P(D) \leq -2n(n+1)(n+2) + n(n+1)^2 < 0.$$

On en déduit alors que f est strictement décroissante sur $[n+2, +\infty[$, il n'y a donc pas de minimum local de l'excès de risque pour le prédicteur moindre carrés.

Illustrons ces trois cas de figure.

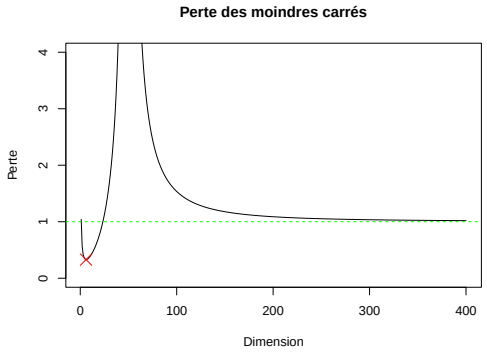

FIGURE 3.1 – Pour les paramètres $\|\theta^*\| = 1$, $n = 50$, $r^2 = 1$.

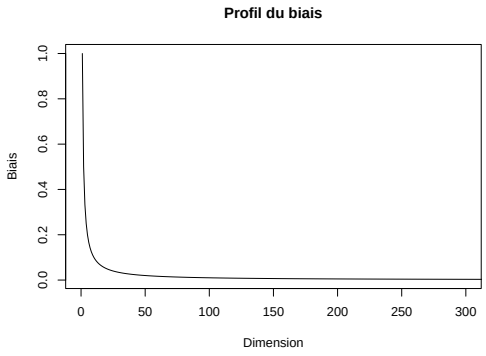

FIGURE 3.2 – Biais : $b^2(D) = D^{-1}$

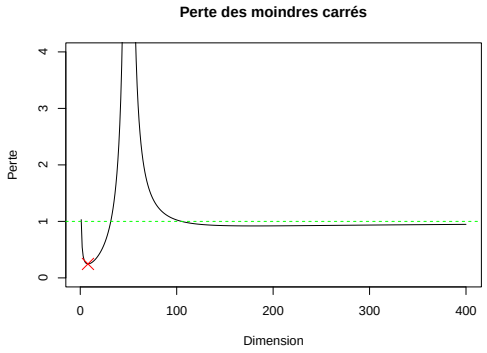

FIGURE 3.3 – Pour les paramètres $\|\theta^*\| = 1$, $n = 50$, $r^2 = 1/2$.

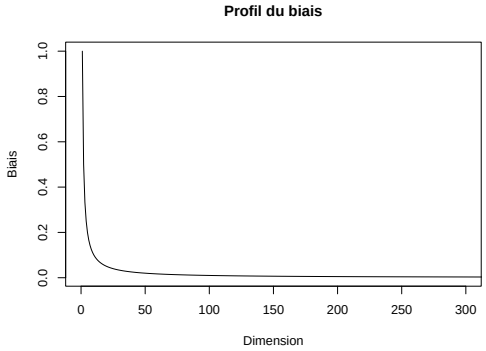

FIGURE 3.4 – Biais : $b^2(D) = D^{-1}$

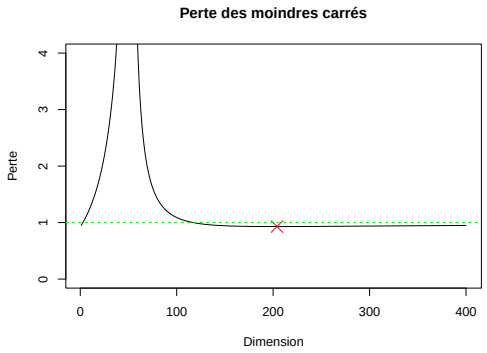

FIGURE 3.5 – Pour les paramètres $\|\theta^*\| = 1$, $n = 50$, $r^2 = 1/2$.

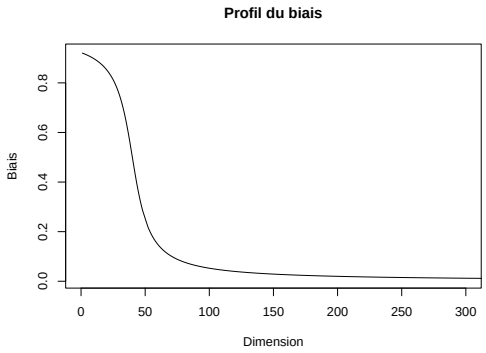

FIGURE 3.6 – Biais : $b^2(D) = \frac{1}{2} - \frac{1}{\pi}\arctan((D-40)/10)$

Dans tous ces cas de figure, on a bien un phénomène de double descente (ou de descente simple dans le régime surparamétré). Pour que l'optimum se trouve dans le régime surparamétré, 2 conditions doivente être réunies :

1. Le ratio bruit/signal doit être strictement inférieur à 1.
2. Le signal ne doit pas être contenu dans les premières coordonnées, i.e. le biais ne doit pas décroître trop vite.

Les choses se compliquent un peu en dehors du cadre Gaussien isotrope.

3.3.3 Cadre Gaussien anisotrope et extensions

Le modèle Gaussien anisotrope s'écrit toujours

$$y = \langle x, \theta^* \rangle + \varepsilon,$$

où $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ est indépendante de x, mais cette fois x est un processus Gaussien de covariance Σ (que l'on supposera bornée en tant qu'opérateur sur $\mathbb{H}$ et inversible). On observe toujours (avec des notations abusives)

$$Y = X\theta^* + \varepsilon,$$

où les lignes sont toujours i.i.d., et dans le cadre de dimension D on suppose que l'on observe X_D. Le prédicteur par moindres carrés est toujours défini par

$$\hat{\theta}_D = \begin{cases} (X_D^T X_D)^{-1} X_D^T Y & \text{si } D \leq n \\ (X_D^T X_D)^{\dagger} X_D^T Y & \text{si } D > n, \end{cases}$$

mais l'excès de risque en prédiction est cette fois-ci donné par

$$\ell(\hat{\theta}_D, \theta^*) = (\hat{\theta}_D - \theta^*)^T \Sigma (\hat{\theta}_D - \theta^*).$$

On peut donner une borne générale sur le risque en prédiction.

Théorème 3.22

Pour le modèle Gaussien anisotrope avec Σ inversible, on décompose Σ en $\Sigma_{D,D}$, $\Sigma_{D,>D}$, $\Sigma_{>D,D}$ et $\Sigma_{>D,>D}$, et on note $\beta_D^* = \Sigma_{D,D}^{1/2}\theta_D^*$, $\beta_{>D}^* = \Sigma_{>D,>D}^{1/2}\theta_{>D}^*$, $D_{D,>D} = \Sigma_{D,D}^{-1/2}\Sigma_{D,>D}\Sigma_{>D,>D}^{-1/2}$.

Si $D \leq n-2$, on a

$$\mathbb{E}\ell(\hat{\theta}_D, \theta^*) = \left(\|\beta_{>D}^*\|^2 - \|C_{D,>D}\beta_{>D}^*\|^2\right)\left(1 + \frac{D}{n-(D+1)}\right) + \sigma^2 \frac{D}{n-(D+1)}.$$

Si $D \geq n+2$, on a

$$\begin{aligned}\mathbb{E}\ell(\hat{\theta}_D, \theta^*) = &\left(\|\beta_{>D}^*\|^2 - \|C_{D,>D}\beta_{>D}^*\|^2\right)\left(1 + \frac{n}{D-(n+1)}\right) + \sigma^2 \frac{n}{D-(n+1)} \\ &+ \left(1 - \frac{n}{D}\right)\left\|\beta_D^* - C_{D,>D}\beta_{>D}^*\right\|^2.\end{aligned}$$

Preuve du Théorème 3.21. **Cas $D \leq n-2$**

Commençons par le cas $D \leq n-2$. Dans ce cas,

$$\begin{aligned}\hat{\theta}_D &= (X_D^T X_D)^{-1} X_D^T Y \\ &= \theta_D^* + (X_D^T X_D)^{-1} X_D^T X_{>D}\theta_{>D}^* + (X_D^T X_D)^{-1} X_D^T \varepsilon.\end{aligned}$$

On en déduit

$$\begin{aligned}(\hat{\theta}_D - \theta^*)^T \Sigma (\hat{\theta}_D - \theta^*) &= (\hat{\theta}_D - \theta_D^*)^T \Sigma_{D,D} (\hat{\theta}_D - \theta_D^*) + (\theta_{>D}^*)^T \Sigma_{>D,>D}\theta_{>D}^* - 2(\theta_{>D}^*)^T \Sigma_{>D,D}(\hat{\theta}_D - \theta_D^*) \\ &:= I + II + III.\end{aligned} \tag{3.18}$$

Pour exprimer le premier et dernier terme, on aura besoin du Lemme suivant.

