APPC
Apprentissage parcimonieux

théorie et algorithmes

A. Rakotomamonjy

29 septembre 2017

Chapitre 1

Apprentissage supervisé

1.1 Introduction

On s'intéresse à un cadre d'apprentissage supervisé *ie* à un problème où on cherche à apprendre une fonction qui estime la dépendance entre deux espaces, à partir de couples d'exemples $\{(\mathbf{x}_i, y_i)\}$ avec $\mathbf{x}_i \in \mathcal{X}$ et $y_i \in \mathcal{Y}$. Dans la plupart des cas qui nous interessent, $\mathcal{X} = \mathbb{R}^d$ et $\mathcal{Y} = \mathbb{R}$.

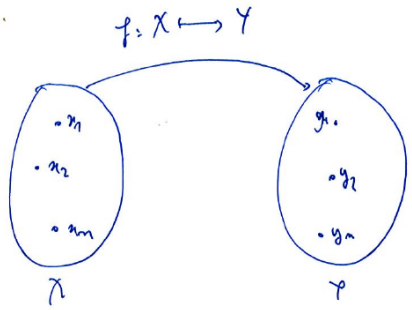

FIGURE 1.1 – Estimation de dépendance entre deux espaces $\mathcal{X}$ et $\mathcal{Y}$ à partir de données.

Le lien qui relie les deux élements du couple $(\mathbf{x}_i, y_i)$ est régi par une loi de probabilité jointe $P(X, Y)$ qui est inconnue. On considère par ailleurs que les couples d'exemples $(\mathbf{x}_i, y_i)$ ont été tirées *iid* selon cette loi.

Notre objectif est de construire ou d'apprendre une fonction $f(\mathbf{x})$ qui "prédit" au mieux les "y" associée pour l'ensemble des couples $(\mathbf{x}, y)$ possibles. La notion de prédire au mieux est subjective et dépend du problème considéré. Typiquement, on utilise des fonctions coûts pénalisant une mauvaise prédiction :

— coût quadratique :
$$\ell(y, \hat{y}) = \|y - \hat{y}\|_2^2$$

— coût Hinge :
$$\ell(y, \hat{y}) = \max(0, 1 - y\hat{y})$$

— coût logistique
$$\ell(y, \hat{y}) = \log(1 + e^{y\hat{y}})$$

avec $\hat{y} = f(x)$.

Une fois que le coût d'une mauvaise prédiction est définie, on peut également formaliser le problème d'optimisation qui nous intéresse :

$$\min_{f \in \mathcal{H}} \int_{\mathcal{X} \times \mathcal{Y}} \ell(y, f(\mathbf{x})) p(\mathbf{x}, y) d\mathbf{x} dy$$

avec $\mathcal{H}$, l'espace d'hypothèses (l'espace où on cherche la fonction f).

Un algorithme d'apprentissage sera donc une "fonction" qui, étant donnée les exemples d'apprentissage $\{(\mathbf{x}_i, y_i)_{i=1}^n\}$ et un espace d'hypothèses $\mathcal{H}$, nous retourne une fonction de $\mathcal{H}$ qui prédit au mieux les futurs y en résolvant ou en approximant une solution du problème ci-dessus. On remarque par ailleurs, que l'on cherche une fonction qui minimise l'ésperance du coût, notée

$$E_{X,Y}[\ell(Y, f(X)] = L(f)$$

comme $P(X, Y)$ est inconnue, ce problème de minimisation est très difficile et on s'interesse plutôt au problème de minimisation du risque empirique (MRE)

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) = L_S(f)$$

1.2 Apprentissage et régularisation

Dans la plupart des méthodes d'apprentissage, on cherche à apprendre une fonction de prédiction $f(\mathbf{x})$ en minimisant non pas un risque empirique mais un risque empirique régularisé. Notre objectif est de comprendre quel est l'intêret de ce terme de régularisation et ensuite d'apprendre, dans la suite du cours, à résoudre des problèmes d'apprentissage où le terme de régularisation prend des formes plus générales.

1.2.1 Etude de cas : régression au sens des moindres carrés

Pour faire émerger l'intêret de la régularisation, on s'interesse au cas de la régression au sens des moindres carrés. On cherche dans ce cas à apprendre

une fonction de prédiction à partir d'un ensemble de données d'apprentissage $D = \{\mathbf{x}_i, \mathbf{y}_i\}$ avec $\mathbf{x}_i \in \mathbb{R}^d$. On suppose que cette fonction est paramétrée par un vecteur $\mathbf{w}$ et appartient à une classe de fonction $\mathcal{H}$ également appelée l'*espace des hypothèses*.

Notons par ailleurs que notre objectif idéal est de trouver la meilleure fonction de prédiction qui est

$$f^\star(\mathbf{x}) = \arg \min_{f:\mathbb{R}^d \mapsto \mathbb{R}} E_{X,Y}[f(X) - Y)^2]$$

alors qu'en pratique on cherche la solution qui minimise le risque empirique dans dans l'espace $\mathcal{H}$. Dans le cas des moindres carrés, on peut montrer que cette fonction "optimale" s'écrit $f^\star(\mathbf{x}) = E(Y|X = x)$.

Supposons maintenant que le modèle que l'on a appris appliqué à une donnée $\mathbf{x}$ s'écrit $\hat{f}(\mathbf{x}) = f(\mathbf{x}; \hat{\mathbf{w}})$ où $\hat{\mathbf{w}}$ a été appris à partir des données d'apprentissage D. On suppose également que le vrai modèle qui a généré les données s'écrit :

$$y = F(\mathbf{x}) + \varepsilon$$

où ε est un bruit centré et indépendant de $\mathbf{x}$. On note également $\bar{f}(\mathbf{x}) = E_D(f(x; \hat{w}))$ l'espérance sur l'ensemble des données d'apprentissage de la prédiction en $\mathbf{x}$.

Pour évaluer la qualité de notre fonction de prédiction, on s'interesse à l'erreur quadratique moyenne en un point $\mathbf{x}$ donnée :

$$E_{D,y|x\sim p(y|x)}((y - \hat{f}(\mathbf{x}))^2)$$

En développant cette erreur, on peut montrer qu'elle se décompose en trois parties

$$E((y - \hat{f}(\mathbf{x}))^2) = E_D((\bar{f}(\mathbf{x}) - \hat{f}(\mathbf{x}))^2) + (F(x) - \bar{f}(x))^2 + E_y((y - F(\mathbf{x}))^2)$$

où le premier terme peut s'interpréter comme la variance de notre estimation au regard des données d'apprentissage, le deuxième s'interprete comme étant un biais au carré de notre estimation par rapport au vrai modèle tandis que le troisième terme est une erreur inhérente aux données observées (elle correspond en fait à la variance du bruit ε).

Autant il est impossible d'influence l'erreur liée au bruit, autant les erreurs dues aux biais et la variance sont liées à notre espace d'hypothèse. La question qui nous interesse est donc : peut on trouver un espace d'hypothèse qui permet de minimiser conjointement le biais et la variance de l'erreur ?

Le biais peut éventuellement être annulé si l'espace d'hypothèse contient le vrai modèle $F(\mathbf{x})$, par exemple dans le cas où le $F(\mathbf{x})$ est une fonction linéaire et l'espace d'hypothèse est lui-même l'espace des fonctions linéaires ou l'espace des polynomes de degré $< d$. Cependant, l'inégalité de Cramer-Rao nous dit que la variance d'un estimateur sans biais est bornée

inférieurement. On ne peut donc pas avoir un biais nulle et une variance arbitrairement faible. Un compromis doit donc être réalisé entre le biais et la variance : On peut accepter un léger biais si elle permet de réduire considérablement la variance.

1.2.2 Compléxité d'un espace d'hypothèses et compromis biais-variance

De manière intuitive, la compléxité d'un espace d'hypothèse ou d'un modèle issu de cet espace d'hypothèse peut être évaluée par son dégré de liberté, une grandeur liée au nombre de paramètres libres à estimer dans le modèle. Ainsi,

- plus un modèle est complexe, plus l'espace de fonctions qu'il peut approximer est grand *i.e* l'espace d'hypothèse est large. Dans ce cas, on peut donc espérer avoir un modèle avec un biais faible mais malheureusement avec une grande variance car la fonction choisie sera très sensible aux données d'apprentissage.
- Si l'espace d'hypothèse est simple (ou peu complexe), les fonctions constituant cet espace seront rigide et s'adapteront moins aux données. Les modèles issus de cet espace auront donc une faible variance mais un fort biais.

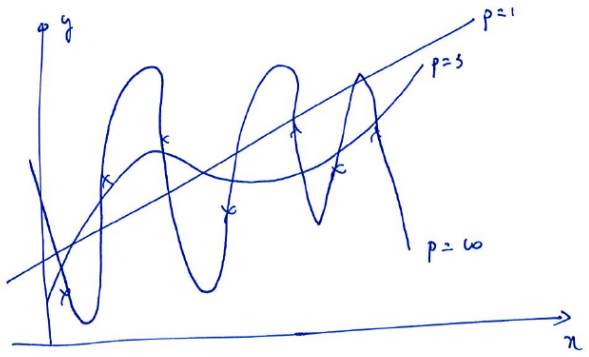

FIGURE 1.2 – Illustration dans le cas des espaces de polynomes. On cherche à estimer la fonction qui a généré les données à l'aide de polynomes de dégré p.

Lorsqu'on cherche à apprendre une fonction de prédiction, un point essentiel est de trouver le bon espace d'hypothèse qui permet d'avoir un bon compromis biais-variance.

1.2.3 Régulariser pour contrôler la compléxité

Une manière de controler le compromis biais-variance est donc de fixer de manière explicite la compléxité de l'espace d'hypothèse. Cela peut se faire par exemple en apprenant un modèle pour chaque espace de complexité croissante et en sélectionnant ensuite celui qui aboutit au plus faible erreur de validation ou de test.

Une autre façon de restreindre l'espace d'hypothèse est la notion de régularisation. Cette technique permet de contraindre la "taille" de l'espace d'hypothèse directement dans le problème de minimisation. Au lieu de minimiser le risque empirique dans l'espace $\mathcal{H}$, on cherche

$$\begin{array}{ll} \min_{f\in\mathcal{H}} & L_S(f) \\ & \|f\|_{\mathcal{H}}^2 \leq A \end{array}$$

De manière équivalente (pour les coûts qui nous intéressent), ce problème équivaut à

$$\min_{f\in\mathcal{H}} L_S(f) + \lambda\|f\|_{\mathcal{H}}^2$$

cette dernière formulation peut être étendu en un cadre général de risque empirique généralisé

$$\min_{f\in\mathcal{H}} L_S(f) + \lambda\Omega(f)$$

où λ est un paramètre de régularisation à définir et $\Omega(f)$ est un terme régularisant permettant de controler le compromis biais-variance et donc de garantir une bonne généralisation. Cette formulation englobe un nombre de problèmes d'apprentissage supervisé.

Exemple 1

- *régression linéaire*
  - *fonction :* $f(\mathbf{x}) = \mathbf{x}^\top\mathbf{w} + b$
  - *coût :* $\ell(y, f(\mathbf{x})) = (y - (\mathbf{x}^\top\mathbf{w} + b))^2$
  - *régularisation :* $\Omega(f) = \|\mathbf{w}\|^2$
- *SVM linéaire*
  - *fonction :* $f(\mathbf{x}) = x^\top\mathbf{w} + b$
  - *coût :* $\ell(y, f(\mathbf{x})) = \max(0, 1 - y(\mathbf{x}^\top\mathbf{w} + b))$
  - *régularisation :* $\Omega(f) = \|\mathbf{w}\|^2$
- *Réseau de neurones à une couche*
  - *fonction :* $f(\mathbf{x}) = \sigma(\mathbf{x}^\top\mathbf{w} + b)$
  - *coût :* $\ell(y, f(\mathbf{x})) = (y - (\mathbf{x}^\top\mathbf{w} + b))^2$
  - *régularisation :* $\Omega(f) = \|\mathbf{w}\|^2$

Au final, notre problème d'apprentissage supervisé se résume à un problème d'optimisation qui, étant donnée les données d'apprentissage, l'espace d'hypothèses et le terme régularisant $\Omega(f)$ et son paramètre λ nous renvoie les

paramètres d'une fonction f. Cette fonction de regression ou de décision nous permet ensuite de prédire les labels associés à un nouveau x.

1.3 Régularisation et parcimonie

Dans le cas de la régression linéaire au sens des moindres carrés, une autre manière de réduire la complexité de l'espace d'hypothèses consiste à utiliser moins de variables explicatives. En effet, dans ce cadre, on peut montrer que la variance $E((\hat{f}(\mathbf{x}) - F(\mathbf{x})^2)$ est proportionnel à $\frac{d}{N}$. Ainsi, réduire le nombre de variables explicatives permet donc de réduire la variance du biais. En pratique, il est donc interessant de faire de la sélection de variables surtout lorsque l'erreur décroit avec le nombre de variables explicatives. Cette selection des variables peut se faire à l'aide de différentes approches, y compris des méthodes exhaustives ou en choisissant un terme regularisant approprié.

Dans le cas linéaire, cadre dans lequel on se placera, on va définir l'espace d'hypothèses tel que :

$$f(x) = \mathbf{x}^\top \mathbf{w} + b \quad \mathbf{x} \in \mathbb{R}^d \text{ et } \mathbf{w} \in \mathbb{R}^d b\mathbb{R}$$

pour ces modèles, le terme regularisant le plus usité est le terme quadratique

$$\Omega(f) = \|\mathbf{w}\|^2$$

Dans ce cours, on s'interessera également à des termes régularisants qui induisent la parcimonie *ie* qui induit le fait que plusieurs composantes de $\mathbf{w}$ seront nulles et qui réalise donc de manière implicite de la sélection de variables.

