2ème partie : Optimisation non-linéaire

Programme mathématique générale

$$
\begin{array}{lll}
\text{minimiser } f(x) & & \text{[fonction objective]} \\
\text{sous contraintes} & & \\
& h_i(x) = 0,\ i = 1, ..., m & \text{[contraintes d'égalité]} \\
& g_j(x) \le 0,\ j = 1, ..., k & \text{[contraintes d'inégalité]} \\
& x \in X & \text{[domaine du problème]}
\end{array}
\qquad \text{(PM)}
$$

Remarque : *les contraintes dans le système sont toujours liées par "et" – une solution réalisable doit satisfaire toutes les contraintes :*

$$\left\{ x = [x_1; ...; x_n] : g_i(x) \begin{array}{c} \geq \\ = \\ \leq \end{array} b_i \text{ pour tout } i = 1, ..., m, \right\}$$

- Ainsi, le problème

$$\min_{x=[x_1;x_2]} \left\{x_1 + x_2 : \underbrace{x_1 - x_2 - 3}_{g_1(x)} \leq 0 \text{ ou } \underbrace{\sin(x_1)}_{g_2(x)} \leq 0\right\}$$

n'est pas dans le format de programme mathématique.

- La forme éligible de ce problème serait, par exemple,

$$\min_{x=[x_1;x_2]} \left\{x_1 + x_2 : \underbrace{\min[x_1 - x_2 - 3, \ \sin(x_1)]}_{g(x)} \leq 0\right\}$$

En effet, dire que

$$g_1(x) \leq b_1 \text{ ou } g_2(x) \leq b_2 \text{ ou ... ou } g_m(x) \leq b_m$$

est exactement le même que de dire

$$g(x) := \min\left[g_1(x) - b_1, g_2(x) - b_2, ..., g_m(x) - b_m\right] \leq 0.$$

Par contre, dire

$$g_1(x) \leq b_1 \text{ et } g_2(x) \leq b_2 \text{ et ... et } g_m(x) \leq b_m$$

est exactement le même que de dire

$$g(x) := \max\left[g_1(x) - b_1, g_2(x) - b_2, ..., g_m(x) - b_m\right] \leq 0.$$

Remarque : (Presque) tout problème en mathématique appliquée peut être exprimée comme un problème de programmation mathématique. $\Rightarrow$ *De façon générale, un problème de programmation non-linéaire est difficile – on ne peut pas espérer de le résoudre en un temps raisonnable.*

Question : *Alors comment peut-on traiter des problems avec des dizaines de milliers de variables et de contraintes avec une grande precision ?*

Réponse : *L'idee serait d'utiliser la structure du problème. Une structure favorable permet d'utiliser l'information local sur l'objectif et les contraintes pour inférer sur une solution globalement optimale.*

Une "structure favorable" standard est celle de *convexité*.

Optimisation convexe

Problème générale de programmation convexe

$$
\begin{array}{lll}
\text{minimiser } f(x) & & \text{[fonction objective]} \\
\text{sous contraintes} & & \\
& g_j(x) \leq 0,\ j = 1, ..., k & \text{[contraintes d'inégalité]} \\
& x \in X & \text{[domaine du problème]}
\end{array}
\qquad \text{(PC)}
$$

où

- $f,\ g_1, ..., g_m$ sont des *fonctions convexes*
- $X \subset \mathbb{R}^n$ est un *ensemble convexe.*

Remarque : il n'y a pas de contraintes d'égalité (!)
Autrement dit, les seules contraintes d'égalité autorisées sont les *contraintes linéaires* $a^T x - b = 0$*,* facilement transformables en contraintes d'inégalité avec des fonctions linéaires (donc convexes)

$$a^T x - b \leq 0, \;\; -a^T x + b \leq 0.$$

Ensembles convexes : définitions

Ensemble $X \subset \mathbb{R}^n$ *est dit* convexe *si avec tout point* x, y*, il contient le segment entier qui les joint :*

$$x, y \in X, \lambda \in [0, 1] \;\Rightarrow\; (1-\lambda)x + \lambda y \in X.$$

Définition équivalente : $X \in \mathbb{R}^n$ *est convexe, si* X *contient* toute combinaison convexe *de ses éléments (i.e.,* combinaison linéaire avec des coefficients non-négatifs dont la somme fait 1*) :*

$$x_1, ..., x_k \in X \;\Rightarrow\; \sum_{i=1}^{k} \lambda_i x_i \in X \;\; \forall \lambda \geq 0 \text{ tel que } \sum_{i=1}^{k} \lambda_i = 1.$$

Exemple : un ensemble polyédrique $X = \{x \in \mathbb{R}^n : \; Ax \leq b\}$ est convexe. $\Rightarrow$ sous-espaces linéaires et affines sont des ensembles convexes.

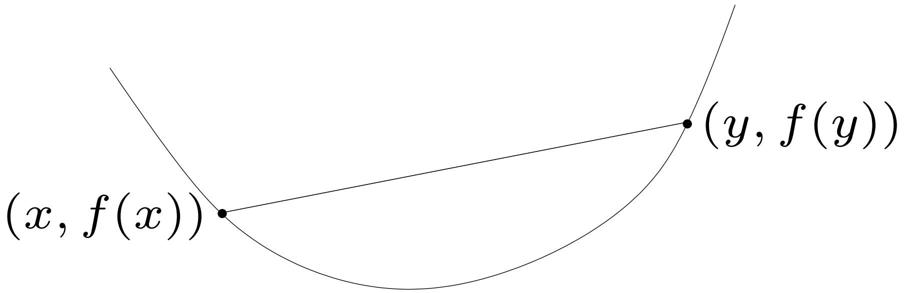
En effet, $x \in X, \;\; y \in X \Leftrightarrow Ax \leq b$, $Ay \leq b$.
Alors pour tout $0 \leq \lambda \leq 1$ et $z = \lambda x + (1-\lambda)y$,

$$Az = A[\lambda x + (1-\lambda)y] = \lambda Ax + (1-\lambda)Ay \leq \lambda b + (1-\lambda)b = b \;\Rightarrow z \in X.$$

Fonctions convexes : définitions

Fonction $f : \mathbb{R}^n \to \mathbb{R}$ *est dite* convexe *si pour tout* x, y *et* $\lambda \in [0, 1]$,

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + f(\lambda y).$$



$f : \mathbb{R}^n \to \mathbb{R}$ *est dite* concave *si* $-f$ est convexe.