Lemme 3.23 : Conditionnement Gaussien

Si x est un processus Gaussien à valeurs dans $\mathbb{H}$ de covariance Σ, on a

$$\mathbb{E}(x_{>D} \mid x_D) = \Sigma_{>D,D}\Sigma_{D,D}^{-1} x_D,$$

et $(x_{>D} - \mathbb{E}(x_{>D} \mid x_D)) \perp\!\!\!\perp x_D$, de covariance $\Sigma_{>D,>D} - \Sigma_{>D,D}\Sigma_{D,D}^{-1}\Sigma_{D,>D}$.

Preuve du Lemme 3.23. Cette preuve est basée sur la propriété d'équivalence $A\Sigma B^T = 0 \quad \Leftrightarrow Ax \perp\!\!\!\perp Bx$, pour des applications linéaires A et B, caractéristique des processus Gaussiens, et est laissée en exercice. □

Revenons à notre expression. On a immédiatement

$$II = \|\beta^*_{>D}\|^2.$$

Regardons le premier terme. On a

$$\begin{aligned}\mathbb{E}(I \mid X) = &\left((X_D^T X_D)^{-1} X_D^T X_{>D} \theta^*_{>D}\right)^T \Sigma_{D,D} \left((X_D^T X_D)^{-1} X_D^T X_{>D} \theta^*_{>D}\right) \\ &+ \sigma^2 \mathrm{Tr}\left(X_D (X_D^T X_D)^{-1} \Sigma_{D,D} (X_D^T X_D)^{-1} X_D^T\right)\end{aligned}$$

Comme $x_D \sim \sqrt{\Sigma_{D,D}}\mathcal{N}(0, I_d)$, on peut écrire

$$X_D = N_D \sqrt{\Sigma_{D,D}},$$

où N_D est à entrées i.i.d. $\mathcal{N}(0,1)$ (comme dans le cas isotrope). On a alors

$$\begin{aligned}(X_D^T X_D)^{-1} &= \Sigma_{D,D}^{-1/2} (N_D^T N_D)^{-1} \Sigma_{D,D}^{-1/2} \\ &= \Sigma_{D,D}^{-1/2} W_D(n)^{-1} \Sigma_{D,D}^{-1/2}.\end{aligned}$$

On en déduit

$$\begin{aligned}\sigma^2 \mathbb{E}\mathrm{Tr}\left(X_D (X_D^T X_D)^{-1} \Sigma_{D,D} (X_D^T X_D)^{-1} X_D^T\right) &= \sigma^2 \mathbb{E}\mathrm{Tr}\left(N_D (N_D^T N_D)^{-1} (N_D^T N_D)^{-1} N_D^T\right) \\ &= \sigma^2 \mathbb{E}\mathrm{Tr}\left(W_D(n)^{-1}\right) \\ &= \sigma^2 \frac{D}{n-(D+1)},\end{aligned}$$

en utilisant la Proposition 3.20. Regardons maintenant l'autre terme de $\mathbb{E}(I \mid X)$. En utilisant le Lemme 3.23, on peut écrire

$$\mathbb{E}(X_{>D} \mid X_D) = X_D \Sigma_{D,D}^{-1} \Sigma_{D,>D},$$

et

$$\left(X_{>D} - \mathbb{E}(X_{>D} \mid X_D)\right) \theta^*_{>D} \sim \mathcal{N}(0, v I_n) \perp\!\!\!\perp X_D,$$

avec

$$\begin{aligned}v &= (\theta^*_{>D})^T \left(\Sigma_{>D,>D} - \Sigma_{>D,D} \Sigma_{D,D}^{-1} \Sigma_{D,>D}\right) \theta^*_{>D} \\ &= \|\beta^*_{>D}\|^2 - \|C_{D,>D} \beta^*_{>D}\|^2.\end{aligned}$$

On peut décomposer

$$\begin{aligned}&\mathbb{E}\left((X_D^T X_D)^{-1} X_D^T X_{>D} \theta^*_{>D}\right)^T \Sigma_{D,D} \left((X_D^T X_D)^{-1} X_D^T X_{>D} \theta^*_{>D}\right) \\ &\quad = \mathbb{E}\left(\left((X_D^T X_D)^{-1} X_D^T \mathbb{E}(X_{>D} \mid X_D) \theta^*_{>D}\right)^T \Sigma_{D,D} \left((X_D^T X_D)^{-1} X_D^T \mathbb{E}(X_{>D} \mid X_D) \theta^*_{>D}\right)\right) \\ &\qquad + \mathbb{E}\left((X_D^T X_D)^{-1} X_D^T \left(X_{>D} - \mathbb{E}(X_{>D} \mid X_D)\right) \theta^*_{>D}\right)^T \Sigma_{D,D} \left((X_D^T X_D)^{-1} X_D^T \left(X_{>D} - \mathbb{E}(X_{>D} \mid X_D)\right) \theta^*_{>D}\right.\end{aligned}$$

Le deuxième terme se traite comme précédemment :

$$\begin{aligned}
&\mathbb{E}\left((X_D^TX_D)^{-1}X_D^T\left(X_{>D}-\mathbb{E}(X_{>D}\mid X_D)\right)\theta^*_{>D}\right)^T\Sigma_{D,D}\left((X_D^TX_D)^{-1}X_D^T\left(X_{>D}-\mathbb{E}(X_{>D}\mid X_D)\right)\theta^*_{>D}\right)\\
&\quad= v\mathbb{E}\mathrm{Tr}\left(X_D(X_D^TX_D)^{-1}\Sigma_{D,D}(X_D^TX_D)^{-1}X_D^T\right)\\
&\quad= v\frac{D}{n-(D+1)}.
\end{aligned}$$

Le premier terme lui s'écrit

$$\begin{aligned}
&\mathbb{E}\left(\left((X_D^TX_D)^{-1}X_D^T\mathbb{E}(X_{>D}\mid X_D)\theta^*_{>D}\right)^T\Sigma_{D,D}\left((X_D^TX_D)^{-1}X_D^T\mathbb{E}(X_{>D}\mid X_D)\theta^*_{>D}\right)\right)\\
&\quad=\mathbb{E}\left[\left(\Sigma_{D,D}^{-1}\Sigma_{D,>D}\theta^*_{>D}\right)^T\Sigma_{D,D}\left(\Sigma_{D,D}^{-1}\Sigma_{D,>D}\theta^*_{>D}\right)\right]\\
&\quad=\|C_{D,>D}\beta^*_{>D}\|^2.
\end{aligned}$$

On en déduit

$$\mathbb{E}(I)=\left(\sigma^2+\|\beta^*_{>D}\|^2-\|C_{D,>D}\beta^*_{>D}\|^2\right)\frac{D}{n-(D+1)}+\|C_{D,>D}\beta^*_{>D}\|^2.$$

Il reste un dernier terme. Rappelons que $III=-2(\theta^*_{>D})^T\Sigma_{>D,D}(\hat\theta_D-\theta^*_D)$. En utilisant le Lemme 3.23 encore, on obtient

$$\begin{aligned}
\mathbb{E}(III\mid X_D)&=-2(\theta^*_{>D})^T\Sigma_{>D,D}\left((X_D^TX_D)^{-1}X_D^T\mathbb{E}(X_{>D}\mid X_D)\theta^*_{>D}\right)\\
&=-2(\theta^*_{>D})^T\Sigma_{>D,D}\Sigma_{D,D}^{-1}\Sigma_{D,>D}\theta^*_{>D}\\
&=-2\|C_{D,>D}\beta^*_{>D}\|^2.
\end{aligned}$$