La notion de parcimonie est importante dans plusieurs domaines scientifiques et se résume globalement à l'intuition suivante : "une explication simple d'un phénomène donné doit être préféré à une explication plus complexe". C'est le principe d'Occam. En apprentissage, ce principe se traduit par la notion de "sélection de variables" dont l'objectif est double :

— dans un modèle linéaire, la sélection de variable peut être réalisé en forçant des coordonnées de $\mathbf{w}$ à être à 0. Cela implique que ces variables ne seront pas utiles pour l'évaluation de $f(x)$. Le temps d'évaluation de $f(x)$ s'en retrouve également réduit.
— le fait d'avoir moins de variables impliquées dans la fonction d'évaluation fait également que le modèle devient plus interprétable. On sait quelles sont les variables qui sont pertinentes pour la généralisation.

Dans les modèles linéaires, la parcimonie peut être induit directement lors de la minimisation du risque empirique, en contraignant le cardinal du support de $\mathbf{w}$: $\|\mathbf{w}\|_0 = card(\{w_k : |w_k| \neq 0, k \in [1, \cdots, d]\})$ ce qui aboutit au problème suivant :

$$\begin{array}{ll} \min_f & L_S(f) \\ st & \|\mathbf{w}\|_0 \leq K \end{array} \qquad \text{ou} \ \min_f L_S(f) + \lambda \|\mathbf{w}\|_0$$

Cela dit, ce problème est un problème difficile car $\|\mathbf{w}\|_0$ est une fonction discontinue et non-convexe. Dans ce cours, on s'interessera à une relaxation convexe de la pseudo-norme $\|\mathbf{w}\|_0$ soit la norme $\|\mathbf{w}\|_1 = \sum_{k=1}^{d} |w_k|$ est convexe mais toujours non-différentiable.

Interprétation de la pénalité de type norme ℓ_1 Considèrons le problème suivant

$$\begin{array}{ll} \min_{\mathbf{w}} & \frac{1}{2}\|\mathbf{w} - \mathbf{w}_0\|^2 \\ st & \|\mathbf{w}\|_1 \leq T \end{array} \qquad \text{ou} \ \min_w \frac{1}{2}\|\mathbf{w} - \mathbf{w}_0\|^2 + \lambda \|\mathbf{w}\|_1$$

où les deux problèmes sont équivalents *i.e* $\forall \lambda, \exists T$ tels que les deux problèmes ont la même solution. A l'optimalité, le gradient de la fonction objective est orthogonale au courbe de niveau de la boule B définissant les contraintes Si

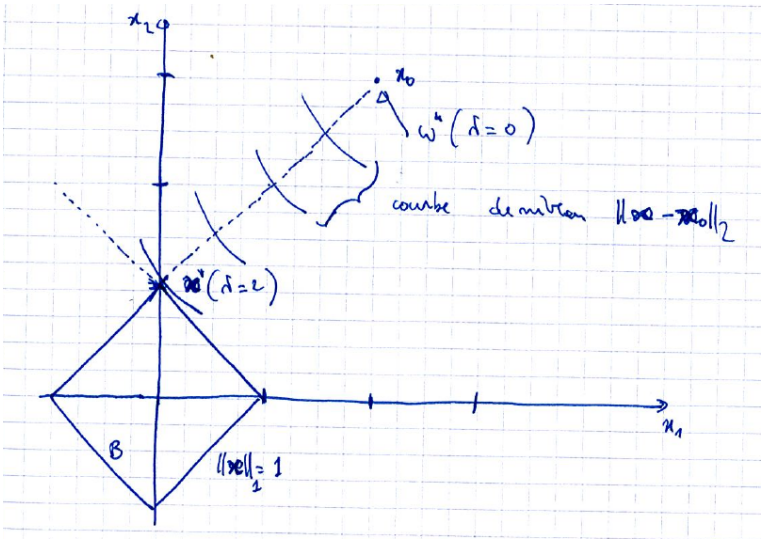

FIGURE 1.3 – Boule de pénalité ℓ_1 et courbe iso-coût. Le point que l'on cherche à approximer est $(2; 3)$ et la boule unité de la norme 1 contraint la solution

les courbes de niveau de la fonction cout sont tangents à ceux de B au point de singularité de B alors la solution est parcimonieuse.

Chapitre 2

Les outils d'optimisation

2.1 Introduction

Pour résoudre un problème d'apprentissage supervisé, on est souvent amené à résoudre un problème d'optimisation et l'algorithme d'apprentissage correspond en grande partie à l'algorithme permettant de résoudre le problème d'optimisation.

De manière générale, dans le cas linéaire, *ie* $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w} + b$, nous sommes amenés à résoudre un problème d'optimisation du type

$$\min_{\mathbf{w},b} L_S(\mathbf{w}, b) + \lambda \Omega(\mathbf{w})$$

où $L_S(\mathbf{w}, b)$ évalue le risque empirique d'attache aux données du couple $(\mathbf{w}, b)$, $\Omega(\mathbf{w})$ un terme régularisant et λ un paramètre de régularisation. Le cadre qui nous intéresse est celui où $L_S(\mathbf{w}, b)$ est convexe et différentiable tandis que $\Omega(\mathbf{w})$ est convexe pouvant être non-différentiable (c'est le cas par exemple pour la norme ℓ_1).

Dans ce qui suit, nous allons revenir succintement sur les outils nécessaires pour résoudre ce genre de problème d'optimisation et décrire un algorithme générique, relativement efficace, et simple à mettre en oeuvre : les méthodes proximales. Bien que pouvant être moins efficace que certaines méthodes spécifiques, la généricité des méthodes proximales en font des outils intéressants.

On note cependant que ce cours n'est pas un cours d'optimisation mais rappelle juste certaines notions importantes avec probablement moins de rigueur que nécessaire. Les lecteurs plus avides de détails mathématiques et de rigueur pourront faire référence aux livres appropriées.

2.2 Convexité

Un ensemble C est convexe si et seulement si pour tout $x, y \in C$ et $\alpha \in [0, 1]$

$$\alpha \mathbf{x} + (1 - \alpha)\mathbf{y} \in C$$

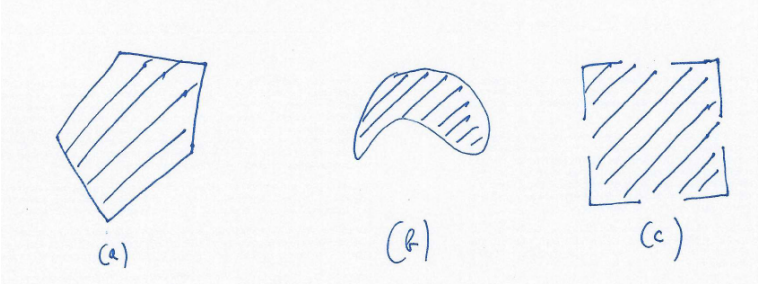

FIGURE 2.1 – Exemples d'ensemble convexe (a) et non-convexes (b) et (c)

Exemple 2 — *l'ensemble des solutions de* $\mathbf{Ax} = \mathbf{b}$
— *les boules de types* $\|\mathbf{x}\| \leq R$

■

Une fonction f est convexe si et seulement si son domaine est convexe et si pour tout $x, y \in \text{dom } f$ et $\alpha \in [0, 1]$

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y)$$

une fonction f est strictement convexe si l'inégalité est stricte.

Exemple 3

— *fonction affine* $f(x) = ax$
— e^x, $\log x$
— $|x|^\alpha, \alpha \geq 1$

■

Pour le cas des fonctions différentiables, on peut avoir d'autres définitions de la convexité.

— Une fonction f différentiable est convexe si et seulement si son domaine est convexe et si pour tout $x, y \in \text{dom } f$

$$f(y) \geq f(x) + \nabla f(x)^t (y - x)$$

ce qui signifie que les tangentes supportent la fonction f.

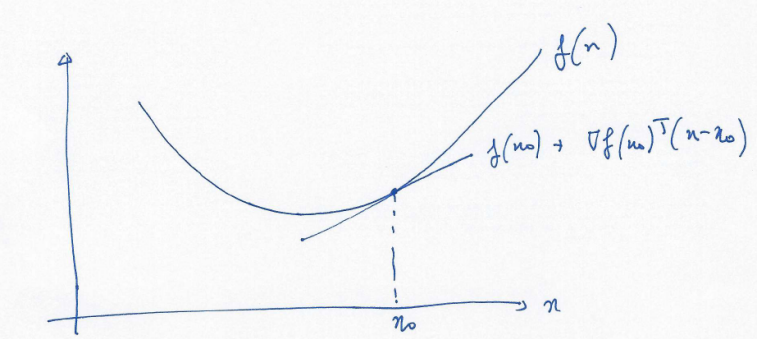

FIGURE 2.2 – Exemple de fonction convexe et sa tangente en un point

— Une fonction f deux fois différentiable est convexe si et seulement si son domaine est convexe et si pour tout $x \in \text{dom } f$

$$\nabla^2 f(x) \succeq 0$$

Parmi les opérations préservant la convexité des fonctions, on notera la somme positive et la composition affine.

2.3 Sous-gradient et non-différentiabilité

Dans ce qui suit, on supposera que les fonctions f sont des fonctions de $\mathbb{R}^d$ dans $\mathbb{R}$.

2.3.1 sous-gradient

Dans le cas des fonctions différentiables et convexes, on a la relation suivante de minoration d'une fonction

$$f(y) \geq f(x) + \nabla f(x)^t (y - x) \quad \forall x, y \in \text{dom } f$$

soit l'approximation d'ordre 1 en un point x donné donne une fonction minorante de $f(\cdot)$.

Cette notion peut être étendue aux cas des fonctions non-différentiables.

Definition 1
g *est le sous-gradient d'une fonction f (qui peut ne pas être convexe) en x si*

$$\forall y, f(y) \geq f(x) + \nabla \mathbf{g}^t (y - x)$$

Exemple 4

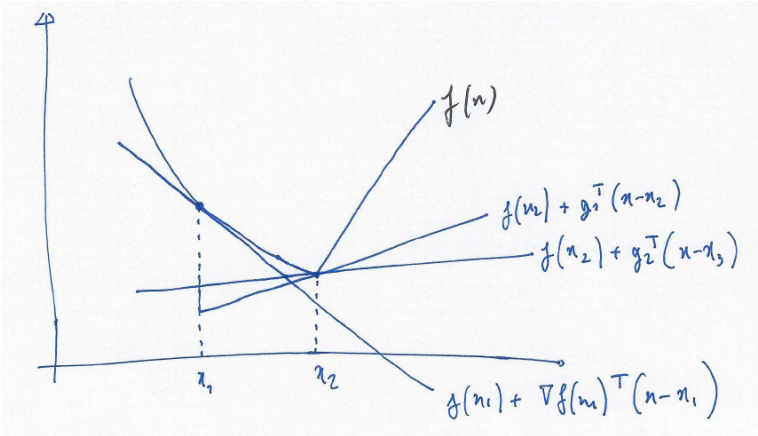

FIGURE 2.3 – fonction avec un point de non-singularité et ses tangentes en ce point

Les propriétés du sous-gradient sont :
- le sous-gradient définit une fonction affine minorante de f
- si f est convexe, il existe au moins un sous-gradient en tout point x à l'intérieur du domaine de f.
- si f est convexe et différentiable, $\nabla f(x)$ est l'unique sous-gradient de f en x.

L'interet pratique des sous-gradients résident dans leur utilité pour la résolution de problème d'optimisation de fonctions convexes et non-différentiables (par descente de sous-gradient) et dans le fait qu'ils permettent de définir les conditions d'optimalité pour ces problèmes non-différentiables.

2.3.2 Sous-différentielle

Definition 2 *On appelle sous-différentiel d'une fonction f en $\mathbf{x}$ l'ensemble des sous-gradients de f en $\mathbf{x}$. On le note $\partial f(\mathbf{x})$*

Quelques propriétés de la sous-différentielle :
- $\partial f(\mathbf{x})$ est un ensemble convexe.
- $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$ si f est différentiable en $\mathbf{x}$.
- si $\partial f(\mathbf{x}) = \{g\}$ et f différentiable en x alors $\nabla f(\mathbf{x}) = g$
- soit $h(\mathbf{x}) = \alpha_1 f_1(\mathbf{x}) + \alpha_2 f_2(\mathbf{x})$ avec α_1, $\alpha_2 \geq 0$ alors

$$\partial h(\mathbf{x}) = \alpha_1 \partial f_1(\mathbf{x}) + \alpha_2 \partial f_2(\mathbf{x})$$

Il est souvent difficile de calculer le sous-différentiel d'une fonction en un point mais bien souvent l'obtention d'un sous-gradient peut être aisé.

Exemple 5

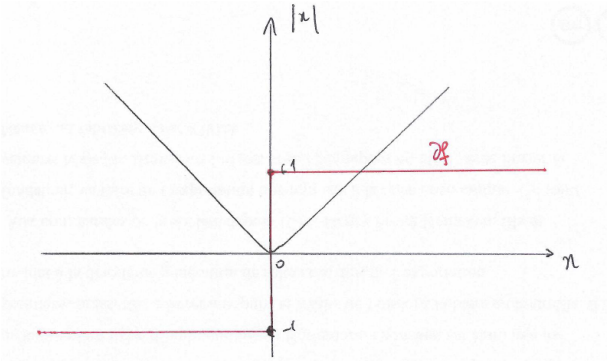

FIGURE 2.4 – $f(x) = |x|$ et ∂f

Optimisation et conditions d'optimalité

Soit $f(x)$ une fonction convexe que l'on cherche à minimiser. $\mathbf{x}^\star$ minimise $f(\mathbf{x})$ si et seulement si

$$0 \in \partial f(\mathbf{x}^\star)$$

Preuve 1

Exemple 6

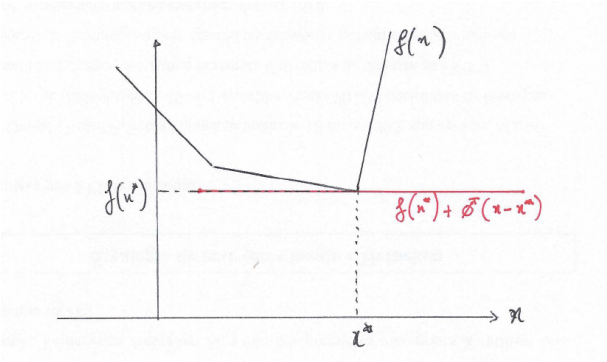

FIGURE 2.5 – Illustration d'une optimalité en un point non-différentiable.