Exemples

— fonction affine $ax + b$ sur $\mathbb{R}$ est convexe (et concave)

— fonction affine $a^T x + b$ sur $\mathbb{R}^n$ est convexe (et concave)

— fonction e^{ax} est convexe pour tout $a \in \mathbb{R}$

— fonction $x \log x$ est convexe sur $\mathbb{R}_{+*}$

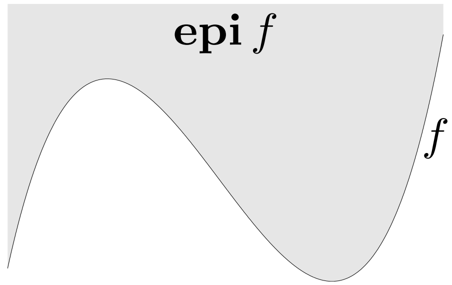
— fonction $\|x\|_2$ est convexe sur $\mathbb{R}^n$

— ...

Épigraphe d'une fonction *Soit* $f : \mathbb{R}^n \to \mathbb{R}$, *l'ensemble*

$$\mathrm{Epi}\ f = \{[x;\tau] \in \mathbb{R}^n :\ f(x) \leq \tau\}$$

s'appelle épigraphe de f.



Définition équivalente : *Une fonction* $f(x) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *est* convexe, si et seulement si son épigraphe $\mathrm{Epi}\ f$ est un ensemble convexe.

Exemple La fonction linéaire par morceaux

$$f(x) = \begin{cases} \max_i [a_i^T x + b_i], & \text{si } Px \leq p \\ +\infty, & \text{sinon} \end{cases}$$

est convexe.

En effet, l'épigraphe de f,

$$\mathrm{Epi}\ f = \{[x;t] \in \mathbb{R}^n :\ Px \leq p,\ t \geq a_i^T x + b_i,\ \forall i\}$$

est un ensemble polyédrique.

Inégalité de Jensen

Convexité :

$$\forall \lambda \in [0,1],\ f((1-\lambda)x + \lambda y) \le (1-\lambda) f(x) + f(\lambda y) \qquad (*)$$

Généralisation : si f est convexe, alors pour tout x et $\lambda_1, ..., \lambda_m$ tels que

$$\lambda_i \ge 0\ \forall i,\ \sum_{i=1}^{m} \lambda_i = 1,$$

nous avons

$$f\left(\sum_{i=1}^{m} \lambda_i x_i\right) \le \sum_{i=1}^{m} \lambda_i f(x_i)$$

(verification en utilisant la caractérisation de convexité par épigraphe).

En particulier, soit f convexe, alors

$$f(\mathbf{E}(Z)) \le \mathbf{E}(f(Z))$$

pour tout vecteur aléatoire Z sur $\mathbb{R}^n$.

L'inégalité $(*)$ "à 2 points" correspond à cas de la loi discrète telle que

$$\mathrm{Prob}\{Z = x\} = \lambda, \quad \mathrm{Prob}\{Z = y\} = 1 - \lambda.$$

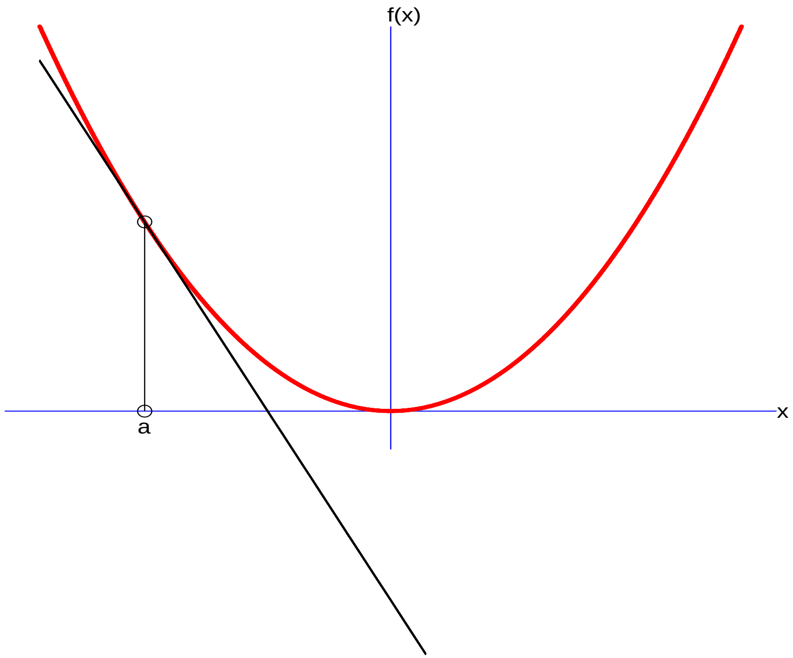
Rôle de la convexité

On considère le problème $\min_{x\in X} f(x)$ de minimisation d'une fonction f *différentiable* sur un domaine *simple*, e.g., une "boite" n-dimensionnelle

$$X = \{x \in \mathbb{R}^n : -1 \le x_i \le 1,\ i = 1, ..., n\}.$$

• *Pour f différentiable*, la convexité est définie comme *la propriété de f de dominer ses linéarisations :*

$$\begin{aligned} f(y) &\ge f(x) + [\nabla f(x)]^T (y - x) \\ &:= f(x) + \sum_{i=1}^n \frac{\partial f(x)}{\partial x_i}(y_i - x_i) \text{ for all } x, y \end{aligned}$$



Soit $f : [-1, 1] \to \mathbb{R}$.