En additionnant les trois termes on se retrouve avec

$$\begin{aligned}
\mathbb{E}\ell(\hat\theta_D,\theta^*)&=\left(\sigma^2+\|\beta^*_{>D}\|^2-\|C_{D,>D}\beta^*_{>D}\|^2\right)\frac{D}{n-(D+1)}+\|C_{D,>D}\beta^*_{>D}\|^2-2\|C_{D,>D}\beta^*_{>D}\|^2+\|\beta^*_{>D}\|^2\\
&=\left(\|\beta^*_{>D}\|^2-\|C_{D,>D}\beta^*_{>D}\|^2\right)\left(1+\frac{D}{n-(D+1)}\right)+\sigma^2\frac{D}{n-(D+1)}
\end{aligned}$$

Cas $D\geq n+2$

On a cette fois ci

$$\hat\theta_D=(X_D^TX_D)^\dagger(X_D^TX_D)\theta^*_D+(X_D^TX_D)^\dagger X_D^TX_{>D}\theta^*_{>D}+(X_D^TX_D)^\dagger X_D^T\varepsilon,$$

et on repart de la décomposition de l'excès de risque (3.18). On a toujours

$$II=\|\beta^*_{>D}\|^2.$$

Pour le terme I, on peut encore écrire

$$\begin{aligned}
&\mathbb{E}(I\mid X)\\
&=\left(\left[(X_D^TX_D)^\dagger X_D^TX_D-I_D\right]\theta^*_D+(X_D^TX_D)^\dagger X_{>D}\theta^*_{>D}\right)^T\Sigma_{D,D}\left(\left[(X_D^TX_D)^\dagger X_D^TX_D-I_D\right]\theta^*_D+(X_D^TX_D)^\dagger X_{>}\right.\\
&\quad+\sigma^2\mathbb{E}\mathrm{Tr}\left(X_D(X_D^TX_D)^\dagger\Sigma_{D,D}(X_D^TX_D)^\dagger X_D^T\right).
\end{aligned}$$

Comme pour le cas $D \leq n-2$, en utilisant $X_D = N_D \Sigma_{D,D}^{1/2}$, on a

$$\begin{aligned}\sigma^2 \mathbb{E}\mathrm{Tr}\left(X_D (X_D^T X_D)^\dagger \Sigma_{D,D} (X_D^T X_D)^\dagger X_D^T\right) &= \sigma^2 \mathbb{E}\mathrm{Tr}\left((N_D^T N_D)^\dagger\right) \\ &= \sigma^2 \frac{n}{D-(n+1)},\end{aligned}$$

en utilisant la Proposition 3.20. Pour le premier terme de $\mathbb{E}(I \mid X)$ on peut encore décomposer en prenant $\mathbb{E}(I \mid X_D)$, ce qui donne

$$\begin{aligned}&\left(\left[(X_D^T X_D)^\dagger X_D^T X_D - I_D\right]\theta_D^* + (X_D^T X_D)^\dagger \mathbb{E}(X_{>D} \mid X_D)\theta_{>D}^*\right)^T \Sigma_{D,D} \left(\left[(X_D^T X_D)^\dagger X_D^T X_D - I_D\right]\theta_D^* + (X_D^T X\right. \\ &\quad + \mathbb{E}\left(\left((X_D^T X_D)^\dagger X_D^T V\right)^T \Sigma_{D,D} \left((X_D^T X_D)^\dagger X_D^T V\right)\right),\end{aligned}$$

où $V \sim \mathcal{N}(0, v I_n)$, avec $v = \|\beta_{>D}^*\|^2 - \|C_{D,>D}\beta_{>D}^*\|^2$. On a comme précédemment

$$\begin{aligned}\mathbb{E}\left(\left((X_D^T X_D)^\dagger X_D^T V\right)^T \Sigma_{D,D} \left((X_D^T X_D)^\dagger X_D^T V\right)\right) &= v \mathbb{E}\mathrm{Tr}\left(X_D (X_D^T X_D)^\dagger \Sigma_{D,D} (X_D^T X_D)^\dagger X_D^T\right) \\ &= v \frac{n}{D-(n+1)}.\end{aligned}$$

Reste à contrôler l'espérance du gros terme

$$\begin{aligned}&\left(\left[(X_D^T X_D)^\dagger X_D^T X_D - I_D\right]\theta_D^* + (X_D^T X_D)^\dagger \mathbb{E}(X_{>D} \mid X_D)\theta_{>D}^*\right)^T \Sigma_{D,D} \left(\left[(X_D^T X_D)^\dagger X_D^T X_D - I_D\right]\theta_D^* + (X_D^T X\right. \\ &= \left\|((N_D^T N_D)^\dagger (N_D^T N_D) - I_D)\beta_D^* + (N_D^T N_D)^\dagger (N_D^T N_D) C_{D,>D}\beta_{>D}^*\right\|^2 \\ &= \left\|\pi_{V_n^\perp}(\beta_D^*) + \pi_{V_n}(C_{D,>D}\beta_{>D}^*)\right\|^2 \\ &= \left\|\pi_{V_n^\perp}(\beta_D^*)\right\|^2 + \left\|\pi_{V_n}(C_{D,>D}\beta_{>D}^*)\right\|^2,\end{aligned}$$

où $V_n \subset \mathbb{R}^D$ est l'espace image de $N_D^T N_D$. En utilisant la Proposition 3.20 comme dans la preuve du Théorème 3.21, on obtient

$$\mathbb{E}\left[\left\|\pi_{V_n^\perp}(\beta_D^*)\right\|^2 + \left\|\pi_{V_n}(C_{D,>D}\beta_{>D}^*)\right\|^2\right] = \left(1 - \frac{n}{D}\right)\|\beta_D^*\|^2 + \frac{n}{D}\|C_{D,>D}\beta_{>D}^*\|^2,$$

d'où on déduit

$$\mathbb{E}(I) = \left(\sigma^2 + \|\beta_{>D}^*\|^2 - \|C_{D,>D}\beta_{>D}^*\|^2\right)\frac{n}{D-(n+1)} + \left(1 - \frac{n}{D}\right)\|\beta_D^*\|^2 + \frac{n}{D}\|C_{D,>D}\beta_{>D}^*\|^2.$$

Reste le terme III. On peut écrire

$$\begin{aligned}\mathbb{E}(III \mid X_D)) &= -2(\theta_{>D}^*)^T \Sigma_{>D,D}\left[(X_D^T X_D)^\dagger X_D^T X_D - I_D\right]\theta_D^* \\ &\quad - 2(\theta_{>D}^*)^T \Sigma_{>D,D}(X_D^T X_D)^\dagger X_D^T X_D \Sigma_{D,D}^{-1}\Sigma_{D,>D}\theta_{>D}^* \\ &= -2(\beta_{>D}^*)^T \Sigma_{>D,>D}^{-1/2}\Sigma_{>D,D}\Sigma_{D,D}^{-1/2}((N_D^T N_D)^\dagger (N_D^T N_D) - I_D)\beta_D^* \\ &\quad - 2(\beta_{>D}^*)^T \Sigma_{>D,>D}^{-1/2}\Sigma_{>D,D}\Sigma_{D,D}^{-1/2}(N_D^T N_D)^\dagger (N_D^T N_D) C_{D,>D}\beta_{>D}^* \\ &= -2\left\langle C_{D,>D}(\beta_{>D}^*), \pi_{V_n^\perp}\beta_D^*\right\rangle - 2\left\|\pi_{V_n} C_{D,>D}\beta_{>D}^*\right\|^2.\end{aligned}$$

On en déduit alors

$$\mathbb{E}(III) = -2\left(1 - \frac{n}{D}\right)\langle C_{D,>D}(\beta_{>D}^*), \beta_D^*\rangle - 2\frac{n}{D}\left\|C_{D,>D}\beta_{>D}^*\right\|^2.$$