2.3.3 Conjugué de Fenchel

Pour simplier le calcul d'un sous-gradient ou pour trouver le sous-gradient d'une fonction en un point, on peut se baser sur la notion de conjugué de Fenchel.

Definition 3
Pour une fonction f, on appelle conjugué de Fenchel de f la fonction $f^\star$ définie comme étant

$$\forall y, \quad f^\star(\mathbf{y}) = \sup_{\mathbf{x}}(\mathbf{x}^t\mathbf{y} - f(\mathbf{x}))$$

On peut remarquer que si $f(x)$ est une fonction convexe et dérivable alors à l'optimum le point $\mathbf{x}'$ satisfait $y - \nabla f(\mathbf{x}') = 0$ soit $\nabla f(\mathbf{x}') = y$. Ainsi, le maximiseur est obtenue pour un point $\mathbf{x}'$ dont la dérivée est égale à $\mathbf{y}$

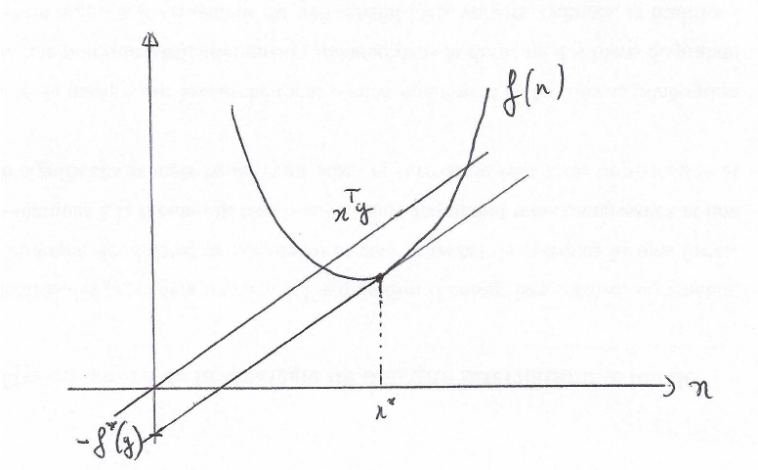

FIGURE 2.6 – Exemple illustrant le conjugué de Fenchel pour une fonction convexe et différentiable.

De même, si $f(x)$ est convexe et non-différentiable alors la condition d'optimalité du conjugué de Fenchel donne, le maximiseur $\mathbf{x}'$ satisfait $y \in \partial f(\mathbf{x}')$ soit

$$\mathbf{y} \in \partial f(\mathbf{x}')$$

ainsi, $\mathbf{x}'$ est un point $\mathbf{x}$ dont $\mathbf{y}$ est sous-gradient.

Exemple 7
- $f(x) = e^x$ *avec* $x \in \mathbb{R}$
- $f(x) = |x|$ *avec* $x \in \mathbb{R}$
- $f(x) = ax + b$ *avec* $x \in \mathbb{R}$

Une inégalité importante permet de faire le lien entre la notion de sous-différentiel et la fonction conjuguée. L'inégalite de Fenchel-Young nous dit que

$$f(\mathbf{x}) + f^\star(\mathbf{y}) \geq \mathbf{x}^t\mathbf{y} \qquad \forall \mathbf{x}, \mathbf{y}$$

On a une égalité si et seulement si $\mathbf{y} \in \partial f(\mathbf{x})$.

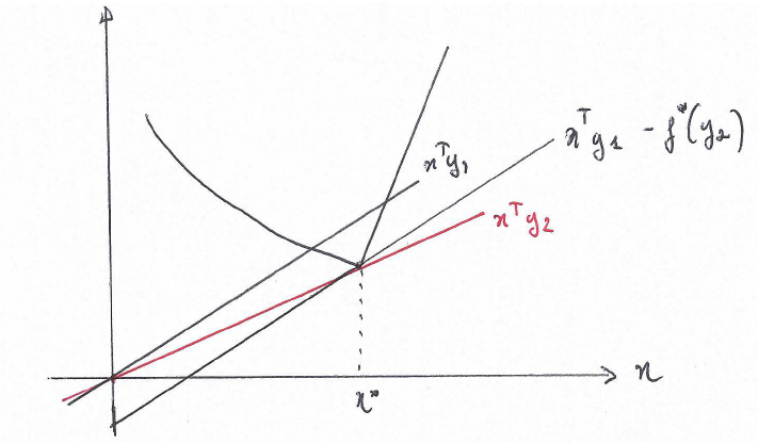

FIGURE 2.7 – Exemple illustrant le conjugué de Fenchel pour une fonction non-différentiable en un point.

Preuve 2

Jusqu'ici, le concept de sous-gradient nous donne une condition d'optimalité d'un problème d'optimisation non-différentiable. On sait relier le concept de sous-différentiel aux fonctions conjugués pour caractériser un sous-gradient.

2.3.4 Sous-gradient et sous-différentiel de normes

Dans notre problème d'apprentissage, ce qui nous pose problème est bien souvent le terme régularisant qui est défini, dans la plupart des cas, comme étant une norme et est non-différentiable. Ce qui nous intéresse est donc de connaitre le sous-différentiel d'une norme $\|\mathbf{w}\|_p$ afin de pouvoir caractériser les conditions d'optimalités d'un problème d'apprentissage avec un coût convexe et différentiable et une pénalité non-différentiable.

Dans cette section, on cherche à caractériser la notion de sous-différentiel d'une norme $\partial\|\mathbf{w}\|_p$. Pour aboutir à ce résultat, on se propose de calculer la fonction conjuguée d'une norme et d'utiliser l'inégalité de Fenchel-Young.

Avant tout, nous définissons la notion de norme duale.

Definition 4
Soit $\|\cdot\|$ une norme définie dans un espace euclidien, on appelle norme duale $\|\cdot\|_\star$ la fonction qui à $\mathbf{y}$ associe :

$$\|\mathbf{y}\|_\star = \sup_{\|\mathbf{x}\|\leq 1} \mathbf{y}^t\mathbf{x}$$

Pour les normes $\|\mathbf{x}\|_p$ de $\mathbb{R}^d$ tels que $\|\mathbf{x}\|_p = \left(\sum_{k=1}^{d} |x_k|^p\right)^{\frac{1}{p}}$ avec $p \in [1, \infty[$, la norme duale est $\|\mathbf{x}\|_q$ avec $\frac{1}{p} + \frac{1}{q} = 1$.

Preuve 3

On est maintenant en mesure de calculer la fonction conjuguée d'une norme $f(x) = \|x\|_p$.

Definition 5
La fonction conjuguée de la norme $\|\mathbf{x}\|_p$ est

$$f^\star(\mathbf{y}) = \sup_{\mathbf{x}}(\mathbf{x}^t\mathbf{y} - \|\mathbf{x}\|_p) = \begin{cases} 0 & si\ \|\mathbf{y}\|_q \leq 1 \\ \infty & sinon \end{cases}$$

On notera que la fonction conjuguée d'une norme n'est pas la norme duale mais l'indicatrice de cette norme duale $I_{\|\mathbf{y}\|_q \leq 1}$.

Preuve 4

Exemple 8 *Exemple de normes duales*

— *pour* $\|\mathbf{x}\|_1$

$$f^\star_{\|\mathbf{x}\|_1}(\mathbf{y}) = \begin{cases} 0 & si\ \|\mathbf{y}\|_\infty \leq 1 \\ \infty & sinon \end{cases}$$

— *pour* $\|\mathbf{x}\|_2$

$$f^\star_{\|\mathbf{x}\|_2}(\mathbf{y}) = \begin{cases} 0 & si\ \|\mathbf{y}\|_2 \leq 1 \\ \infty & sinon \end{cases}$$

Nous avons maintenant tout les outils pour obtenir l'expression de la sous-différentielle d'une norme. En effet, l'inégalité de Fenchel-Young nous dit que

$$f(\mathbf{x}) + f^\star(\mathbf{y}) = \mathbf{x}^t\mathbf{y} \Leftrightarrow \mathbf{y} \in \partial f(\mathbf{x})$$

Dans le cas des normes, cela signifie que

$$\begin{aligned} \mathbf{y} \in \partial f(\mathbf{x}) &\Leftrightarrow \|\mathbf{x}\|_p + f^\star_{\|\mathbf{x}\|_p}(y) = \mathbf{x}^t\mathbf{y} && (2.1) \\ &\Leftrightarrow \|\mathbf{x}\|_p + 1_{\|\mathbf{y}\|_q \leq 1} = \mathbf{x}^t\mathbf{y} && (2.2) \\ & && (2.3) \end{aligned}$$

soit, l'égalite suivante doit être respectée

$$\mathbf{x}^t\mathbf{y} - \|\mathbf{x}\|_p = 1_{\|\mathbf{y}\|_q \leq 1}$$

Pour avoir une égalité finie, cette équation n'est valable que si $\|\mathbf{y}\|_q \leq 1$, ainsi :

— on remarque que si $\mathbf{x} = 0$, tout vecteur $\mathbf{y}$ tel que $\|\mathbf{y}\|_q \leq 1$ satisfait l'égalité

— si $\mathbf{x} \neq 0$, on doit avoir $\mathbf{x}^t\mathbf{y} = \|\mathbf{x}\|_p$.

soit

$$\partial\|\mathbf{x}\|_p = \begin{cases} y : \|\mathbf{y}\|_q \leq 1 & \text{si } \mathbf{x} = 0 \\ y : \|\mathbf{y}\|_q \leq 1 \text{ et } \mathbf{x}^t\mathbf{y} = \|\mathbf{x}\|_p & \text{si } \mathbf{x} \neq 0 \end{cases}$$

Quant à l'égalité $\mathbf{x}^t\mathbf{y} = \|\mathbf{x}\|_p$, pour p tel que $1 < p < \infty$, pour la satisfaire, on peut choisir

$$y_k = |x_k|^{p-1}\frac{\text{sign}(x_k)}{\|x_k\|_p^{p-1}}$$

par exemple, pour $p = 2$, on a

$$y_k = |x_k|\frac{\text{sign}(x_k)}{\|x_k\|_2} = \frac{x_k}{\|x_k\|_2}$$

Dans le cas de $p = 1$, on cherche à satisfaire la condition $\mathbf{x}^t\mathbf{y} = \sum_k |x_k|$ qui peut être satisfait par un vecteur $\mathbf{y}$ tel que

$$y_k = \begin{cases} \text{sign}(x_k) & \text{si } x_k \neq 0 \\ [-1, 1] & \text{si } x_k = 0 \end{cases}$$

Grâce à l'ensemble de ces outils, on est maintenant en mesure de caractériser les conditions d'optimalités d'un problème d'optimisation convexe et non-différentiable, utile en apprentissage statistique pour résoudre les problèmes de sélection de variables.

2.3.5 Exercices d'application

Résoudre les problèmes suivants

1. $$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

2. $$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_2$$

2.4 Dualité lagrangienne

Dans certains problèmes d'apprentissage supervisé, le problème d'optimisation s'écrit comme un problème d'optimisation sous constraintes. Par exemple, dans les SVMs, on cherche à maximiser une marge sous les contraintes que l'ensemble des points d'apprentissage soient bien classés.
Dans cette section, on rappelle le principe de la dualité Lagrangienne qui permet d'établir les conditions d'optimalités d'un problème d'optimisation sous contraintes convexes.

2.4.1 Le Lagrangien

Soit le problème d'optimisation générique suivant

$$\begin{array}{ll} \min_{\mathbf{x}\in\mathbb{R}^d} & f(\mathbf{x}) \\ & g_i(\mathbf{x}) \leq 0 \quad i = 1, \cdots, n \\ & h_i(\mathbf{x}) = 0 \quad i = 1, \cdots, m \end{array}$$

A partir de ce problème, il est possible de construire une fonction que l'on appelle le Lagrangien qui s'écrit

$$\mathcal{L}(\mathbf{x}, \alpha, \beta) = f(\mathbf{x}) + \sum_i \alpha_i g_i(\mathbf{x}) + \sum_i \beta_i h_i(\mathbf{x})$$

Les variables $\alpha \geq 0$ et β sont appellés les variables duales en opposition à $\mathbf{x}$, la variable primale et sont également appelés les multiplicateurs de Lagrange.

Le Lagrangien peut être interprété comme étant une version "modifiée" du problème d'optimisation originale qui tient compte des contraintes. En effet, on va considérer α_i et β_i comme étant les "coûts" que l'on paye lorsque les contraintes sont violées. Par ailleurs, quelque soit le problème d'optimisation considérée, il existe des valeurs de α_i et β_i telles que le minimum du Lagrangien par rapport aux variables primales coincide avec la solution du problème originale.

2.4.2 Problème primal

Pour comprendre le lien entre le Lagrangien et le problème original, on va introduire la notion de problème primal et problème dual liée au Lagrangien.