• Si nous avons calculé f and f' en $a \in [-1, 1]$, et $f'(a) < 0$

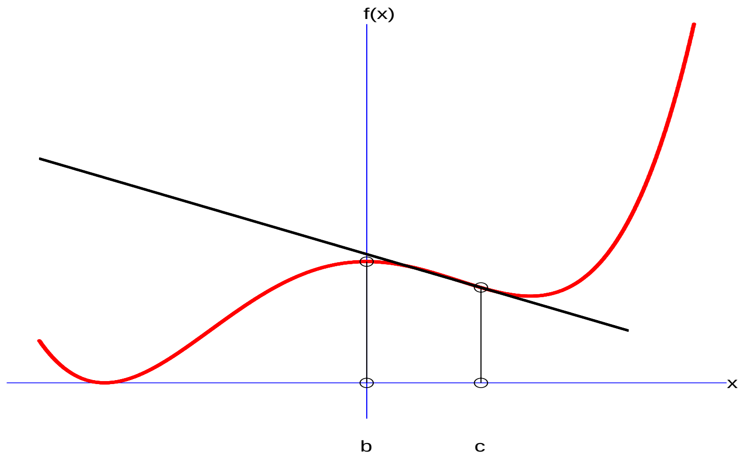
$\Rightarrow$ à gauche de a, la linéarisation de f est $> f(a)$

$\Rightarrow$ *a gauche de a, f elle-même est $> f(a)$*

$\Rightarrow$ *on peut réduire le domaine du problème en éliminant tous les points $< a$!*

• Le schéma des "coupes" peut être généralisé aux problèmes convexes multi-dimensionnels (i.e., avec l'objectif et les contraintes convexes).

Remarque : la convexité de f est cruciale dans ce cas. Par exemple, en cas de la fonction f non convexe



l'information locale autour de c *ne dit rien* sur la position du minimum *global* et *ne permet pas* d'éliminer une partie "massive" du domaine.

Reconnaître fonctions convexes I

- Critère différentiel, fonctions d'une variable
  - *fonction differentiable* $f : \mathbb{R} \to \mathbb{R}$ est convexe *ssi* sa dérivée $f'(x)$ *est monotone non-décroissante :* $x_1 \leq x_2 \Rightarrow f'(x_1) \leq f'(x_2)$
  - *fonction 2 fois differentiable* $f : \mathbb{R} \to \mathbb{R}$ est convexe ssi sa dérivée seconde $f''(x)$ est non-négative : $f''(x) \geq 0 \; \forall x \in \mathbb{R}$.

- Fonctions de n variables :

  fonction $f : \mathbb{R}^n \to \mathbb{R}$ *2 fois differentiable* est convexe *ssi sa matrice hessienne est semi-définie positive pour tout* x *:* $\nabla^2 f(x) \succeq 0, \; \forall x \in \mathbb{R}^n$ (toutes les valeurs propres de $\nabla^2 f(x)$ sont non négatives).

Exemples

- *Fonction quadratique* $f(x) = \frac{1}{2}x^T Px + q^T x + r$ avec

$$\nabla f = Px + q, \quad \nabla^2 f = P,$$

est convexe sur $\mathbb{R}^n$ ssi $P \succeq 0$ (P est semi-définie positive)

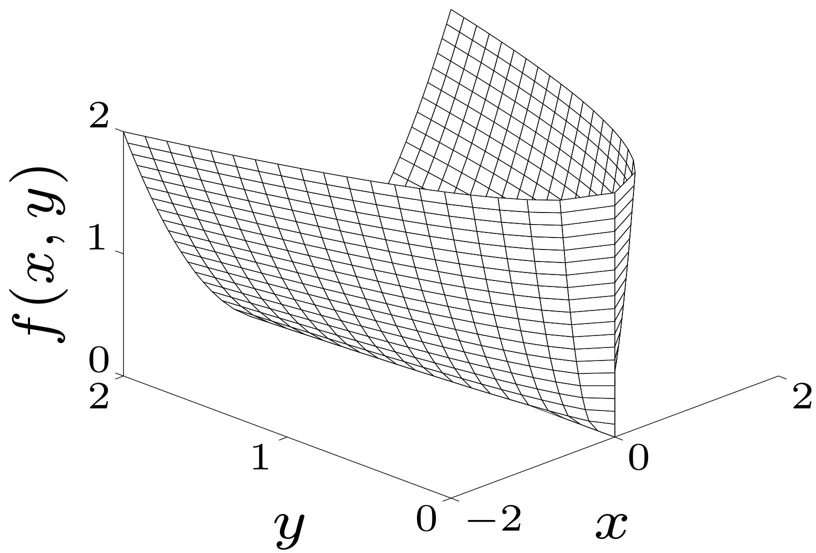
- *Fonction quadratique-sur-linéaire*

$$f(x,y) = x^2/y,$$

$$\nabla f(x,y) = \frac{1}{y^2}\begin{bmatrix} 2xy \\ -x^2 \end{bmatrix},$$

$$\nabla^2 f(x,y) = \frac{2}{y^3}\begin{bmatrix} y \\ -x \end{bmatrix}\begin{bmatrix} y \\ -x \end{bmatrix}^T \succeq 0$$

est convexe pour $x \in \mathbb{R}$ et $y > 0$



Reconnaître fonctions convexes II : opérations qui préservent la convexité

• *multiplication par un réel non-négatif :* si f est convexe, $\alpha \geq 0$, alors αf est convexe
• *somme :* si $f_1,\ f_2$ sont convexes, $f_1 + f_2$ est convexe (ainsi que $\alpha_1 f_1 + \alpha_2 f_2$ pour $\alpha_1, \alpha_2 \geq 0$)
• *composition avec une fonction affine :* si f est convexe, $f(Ax + b)$ l'est aussi

Exemples

— fonction $\|Ax + b\|_2$
— fonction $\sum_i \exp(a_i^T x + b_i)$
— fonction "barrière"

$$f(x) = -\sum_{i=1}^{m} \log(b_i - a_i^T x)$$

définie sur $\mathrm{Dom} f = \{x \in \mathbb{R}^n :\ Ax < b\}$

— ...