En recollant les trois morceaux on arrive à

$$\begin{aligned}
\mathbb{E}\ell(\hat{\theta}_D, \theta^*) &= \left(\sigma^2 + \|\beta^*_{>D}\|^2 - \|C_{D,>D}\beta^*_{>D}\|^2\right)\frac{n}{D-(n+1)} + \left(1-\frac{n}{D}\right)\|\beta^*_D\|^2 + \frac{n}{D}\|C_{D,>D}\beta^*_{>D}\|^2 \\
&\quad - 2\left(1-\frac{n}{D}\right)\langle C_{D,>D}(\beta^*_{>D}), \beta^*_D\rangle - 2\frac{n}{D}\|C_{D,>D}\beta^*_{>D}\|^2 + \|\beta^*_{>D}\|^2 \\
&= \left(\sigma^2 + \|\beta^*_{>D}\|^2 - \|C_{D,>D}\beta^*_{>D}\|^2\right)\frac{n}{D-(n+1)} + \left(1-\frac{n}{D}\right)\|\beta^*_D - C_{D,>D}\beta^*_{>D}\|^2 \\
&\quad + \|\beta^*_{>D}\|^2 - \|C_{D,>D}\beta^*_{>D}\|^2 \\
&= \left(\|\beta^*_{>D}\|^2 - \|C_{D,>D}\beta^*_{>D}\|^2\right)\left(1 + \frac{n}{D-(n+1)}\right) + \sigma^2\frac{n}{D-(n+1)} \\
&+ \quad \left(1-\frac{n}{D}\right)\|\beta^*_D - C_{D,>D}\beta^*_{>D}\|^2.
\end{aligned}$$

□

Si on se place dans le cas où Σ est un opérateur diagonal, on se retrouve avec les bornes

$$\mathbb{E}\ell(\hat{\theta}_D, \theta^*) = \begin{cases} \|\beta^*_{>D}\|^2\left(1+\frac{D}{n-(D+1)}\right) + \sigma^2\frac{D}{n-(D+1)} & \text{si } n \leq D-2, \\ \|\beta^*_{>D}\|^2\left(1+\frac{n}{D-(n+1)}\right) + \sigma^2\frac{n}{D-(n+1)} + \left(1-\frac{n}{D}\right)\|\beta^*_D\|^2 & \text{si } D \geq n+2. \end{cases}$$

On retrouve les mêmes expressions que dans le cadre isotrope (en remplaçant les θ^* par des β^*). On a alors la même heuristique, à savoir que la double-descente va fréquemment arriver, mais pour qu'elle donne un minimum intéressant dans le régime de surparamétrisation il faudra encore

1. $\sigma^2 < \|\beta^*\|^2$,
2. β^* conservant du signal dans les grandes coordonnées ($\|\beta^*_{>D}\|^2$ ne décroît pas trop vite). On peut remarquer que lorsque on fixe $\|theta^*\|^2$, la décroissance pas trop rapide de $\|\beta^*_{>D}\|^2$ impose une décroissance lente des valeurs propres de Σ, ce qui est l'esprit des conditions formulées dans Tsigler and Bartlett [2023] pour caractériser ce phénomène de double descente.

Enfin, toujours dans le registre Gaussien, plutôt que de considérer des augmentations successives de l'ensemble de variables de départ, on peut considérer des projections aléatoires sur des espaces de dimension D comme dans Bach [2023]. Les résultats sont sensiblement les mêmes.

En dehors du cadre Gaussien

Des bornes non asymptotiques dans le cadre de variables prédictives sous-Gaussiennes peuvent être formulées via Tsigler and Bartlett [2023]. Dans ce cadre précis, les conditions de double descente sont exprimées en termes de décroissances des valeurs propres de Σ (pas trop rapide). Des théorèmes limites (dans le cadre où on autorise D et n qui tendent vers $+\infty$, avec $D/n = \gamma$ constant) sont montrés dans Hastie et al. [2022] sous des hypothèses plus faibles de type $\mathbb{E}((x^(j))^{4+\eta}) < +\infty$ (conditions de moment des variables prédictives).

On peut se faire une idée heuristique du phénomène de la manière suivante, dans le cadre $D/n = \gamma$, $n \to +\infty$. Le terme de variance d'un estimateur par moindre carrés s'écrit

$$\sigma^2\mathbb{E}\mathrm{Tr}((X^TX)^\dagger).$$

Si on suppose les $x_{i,j}$ i.i.d. de variance v (non nécessairement Gaussiens), le Théorème 3.5 donne

$$\Lambda_{\frac{1}{n}(X^TX)} \rightsquigarrow \Lambda_{MP}.$$

Comme

$$\mathrm{Tr}((X^TX)^\dagger) = \int_{u>0} \frac{1}{u}\Lambda_{X^TX}(du) = \int_{u>0} \frac{1}{nu}\Lambda_{\frac{1}{n}X^TX}(du).$$

Sous réserve qu'on puisse intervertir intégration et limite en loi, on a

$$\mathbb{E}\mathrm{Tr}((X^TX)^\dagger) \underset{D\to+\infty}{\sim} \frac{\gamma}{D}\int_{u>0} \frac{1}{u}\Lambda^{MP}(du),$$

on s'attend donc à ce que la variance diminue bien avec la dimension dans ce régime. Ces interversions sont loisibles dans bien des cas, et on peut aussi au prix de plus de technique traiter le terme de biais. Un lecteur intéressé trouvera un exemple dans Hastie et al. [2022], où la morale reste la même : pour que double-descente intéressante il y ait, le SNR doit être > 1 et le biais ne doit pas trop fortement décroître vers 0.

Bibliographie

Entropy and the combinatorial dimension. 152(1) :37–55, 2003. doi : 10.1007/s00222-002-0266-3.

E. Aamari and C. Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, 47(1) :177 – 204, 2019. doi : 10.1214/18-AOS1685. URL https://doi.org/10.1214/18-AOS1685.

T. W. Anderson. *An introduction to multivariate statistical analysis*. John Wiley & Sons, Inc., New York ; Chapman & Hall, Ltd., London, 1958.

M. Anthony and P. L. Bartlett. *Neural Network Learning : Theoretical Foundations*. Cambridge University Press, 1999. doi : 10.1017/CBO9780511624216.

F. Bach. High-dimensional analysis of double descent for linear regression with random projections, 2023.

Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, New York, second edition, 2010. ISBN 978-1-4419-0660-1. doi : 10.1007/978-1-4419-0661-8. URL https://doi.org/10.1007/978-1-4419-0661-8.

P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63) :1–17, 2019. URL http://jmlr.org/papers/v20/17-612.html.

P. C. Bellec, G. Lecué, and A. B. Tsybakov. Slope meets Lasso : Improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B) :3603 – 3642, 2018. doi : 10.1214/17-AOS1670. URL https://doi.org/10.1214/17-AOS1670.

P. J. Bickel and B. Li. Local polynomial regression on unknown manifolds. In *Complex datasets and inverse problems*, volume 54 of *IMS Lecture Notes Monogr. Ser.*, pages 177–186. Inst. Math. Statist., Beachwood, OH, 2007. ISBN 978-0-940600-70-6 ; 0-940600-70-6. doi : 10.1214/074921707000000148. URL https://doi.org/10.1214/074921707000000148.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 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. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

Boucheron, Stéphane, Bousquet, Olivier, and Lugosi, Gábor. Theory of classification : a survey of some recent advances. *ESAIM : PS*, 9 :323–375, 2005. doi : 10.1051/ps:2005018. URL https://doi.org/10.1051/ps:2005018.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. doi : 10.1017/CBO9780511804441.

O. Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004. ISBN 3-540-22572-2. doi : 10.1007/b99352. URL https://doi.org/10.1007/b99352. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.

L. Chizat and F. R. Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Annual Conference Computational Learning Theory*, 2020. URL https://api.semanticscholar.org/CorpusID:211076120.

A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 97–111. Springer, Berlin, 2007. ISBN 978-3-540-72925-9 ; 3-540-72925-9. doi : 10.1007/978-3-540-72927-3_9. URL https://doi.org/10.1007/978-3-540-72927-3_9.

L. Devroye and G. Lugosi. Lower bounds in pattern recognition and learning. *Pattern Recognition*, 28(7) :1011–1018, 1995. ISSN 0031-3203. doi : https://doi.org/10.1016/0031-3203(94)00141-8. URL https://www.sciencedirect.com/science/article/pii/0031320394001418.

C. Giraud. *Introduction to high-dimensional statistics*, volume 168 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, second edition, 2022. ISBN 978-0-367-71622-6 ; 978-0-367-74621-6 ; 978-1-003-15874-5. doi : 10.1201/9781003158745. URL https://doi.org/10.1201/9781003158745.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Ann. Statist.*, 50(2) :949–986, 2022. ISSN 0090-5364,2168-8966. doi : 10.1214/21-aos2133. URL https://doi.org/10.1214/21-aos2133.

M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 2011. ISBN 978-3-642-20211-7. Isoperimetry and processes, Reprint of the 1991 edition.