Le problème primal est défini comme étant

$$\min_{\mathbf{x}} \underbrace{\max_{\alpha\geq 0,\beta} \mathcal{L}(\mathbf{x}, \alpha, \beta)}_{\text{fonction objective primale}} = \underbrace{\min_{\mathbf{x}} \theta_p(\mathbf{x})}_{\text{problème primal}} = p^\star$$

on dit que $\mathbf{x}$ est un point primal faisable si $\forall i, g_i(\mathbf{x}) \leq 0$ et $\forall i\, h_i(\mathbf{x}) = 0$. On appelle $\mathbf{x}^\star$ la solution de ce problème primal.

On peut ré-interpréter la fonction objective primale de la façon suivante. Tout d'abord, on note que $\theta_p(\mathbf{x})$ est une fonction convexe en $\mathbf{x}$ et on peut le re-écrire :

$$\begin{aligned} \theta_p(\mathbf{x}) &= \max_{\alpha \geq 0, \beta} \mathcal{L}(\mathbf{x}, \alpha, \beta) && (2.4) \\ &= \max_{\alpha \geq 0, \beta} f(\mathbf{x}) + \sum_i \alpha_i g_i(\mathbf{x}) + \sum_i \beta_i h_i(\mathbf{x}) && (2.5) \\ &= f(\mathbf{x}) + \max_{\alpha \geq 0, \beta} [\sum_i \alpha_i g_i(\mathbf{x}) + \sum_i \beta_i h_i(\mathbf{x})] && (2.6) \end{aligned}$$

Analysons le deuxième terme de $\theta_p(\mathbf{x})$:
— si $\mathbf{g}_i(\mathbf{x}) > 0$ alors on peut choisir un $\alpha_i > 0$ tel que $\alpha_i g_i(x) \longrightarrow \infty$.
— si $\mathbf{g}_i(\mathbf{x}) \leq 0$, le α_i qui maximise $\alpha_i g_i(x)$ est $\alpha_i = 0$
— si $h_i(\mathbf{x}) \neq 0$ alors on peut choisir un $\beta_i \neq 0$ tel que $\beta_i h_i(x) \longrightarrow \infty$
— si $h_i(\mathbf{x}) = 0$, on a $\beta_i h_i(x) = 0$ quelque soit β_i.

donc si $\mathbf{x}$ est primal faisable, la valeur optimale du max dans $\theta_p(\mathbf{x})$ est 0 sinon $\theta_p(\mathbf{x}) \longrightarrow \infty$. Ainsi, on a :

$$\theta_p(\mathbf{x}) = f(x) + \begin{cases} 0 & \text{si } \mathbf{x} \text{ est primal faisable} \\ \infty & \text{sinon} \end{cases}$$

$\theta_p(\mathbf{x})$ est donc une version amélioré de $f(\mathbf{x})$ dans la mesure où cette fonction prend en compte les contraintes car si $\mathbf{x}$ n'est pas primal faisable $\theta_p(\mathbf{x})$ prend une valeur ∞. On a un effet "barrière". On note également que l'optimum du problème original coincide avec l'optimal du problème primal.

2.4.3 Le problème dual

En intervertissant le min et le max du problème primal, on obtient un problème différent (D)

$$\max_{\alpha \geq 0, \beta} \underbrace{\min_x \mathcal{L}(\mathbf{x}, \alpha, \beta)}_{\text{fonction objective duale}} = \underbrace{\max_{\alpha \geq 0, \beta} \theta_D(\mathbf{x})}_{\text{problème dual}} = d^\star$$

On dit que α et β sont dual faisable si $\alpha \geq 0$ et on note $\alpha^\star$,$\beta^\star$ le couple solution de (D). On peut également re-interpréter ce problème dual.

On note d'abord que $\theta_D(\alpha, \beta)$ est une fonction concave en ses paramètres (car elle est affine en fonction de α et β) et que

$$\theta_D(\alpha, \beta) \leq p^\star$$

.

Preuve 5

Ce que nous dit cette propriété est que $\theta_D(\alpha, \beta)$ est une fonction dont les valeurs minorent $p^\star$. Comme le problème dual maximise $\theta_D(\alpha, \beta)$, on peut dire que le problème duale cherche le plus grand minorant de $p^\star$. On appelle cette propriété la dualité faible :

$$d^\star \leq p^\star$$

Pour montrer cette propriété, il suffit de prendre $\alpha^\star$ et $\beta^\star$.

Pour certaines situations, on a un résultat plus fort que l'on appelle la dualité forte. En effet, pour tout paire de problème primal/dual satisfaisant certaines contraintes de qualification, on a

$$d^\star = p^\star$$

Un exemple classique de contrainte de qualification que l'on considère souvent sont les conditions de Slater. On dit d'un problème qu'il satisfait les conditions de Slater si il existe un $\mathbf{x}$ pour lequel $g_i(\mathbf{x}) < 0, \forall i$.

Une conséquence intéressante de la dualité forte est ce qu'on appelle les conditions de complémentarité :

En cas de dualité forte on a alors $\alpha_i^\star \mathbf{g}_i(\mathbf{x}^\star) = 0, \forall i$

Preuve 6

L'ensemble des inégalités sont en fait des égalités car $p^\star = d^\star$. Donc, $\sum_i \underbrace{\alpha_i^\star g_i(\mathbf{x}^\star)}_{\leq 0} + \sum_i \underbrace{\beta_i^\star h_i(\mathbf{x}^\star)}_{=0} = 0$. Or comme la première somme est une somme de termes négatives dont la somme fait 0, cela implique que individuellement $\alpha_i^\star g_i(\mathbf{x}^\star) = 0, \forall i$. On peut également déduire de ces égalités que $\mathbf{x}^\star$ minimise $\mathcal{L}(x, \alpha^\star, \beta^\star)$.

2.4.4 Conditions d'optimalité KKT

Dans le cas convexe, si on a la dualité forte alors le triplet $\mathbf{x}^\star, \alpha^\star, \beta^\star$ sont optimales pour le problème primal et duale si et seulement si

$$\begin{aligned}
g_i(\mathbf{x}^\star) &\leq 0 \quad \forall i && (2.7)\\
h_i(\mathbf{x}^\star) &= 0 \quad \forall\, i && (2.8)\\
\alpha_i^\star &\geq 0 \quad \forall i && (2.9)\\
\alpha_i^\star g_i(\mathbf{x}^\star) &= 0 \quad \forall i && (2.10)\\
\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^\star, \alpha^\star, \beta^\star) &= 0 && (2.11)
\end{aligned}$$

Les conditions de KKT caractérisent la solution du problème d'optimisation à travers les solutions du problème primal et dual. On peut résoudre le problème en trouvant une solution aux conditions KKT ou en résolvant le problème primal/dual et en vérifiant que toutes les conditions sont vérifiées.

Exemple 9 *SVM séparable*

$$\begin{array}{ll} \min_{\mathbf{x}} & \frac{1}{2}\|\mathbf{w}\|^2 \\ & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \forall i \end{array}$$

Exemple 10 *Projection sur le cône positif*

$$\begin{array}{ll} \min_{\mathbf{x}} & \frac{1}{2}\|\mathbf{x} - \mathbf{v}\|^2 \\ & x_k \geq 0 \quad \forall k \end{array}$$

2.4.5 Exercices d'application

1. $$\begin{array}{ll} \min_{\mathbf{x}} & \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \\ \text{avec} & \|\mathbf{x}\|_2 \leq \tau \end{array}$$

2. Projection sur un espace affine

$$\begin{array}{ll} \min_{\mathbf{x}} & \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \\ \text{avec} & \mathbf{A}\mathbf{x} = \mathbf{b} \end{array}$$

Chapitre 3

Apprentissage parcimonieux

3.1 Introduction

On s'interesse à un problème d'apprentissage où les données $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ sont telles que $\mathbf{x} \in \mathbb{R}^d$ avec d typiquement grand par rapport à n et $\mathbf{y}_i \in \mathbb{R}$. On est donc dans le cas d'un problème de regression et on cherche à apprendre la relation de dépendance existant entre les $\mathbf{x}$ et les $\mathbf{y}$.

On se limite à un cadre linéaire et on cherche donc une fonction $f(\mathbf{x})$ telle que

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + b$$

Pour simplifier le problème, on suppose que la matrice des données $\mathbf{X} \in \mathcal{M}_{n,d}$ est centrée et réduite *i.e* :

$$\sum_i X_{i,j} = 0 \text{ et } \sum_i X_{i,j}^2 = 1$$

on fixe $b = \frac{1}{n}\sum_i y_i$ et on remplace $y_i \leftarrow y_i - \frac{1}{n}\sum_i y_i$. Ce pré-traitement des données simplifie le problème en éliminant le biais b, ce qui nous permet de focaliser seulement sur la recherche d'une fonction $f(x) = \mathbf{x}^T\mathbf{w}$.

3.2 Moindres carrées et régression ridge

Usuellement, un problème de régression se résoud en cherchant la fonction de regression qui minimise un coût quadratique. Dans le cas de fonctions linéaires, on cherche donc $f(\mathbf{x}) = \mathbf{x}^T\mathbf{w}$ tel que

$$\min_{\mathbf{w}} J(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

La solution $\mathbf{w}^\star$ minimisant ce problème s'obtient tout simplement en annulant la dérivée car $J(\mathbf{w})$ est une fonction convexe et différentiable. Cela

donne :

$$\begin{aligned} \nabla_{\mathbf{w}} J &= -\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) = 0 \\ &\Leftrightarrow \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\mathbf{w} = 0 \end{aligned}$$

soit

$$\mathbf{w}^\star = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{y})$$

On notera que cette solution optimale et unique s'obtient sous réserve que la matrice $(\mathbf{X}^T\mathbf{X})$ soit inversible. Cependant, si $d >> n$, cette condition n'est plus verifiée car la matrice $(\mathbf{X}^T\mathbf{X})$ n'est pas de rang pleine.

Pour rendre le problème stable et lui assuré une solution unique, une solution est de régulariser le problème en forçant la norme $\|\mathbf{w}\|_2$ à être faible. le problème de moindre carrées régularisées est communément appelé ridge régression :

$$\min_{\mathbf{w}} J(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

De même manière que précedemment, la résolution de ce problème peut s'obtenir analytiquement.

$$\begin{aligned} \nabla_{\mathbf{w}} J &= -\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda\mathbf{w} = 0 \\ &\Leftrightarrow \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda\mathbf{w} = 0 \end{aligned}$$

soit

$$\mathbf{w}^\star = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}(\mathbf{X}^t\mathbf{y})$$

Dans ce dernier cas, le fait de rajouter un terme λ sur la diagonale de $\mathbf{X}^T\mathbf{X}$ permet d'assurer que la matrice $(\mathbf{X}^T\mathbf{X} + \lambda I)$ est inversible et donc que le problème admet une solution unique. Cependant, cette approche peut présenter plusieurs inconvénients notamment losrque d est grand :

- on doit inverser une matrice de taille $d \times d$ ou résoudre un système linéaire de même taille. En pratique, cela implique un calcul de compléxité d^3, ce qui est vite prohibitif pour d grand.

- La solution $\mathbf{w}^\star$ est dense, c'est à dire que toutes ou une grande majorité des coordonnées de ce vecteur sont non-nulles Une conséquence de cette densité est que l'ensemble des variables sont impliquées dans la fonction de régression. Le modèle est donc difficement interprétable.

3.3 Régression Lasso

Afin de pallier aux inconvénients de la régression ridge lorsque la dimension du problème devient grand ($d >> n$), il peut être intéressant de faire de

la sélection de variables, tout en apprenant le modèle et son vecteur de paramètres $\mathbf{w}$. Dans ce cas, on cherche à résoudre le problème d'optimisation suivant :

$$\min_{\mathbf{w}} J_L(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1$$

Idéalement, pour induire une sélection de variables, on aurait considéré une pénalité de type ℓ_0, du genre $\|\mathbf{w}\|_0$, mais celle-ci est non-convexe et non-différentiable ce qui rend l'ensemble du problème difficile à optimiser.

3.3.1 Conditions d'optimalité

Pour résoudre ce problème et pour obtenir ses conditions d'optimalités, dans la mesure où la $\|\mathbf{w}\|_1$ n'est pas différentiable en 0, ces conditions nécessitent donc l'utilisation des sous-gradients. On rappelle que le sous-différentiel $\partial\|\mathbf{w}\|_1$ s'écrit :

$$\partial\|\mathbf{w}\|_1 = \begin{cases} \mathbf{g} : \|\mathbf{g}\|_\infty \leq 1 & \text{si } \mathbf{w} = 0 \\ \mathbf{g} : \|\mathbf{g}\|_\infty \leq 1 \text{ et } \mathbf{g}^T\mathbf{w} = \|\mathbf{w}\|_1 & \text{si } \mathbf{w} \neq 0 \end{cases}$$

Pour le deuxième cas, on prendra $\mathbf{g}$ tel que $g_k = \text{sign}(w_k)$ si $w_k \neq 0$ ou $g_k \in [-1, 1]$ sinon.