- *Maximum "point par point :"* si $f_1, \ldots f_m$ sont convexes, alors la fonction

$$\bar{f}(x) = \max\{f_1(x), \ldots, f_m(x)\}$$

est convexe.

Exemples

— fonction linéaire par morceaux $f(x) = \max_i(a_i^T x + b_i)$, et donc la fonction (valeur absolue) $|x| = \max\{x, -x\}$

— la norme $\|x\|_\infty = \max_i |x_i|$

— la norme $\|x\|_1 = \sum_{i=1}^n |x_i|$

- *Supremum par point :* si $f_\alpha(x)$ est convexe en x pour tout $\alpha \in \mathcal{A}$, la fonction

$$\bar{f}(x) = \sup_{\alpha \in \mathcal{A}} f_\alpha(x)$$

est convexe.

Exemple : la plus grande valeur propre $\lambda_{\max}(A)$ d'une matrice symétrique A,

$$\lambda_{\max}(A) = \sup_{y:\, \|y\|_2 = 1} y^T A y$$

- *Superposition convexe-monotone :* Soit
  - — $g_i(x) :\ \mathbb{R}^n \to \mathbb{R}$ fonctions convexes
  - — $F(y)\ :\ \ \mathbb{R}^m \to \mathbb{R}$ fonction convexe et monotone non-décroissante en tout $y_1, ..., y_m$:

$$y^1 \leq y^2 \ \Rightarrow\ F(y^1) \leq F(y^2)$$

Alors, la fonction composée (la superposition de F et $g_1, ..., g_m$)

$$f(x) = F(g_1(x), ..., g_m(x))$$

est convexe.

- ...

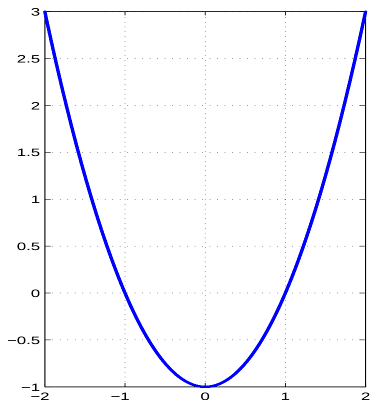
Illustration : soit $g_1, \ldots, g_m$ fonctions convexes *non-négatives*, et soit $F(y_1, \ldots, y_m) = \sum_{i=1}^m y_i^2$.

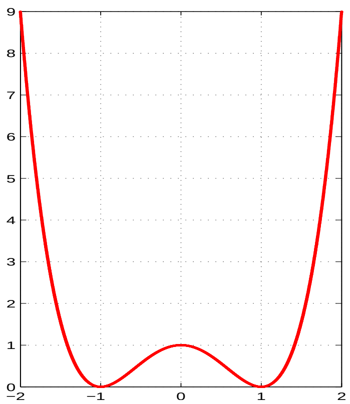
Fonction $f(x) = F(g_1(x), \ldots, g_m(x)) = \sum_{i=1}^m g_i^2(x)$ est-elle convexe ?

• La propriété de superposition n'est pas applicable directement, car F n'est pas monotone.

• Néanmoins, *sur l'orthant non-négatif* $Q = \{y : y \geq 0\}$, F *est monotone,* et *comme toutes les* g_i *sont non-négatives,* on peut appliquer ce résultat pour montrer que f *est convexe*.

Remarque : la non-négativité des g_i est importante. *Le carré d'une fonction convexe n'est pas forcement convexe.*





à gauche : x^2, à droite : $(x^2 - 1)^2$

D'habitude, le "calcul de convexité" avec le critère différentiel suffisent pour verifier la convexité des fonctions multi-variées.

Exemple. Soit

$$f(x) = \log\Big(\exp(a_1^T x + b_1) + \ldots + \exp(a_m^T x + b_m)\Big)$$

1^o. Fonction lisse $g(y) = \log(1 + e^y) : \mathbb{R} \to \mathbb{R}_+$, est convexe, avec

$$g'(x) = \frac{e^y}{1 + e^y}, \quad g''(y) = \frac{e^y}{(1 + e^y)^2} \geq 0$$

2^o. Fonction

$$h(y_1, y_2) = \log\Big(e^{y_1} + e^{y_2}\Big) = \log(1 + e^{y_1 - y_2}) + y_2 = g(y_1 - y_2) + y_2$$

est convexe (transformation linéaire d'argument et somme de fonctions convexes) $\Rightarrow$ fonction

$$\ell(y) = \log\Big(e^{y_1} + \ldots + e^{y_m}\Big) : \mathbb{R}^m \to \mathbb{R}_+$$

est convexe

3^o. Finalement, fonction $f(x) = \ell(Ax + b)$ est convexe aussi (transformation affine d'argument).

Et ainsi de suite...