K.-C. Li. Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, 86(414) :316–342, 1991. ISSN 0162-1459,1537-274X. URL http://links.jstor.org/sici?sici=0162-1459(199106)86:414<316:SIRFDR>2.0.CO;2-V&origin=MSN. With discussion and a rejoinder by the author.

P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. ISBN 978-3-540-48497-4 ; 3-540-48497-3. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

P. Massart and É. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5) :2326 – 2366, 2006. doi : 10.1214/009053606000000786. URL https://doi.org/10.1214/009053606000000786.

J. Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics*, 50(4) :2157–2178, 2022. URL https://doi.org/10.1214/22-AOS2181.

B. Pelletier. Non-parametric regression estimation on closed Riemannian manifolds. *J. Nonparametr. Stat.*, 18(1) :57–67, 2006. ISSN 1048-5252,1029-0311. doi : 10.1080/10485250500504828. URL https://doi.org/10.1080/10485250500504828.

B. Raducanu and F. Dornaika. A supervised non-linear dimensionality reduction approach for manifold learning. *Pattern Recognition*, 45(6) :2432–2444, 2012. ISSN 0031-3203. doi : https://doi.org/10.1016/j.patcog.2011.12.006. URL https://www.sciencedirect.com/science/article/pii/S0031320311005073. Brain Decoding.

A. Tsigler and P. L. Bartlett. Benign overfitting in ridge regression. *J. Mach. Learn. Res.*, 24 :Paper No. [123], 76, 2023. ISSN 1532-4435,1533-7928.

A. Tsybakov. Aggregation and high-dimensional statistics (preliminary notes of saint-flour lectures, july 8-20, 2013). 2013. URL https://api.semanticscholar.org/CorpusID:125347443.

A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. ISBN 978-0-387-79051-0. doi : 10.1007/b13794. URL https://doi.org/10.1007/b13794. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

S. van de Geer. *Estimation and testing under sparsity*, volume 2159 of *Lecture Notes in Mathematics*. Springer, [Cham], 2016. ISBN 978-3-319-32773-0 ; 978-3-319-32774-7. doi : 10.1007/978-3-319-32774-7. URL https://doi.org/10.1007/978-3-319-32774-7. Lecture notes from the 45th Probability Summer School held in Saint-Four, 2015, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

S. A. van de Geer. Empirical processes in m-estimation. 2000.

A. Van Der Vaart and J. A. Wellner. A note on bounds for vc dimensions. *Institute of Mathematical Statistics collections*, 5 :103, 2009.

N. Verzelen. Minimax risks for sparse regressions : ultra-high dimensional phenomenons. *Electron. J. Stat.*, 6 :38–90, 2012. ISSN 1935-7524. doi : 10.1214/12-EJS666. URL https://doi.org/10.1214/12-EJS666.

J. Wellner et al. *Weak convergence and empirical processes : with applications to statistics*. Springer Science & Business Media, 2013.

B. Yu. Assouad, fano and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

Y. Zhang, M. J. Wainwright, and M. I. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In M. F. Balcan, V. Feldman, and C. SzepesvÃ¡ri, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 921–948, Barcelona, Spain, 13–15 Jun 2014. PMLR.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(90) :2541–2563, 2006. URL http://jmlr.org/papers/v7/zhao06a.html.

Annexe A

Méthodes pour les bornes inférieures

On rappelle que le point de départ en apprentissage statistique (du point de vue théorique) est de déterminer la difficulté d'un problème via sa vitesse d'apprentissage

$$\inf_{\hat{f}_n} \sup_{P \in \mathcal{P}} E_{D_n \sim P^{\otimes n}}(R_P(\hat{f}_n) - R_P^*) \sim_n a_n,$$

où la classe $\mathcal{P}$ est l'ensemble des configurations possibles pour la loi générant les données, correspondant à la classe de problèmes que l'on souhaite résoudre. Pour rappel, on ne peut pas prendre $\mathcal{P} = \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ (ensemble des mesures de probas sur le produit espace des variables prédictives $\times$ espace des valeurs à prédire) d'après le No-Free Lunch Theorem. On a successivement vu

— en classification le cas où $f_P^* \in \mathcal{F}$, où $\mathcal{F}$ est de dimension de Vapnik finie,

— en régression le cas où $f^* = \langle x, \theta^* \rangle$ pour un certain θ^*, puis des problèmes avec des hypothèses de parcimonie sur θ^* (ensemble des problèmes tels que $\|\theta^*\|_0 \leq s$ par exemple).

Dans le cours ces cas étaient présentés comme l'erreur d'estimation (pour une classe de problème plus générale). Dans le cas où l'erreur d'approximation est nulle, cette erreur d'estimation est confondue avec la vitesse d'apprentissage.

Par ailleurs, on s'est concentré sur la majoration des ces vitesses d'apprentissage : on a, suivant les situations, trouvé des méthodes $\hat{f}_n$ particulières, pour lesquelles on a étudié $E_{D_n}(R_P(\hat{f}_n) - R_P^*)$ ce dont on a déduit une majoration de a_n (en essayant de faire ressortir l'influence de la dimension).

Pour attester de l'optimalité d'une procédure de prédiction, il reste à vérifier que l'ordre de grandeur de la vitesse précédemment obtenue ne peut pas être dépassé uniformément sur $\mathcal{P}$, c'est à dire qu'il faut prouver des résultats du type

$$\inf_{\hat{f}_n} \sup_{P \in \mathcal{P}} E_{D_n \sim P^{\otimes n}}(R_P(\hat{f}_n) - R_P^*) \gtrsim a_{n,-},$$

c'est l'objet des méthodes de **bornes inférieures** (sur l'excès de risque en prédiction).

A.1 Le Bayésien comme minorant du minimax

On l'a déjà fait informellement pour le No-Free Lunch Theorem : "l'astuce" est de minorer

$$\sup_{P\in\mathcal{P}} E_{D_n\sim P^{\otimes n}}(R_P(\hat{f}_n) - R_P^*) \geq \int_\Theta \pi(d\theta) E_{D_n\sim P_\theta^{\otimes n}}(R_{P_\theta}(\hat{f}_n) - R_\theta^*),$$

c'est à dire de minorer un sup d'excès de risque sur une classe $P_\theta, \theta \in \Theta$ inclue dans $\mathcal{P}$ par une intégrale de ces excès de risque. Pourquoi calculer une telle intégrale serait plus simple ? La réponse est que, lorsque l'on met une répartition de probabilité sur l'ensemble des distributions source, on tombe dans le cadre général de la Statistique Bayésienne, pour laquelle on sait faire certaines choses (comme déterminer facilement les prédicteurs optimaux).

A.1.1 Le minimum de Statistique Bayésienne

On repart dans le domaine de la statistique inférentielle classique, et on suppose que l'on veut estimer $q(\theta) \in \mathbb{R}^k$, pour $\theta \in \Theta$, indexant aussi le modèle $(P_\theta)_{\theta\in\Theta}$ de lois sur $\mathcal{X}$ (espace d'observation). Les performances d'un estimateur $T : \mathcal{X} \to \mathbb{R}^k$ sont mesurées avec une fonction de perte

$$\ell : \begin{cases} \mathbb{R}^k \times \mathbb{R}^k & \to & [0,+\infty] \\ (y',y) & \mapsto & \ell(y',y). \end{cases}$$

Le risque d'un estimateur T est défini par

$$R_T : \begin{cases} \Theta & \to & [0,+\infty] \\ \theta & \mapsto & E_\theta(\ell(q(\theta),T(X)), \end{cases}$$

ce qui correspond à la fonction qui à un θ associe sa perte moyenne si les données X sont tirées suivant P_θ (ce qui correspond au cadre **fréquentiste**, θ est fixé et on observe $X \sim P_\theta$). Pour un tel estimateur, le risque maximal est alors défini par

$$\bar{R}_T = \sup_{\theta\in\Theta} R_T(\theta),$$

et, si π est une loi sur Θ, le risque intégré (ou bayésien) de T est défini par

$$\rho(T,\pi) = \int_\Theta \pi(d\theta) R_T(\theta),$$

sous les conditions de mesurabilités pour lesquelles cela à un sens (légères). On a alors évidemment

$$\bar{R}_T \geq \rho(T,\pi) \geq \inf_T \rho(T,\pi),$$

où le dernier terme est appelé **Risque Bayésien** (tout court), pour toute loi π sur Θ, et l'idée pour les minorations de vitesses d'apprentissage va être de relier les risques maximaux de prédicteurs à un problème Bayésien (cf les exemples qui suivent).