Dans le cas général, on écrit le coût d'attache aux données comme $L_S(\mathbf{w})$, et les conditions d'optimalités s'écrivent donc :

$$\nabla_{\mathbf{w}} L_S(\mathbf{w}^\star) + \lambda\mathbf{g} = 0 \quad \text{avec } \mathbf{g} \in \partial\|\mathbf{w}^\star\|_1$$

Ainsi, un vecteur $\mathbf{w}^\star$ est optimal si et seulement si

$$\begin{array}{lll} \text{sign } (w_k^\star) \neq 0 & \Rightarrow & \nabla_{\mathbf{w}} L_S|_k + \lambda\text{sign}(w_k^\star) = 0 \\ \text{sign } (w_k^\star) = 0 & \Rightarrow & |\nabla_{\mathbf{w}} L_S|_k| \leq \lambda \end{array}$$

Dans le cas d'un coût de type moindres carrés utile dans le cadre des problèmes de regression, ces conditions d'optimalités deviennent :

$$\begin{array}{lll} \text{sign } (w_k) \neq 0 & \Rightarrow & -\mathbf{X}_{\cdot,k}^T(\mathbf{y} - X\mathbf{w}^\star) + \lambda\text{sign}(w_k^\star) = 0 \\ \text{sign } (w_k) = 0 & \Rightarrow & |\mathbf{X}_{\cdot,k}^T(\mathbf{y} - X\mathbf{w}^\star)| \leq \lambda \end{array}$$

3.3.2 Interprétation dans le cas de la régression Lasso

On rappelle que dans le cadre qui nous intéresse, les vecteurs de variables explicative $\|\mathbf{X}_{\cdot,k}\|$ sont normalisées donc, on peut interpréter

$$\mathbf{X}_{\cdot,k}^T(\mathbf{y} - \mathbf{X}\mathbf{w}^\star) = \|\mathbf{X}_{\cdot,k}\|\|\mathbf{y} - \mathbf{X}\mathbf{w}^\star\|\cos(\theta)$$

où θ est l'angle entre la variable explicative et le vecteur des résidus à l'optimalité. A l'optimalité, pour les composantes $\mathbf{w}^\star$ nulles, cela signifie que

$\|\mathbf{y}\|\cos(\theta)$ est plus faible que θ. Ainsi, lorsque λ est faible, il y aura plus de variables impliquées dans le modèle.

A l'extrême, la solution $\mathbf{w}^\star$ est le vecteur nul si pour tout k, $|\mathbf{X}_{\cdot,k}^T\mathbf{y}|$ est plus faible que λ *i.e* la corrélation entre $\mathbf{y}$ et toutes les variables explicatives sont plus faibles que λ.

Ainsi, si on pose λ_m tel que $\lambda_m = \max_k(|\mathbf{X}_{\cdot,k}^T\mathbf{y}|)$ et on prend $\lambda = \lambda_m - \epsilon$ avec $\epsilon > 0$, les conditions d'optimalité nous disent alors que

$$\begin{array}{lll} \text{sign } (w_k^\star) \neq 0 & \Rightarrow & -\mathbf{X}_{\cdot,k}^T(\mathbf{y} - X\mathbf{w}^\star) + \lambda\text{sign}(w_k^\star) = 0 \\ \text{sign } (w_k^\star) = 0 & \Rightarrow & |\mathbf{X}_{\cdot,k}^T(\mathbf{y} - X\mathbf{w}^\star)| \leq \lambda \end{array}$$

pour $\lambda = \lambda_m$, il ne peut y avoir qu'une solution possible avec une variable active possible(pour lequel w_k est potentiellement non-nulle). Cette variable est la variable qui maximise $|\mathbf{X}_{\cdot,k}^T\mathbf{y}|$. Au fur et à mesure que λ_m décroit, plusieurs variables peuvent devenir actives sous réserve que les conditions d'optimalité soient respectées. On notera que les variables actives sont equi-corrélés au résidu $\mathbf{y} - \mathbf{X}\mathbf{w}^\star$.

On peut maintenant se poser la question de comment évolue le vecteur $\mathbf{w}^\star$ lorsque λ change et que l'ensemble des variables actives restent identiques. Pour ces variables actives $\mathcal{A}$ et pour deux différentes valeurs de λ, λ_1 et $\lambda_1 + \epsilon$, les conditions d'optimalités sont :

$$\mathbf{X}_{\mathcal{A}}^T(\mathbf{y} - \mathbf{X}\mathbf{w}_1^\star) = \lambda_1 \quad \text{et} \quad \mathbf{X}_{\mathcal{A}}^T(\mathbf{y} - \mathbf{X}\mathbf{w}_2^\star) = \lambda_1 + \epsilon$$

si on soustrait les deux équations, on obtient : soit $\mathbf{w}_2^\star = -(\mathbf{X}_{\mathcal{A}}^T\mathbf{X})^{-1}\epsilon + \mathbf{w}_1^\star$. On en déduit donc que les coefficients associées aux variables actives évoluent linéairement en fonction du paramètre de régularisation λ.

Une autre propriété importante de la régression Lasso provient du nombre maximal de variables non-nulles dans le vecteur $\mathbf{w}^\star$. En effet, la régression Lasso intégrera au plus n variables actives. Il existe une preuve théorique de cette affirmation que l'on ne développera pas, et qui se base sur le théorème de Carathéodory. L'intuition que l'on peut avoir de cette propriété est que parmi les d, $d > n$ variables explicatives, sauf cas totalement dégénéré, on a n variables qui forme une famille libre de $\mathbb{R}^n$, et qui peut donc approximé arbitrairement bien tout vecteur de $\mathbb{R}^n$.

Cette propriété de parcimonie est clairement un avantage, mais peut posé problème lorsque le déséquilibre entre d et n (dans une problème de sélection de gènes dans les puces biologiques, on a souvent $p \approx 10000$ et $n \approx 100$).

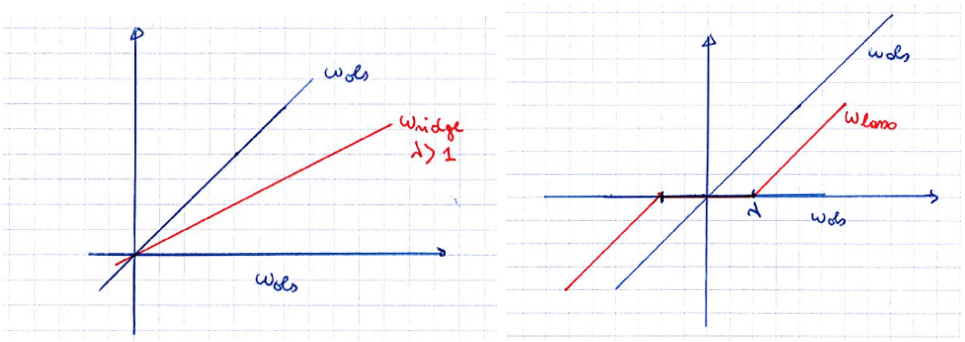

FIGURE 3.1 – Géometrie des solutions par rapport à la solution des moindres de carrés dans le cas d'une matrice $\mathbf{X}$ orthogonale. (gauche) regression ridge. (droite) regression Lasso.

3.3.3 Géométrie de la solution dans le cas orthogonal

Pour une régularisation de type moindres carrées, la solution dans le cas orthogonal peut s'écrire

$$\begin{aligned} \mathbf{w}^\star_{ridge} &= (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}(\mathbf{X}^t\mathbf{y}) \\ &= \tfrac{1}{\lambda}\mathbf{X}^T\mathbf{y} \\ &= \tfrac{1}{\lambda}\mathbf{w}_{ols} \end{aligned}$$

pour une régularisation $\|\mathbf{w}\|_1$, on a pour les solutions non-nulles

$$\begin{aligned} -X^T_{\cdot,k}(\mathbf{y} - \mathbf{X}\mathbf{w}^\star) + \lambda \text{sign}(w^\star_k) &= 0 \\ -w_{ols,k} + w^\star_k + \lambda \text{sign}(w^\star_k) &= 0 \end{aligned}$$

or comme $w^\star_k = |w^\star_k|\text{sign}(w^\star_k)$, on a donc $w_{ols,k} = w^\star_k + \lambda\text{sign}(w^\star_k) = \text{sign}(w^\star_k)(|w^\star_k| + \lambda)$, ce qui signifie que $w_{ols,k}$ et $w^\star_k$ sont de même signe. On peut donc réécrire

$$\begin{aligned} w^\star_k &= w_{ols,k} - \lambda\text{sign}(w_{ols,k}) \\ &= \text{sign}(w_{ols,k})(|w_{ols,k}| - \lambda) \end{aligned}$$

La solution de la regression Lasso consiste donc en un rétrecissement de la solution dans le cas moindres carrés. Pour les solutions $w^\star_k$ nulles, la condition d'optimalité devient

$$\begin{aligned} |-\mathbf{X}^T_{\cdot,k}(\mathbf{y} - \mathbf{X}\mathbf{w}^\star)| &\leq \lambda \\ |\mathbf{X}^T_{\cdot,k}\mathbf{y}| &\leq \lambda \end{aligned}$$

3.3.4 Implémentation descente de coordonnées

Le problème du Lasso peut se résoudre de différentes manières, notamment en le mettant sous la forme d'un problème quadratique (QP). Cependant, il peut se résoudre sans une machinerie compliquée en utilisant une

méthode de type descente de coordonnées. Cette approche consiste à optimiser un problème par bloc de coordonnées ou coordonnées par coordonnées, comme nous allons le faire ici. Dans le cas du Lasso,on peut démontrer que cette algorithme converge vers la solution du problème.

Pour une coordonnée k donnée, le problème du Lasso optimisé que sur cette variable k peut se ré-écrire :

$$\begin{array}{ll} & \min_{w_k} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1 \\ \Leftrightarrow & \min_{w_k} \frac{1}{2}\sum_i (y_i - \sum_j X_{i,j} w_j)^2 + \lambda|w_k| \\ \Leftrightarrow & \min_{w_k} \frac{1}{2}\sum_i (y_i - \sum_{j\neq k} X_{i,j} w_j - X_{i,k} w_k)^2 + \lambda|w_k| \\ \Leftrightarrow & \min_{w_k} \frac{1}{2}\sum_i (s_i - X_{i,k} w_k)^2 + \lambda|w_k| \\ \Leftrightarrow & \min_{w_k} \frac{1}{2}\|\mathbf{s} - \mathbf{X}_{\cdot,k} w_k\|^2 + \lambda|w_k| \end{array}$$

avec $\mathbf{s}$ le vecteur de coordonnées $s_i = y_i - \sum_{j\neq k} X_{i,j} w_j$. Le dernier problème est un problème de régression Lasso à une seule variable dont il est facile de montrer, étant donnée les cours précédents que

$$w_k^\star = \text{sign}(\mathbf{X}_{\cdot,k}^T \mathbf{s})(|\mathbf{X}_{\cdot,k}^T \mathbf{s}| - \lambda)_+$$

Le principe de l'algorithme par descente de coordonnées consiste donc à appliquer cette équation itérativement à toutes les coordonnées en recalculant $\mathbf{s}$ à chaque fois.

3.4 Régression Elastic net

Comme nous l'avons souligné, dans un premier cas, utiliser un terme de régularisation de type $\|\mathbf{w}\|_2^2$ a tendance à induire une solution $\mathbf{w}^\star$ dense, donc à tendance à construire un modèle qui utilise toutes les variables. Dans un deuxième cas, la régularisation $\|w^\star\|_1$ sélectionne des variables mais si $d >> n$ alors le nombre maximal de variables sélectionnées est de n.

Un autre désavantage du Lasso est que l'approche ne permet pas de faire de la sélection par groupe : si il existe un groupe de variables à forte corrélation, alors le Lasso aura tendance à n'en sélectionner qu'un seul parmi le groupe : la sélection par groupe pouvant être utiles dans certains problèmes spécifiques (sélection de gènes).

Pour résoudre ces deux problèmes, on peut réaliser un compromis entre ces deux termes régularisants en utilisant une régularisation mixte que l'on appelle la régression Elastic-net.

3.4.1 Cadre

On s'interesse donc au problème à régularisation mixte suivant :

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \frac{\lambda_2}{2}\|\mathbf{w}\|_2^2 + \lambda_1\|\mathbf{w}\|_1$$

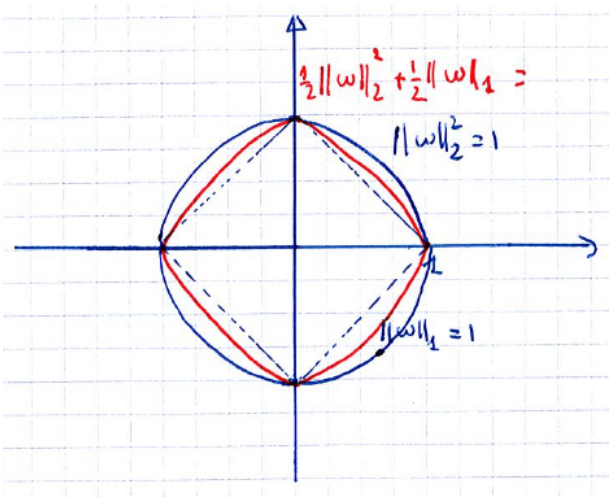

FIGURE 3.2 – Isocontour de la pénalité Elastic Net

qui généralise la régression Lasso ($\lambda_2 = 0$) et la régression ridge ($\lambda_1 = 0$).

Si on développe la fonction objective $J(\mathbf{w})$, on obtient

$$J(\mathbf{w}) = \frac{1}{2}\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} + \frac{1}{2}\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} + \frac{\lambda_2}{2}\mathbf{w}^T\mathbf{w} + \lambda_1\|\mathbf{w}\|_1$$

et donc on peut ré-écrire le problème comme

$$\frac{1}{2}\left\|\begin{pmatrix}\mathbf{y}\\0\end{pmatrix} - \begin{pmatrix}\mathbf{X}\\\sqrt{\frac{\lambda_2}{2}}I_d\end{pmatrix}\mathbf{w}\right\|_2^2 + \lambda_1\|\mathbf{w}\|_1$$

où I_d est une matrice identité de taille $d \times d$. On peut donc constater que le problème de régression Elastic Net se réduit à une régression Lasso avec un nombre d'observations augmentés : en effet, on a dans ce problème, toujours d variables mais $n + d$ observations. Dans ce contexte, la régression Lasso peut donc potentiellement sélectionner les d variables.

3.4.2 Géométrie de la régularisation mixte et de la solution quand $\mathbf{X}^T\mathbf{X} = I_d$

Afin d'avoir une idée sur la façon dont agit la régularisation mixte de type Elastic Net, on peut s'interesser à la forme de la boule induite par le terme de régularisation et à la forme de la solution dans un cas orthogonal.