Quiz : *Lesquelles parmi les fonctions suivantes sont convexes ?*

- $\ln(e^{2x+3y} + 2e^{y-x})$
- $\ln(e^{x^2} + e^{y^2})$
- $\ln(e^{-x^2} + e^{y^2})$
- $\ln(e^{x^2} + 2e^{-3x^2})$
- $\ln(e^{x^2} + e^{-x^2})$

- $\ln(e^{2x+3y} + 2e^{y-x})$ – convexe avec $\ln(e^{x_1} + e^{x_2})$ (substitution affine d'argument)

- $\ln(e^{x^2} + e^{y^2})$ – convexe avec $\ln(e^{x_1} + e^{x_2})$ et x^2, y^2 (superposition monotone, notez que $\ln(e^{x_1} + e^{x_2})$ est non-décroissante en x_1 et x_2)

- $\ln(e^{-x^2} + e^{y^2})$ – non-convexe : regardez ce qui se passe quand $y = 0$: $\frac{d}{dx}f(x,0) = -\frac{2xe^{-x^2}}{e^{-x^2}+1}$, et la dérivée *n'est pas non-décroissante en x*

- $\ln(e^{x^2} + 2e^{-3x^2})$ – non-convexe : $\frac{d}{dx}f(x) = -\frac{x(6e^{-3x^2}-2e^{x^2})}{e^{x^2}+e^{-3x^2}}$, et la dérivée *n'est pas non-décroissante autour de $x = 0$*

- $\ln(e^{x^2} + e^{-x^2})$ – convexe car fonction $\ln(e^{s} + e^{-s})$ est convexe *et non-décroissante* pour $s \geq 0$, et x^2 est convexe *et non-négative*

Minima des fonctions convexes

Soit X ensemble convexe dans $\mathbb{R}^n$, et f une fonction convexe sur $\mathbb{R}^n$. On considère le problème d'optimisation

$$\mathrm{Opt} = \min_{x \in X} f(x)$$

- *Tout minimiseur local x_* de f sur X est un minimiseur global de f sur X :*
  - *si $x_* \in X$ est tel que pour un $r > 0$, $f(x) \geq f(x_*)$ pour tout $x \in X$ et $\|x - x_*\|_2 \leq r$,*
  - *alors $f(x) \geq f(x_*)$ pour tout $x \in X$.*

Soit x_* un minimiseur local de f sur X ; et soit $x \neq x_*$, $x \in X$. Dans ce cas,

$$\frac{f(x_* + \lambda[x - x_*]) - f(x_*)}{\lambda\|x - x_*\|_2} \leq \frac{f(x) - f(x_*)}{\|x - x_*\|_2}$$

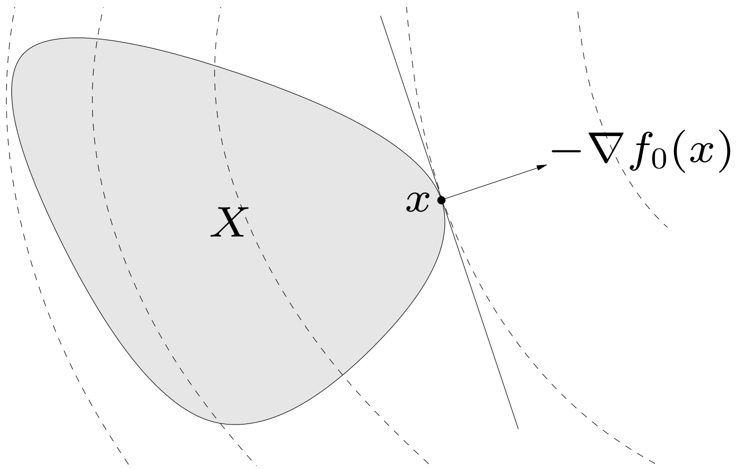
pour tout $\lambda \in (0, 1)$. Comme x_* est le minimiseur local de f, nous avons $f(x_* + \lambda[x - x_*]) \geq f(x_*)$ pour λ petit
$\Rightarrow$ le ratio à droite est non-négatif $\Rightarrow f(x) \geq f(x_*)$. $\square$

Question

Soit X un ensemble convexe dans $\mathbb{R}^n$, f fonction convexe, et soit $x_* \in X$ un point tel que f est derivable en x_*. Quand est-ce que x_* est un minimiseur global de f sur X ?

Réponse : *c'est le cas si et seulement si*

$$\forall (x \in X) : \nabla f(x_*)^T (x - x_*) \geq 0$$



Géométriquement : X appartient au demi-espace

$$H = \{x \in \mathbb{R}^n : \nabla f(x_*)^T x \geq b := \nabla f(x_*)^T x_*\}.$$

Autrement dit, le hyperplan

$$\Pi = \{x \in \mathbb{R}^n : \nabla f(x_*)^T x = b\}$$

est "tangent" à X en x_*.

Nécessité (seulement si) : pour tout $x \in X$ et $0 \le \lambda \le 1$, nous devons avoir

$$g(\lambda) := f(x_* + \lambda(x - x_*)) \ge f(x_*) = g(0);$$

ainsi

$$0 \le g'(0) = \nabla f(x_*)^T (x - x_*),$$

et ceci pour tout $x \in X$.

Suffisance (si) : nous savons que, $f(x) \ge f(x_*) + \nabla f(x_*)^T (x - x_*)$ pour tout x, donc $f(x) \ge f(x_*)$ quand $x_* \in X$ et $\nabla f(x_*)^T (x - x_*) \ge 0$ pour tout $x \in X$.

Remarque : *Quand x_* se trouve dans l'intérieur de X (c.-à-d. que pour un $r > 0$ toute la boule $\{x : \|x - x_*\|_2 \le r\} \subset X$, la condition ci-dessus devient la règle de Fermat : $\nabla f(x_*) = 0$.*

Fonction de Lagrange et dualité de Lagrange

On considère le problème de programmation mathématique

$$\mathrm{Opt}(P) = \min_{x \in X \subset \mathbb{R}^n} \left\{ f(x) : \; g_i(x) \leq 0, i = 1, ..., m \right\} \qquad (P)$$

- *La fonction de Lagrange du problème* (P) *est la fonction*

$$L(x, \lambda) := f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) : \; X \times \mathbb{R}_+^m \to \mathbb{R}$$

Remarque : *quand on parle de la fonction de Lagrange,*

— *variable* x varie dans X

— *variable* λ varie dans $\mathbb{R}_+^m$

on veut que les multiplicateurs de Lagrange $\lambda_1, ..., \lambda_m$ *soient non-négatives.*