Reste à calculer ces risques Bayésiens. Pour cela, on se place du *point de vue Bayésien*, c'est à dire que l'on suppose observer non plus des données X de loi P_θ à θ fixé, mais plutôt que la "nature" tire un θ suivant la loi π maintenant appelée **loi a priori** (on notera cette variable aléatoire $\tilde{\theta} \sim \pi$), et que l'on observe $\tilde{X}$ de loi $P_{\tilde{\theta}}$. En d'autres termes, on observe $\tilde{X}$ défini par $\tilde{X} \mid \tilde{\theta} \sim P_{\tilde{\theta}}$, où $\tilde{\theta} \sim \pi$. Formellement il y a deux points fondamentaux qui diffèrent du point de vue fréquentiste :

1. la loi de $\tilde{X}$ n'est plus un élément de $(P_\theta)_{\theta\in\Theta}$ mais un mélange d'éléments de cet ensemble de lois. Plus précisément, si $P_\theta \sim f_\theta \mu$ (modèle dominé par μ), $\tilde{X}$ aura pour densité

$$f(x) = \int_{t\in\Theta} f_t(x)\pi(dt).$$

2. Si on observe un n-échantillon $\tilde{X}_1, \dots, \tilde{X}_n$ suivant le point de vue Bayésien, $(\tilde{X}_1, \dots, \tilde{X}_n)$ **n'est plus indépendant** : ils dépendent les uns des autres via $\tilde{\theta}$. En revanche, les $\tilde{X}_i \mid \tilde{\theta}$ eux sont indépendants.

On sent bien que, techniquement parlant, le Bayésien se résume souvent à un jeu de conditionnement. L'élément clé de l'estimation Bayésienne est la **loi a posteriori**, c'est à dire la loi de $\tilde{\theta} \mid \tilde{X}$, ou encore la loi du paramètre sachant les observations. Par exemple, dans le cas dominé précédent, la formule de Bayes stipule que $\tilde{\theta} \mid \tilde{X}$ a une densité par rapport à la loi a priori π, donnée par

$$\frac{f_t(\tilde{X})}{\int_{t\in\Theta} f_t(\tilde{X})}.$$

Exemple A.1 *: Cas Gaussien.* Si $\tilde{\theta} \sim \mathcal{N}(0, v)$, et $\tilde{X} \mid \tilde{\theta} \sim \mathcal{N}(\tilde{\theta}, \sigma^2)$, la formule de Bayes donne

$$\begin{aligned}(\tilde{\theta} \mid \tilde{X})(d\theta) &\propto e^{-(\tilde{X}-\theta)^2/2\sigma^2} e^{-\theta^2/2v} d\theta \\ &\propto e^{\left[\frac{1}{2}\left(\frac{1}{v}+\frac{1}{\sigma^2}\right)(\theta - \frac{v}{\sigma^2+v}\tilde{X})^2\right]},\end{aligned}$$

où les $\propto$ désignent proportionnalité à un terme **ne dépendant pas de** θ près (il peut y avoir du $\tilde{X}$ dedans mais dans ces calculs il est considéré comme fixe). On en déduit que

$$\tilde{\theta} \mid \tilde{X} \sim \mathcal{N}\left(\frac{v}{\sigma^2+v}\tilde{X}, \frac{\sigma^2 v}{\sigma^2+v}\right).$$

Cette loi a posteriori reflète (et c'est un schéma général) un compromis entre la loi a priori et l'influence des observations :

— si $v \to 0$, l'information a priori "θ est proche de 0" est de plus en plus certaine. Cela se traduit pas $\tilde{\theta} \mid \tilde{X} \rightsquigarrow \delta_0$, l'information a priori "proche de 0" prend le pas sur les observations.

— si $v \to +\infty$, l'information a priori "θ est proche de 0" est de moins en moins certaine. Cela se traduit alors par $\tilde{\theta} \mid \tilde{X} \rightsquigarrow \mathcal{N}(\tilde{X}, \sigma^2)$, soit une loi a posteriori 'sans information a priori'.

Une fois la loi a posteriori définie, on peut minimiser (au moins formellement) le risque Bayésien assez facilement. On commence par remarquer que si T est un estimateur de $q(\theta)$, alors

$$\begin{aligned}\rho(T,\pi) &= \mathbb{E}(\ell(q(\tilde{\theta}), T(\tilde{X}))) \\ &= \mathbb{E}\left[\mathbb{E}\left(\ell(q(\tilde{\theta}), T(\tilde{X}))) \mid \tilde{X}\right)\right],\end{aligned}$$

où la dernière quantité $\mathbb{E}\left(\ell(q(\tilde{\theta}), T(\tilde{X}))) \mid \tilde{X}\right)$ est appelée **risque a posteriori** de T. Un estimateur T est alors optimal au sens du risque intégré (on dit qu'il est **bayésien**) s'il minimise le risque a posteriori. On définit alors

$$T_b(\tilde{X}) \in \arg\min_y \mathbb{E}\left(\ell(q(\tilde{\theta}), y) \mid \tilde{X}\right),$$

lorsque bien défini, et on a immédiatement que $T_b \in \arg\min_T \rho(T, \pi)$, et donc, que pour tout estimateur T,

$$\bar{R}_T \geq \rho(T, \pi) \geq \rho(T_b, \pi).$$

Exemple A.2 *: Perte quadratique et test bayésien.* Les deux cas les plus courants de fonctions de perte (et les estimateurs bayésiens correspondant) sont

1. la perte quadratique $\ell(y, y') = \|y - y'\|^2$. Un estimateur bayésien minimise donc en y la fonction

   $$\mathbb{E}\left(\|y - \tilde{\theta}\|^2 \mid \tilde{X}\right),$$

   et sous conditions d'intégrabilité on a immédiatement $T_b(\tilde{X}) = \mathbb{E}(\tilde{\theta} \mid \tilde{X})$.
2. La perte 0/1, $\ell(y, y') = \mathbb{1}_{y \neq y'}$ (pour $y, y' \in \{0, 1\}$, correspondant en fait à un problème où $q(\theta) \in \{0, 1\}$, et donc a un problème de test). Si on note Θ_j les zones de Θ correspondant aux deux hypothèses de test, on a

   $$\mathbb{E}\ell(q(\tilde{\theta}), y) \mid \tilde{X}) = \mathbb{1}_{y=1}\mathbb{P}\left(\tilde{\theta} \in \Theta_0 \mid \tilde{X}\right) + \mathbb{1}_{y=0}\mathbb{P}\left(\tilde{\theta} \in \Theta_1 \mid \tilde{X}\right),$$

   qui est alors minimale pour $T_b(\tilde{X}) \in \arg\max_j \mathbb{P}\left(\tilde{\theta} \in \Theta_j \mid \tilde{X}\right)$, soit l'hypothèse la plus chargée par la loi a posteriori (la plus vraisemblable a posteriori en somme).

Une fois que l'on a déterminé notre estimateur bayésien, et le risque bayésien $\rho(\pi)$ correspondant, on rappelle qu'on aura toujours

$$\inf_T \sup_\theta R_T(\theta) \geq \rho(\pi).$$

Il s'agit alors de choisir une bonne (suite de) loi a priori π_k, pour avoir une minoration du risque minimax en

$$\inf_T \sup_\theta R_T(\theta) \geq \limsup_k \rho(\pi_k).$$

Pour revenir au problème initial de minoration d'une vitesse d'apprentissage, le cahier des charges technique est donc :

1. Exprimer votre problème d'apprentissage sous un jour Bayésien (avec une fonction de perte appropriée).
2. Construire une suite (des fois une seule suffit) de lois a priori telle que $\rho(\pi_k) \to a_{n,-}$.