Etudions maintenant la forme de la solution dans le cas orthogonal. En dérivant la fonction objective $J(\mathbf{w})$ et en simplifiant car $\mathbf{X}^T\mathbf{X} = I_p$, on a, à l'optimalité,

$$-\mathbf{X}^T\mathbf{y} + \mathbf{w}^\star + \lambda_2\mathbf{w}^\star + \lambda_1\mathbf{g} = 0$$

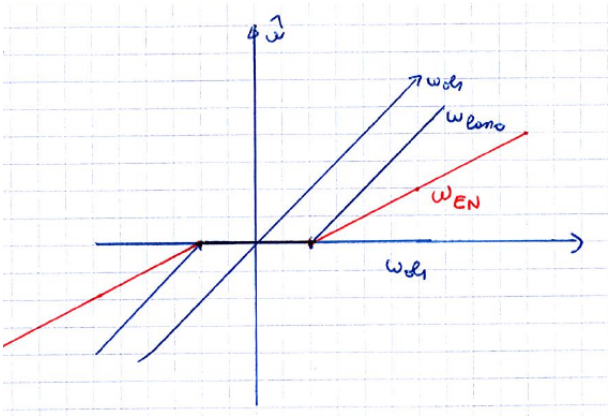

FIGURE 3.3 – Forme du retrécissement pour la pénalité Elastic Net dans le cas orthogonal.

avec $\mathbf{g} \in \partial\|\mathbf{w}^\star\|_1$. Et donc, on a

$$\mathbf{w}^\star(1+\lambda_2) + \lambda_1\mathbf{g} - \mathbf{w}_{ols} = 0$$

En tenant le même raisonnement que pour la régression Lasso, on obtient

$$\mathbf{w}^\star = \frac{1}{1+\lambda_2}\text{sign}(\mathbf{w}_{ols})(|\mathbf{w}_{ols}| - \lambda_1)_+$$

3.4.3 Propriété

En introduisant la régression Elastic Net, nous avons dit que cette approche permettait de sélectionner des variables corrélées. La véracité de cette assertion est justifiée par la proposition suivante :

Theorème 1 *Soit des données $\{\mathbf{x}_i, y_i\}_i$ des données telles que la matrice $\mathbf{X}$ soit centrée et normalisée et soit $\hat{\mathbf{w}}$, la solution du problème de régression Elastic Net, on définit*

$$D(j,k) = \frac{1}{\|\mathbf{y}\|_2}|\hat{w}_j - \hat{w}_k|$$

Si $\hat{w}_j \cdot \hat{w}_k > 0$ alors, on a

$$D(j,k) \leq \frac{1}{\lambda_2}\sqrt{2(1-\rho)}$$

où $\rho = \mathbf{X}_{\cdot,j}^T\mathbf{X}_{\cdot,k}$ (on notera que $|\rho| \leq 1$).

Preuve : Tout d'abord, $\hat{w}_j \cdot \hat{w}_k > 0$ signifie que $\hat{w}_j$ et $\hat{w}_k$ sont non-nuls et de meme signe. Par ailleurs, pour ces deux composantes, la condition d'optimalité nous dit que

$$\begin{aligned} -X_{.,j}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda_2 w_j + \lambda_1 \text{sign}(w_j) = 0 \\ -X_{.,k}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda_2 w_k + \lambda_1 \text{sign}(w_k) = 0 \end{aligned}$$

En soustrayant les deux équations, on obtient :

$$(X_{.,k} - X_{.,j})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda_2(w_j - w_k) = 0$$

Soit, à l'optimalité, on a donc l'expression suivante qui est vérifiée :

$$\hat{w}_j - \hat{w}_k = \frac{1}{\lambda_2}(X_{.,j} - X_{.,k})^T r$$

r étant le vecteur de résidus. Par ailleurs, comme $\hat{\mathbf{w}}$ est optimal, il est associé au cout minimal et donc

$$J(\hat{\mathbf{w}}) \leq J(0)$$

soit

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|_2^2 + \underbrace{\lambda_2\|\hat{\mathbf{w}}\|^2 + \lambda_1\|\hat{\mathbf{w}}\|_1}_{\geq 0} \leq \frac{1}{2}\|\mathbf{y}\|^2$$

et $\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|_2^2 \leq \|\mathbf{y}\|^2$. A partir des relations ci-dessus, on a

$$\begin{array}{rcl} |\hat{w}_j - \hat{w}_k| & = & \frac{1}{\lambda_2}|(X_{.,j} - X_{.,k})^T r| \\ & \leq & \frac{1}{\lambda_2}\|(X_{.,j} - X_{.,k})\|\|r\| \\ \frac{1}{\|\mathbf{y}\|}|\hat{w}_j - \hat{w}_k| & \leq & \frac{1}{\lambda_2}\|(X_{.,j} - X_{.,k})\|\frac{\|r\|}{\|\mathbf{y}\|} \\ \frac{1}{\|\mathbf{y}\|}|\hat{w}_j - \hat{w}_k| & \leq & \frac{1}{\lambda_2}\|(X_{.,j} - X_{.,k})\| \end{array}$$

or $\|(X_{.,j} - X_{.,k})\|^2 = \|X_{.,j}^T X_{.,j}\|^2 + \|X_{.,k}^T X_{.,k}\|^2 - 2X_{.,j}^T X_{.,k} = 2 - 2\rho$ ce qui permet de conclure.

On peut intepréter ce théorème de la façon suivante. En fonction des paramètres λ_1 et λ_2, $D_{j,k}$ décrit la différence entre $\hat{w}_j$ et $\hat{w}_k$ pour deux variables d'indices j et k données. Si les variables associées sont très corréles, le théoreme nous garantit que $\hat{w}_j$ et $\hat{w}_k$ auront des valeurs très similaires à l'optimalité. On notera qu'un théorème similaire peut être obtenu dans le cadre de la classification *i.e* dans le cas des problèmes suivants :

$$\min_{\mathbf{w}} \sum_i \phi(y_i w^T X_{.,i}) + \lambda_2\|\mathbf{w}\|^2 + \|\mathbf{w}\|_1$$

sous certaines conditions de régularités de la fonction ϕ.

3.4.4 Elastic Net renormalisé

Dans certains cas pratiques, on remarque que les performances de régressions ne sont pas satisfaisantes pour la régression Elastic Net. Cela est dû à la double régularisation introduit par le Lasso et la régression Ridge. Pour atténuer ce problème, on peut mettre en oeuvre l'heuristique suivante et définir une solution Elastic Net renormalisé

$$\hat{\mathbf{w}}_{EN-N} = (1+\lambda_2)\hat{\mathbf{w}}_{EN}$$

où $\mathbf{w}_{EN}$ est la solution de la régression Elastic-Net. Une justification de cette approche peut être :

- la multiplication par $1+\lambda_2$ permet d'annuler l'atténuation due à la régression ridge dans le cas orthogonale
- on peut trouver un problème d'optimisation de type Regression Lasso dont la solution est $\mathbf{w}_{EN-N}$. Ce problème correspond à une regression Lasso où la matrice $\mathbf{X}^T\mathbf{X}$ aurait été stabilisé :

$$\mathbf{X}^T\mathbf{X} \leftarrow \frac{\mathbf{X}^T\mathbf{X} + \lambda_2/2 I_p}{1+\lambda_2}$$

car $\mathbf{w}_{EN-N}$ serait solution de

$$\min \frac{1}{2} \left\| \underbrace{\begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix}}_{\mathbf{y}^\star} - \underbrace{\begin{pmatrix} \mathbf{X} \\ \sqrt{\frac{\lambda_2}{2}} I_p \end{pmatrix}}_{\mathbf{X}_\star} \frac{\mathbf{w}}{1+\lambda_2} \right\|_2^2 + \frac{\lambda_1}{1+\lambda_2}\|\mathbf{w}\|_1$$

obtenu en remplacant $\mathbf{w}$ par $\frac{\mathbf{w}}{1+\lambda_2}$ dans le problème de la régression Lasso. En développant la fonction objective, on aurait $\mathbf{y}^{\star T}\mathbf{y} = \mathbf{y}^T\mathbf{y}$, $\mathbf{X}_\star^T\mathbf{X} = \mathbf{X}^T\mathbf{X} + \frac{\lambda_2}{2} I_p$, et $\mathbf{y}^{\star T}\mathbf{X}_\star = \mathbf{y}^T\mathbf{X}$, donc :

$$\min_{\mathbf{w}} \frac{1}{1+\lambda_2}[\frac{1}{2}\mathbf{w}^T \frac{\mathbf{X}^T\mathbf{X} + \lambda_2/2 I_p}{1+\lambda_2}\mathbf{w} - \mathbf{y}^T\mathbf{X}\mathbf{w} + \lambda_1\|\mathbf{w}\|_1]$$

ce qui correspond à un Lasso avec une matrice $\mathbf{X}^T\mathbf{X}$ spéciale puisque le problème de la régression Lasso est

$$\min_{\mathbf{w}}[\frac{1}{2}\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{y}^T\mathbf{X}\mathbf{w} + \lambda_1\|\mathbf{w}\|_1]$$

Chapitre 4

Algorithme de gradient proximale

4.1 Introduction

Dans le chapitre précédent, nous avons introduit un cadre méthodologique pour l'apprentissage parcimonieux utilisant des termes de régularisation spécifique de type norme $\|\mathbf{w}\|_1$ ou de type mixte mélangeant la norme ℓ_1 et la norme ℓ_2. Pour une fonction de coût de type "coût quadratique" $\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$ et les termes de régularisations pré-citées, on peut construire des algorithmes dédiées de manière relativement simples. Cependant, ces approches ne s'étendent pas aisément à d'autres fonctions de coût ou à d'autres types de termes régularisants.

Les algorithmes proximales que nous présentons dans ce chapitre permet de contourner ces difficultés. Ce sont des méthodes qui sont génériques et qui s'adaptent aisément à une large classe de problème d'apprentissage statistique.

4.2 Opérateur proximal

On appelle opérateur proximal d'une fonction convexe $\Omega(\cdot)$ l'opérateur qui a un vecteur $\mathbf{x} \in \mathbb{R}^d$ donné renvoie $\text{prox}_\Omega(\mathbf{x}) \in \mathbb{R}^d$ tel que

$$\text{prox}_\Omega(\mathbf{x}) = \arg\min_{\mathbf{u}} \frac{1}{2}\|\mathbf{x} - \mathbf{u}\|_2^2 + \Omega(\mathbf{u})$$

Des exemples d'opérateurs proximals sont
— Pour $\Omega(\cdot) = 0$, l'opérateur $\text{prox}_\Omega(\mathbf{x}) = \mathbf{x}$
— Pour $\Omega(\cdot) = I_C(\cdot)$, où I_c est l'indicateur d'un ensemble convexe C *i.e*, il vaut 0 si son argument appartient à l'ensemble C et l'infini

sinon. Dans ce cas, l'operateur proximal $\text{prox}_\Omega(\mathbf{x})$ correspond à la projection orthogonale de $\mathbf{x}$ sur l'ensemble C.

$$\text{prox}_\Omega(\mathbf{x}) = \arg\min_{\mathbf{u}\in C} \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|_2^2$$

— Pour $\Omega(\cdot) = \|\cdot\|_2^2$, on peut obtenir l'expression analytique de l'opérateur

$$\text{prox}_\Omega(\mathbf{x}) = \arg\min_{\mathbf{u}} \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|_2^2 + \|\mathbf{u}\|_2^2 = \frac{1}{3}\mathbf{x}$$

— Pour $\Omega(\cdot) = \|\cdot\|_1$, on peut procéder de la même manière pour obtenir la solution analytique de

$$\text{prox}_\Omega(\mathbf{x}) = \arg\min_{\mathbf{u}} \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|_2^2 + \|\mathbf{u}\|_1$$

qui correspond à un seuillage.

4.2.1 Propriétés

On remarque tout d'abord que si $\Omega(\cdot)$ est convexe alors l'ensemble du problème est strictement convexe et admet donc une solution unique.

On peut également caractériser la solution du problème lié à l'opérateur proximal. En effet, on peut noter $\mathbf{u}^\star = \text{prox}_\Omega(\mathbf{x})$ le minimiseur du problème

$$\min_{\mathbf{u}} \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|_2^2 + \Omega(\mathbf{u})$$

ainsi, de par la strict convexité du problème, on sait que la solution $\mathbf{u}^\star = \text{prox}_\Omega(\mathbf{x})$ satisfait à la condition d'optimalité suivante :

$$-\mathbf{x} + \mathbf{u}^\star + \partial\Omega(\mathbf{u}^\star) = 0 \quad \Leftrightarrow \mathbf{x} - \mathbf{u}^\star \in \partial\Omega(\mathbf{u}^\star)$$

où $\partial\Omega(\mathbf{u}^\star)$ désigne le sous-différentiel de Ω en $\mathbf{u}^\star$. De cette propriété, on peut donc également déduire que si on a $\mathbf{u}^\star = \text{prox}_\Omega(\mathbf{u}^\star)$ alors $\mathbf{u}^\star$ est un point minimisant la fonction $\Omega(\cdot)$. Si on peut montrer la propriété suivante :

$$\|\text{prox}_\Omega(\mathbf{y}) - \text{prox}_\Omega(\mathbf{x})\| \leq \|\mathbf{y}-\mathbf{x}\| \qquad \forall \mathbf{y}, \mathbf{x}$$

alors l'algorithme itératif $\mathbf{x}_{k+1} = \text{prox}_\Omega(\mathbf{x}_k)$ converge vers ce minimum pour un $\mathbf{x}_0$ donné, or cette propriété est vrai pour tout Ω.