Plus généralement,

- si problème de *minimisation*

— contrainte $g(x) \leq 0 \Rightarrow \lambda$ correspondant est ≥ 0

— contrainte $g(x) \geq 0 \Rightarrow \lambda$ correspondant est ≤ 0

- si problème de *maximisation*,

— contrainte "$\leq$" $\Rightarrow \lambda$ correspondant est ≤ 0

— contrainte "$\geq$" $\Rightarrow \lambda$ correspondant est ≥ 0

$$\begin{array}{l} \mathrm{Opt}(P) = \min_{x \in X \subset \mathbb{R}^n} \left\{ f(x) : \ g_i(x) \leq 0, i = 1, ..., m \right\} \qquad (P) \\ L(x, \lambda) := f(x) + \sum_{i=1}^m \lambda_i g_i(x) : \ X \times \mathbb{R}^m_+ \to \mathbb{R} \end{array}$$

Remarque : Nous avons deja rencontré la fonction de Lagrange dans le cas OL, où $X = \mathbb{R}^n$, f est linéaire, et $g_1, ..., g_m$ sont affines (dans le cas OL, il s'agissait d'un programme de maximisation, tandis qu'ici on s'interesse au problème de minimisation).

Observation : *pour tout* $\lambda \geq 0$, fonction de Lagrange sous-estime $f(x)$ *en tout* x *realisable. Ainsi,* pour tout $\lambda \geq 0$, *la function*

$$\underline{L}(\lambda) = \inf_{x \in X} L(x, \lambda) : \mathbb{R}^m_+ \to \mathbb{R} \cup \{-\infty\}$$

satisfait $\underline{L}(\lambda) \leq \mathrm{Opt}(P)$.

- *Le problème de programmation mathématique*

$$\begin{array}{rcl} \mathrm{Opt}(D) & = & \max_{\lambda \geq 0} \underline{L}(\lambda) \\ & = & \max_{\lambda \geq 0} \left[\inf_{x \in X} L(x, \lambda) \right] \end{array} \qquad (D)$$

s'appelle problème dual de Lagrange *de* problème primal (P).

$$
\begin{array}{rcll}
\mathrm{Opt}(P) & = & \min\limits_{x\in X\subset\mathbb{R}^n} \{f(x):\ g_i(x)\le 0, i=1,...,m\} & (P)\\
L(x,\lambda) & = & f(x)+\sum\limits_{i=1}^m \lambda_i g_i(x):\ X\times\mathbb{R}^m_+\to\mathbb{R} & \\
\underline{L}(\lambda) & = & \inf\limits_{x\in X} L(x,\lambda):\mathbb{R}^m_+\to\mathbb{R}\cup\{-\infty\} & \\
\mathrm{Opt}(D) & = & \max\limits_{\lambda\ge 0}\underline{L}(\lambda), & (D)\\
 & = & \max\limits_{\lambda\ge 0}\left[\inf\limits_{x\in X} L(x,\lambda)\right] &
\end{array}
$$

[Dualité faible] : *par construction,*

$$\mathrm{Opt}(D)\le\mathrm{Opt}(P).$$

Remarque : ici la convexité n'est pas importante.

- Sous hypothèses supplémentaires “peu contraignantes,” dans le cas convexe,

$$\mathrm{Opt}(D)=\mathrm{Opt}(P).$$

$$\begin{array}{rcll}
\mathrm{Opt}(P) & = & \min\limits_{x\in X\subset\mathbb{R}^n} \{f(x):\ g_i(x)\le 0, i=1,...,m\} & (P)\\
L(x,\lambda) & = & f(x)+\sum\limits_{i=1}^m \lambda_i g_i(x):\ X\times\mathbb{R}^m_+\to\mathbb{R} & \\
\underline{L}(\lambda) & = & \inf\limits_{x\in X} L(x,\lambda):\mathbb{R}^m_+\to\mathbb{R}\cup\{-\infty\} & \\
\mathrm{Opt}(D) & = & \max\limits_{\lambda\ge 0}\underline{L}(\lambda), & (D)\\
 & = & \max\limits_{\lambda\ge 0}\left[\inf\limits_{x\in X} L(x,\lambda)\right] &
\end{array}$$

Condition de Slater : (P) *admet une solution* strictement réalisable $\bar{x}$, c.-à-d. telle que $\bar{x}\in X$ and $g_i(\bar{x})< 0$ *pour tout* $i=1,...,m$.

Condition de Slater relaxée : (P) *admet une solution réalisable* $\bar{x}$ *dans l'intérieur de* X, *telle que toutes contraintes non-affines sont satisfaites comme inégalités strictes en* $\bar{x}$.

Pour (P) *convexe, condition de Slater relaxée est plus "légère" que la condition de Slater.*

$$\begin{aligned}
\mathrm{Opt}(P) &= \min_{x\in X\subset\mathbb{R}^n} \{f(x):\ g_i(x)\le 0, i=1,...,m\} \quad (P)\\
L(x,\lambda) &= f(x)+\sum_{i=1}^m \lambda_i g_i(x):\ X\times\mathbb{R}^m_+\to\mathbb{R}\\
\underline{L}(\lambda) &= \inf_{x\in X} L(x,\lambda) : \mathbb{R}^m_+\to\mathbb{R}\cup\{-\infty\}\\
\mathrm{Opt}(D) &= \max_{\lambda\ge 0} \underline{L}(\lambda), \quad (D)\\
&= \max_{\lambda\ge 0}\left[\inf_{x\in X} L(x,\lambda)\right]
\end{aligned}$$

Théorème de dualité de Lagrange *Sous la condition de convexité de* (P) et la condition relaxée de Slater, *(D) est soluble, et*

$$\mathrm{Opt}(D)=\mathrm{Opt}(P)$$

Remarque : le problème primal (P) peut être aussi obtenu à partir de la fonction de Lagrange $L(x,\lambda)$: on remarque que

$$\overline{L}(x)=\sup_{\lambda\ge 0} L(x,\lambda)=\begin{cases} f(x), & g_i(x)\le 0\ \forall i\\ +\infty, & \text{sinon}\end{cases}$$

et (P) s'écrit de façon équivalente $\min_{x\in X}\left\{\overline{L}(x)=\sup_{\lambda\ge 0} L(x,\lambda)\right\}$.