A.1.2 Borne inférieure en régression linéaire

On se place dans le cadre de la régression linéaire à design fixe, pour la classe de problème du type

$$Y = X\theta + \varepsilon,$$

où $Y \in \mathbb{R}^n$, les ε_i sont i.i.d., centrés, de variance σ^2, et pour simplifier $X^T X = I_d$. On a vu dans ce cas que, pour le problème de prédiction,

$$\begin{aligned} R(\hat{\theta}_{LS}) - R(\theta^*) &= \|\hat{\theta}_{LS} - \theta^*\|^2 \\ &= d\sigma^2. \end{aligned}$$

On a donc, au sens des vitesses d'apprentissage,

$$\inf_{\hat{f}} \sup_{\theta \in \mathbb{R}^d, \varepsilon | Y \sim X\theta + \varepsilon} \mathbb{E}\|\hat{f} - \theta\|^2 \leq d\sigma^2.$$

On remarque que le supremum est pris sur toutes les lois d'erreurs possibles. Essayons de calculer la borne inférieure correspondante. On a directement

$$\inf_{\hat{f}} \sup_{\theta \in \mathbb{R}^d, \varepsilon | Y \sim X\theta + \varepsilon} \mathbb{E}\|\hat{f} - \theta\|^2 \geq \inf_{\hat{f}} \sup_{\theta \in \mathbb{R}^d, Y \sim \mathcal{N}(X\theta, \sigma^2 I_n)} \mathbb{E}\|\hat{f} - \theta\|^2,$$

où on a restreint le problème à droite à des erreurs Gaussiennes, et donc à un modèle Gaussien standard. Si on se place dans un cadre Bayésien en mettant une loi a priori π sur θ, on aura alors immédiatement

$$\inf_{\hat{f}} \sup_{\theta \in \mathbb{R}^d, \varepsilon | Y \sim X\theta + \varepsilon} \mathbb{E}\|\hat{f} - \theta\|^2 \geq \mathbb{E}\|\tilde{\theta} - T_b(\tilde{Y})\|^2 = \rho(\pi).$$

Prenons par exemple une loi a priori $\tilde{\theta} \sim \mathcal{N}(0_d, vI_d)$. Alors un calcul simple montre que

$$\tilde{\theta} \mid \tilde{Y} \sim \mathcal{N}\left(\frac{v}{v + \sigma^2} X^T \tilde{Y}, \frac{\sigma^2 v}{\sigma^2 + v} I_d\right).$$

L'estimateur bayésien est donc donné par $T_b(\tilde{Y}) = \frac{v}{v+\sigma^2} X^T \tilde{Y}$, et son risque a posteriori par

$$\mathbb{E}\left\|T_b(\tilde{Y}) - \tilde{\theta}\right\|^2 \mid \tilde{Y} = \frac{v\sigma^2 d}{\sigma^2 + v}.$$

On en déduit alors que, pour tout v,

$$\inf_{\hat{f}} \sup_{\theta \in \mathbb{R}^d, \varepsilon | Y \sim X\theta + \varepsilon} \mathbb{E}\|\hat{f} - \theta\|^2 \geq \frac{v\sigma^2 d}{\sigma^2 + v},$$

et en faisant tendre v vers $+\infty$ on en déduit que

$$\inf_{\hat{f}} \sup_{\theta \in \mathbb{R}^d, \varepsilon | Y \sim X\theta + \varepsilon} \mathbb{E}\|\hat{f} - \theta\|^2 \geq d\sigma^2,$$

et donc que l'estimateur par moindre carrés est optimal. Pour le design aléatoire, les calculs sont sensiblement les mêmes. Toujours est-il que c'est comme cela que l'on prouve l'ordre de grandeur en $d\sigma^2/n$ pour le problème de régression linéaire.

A.2 Arsenal technique : réduction du problème Bayésien

La plupart des temps il est compliqué d'obtenir l'expression exacte d'un risque Bayésien $\rho(\pi)$, et donc d'en déduire une minoration pertinente d'une vitesse d'apprentissage. Même si le principe Bayésien reste valable, on l'utilise au travers d'un schéma de réduction à des hypothèses à deux (ou plusieurs) points.

Théorème A.3 : Réduction à deux points : Lemme de Le Cam

Si $\ell(y, y') = \mathrm{d}(y, y')$, où d est une distance. Soient $\mathcal{P}_1$, $\mathcal{P}_2$ deux sous-ensembles de $\mathcal{P}$ tels que, pour tout $P_j \in \mathcal{P}_j$,

$$\mathrm{d}(q(P_1), q(P_2)) \geq \delta,$$

où $q(P)$ est le paramètre à estimer. Alors, pour tout estimateur T de $q(P)$, on a

$$\sup_P E_P \mathrm{d}(T(X), q(P)) \geq \frac{\delta}{4}(1 - \sup_{P_j \in \mathcal{P}_j} \mathrm{d}_{TV}(P_1, P_2)).$$

Le plus souvent ce Lemme est utilisé pour des n-uplets avec des hypothèses simples, c'est à dire $\mathcal{P}_j = P_j^{\otimes n}$. L'heuristique est alors simple : si on arrive à trouver deux lois "similaires" (ici mesuré en termes de distance en variation totale) dont les paramètres associés sont loin, alors une indécidabilité dans le problème de test entre ces deux lois sources induite par la proximité des lois résultera en un gros écart pour la procédure d'estimation (et pour notre vitesse d'apprentissage).

Preuve du Théorème A.3. On rappelle que la distance en variation totale entre deux lois est définie par

$$\begin{aligned}
\mathrm{d}_{TV}(P, Q) &= \sup_{A \in \mathcal{A}} |P(A) - Q(A)| \\
&= \frac{1}{2} \int |p - q| d\mu \\
&= \frac{1}{2} \int (p + q - 2(p \wedge q)) d\mu \\
&= 1 - \|P \wedge Q\|,
\end{aligned}$$

où $\|P \wedge Q\|$ est appelée *affinité* entre P et Q.

Pour faire simple on suppose que l'infimum est réalisé par P_1 et P_2, et on se

donne une loi a priori $\pi \sim \frac{1}{2}(\delta_1 + \delta_2)$. On a alors

$$\begin{aligned}
\sup_P E_P \mathrm{d}(T(X), q(P)) &\geq \sup_{P \in \{P_1, P_2\}} E_P \mathrm{d}(T(X), q(P)) \\
&\geq \mathbb{E}\mathrm{d}(T(\tilde{X}), q(P_{\tilde{\theta}})) \\
&= \mathbb{E}\left(\mathbb{E}\left[\mathrm{d}(T(\tilde{X}), q(P_{\tilde{\theta}})) \mid \tilde{X}\right]\right) \\
&\geq \frac{\delta}{2}\mathbb{E}\left(\mathbb{P}(\tilde{\theta} = 1 \mid \tilde{X}) \wedge \mathbb{P}(\tilde{\theta} = 2 \mid \tilde{X})\right) \\
&= \frac{\delta}{2}\mathbb{E}\left[\frac{p_1}{p_1 + p_2}(\tilde{X}) \wedge \frac{p_2}{p_1 + p_2}(\tilde{X})\right] \\
&= \frac{\delta}{4}\int p_1 \wedge p_2 d\mu,
\end{aligned}$$

où on a utilisé $\tilde{X} \sim \frac{1}{2}(p_1 + p_2)d\mu$. □

L'heuristique de lois similaires à paramètres lointains se traduit techniquement par une minoration pour le problème de test Bayésien correspondant. La réduction à deux points permet d'avoir la bonne vitesse en taille d'échantillon, en utilisant

$$\|P_1^{\otimes n} \wedge P_2^{\otimes n}\| \geq \|P_1 \wedge P_2\|^n = (1 - \mathrm{d}_{TV}(P_1, P_2))^n$$

par exemple, ou d'autres manières de minorer l'affinité entre n-uplets (on peut passer par la distance de Hellinger, Kullback, bref la distance qui arrange, pour plus de détails voir Tsybakov [2009]).

Si on veut faire apparaître la dimension dans les bornes inférieures, il faut souvent aller plus loin que la réduction à deux points en utilisant une réduction à disons m points. On peut par exemple utiliser le Lemme d'Assouad.