4.3 Algorithme de gradient proximal

L'algorithme de gradient proximal a pour but de résoudre un problème d'optimisation de la forme

$$\min_{\mathbf{w}} f(\mathbf{w}) + \lambda\Omega(\mathbf{w})$$

où $f(\mathbf{x})$ est une fonction convexe et différentiable et $\Omega(\mathbf{x})$ une fonction convexe. On note que beaucoup de problème d'apprentissage statistique rentre dans ce cadre. Par exemple :

— regression linéaire

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

— SVM linéaire avec un cout Hinge au carré

$$\min_{\mathbf{w}} \frac{1}{2}\sum_i \max(0, 1 - \mathbf{y}_i\mathbf{x}_i^\top\mathbf{w})^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

4.3.1 Point fixe

L'algorithme de gradient proximal peut de construire comme un algorithme de point fixe comme décrit précédemment pour la minimisation de $\Omega(\cdot)$. En effet, on peut affirmer que

Theorème 2 $\mathbf{w}^\star$ *est solution de*

$$\min_{\mathbf{w}} f(\mathbf{w}) + \lambda\Omega(\mathbf{w})$$

si et seulement si

$$\mathbf{w}^\star = prox_{\nu\lambda\Omega}(\mathbf{w}^\star - \nu\nabla f(\mathbf{w}^\star)) \qquad \forall\nu > 0$$

La preuve s'écrit comme suit :

L'algorithme proximal peut donc se déduire d'après cette équation de point fixe. C'est un algorithme itératif qui s'écrit comme suit à l'itération k.

$$\mathbf{w}_{k+1} = \text{prox}_{\alpha_k\lambda\Omega}(\mathbf{w}_k - \alpha_k\nabla f(\mathbf{w}_k))$$

il se compose donc à chaque itération d'un pas de gradient par rapport à ∇f suivi d'une "projection" par l'opérateur proximal de $\alpha_k\lambda\Omega$. α_k est le pas de descent, qui peut être constant ou déterminé à chaque itération suivant une recherche de pas optimal.

4.4 Ré-interprétation de l'algorithme de gradient proximal

L'algorithme présenté ci-dessus peut se ré-intepréter comme un algorithme minimisant $f(\mathbf{w}) + \lambda\Omega(\mathbf{w})$ par approximation-minimisation successive de la fonction objective. Cette interprétation s'obtient en définissant la fonction $Q(\mathbf{v}, \mathbf{w})$ suivant, pour une constante $L > 0$ donnée :

$$Q_L(\mathbf{v}, \mathbf{w}) = f(\mathbf{w}) + \nabla f(\mathbf{w})^\top (\mathbf{v} - \mathbf{w}) + \frac{L}{2}\|\mathbf{v} - \mathbf{w}\|_2^2 + \lambda\Omega(\mathbf{w})$$

On note que pour un $\mathbf{w}_k$ donné, $Q_L(\mathbf{v}, \mathbf{w}_k)$, les deux premiers termes du coté droit de l'égalité forme une approximation linéaire de $f(\mathbf{w})$ en $\mathbf{w}_k$ et que les 3 premiers termes forment donc une approximation quadratique.

Maintenant, si on définit

$$\mathbf{v}^\star = \arg\min_{\mathbf{v}} Q(\mathbf{v}, \mathbf{w})$$

alors on peut montrer que

$$\begin{aligned} \mathbf{v}^\star &= \arg\min_v \tfrac{L}{2}\|\mathbf{v} - (\mathbf{w} - \tfrac{1}{L}\nabla f(\mathbf{w}))\|^2 + \lambda\Omega(\mathbf{v}) \\ &= \operatorname{prox}_{\frac{\lambda}{L}\Omega}(\mathbf{w} - \tfrac{1}{L}\nabla f(\mathbf{w})) \end{aligned}$$

La preuve de cette propriété s'obtient très simplement en développant les termes de la fonction objective $Q(\mathbf{v}, \mathbf{w})$

Cette ré-écriture permet donc d'interpréter un pas de gradient proximal comme étant la minimisation d'une approximation quadratique de la fonction objective $f(\mathbf{w}) + \lambda\Omega(\mathbf{w})$: l'approximation étant une approximation linéaire de la partie différentiable de la fonction objective et le terme

$$\frac{L}{2}\|\mathbf{v} - \mathbf{w}\|^2$$

permet de controler la proximité (d'où le nom proximal) du minimum de l'approximation.

4.4.1 Le rôle de L

Lorqu'on analyse les deux interprétations de l'algorithme de gradient proximal, on constate que L ou ν contrôle la longueur du pas de descente avant l' opérateur proximal. Son choix peut donc être prépondérant, surtout si il est fixe.

En théorie, afin de garantir la convergence de l'algorithme, les algorithmes de gradients proximales sont appliquées à des cas où la fonction

$f(\mathbf{w})$ en outre d'être différentiable est une fonction à gradient Lipschitzienne de constante L_f, c'est à dire que

$$\forall \mathbf{x}, \mathbf{x}' \qquad \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\|_2 \leq L_f \|\mathbf{x}' - \mathbf{x}\|$$

Intuitivement, cette propriété signifie que la fonction f est une fonction relativement régulière et est à variation lente : pour deux points proches, les gradients de la fonction en ces points ne peuvent pas être très "différents".

Par ailleurs, pour les fonctions à gradient Lipschitzienne, la propriété suivante s'applique également :

$$\forall \mathbf{x}, \mathbf{x}' \quad \text{et } L \geq L_f \qquad f(\mathbf{x}') \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{x}' - \mathbf{x}) + \frac{L}{2}\|\mathbf{x}' - \mathbf{x}\|_2^2$$

et de par cette propriété, on constate que l'approximation quadratique Q_L de la fonction $f(\cdot)$ est alors une approximation majorante $i.e$

$$f(\mathbf{v}) + \lambda\Omega(\mathbf{v}) \leq f(\mathbf{w}) + \nabla f(\mathbf{v})^T(\mathbf{v} - \mathbf{w}) + \frac{L}{2}\|\mathbf{v} - \mathbf{w}\|_2^2 + \lambda\Omega(\mathbf{v})$$

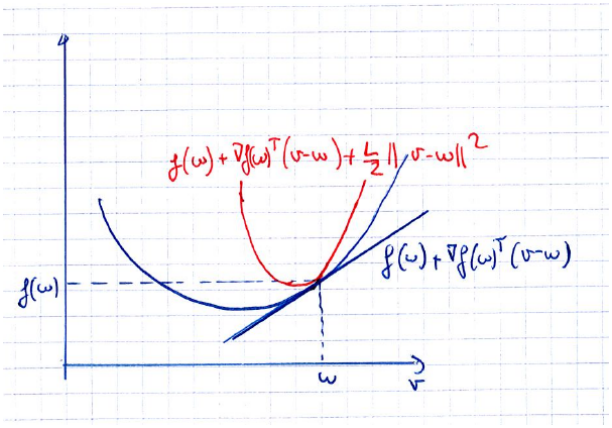

FIGURE 4.1 – Illustration de l'approximation linéaire et l'approximation quadratique.

Cette dernière nous garantit par ailleurs une décroissance de la valeur objective du problème d'optimisation originale à chaque itération de l'algorithme de gradient proximale. En effet, on a

$$f(\mathbf{w}_k) + \lambda\Omega(\mathbf{w}_k) = Q_L(\mathbf{w}_k, \mathbf{w}_k) \geq Q(\mathbf{w}_{k+1}, \mathbf{w}_k) \geq f(\mathbf{w}_{k+1}) + \lambda\Omega(\mathbf{w}_{k+1})$$

où la première inégalité s'obtient grâce à la minimisation et la deuxième par la propriété de majoration de l'approximation.

En pratique, pour bien choisir L, il faut donc connaître L_f. Pour cela, il faut soit démontrer l'existence de ce L_f tel que

$$\forall \mathbf{x}, \mathbf{y} \qquad \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L_f \|\mathbf{y} - \mathbf{x}\|$$

où de manière équivalente, si la fonction f est deux fois différentiable

$$\|\nabla^2 f\|_2 \leq L_f$$

soit la plus grande valeur propre de la Hessienne de f doit majorée L_f.

On peut s'affranchir de ce calcul algorithmiquement en choisissant le pas optimal L à chaque itération de sorte que

$$Q(\mathbf{w}_{k+1}, \mathbf{w}_k) \geq f(\mathbf{w}_{k+1}) + \lambda\Omega(\mathbf{w}_{k+1})$$

4.5 Exercices d'application

1. Montrer que les opérateurs proximaux $t\Omega(\cdot)$ pour les fonctions suivantes :
   (a) $\Omega(\mathbf{x}) = \|\mathbf{x}\|_2$
   (b) $\Omega(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} + b^T\mathbf{x} + c$
   (c) $\Omega(\mathbf{x}) = -\sum_{i=1}^{d} \log x_i$
   sont
   (a) $\text{prox}_{t\Omega}(\mathbf{x}) = \begin{cases} (1 - \frac{t}{\|\mathbf{x}\|_2})\mathbf{x} & \|\mathbf{x}\|_2 \geq t \\ 0 & \|\mathbf{x}\|_2 \leq t \end{cases}$
   (b) $\text{prox}_{t\Omega}(\mathbf{x}) = (I + tA)^{-1}(\mathbf{x} - tb)$
   (c) $\text{prox}_{t\Omega}(\mathbf{x})_i = \frac{x_i + \sqrt{x_i^2 + 4t}}{2}$

2. Calculer l'opérateur proximal associé à la régularisation Elastic-Net :
$$\Omega(\mathbf{x}) = \lambda\|\mathbf{x}\|_1 + \alpha\|\mathbf{x}\|_2^2$$

3. Fonction séparable. Montrer que si $\Omega(\mathbf{x}_1, \mathbf{x}_2)$ (où $x = [x_1^T \; x_2^T]^T$ peut s'écrire comme $h_1(\mathbf{x}_1) + h_2(\mathbf{x}_2)$ alors $\text{prox}_h(\mathbf{x}) = [\text{prox}_{h_1}(x_1)^T \; \text{prox}_{h_2}(x_2)^T]^T$

4. Montrer que si $\Omega(\mathbf{x}) = f(\lambda\mathbf{x} + a)$ alors
$$\text{prox}_\Omega(\mathbf{x}) = \frac{1}{\lambda}(\text{prox}_{\lambda^2 f}(\lambda\mathbf{x} + a) - a)$$

5. Décomposition de Moreau : Soit une fonction Ω telle que $\Omega^{**} = \Omega$. Montrer que pour cette fonction, on a la propriété suivante
   (a) $\mathbf{y} \in \partial\Omega(\mathbf{x})$ si et seulement si $\mathbf{x} \in \partial\Omega^*(\mathbf{y})$. Indication : Montrer tout d'abord que, pour tout f convexe, si $\hat{\mathbf{x}}$ maximise $\mathbf{x}^T\hat{\mathbf{y}} - f(\mathbf{x})$ alors $\hat{\mathbf{x}} \in \partial f^\star(\hat{\mathbf{y}})$.
   (b) $\mathbf{x} = \text{prox}_\Omega(\mathbf{x}) + \text{prox}_{\Omega^\star}(\mathbf{x})$
   (c) pour $t > 0$, $\text{prox}_{t\Omega^\star}(\mathbf{x}) = \mathbf{x} - t\text{prox}_{\Omega/t}(\frac{\mathbf{x}}{t})$

Chapitre 5

Apprentissage de dictionnaires

5.1 Introduction

Un problème de regression vise à approximer un vecteur (représentant lui-même un signal ou une image) comme étant une combinaison linéaire de fonctions de bases, typiquement en résolvant un problème de la forme suivante

$$\min_{\mathbf{w}} \|\mathbf{s} - \mathbf{X}\mathbf{w}\|_2^2$$

$\mathbf{s}$ étant le vecteur à approximer et $\mathbf{X}$ les fonctions de base.

Dans le contexte de la représentation d'un signal ou une image, ce problème est un problème à la fois très ancien mais toujours d'actualité. En effet, la transformée de Fourier visait déjà à cet objectif de représentation de signaux, puisque l'équation

$$s(n) = \sum_{k=0}^{N-1} X_k e^{i2\pi \frac{k}{N} n}$$

décrit un signal $s(n)$ comme étant une combinaison linéaire de signaux complexes à fréquences pré-définies. De la même manière, la transformée en ondelettes cherche à représenter un signal comme étant une combinaison linéaire de signaux localisée en temps et en fréquence :

$$s(n) = \sum_{j,k} a_{j,k} \Psi_{j,k}(n)$$

où $\Psi_{j,k}$ sont les fonctions en ondelettes.

Idéalement, on cherche à représenter un signal sous la forme la plus compacte possible afin d'en extraire une information vraisemblablement pertinente. Cependant, des dictionnaires (les fonctions sur laquelle on cherche à représenter un signal) comme la base canonique de $\mathbb{R}^n$ ou les fonctions de Fourier permettent de représenter de manière compactes qu'une certaines classes de signaux (respectivement les signaux composés d'impulsions ou les signaux sinusoidaux). Typiquement, ces deux familles de dictionnaires ne peuvent pas représenter de manière compacte un signal de type

$$s(n) = \delta_k(n) + \cos(\frac{2\pi k}{N} n)$$

La question que l'on peut se poser alors, et auquelle on s'interessera ici et est il possible d'apprendre, à partir de données, le dictionnaire sur laquelle représenter un ensemble de signaux de manière compacte.