$$
\begin{array}{rcll}
\mathrm{Opt}(P) & = & \min_{x \in X \subset \mathbb{R}^n} \{f(x):\; g_i(x) \le 0, i = 1, ..., m\} & (P) \\
L(x, \lambda) & = & f(x) + \sum_{i=1}^{m} \lambda_i g_i(x):\; X \times \mathbb{R}^m_+ \to \mathbb{R} & \\
\underline{L}(\lambda) & = & \inf_{x \in X} L(x, \lambda) : \mathbb{R}^m_+ \to \mathbb{R} \cup \{-\infty\} & \\
\mathrm{Opt}(D) & = & \max_{\lambda \ge 0} \underline{L}(\lambda) = \max_{\lambda \ge 0} \left[\inf_{x \in X} L(x, \lambda)\right] & (D) \\
\mathrm{Opt}(P) & = & \min_{x \in X} \overline{L}(x) = \min_{x \in X} \left[\sup_{\lambda \ge 0} L(x, \lambda)\right] & (P')
\end{array}
$$

Illustration :

- Soit (P) le problème

$$
\mathrm{Opt}(P) = \min_{x \in X = [0, \infty)} \left\{ f(x) = \frac{1}{1+x} :\; g_1(x) := 20 - x \le 0 \right\}. \qquad (P)
$$

Ici $\mathrm{Opt}(P) = \inf_x \{\frac{1}{1+x} : x \ge 20\} = 0$, mais (P) est *insoluble*.
Néanmoins, le problème est convexe et satisfait la condition de Slater. Nous avons

$$
\underline{L}(\lambda) = \inf_{x \ge 0} \left\{ \frac{1}{1+x} + \lambda(20 - x) \right\} = \begin{cases} 0, & \lambda = 0 \\ -\infty, & \lambda > 0 \end{cases}
$$

et (D) est soluble avec solution optimale $\lambda = 0$ et valeur optimale $\mathrm{Opt}(D) = 0 = \mathrm{Opt}(P)$.

• Toutes les hypothèses du théorème de dualité sont essentielles. Par exemple, le problème

$$\mathrm{Opt}(P) = \min_{x \in X = \mathbb{R}} \left\{ x : g_1(x) := \tfrac{1}{2} x^2 \leq 0 \right\}, \qquad (P)$$

est convexe et soluble avec $\mathrm{Opt}(P) = 0$. Il *ne satisfait pas* la condition de Slater. Nous avons

$$\underline{L}(x) = \min_x \left\{ x + \frac{\lambda}{2} x^2 \right\} = \begin{cases} -\infty, & \lambda = 0 \\ -\frac{1}{2\lambda}, & \lambda > 0 \end{cases}$$

Et nous avons ("par chance") $\mathrm{Opt}(D) = 0 = \mathrm{Opt}(P)$, mais le problème dual n'a pas de solution.

Conditions d'optimalité en optimisation convexe

On considère le problème

$$\mathrm{Opt}(P) = \min_{x \in \mathbb{R}^n} \left\{ f(x) : \; g_j(x) \leq 0, \, j = 1, ..., m \right\} \qquad (P)$$

avec $f, , g_1, ..., g_m$ convexes.

Théorème [conditions de Karush-Kuhn-Tucker] *Soit x_* une solution realisable du problème convexe (P), et soit f, $g_1, ..., g_m$ différentiables en x_*.*

[i] *Soit x_* un point KKT de (P), c.-à-d. que x_* peut être augmenté par un $\lambda^* \geq 0$ pour satisfaire*

- [complémentarité] $\lambda_j^* g_j(x_*) = 0 \, \forall j$
- [équation KKT] $\nabla f(x_*) + \sum_{j=1}^m \lambda_j^* \nabla g_j(x_*) = 0.$

Alors, x_* *est une solution optimale de (P) (et, au fait, λ^* est une solution optimale de (D)).*

[ii] *Supposons que, en plus, (P) satisfait la condition relaxée de Slater. Alors x_* est une solution de (P)* si et seulement si *x_* est un point KKT de (P).*

$$\begin{array}{rcl}\mathrm{Opt}(P) & = & \min\limits_{x\in X\subset\mathbb{R}^n}\{f(x):\ g_i(x)\le 0, i=1,...,m\} \quad (P)\\ L(x,\lambda) & = & f(x)+\sum\limits_{i=1}^m \lambda_i g_i(x) : X\times\mathbb{R}^m_+\to\mathbb{R}\end{array}$$

Explication, [i] – si x_* est un *point KKT* et $\lambda^*\ge 0$ est le vecteur de multiplicateurs de Lagrange associé, alors

• x_* est admissible pour (P), et x_*, λ^* satisfont la condition de complémentarité
$\Rightarrow$ la fonction $L(x_*, \lambda)$ de λ *atteint son maximum sur $\lambda\ge 0$ en λ^** (pourquoi ?) et nous avons

$$L(x_*,\lambda^*)=f(x_*)$$

• La fonction

$$h(x)=f(x)+\sum_i \lambda_i^* g_i(x)$$

est convexe et différentiable en x_* et satisfait $\nabla h(x_*)=0$
$\Rightarrow$ la fonction $h(x)=L(x,\lambda^*)$ de x *atteint son minimum en x_** et

$$h(x_*)=L(x_*,\lambda^*)=f(x_*).$$

Mais pour tout x réalisable, $f(x)\ge h(x)\ge h(x_*)=f(x_*)$.
$\Rightarrow$ *x_* est une solution optimale de (P).* $\square$

Explication, [ii] – on doit verifier que

> “*si* (P) *est convexe et satisfait la condition de Slater relaxée,* f, g_i *sont différentiables en* x_**, et* x_* *est une solution optimale de* (P)*, alors* x_* *est un point KKT de* (P)*.*”