Théorème A.4 : Réduction à m points, Lemme d'Assouad

Si $\mathcal{P}$ contient $\mathcal{P}_m = \{P_\tau \mid \tau \in \{-1, 1\}^m\}$, et la fonction de perte d d'écrit

$$\mathrm{d}(x, y) = \sum_{j=1}^{m} \mathrm{d}_j(x, y),$$

où les d_j sont des pseudo-distances, et si de plus, pour tout $\tau, \tau' \in \{-1, 1\}^m\}$,

$$\tau \sim_j \tau' \quad \Rightarrow \quad \mathrm{d}_j(q(\tau), q(\tau')) \geq \delta_m,$$

où $\tau \sim_j \tau'$ si τ et τ' diffèrent uniquement sur la j-ème coordonnée, alors, on a, pour tout estimateur T de $q(P)$,

$$\sup_{P \in \mathcal{P}} E_P \mathrm{d}(q(P), T(X)) \geq \frac{m\delta_m}{4}(1 - \max_{\tau \sim \tau'} \mathrm{d}_{TV}(P_\tau, P'_\tau)).$$

On voit alors qu'il faudra construire un sous-ensemble de lois dont la différence sur une "coordonnée" impliquera une distance entre lois contrôlée, et un apport à la distance totale conséquent. La preuve est encore basée sur une histoire de test Bayésien.

Preuve du Théorème A.4. En se donnant une loi a priori $\tilde{\tau}$ sur τ uniforme encore, on a

$$\begin{aligned}
\sup_{P\in\mathcal{P}} E_P \mathrm{d}(q(P), T(X)) &\geq \sup_{\tau\in\{-1,1\}^m} E_\tau \mathrm{d}(q(\tau), T(X)) \\
&\geq \mathbb{E}\mathrm{d}(q(\tilde{\tau}, T(\tilde{X})) \\
&= \sum_{j=1}^m \mathbb{E}\mathrm{d}_j(q(\tilde{\tau}), T(\tilde{X})) \\
&\geq \sum_{j=1}^m \frac{\delta_m}{2} \mathbb{E}\left(\mathbb{P}(\tilde{\tau}_j = -1 \mid \tilde{X}) \wedge \mathbb{P}(\tilde{\tau}_j = 1 \mid \tilde{X})\right) \\
&\geq \frac{m\delta_m}{4} \min_{\tau\sim\tau'} \|P_\tau \wedge P'_\tau\|,
\end{aligned}$$

avec les mêmes calculs que dans le cas à deux points précédent. □

L'inconvénient technique du Lemme d'Assouad est la structure d'hypercube à imposer à l'ensemble des lois d'intérêt. On peut trouver d'autres méthodes pour des hypothèses à m points ne faisant pas forcément appel à une structure d'hypercubes, comme le Lemme de Fano-Birgé. Un lecteur intéressé pourra se référer à Yu [1997].

A.2.1 Borne inférieure (simple) en classification

On conclut sur la borne inférieure en classification, sous l'hypothèse que le classifieur optimal appartient à une classe $\mathcal{F}$ de VC-dimension d : on voudrait attraper une borne en $\sqrt{d/n}$ dans le cas général.

Pour fixer les notations, on note $\mathcal{X}$ l'espace des variables prédictives, $\mathcal{F}$ un ensemble de classifieurs de VC-dimension d, et $\mathcal{P}$ l'ensemble des lois sur $\mathcal{X} \times \{0,1\}$ telles que le classifieur de Bayes associé est dans $\mathcal{F}$ (dans ce cas l'excès de risque sur cette classe correspond à l'erreur d'estimation dans le cas général).

On suit l'approche de Devroye and Lugosi [1995] : on se donne $x_1, \ldots, x_d$ pulvérisé par par $\mathcal{F}$, et, pour $\theta \in \{0,1\}^d$, on note P_θ la distribution suivante sur $\mathcal{X} \times \{0,1\}$:

$$\begin{aligned}
&\forall j = 1, \ldots, d \quad P_\theta(X = x_j) = p = \frac{1}{d} \\
&\forall j = 1, \ldots, d \quad P_\theta(Y = 1 \mid X = x_j) = \frac{1}{2} + c(2\theta_j - 1),
\end{aligned}$$

où j est un paramètre à calibrer ultérieurement. On va raisonner pour n assez grand, à calibrer à chaque étape.

Le début est standard :

$$\inf_{\hat{f}} \sup_{P\in\mathcal{P}} E_{D_n\sim P^n}(R_P(\hat{f}) - R_P(f_P^*)) \geq \inf_{\hat{f}} \sup_{\theta\in\{0,1\}^d} E_{D_n\sim P_\theta^n}(R_\theta(\hat{f}) - R_\theta(f_\theta^*)).$$

Maintenant, pour un $\theta \in \{0,1\}^j$, un classifieur f est uniquement déterminé (P_θ presque sûrement) par ses valeurs $\hat{f}_j := \hat{f}(x_j)$. Par ailleurs, $f_\theta^* = \theta$ (en notation vectorielle). On peut alors écrire

$$\begin{aligned}
(R_\theta(\hat{f}) - R_\theta(f_\theta^*)) &= \sum_{j=1}^d 2pc \mathbb{1}_{\hat{f}_j \neq \theta_j} \\
&= \sum_{j=1}^d \mathrm{d}_j(\hat{f}, \theta),
\end{aligned}$$

avec les notations du Théorème A.4. On remarque alors que, si $\theta \sim_j \theta'$, $\mathrm{d}_j(\theta, \theta') = 2pc$, et le Lemme d'Assouad donne alors

$$\inf_{\hat{f}} \sup_{\theta \in \{0,1\}^d} E_{D_n \sim P_\theta^n}(R_\theta(\hat{f}) - R_\theta(f_\theta^*)) \geq \frac{dpc}{2} \min_{\theta \sim \theta'} \|P_\theta^n \wedge P_{\theta'}^n\|.$$

Reste à contrôler $\|P_\theta^n \wedge P_{\theta'}^n\|$. Malheureusement la borne $\|P_\theta^n \wedge P_{\theta'}^n\| \geq \|P_\theta \wedge P_{\theta'}\|^n$ donne un résultat sous-optimal. On est obligés de passer par une distance intermédiaire (distance de Hellinger), au travers du Lemme suivant.

Lemme A.5 : Une inégalité de Le Cam

$$\|P \wedge Q\| \geq \frac{1}{2}\rho(P, Q)^2,$$

où $\rho(P, Q) = \int \sqrt{pq}$ *est l'affinité de Hellinger.*

Cette inégalité est une des inégalités de Le Cam, dont on peut trouver une liste dans [Tsybakov, 2009, Lemme 2.3].

Une fois ce lemme en poche, on a, pour $\theta \sim \theta'$,

$$\begin{aligned}\|P_\theta^n \wedge P_{\theta'}^n\| &\geq \frac{1}{2}\rho(P_\theta^n, P_{\theta'}^n)^2 \\ &\geq \frac{1}{2}\rho(P_\theta, P_{\theta'})^{2n},\end{aligned}$$

ce qui est tout l'avantage de passer par Hellinger. Maintenant,

$$\rho(P_\theta, P_{\theta'}) = 1 - p + \sqrt{p^2(1/2 - c)(1/2 + c)} + \sqrt{p^2(1/2 + c)(1/2 - c)} = 1 - p + p\sqrt{1 - 4c^2}.$$

Maintenant, si on pose $c = \frac{\sqrt{d}}{2\sqrt{n}}$, on a

$$\begin{aligned}\inf_{\hat{f}} \sup_{\theta \in \{0,1\}^d} E_{D_n \sim P_\theta^n}(R_\theta(\hat{f}) - R_\theta(f_\theta^*)) &\geq \frac{1}{8}\sqrt{\frac{d}{n}}\left(1 + \frac{1}{d}\left(\sqrt{1 - (d/n)} - 1\right)\right)^{2n} \\ &\geq \frac{1}{8}\sqrt{\frac{d}{n}} \exp\left[2n\left(\frac{1}{d}(1 - d/(2n) + o(1/n) - 1)\right)\right] \\ &\geq \frac{e^{-1}}{8}\sqrt{\frac{d}{n}} e^{o(1)}.\end{aligned}$$

On en déduit alors que, pour n assez grand (constante dépendant des développements limités mais universelle),

$$\inf_{\hat{f}} \sup_{\theta \in \{0,1\}^d} E_{D_n \sim P_\theta^n}(R_\theta(\hat{f}) - R_\theta(f_\theta^*)) \geq \frac{e^{-1}}{16}\sqrt{\frac{d}{n}},$$

et on a la vitesse voulue. Pour des bornes plus précises en la taille d'échantillon (suivant n grand ou petit), on peut se référer à Devroye and Lugosi [1995].