5.2 Formalisation du problème

Soit un ensemble de M signaux $\{\mathbf{x}_i\}_{i=1}^M$ où chaque signal est représentée sous la forme d'un vecteur de $\mathbb{R}^d$. On cherche à représenter chacun de ces signaux comme étant la combinaison linéaire des élements d'un dictionnaire $\mathbf{D} = \{\mathbf{d}_j\}_{j=1}^N$. On rappelle que $\mathbf{D}$ forme une base si le cardinal de $\mathbf{D}$ est égal à M et que les élements (également appelé atomes) de $\mathbf{D}$ forment une famille libre. $\mathbf{D}$ est un dictionnaire redondant si $\mathbf{D}$ admet plus de M éléments et un sous-ensemble de $\mathbf{D}$ forme une famille libre de $\mathbb{R}^N$. Dans la suite, on note $\mathbf{D}$ la matrice formée par les élements du dictionnaires concaténée en colonnes et $\mathbf{X}$ celle formée par les signaux construites de la même manière.

Maintenant, supposons tout d'abord que le dictionnaire est fixe, le problème d'approximation des signaux $\{\mathbf{x}_i\}$ sur les éléments du dictionnaire $\mathbf{D}$ s'écrit alors

$$\min_{\mathbf{a}_i} \sum_{i=1}^{M} \|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2^2 + \lambda\Omega_1(\mathbf{a}_i)$$

où $\Omega(\cdot)$ est un terme de régularisation induisant la parcimonie dans la représentation de chaque signal : typiquement $\Omega_1(\mathbf{a}_i) = \sum_{j=1}^N |a_{j,i}|$. Ce problème peut être traité signal par signal de manière indépendante mais on peut également l'écrire de manière plus compacte sous forme matricielle

$$\min_{\mathbf{A}} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \lambda\Omega(\mathbf{A})$$

car $\|\mathbf{X}\|_F^2 = \sum_{i,j} x_{i,j}^2 = tr(\mathbf{X}^T\mathbf{X})$, $(\mathbf{D}\mathbf{A})_{i,j} = (\mathbf{D}\mathbf{a}_i)_j$ et $\Omega(\mathbf{A}) = \sum_{j,i} |a_{j,i}|$.

Maintenant, si on cherche également à apprendre le dictionnaire, le problème devient donc

$$\min_{\mathbf{A},\mathbf{D}\in C} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \lambda\Omega(\mathbf{A})$$

où C définit des contraintes de norme unitaire sur chaque atome du dictionnaire. Ces contraintes permettent de rendre le problème mieux posé car si elles n'existaient pas, on pourraient obtenir une valeur objectif arbitrairement petit en divisant $\mathbf{A}$ par un facteur donné tout en multipliant $\mathbf{D}$ par ce meme facteur.

On remarquera que ce problème est un problème est un problème convexe en chacune de ces variables mais qui n'est pas conjointement convexe (on peut le vérifie à partir d'une version simplifiée de cette fonction cout $(uv)^2$ n'est pas une fonction convexe en u et v).

5.3 Méthode des directions optimales

Cette méthode cherche à minimiser le problème

$$\min_{\mathbf{A},\mathbf{D}\in C} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \lambda\Omega(\mathbf{A})$$

en optimisant alternativement sur $\mathbf{A}$ et $\mathbf{D}$ tout en fixant l'autre variable.

5.3.1 Optimisation sur A

Ainsi, si on fixe $\mathbf{D}$, le probleme d'optimisation sur $\mathbf{A}$ est :

$$\min_{\mathbf{A}} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \lambda\Omega(\mathbf{A})$$

le cas où $\Omega(\mathbf{A}) = 0$

Dans ce cas, on a un problème simple qu'il est intéressant d'analyser. Le problème

$$\min_{\mathbf{A}} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2$$

peut également s'écrire comme

$$\min_{\{\mathbf{a}_i\}} \sum_i \|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|^2$$

qui, étant séparable, admet les solutions $\mathbf{a}_i = (\mathbf{D}^T\mathbf{D})^{-1}(\mathbf{D}^T\mathbf{x}_i)$ si $\mathbf{D}^T\mathbf{D}$ est une matrice inversible (ce qui n'est pas le cas pour un dictionnaire redondant). On peut également résoudre de manière directe le problème utilisant la norme de Frobenius. En effet, comme on a

$$\|\mathbf{X}-\mathbf{D}\mathbf{A}\|_F^2 = tr((\mathbf{X}-\mathbf{D}\mathbf{A})^T(\mathbf{X}-\mathbf{D}\mathbf{A})) = tr(\mathbf{X}^T\mathbf{X})-2tr(\mathbf{X}^T\mathbf{D}\mathbf{A})+tr(\mathbf{A}^T\mathbf{D}^T\mathbf{D}\mathbf{A})$$

pour minimiser cette fonction, il faut calculer la dérivée de cette fonction par rapport à $\mathbf{A}$. Calculons alors $\nabla_{\mathbf{V}} tr(\mathbf{UV})$. On rappelle tout d'abord que $tr(\mathbf{UV}) = \sum_j \sum_i u_{j,i} v_{i,j} = tr(\mathbf{VU})$. A partir de cette formule, on obtient aisément que

$$\frac{\partial tr(\mathbf{UV})}{\partial v_{m,n}} = U_{n,m} \quad \text{soit} \quad \nabla_{\mathbf{V}} tr(\mathbf{UV}) = \mathbf{U}^T$$

d'après cette equation, on peut en déduire que

$$\nabla_{\mathbf{A}} tr(\mathbf{X}^T \mathbf{DA}) = \mathbf{D}^T \mathbf{X}$$

On s'interesse maintenant à $\nabla_{\mathbf{V}} tr(\mathbf{VU}^T \mathbf{UV})$. On note que l'on a une forme quadratique et donc on a

$$\nabla_{\mathbf{V}} tr(\mathbf{V}^T \mathbf{U}^T \mathbf{UV}) = 2\mathbf{U}^T \mathbf{UV}$$

d'après cette équation on a donc

$$\nabla_{\mathbf{V}} tr(\mathbf{A}^T \mathbf{D}^T \mathbf{DA}) = 2\mathbf{D}^T \mathbf{DA}$$

et on peut montrer que la condition d'optimalité de notre problème original est

$$\nabla_{\mathbf{A}} \|\mathbf{X} - \mathbf{DA}\|_F^2 = -\mathbf{D}^T \mathbf{X} + \mathbf{D}^T \mathbf{DA} = 0$$

dont la solution est

$$\mathbf{A} = (\mathbf{D}^T \mathbf{D})^{-1} (\mathbf{D}^T \mathbf{X})$$

si $(\mathbf{D}^T \mathbf{D})$ est inversible.

le cas où $\Omega(\mathbf{A}) = \sum_{i,j} |a_{i,j}|$

Dans ce cas, le problème devient

$$\min_{\{\mathbf{a}_i\}} \sum_i \|\mathbf{x}_i - \mathbf{Da}_i\|^2 + \lambda \sum_{i,j} |a_{i,j}|$$

et peut se résoudre en appliquant un algorithme de régression Lasso pour chaque $\mathbf{x}_i$ ou alors, en développant un algorithme spécifique au problème

$$\min_{\mathbf{A}} \|\mathbf{X} - \mathbf{DA}\|_F^2 + \lambda \sum_{i,j} |a_{i,j}|$$

5.3.2 Optimisation sur D

On a maintenant à résoudre le problème suivant

$$\min_{\mathbf{D}\in C} \|\mathbf{X} - \mathbf{DA}\|_F^2$$

L'approche suggérée par les inventeurs de la méthode MOD est de résoudre par rapport à $\mathbf{D}$ sans tenir compte des contraintes puis de normaliser chaque colonne de $\mathbf{D}$. On s'interesse donc à minimiser

$$\|\mathbf{X}-\mathbf{DA}\|_F^2 = tr((\mathbf{X}-\mathbf{DA})^T(\mathbf{X}-\mathbf{DA})) = tr(\mathbf{X}^T\mathbf{X})-2tr(\mathbf{X}^T\mathbf{DA})+tr(\mathbf{A}^T\mathbf{D}^T\mathbf{DA})$$

par rapport à $\mathbf{D}$.

Comme on a $tr(\mathbf{X}^T\mathbf{DA}) = tr(\mathbf{AX}^T\mathbf{D})$, on en déduit que

$$\nabla_{\mathbf{D}} tr(\mathbf{X}^T\mathbf{DA}) = \mathbf{XA}^T$$

quant à la forme quadratique, on sait que $tr(\mathbf{A}^T\mathbf{D}^T\mathbf{DA}) = tr(\mathbf{AA}^T\mathbf{D}^T\mathbf{D}) = tr(\mathbf{DAA}^T\mathbf{D}^T)$ on en déduit donc que

$$\nabla_{\mathbf{D}} tr(\mathbf{A}^T\mathbf{D}^T\mathbf{DA}) = \mathbf{DAA}^T + \mathbf{DAA}^T$$

donc la condition d'optimalité du problème est :

$$-\mathbf{XA}^T + \mathbf{DAA}^T = 0$$

soit

$$\mathbf{D} = (\mathbf{XA}^T)(\mathbf{AA}^T)^{-1}$$

5.3.3 Algorithme proximale alternée

Si on considère le cas plus général où des contraintes plus complexes sur les atomes du dictionnaire peuvent être mise en oeuvre, le problème devient alors

$$\min_{\mathbf{A},\mathbf{D}} \|\mathbf{X} - \mathbf{DA}\|_F^2 + \lambda\Omega(\mathbf{A}) + \lambda_D\Omega_D(\mathbf{D})$$

Si on conserve l'idée de l'optimisation alternée sur $\mathbf{A}$ et $\mathbf{D}$, on a à minimiser alternativement les problèmes

$$\min_{\mathbf{A}} \|\mathbf{X} - \mathbf{DA}\|_F^2 + \lambda\Omega(\mathbf{A})$$

et

$$\min_{\mathbf{D}} \|\mathbf{X} - \mathbf{DA}\|_F^2 + \lambda_D\Omega_D(\mathbf{D})$$

si $\Omega(\mathbf{A}) = \sum_{i,j} |a_{i,j}|$, le problème sur A est toujours un problème de type Lasso, si $\Omega(\mathbf{A})$ a une autre forme, on peut éventuellement utiliser un algorithme proximale si la forme proximale

$$\text{prox}_\Omega(\mathbf{V}) = \arg\min_{\mathbf{U}} \frac{1}{2}\|\mathbf{U} - \mathbf{V}\|_F^2 + \Omega(\mathbf{U})$$

est simple. On peut également procéder de la même manière pour l'optimisation sur $\mathbf{D}$. La question reste alors celui du calcul de l'opérateur proximal associée à $\Omega(\mathbf{D})$. Dans le cas où $\Omega(\mathbf{D}) = I_{\|\mathbf{d}_i\|_2^2 \leq 1, \forall i}$, on s'interesse à l'opérateur proximal défini comme

$$\begin{array}{ll} \min_{\mathbf{U}} & \frac{1}{2}\|\mathbf{U} - \mathbf{V}\|_F^2 \\ sc & \|\mathbf{u}_i\|_2^2 \leq 1 \end{array}$$

que l'on peut calculer séparement sur chaque colonne $\mathbf{u}_i$.

5.4 K-SVD

Cette algorithme cherche également à résoudre le problème

$$\min_{\mathbf{A},\mathbf{D}\in C} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \lambda\Omega(\mathbf{A})$$

où $\Omega(\mathbf{A})$ est de type $\sum_{i,j} |a_{i,j}|$ en utilisant encore une approche alternée. Cependant, la méthode utilisée pour optimiser le dictionnaire est radicalement différent. En effet, on peut remarquer que l'on peut isoler l'influence d'un vecteur $\mathbf{d}_k$ donné dans la fonction objective $\|\mathbf{X}-\mathbf{D}\mathbf{A}\|_F^2 = \|\mathbf{X}-\sum_{k=1}^N \mathbf{d}_k\mathbf{a}'_k\|$ où $\mathbf{a}'_k$ est un vecteur ligne correspond à la ligne k de $\mathbf{A}$, cette ligne correspondant à l'ensemble des coefficients qui pondèrent le dictionnaire $\mathbf{d}_k$ dans l'ensemble des signaux. Ainsi, pour un élement j donné, on peut écrire

$$\begin{aligned} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 &= \|\mathbf{X} - \textstyle\sum_{k=1}^N \mathbf{d}_k\mathbf{a}'_k\|_F^2 = \|\underbrace{\mathbf{X} - \sum_{k=1, j\neq k}^{N} \mathbf{d}_k\mathbf{a}'_k}_{\mathbf{E}_j} - \mathbf{d}_j\mathbf{a}'_j\|_F^2 \\ &= \|\mathbf{E}_j - \mathbf{d}_j\mathbf{a}'_j\|_F^2 \end{aligned}$$

Dans ce contexte, on peut optimiser chaque élement (ici $\mathbf{d}_j$) du dictionnaire en considérant les autres fixes, ce qui équivaut à résoudre le problème suivant

$$\min_{\mathbf{d}_j} \|\mathbf{E}_j - \mathbf{d}_j\mathbf{a}'_j\|_F^2$$

Ce problème correspond à approximer une matrice par une autre matrice de rang 1. On peut résoudre ce problème à l'aide d'une décomposition SVD $E_j = \mathbf{U}\mathbf{S}\mathbf{V}^T$ ce qui permettrait également de mettre à jour les $\mathbf{a}_j$ par $\mathbf{d}_j = \mathbf{u}_1$ et $\mathbf{a}'_j = S_{1,1}\mathbf{v}_1^T$.

L'inconvénient de cette approche est que cette mise à jour des a'_j ne garantit pas que la parcimonie soit conservée. Pour contourner ce problème, on cherche plutot à approximer

$$\min_{\mathbf{d}_j,\mathbf{a}_j} \|\tilde{\mathbf{E}}_j - \mathbf{d}_j\tilde{\mathbf{a}}'_j\|_F^2$$

où $\tilde{\mathbf{E}}_j$ et $\tilde{\mathbf{a}}'_j$ sont restreint aux signaux dont les coefficients $\mathbf{a}_{j,k}$ donnés sont non-nulles.