Soit $\lambda^* \geq 0$ une solution optimale du problème dual. Par le théorème de dualité, nous avons $\forall x \in \mathbb{R}^n,\ \lambda \geq 0$,

$$\begin{array}{rcl} L(x,\lambda^*) & \geq & \inf_x L(x,\lambda^*) = \underline{L}(\lambda^*) \\ & = & \mathrm{Opt}(D) = \mathrm{Opt}(P) = f(x_*) \\ & = & \overline{L}(x_*) = \sup_{\lambda \geq 0} L(x_*,\lambda) \geq L(x_*,\lambda). \end{array}$$

- Nous avons $L(x_*,\lambda^*) \geq \underline{L}(\lambda^*) = f(x_*)$, et, puisque x_* est réalisable,

$$\lambda_j^* g_j(x_*) = 0 \ \ \forall j \ \ \text{(complémentarité)}$$

- La fonction $L(x,\lambda^*) = f(x) + \sum_j \lambda_j^* g_j(x)$ est convexe and différentiable en $x_* \in X$ et atteint en x_* son minimum.

$$\Rightarrow \ \nabla_x L(x,\lambda^*) = \nabla f(x) + \sum_j \lambda_j^* \nabla g_j(x) = 0. \qquad \square$$

Exemples

• **Dualité linéaire :** soit

$$\min_x \{c^T x : \; b - Ax \leq 0\} \qquad \text{[réalisable, borné]}$$

Fonction de Lagrange $L(x,\lambda) = c^T x + \lambda^T(b - Ax)$, mais

$$\inf_x [c^T x + \lambda^T(b - Ax)] = \begin{cases} -\infty & \text{si} \quad c \neq A^T\lambda \\ b^T\lambda & \text{si} \quad c = A^T\lambda \end{cases}$$

$\Rightarrow$ problème dual : $\max_\lambda \left\{ b^T\lambda : \; A^T\lambda = c, \; \lambda \geq 0 \right\}$

• **Système linéaire, moindres carrés :** soit

$$\min_x \{\tfrac{1}{2} x^T x : \; Ax = b\} \qquad \text{[réalisable]}$$

Fonction de Lagrange $L(x,\lambda) = \frac{1}{2} x^T x + \lambda^T(Ax - b)$,

$$\nabla_x L(x,\lambda) = x + A^T\lambda, \;\Rightarrow\; x(\lambda) = -A^T\lambda$$

$\Rightarrow$ objectif dual $\underline{L}(\lambda) = L(A^T\lambda, \lambda) = -\frac{1}{2}\lambda^T A A^T \lambda - b^T\lambda$
$\Rightarrow$ problème dual $\max_\lambda -\frac{1}{2}\lambda^T A A^T \lambda - b^T\lambda$

• **Moindres carrés** (à nouveau) : soit $\min_x\{\|x\|_2 : \; Ax = b\}$.
Fonction de Lagrange $L(x,\lambda) = \|x\|_2 - \lambda^T(Ax-b)$, nous avons

$$\underline{L}(\lambda) = \inf_x [\|x\|_2 - \lambda^T(Ax-b)] = \begin{cases} b^T\lambda & \text{si } \|A^T\lambda\|_2 \leq 1 \\ -\infty & \text{sinon} \end{cases}$$

$\Rightarrow$ problème dual $\max_\lambda\{b^T\lambda : \; \|A^T\lambda\|_2 \leq 1\}$

• **Optimisation quadratique :** soit

$$\min_x \left\{\tfrac{1}{2}x^TPx + q^Tx : \; Ax \leq b, \; Cx = d\right\} \qquad [\text{réalisable, avec } P = P^T \succ 0]$$

Fonction de Lagrange $L(x,\lambda) = \frac{1}{2}x^TPx + q^Tx + \lambda^T(Ax-b) + \nu^T(d-Cx))$,

$$\nabla_x L(x,\lambda) = Px + q + A^T\lambda - C^T\nu, \quad x(\lambda) = P^{-1}(C^T\nu - A^T\lambda - q)$$

$\Rightarrow$ objectif dual

$$\underline{L}(\lambda) = -\tfrac{1}{2}(A^T\lambda - C^T\nu - q)^TP^{-1}(A^T\lambda - C^T\nu - q) - b^T\lambda + d^T\nu$$

$\Rightarrow$ problème dual

$$\max_{\lambda,\nu}\left\{-\tfrac{1}{2}(A^T\lambda - C^T\nu - q)^TP^{-1}(A^T\lambda - C^T\nu - q) - b^T\lambda + d^T\nu : \; \lambda \geq 0\right\}$$

ou encore $\max_{\lambda,\nu,t}\left\{-\frac{1}{2}t^TPt - b^T\lambda + d^T\nu : \; Pt = A^T\lambda - C^T\nu - q, \; \lambda \geq 0\right\}$

• **Problème de répartition :** soit

$$\mathrm{Opt}(P) = \min_x \left\{ x^T W x : \; x_i^2 = 1, \; i = 1, ..., n \right\}$$

— *problème non-convexe,* ensemble réalisable contient 2^n points $\{-1, 1\}^n$

— *interprétation :* répartir les elements de l'ensemble $\{1, ..., n\}$ en 2 sous-ensembles, W_{ij} étant le coût de mettre "i" et "j" dans le même ensemble, avec les coût $-W_{ij}$ de mettre "i" et "j" dans les ensembles différents

Fonction de Lagrange $L(x, \lambda) = x^T W x + \sum_i \lambda_i (x_i^2 - 1)$

$\Rightarrow$ objectif dual

$$\underline{L}(\lambda) = \inf_x \left[x^T (W + \mathrm{Diag}(\lambda)) x - \mathbf{1}^T \lambda \right] = \begin{cases} -1^T \lambda & \text{si } W + \mathrm{Diag}(\lambda) \succeq 0 \\ -\infty & \text{sinon} \end{cases}$$

$\Rightarrow$ problème dual

$$\mathrm{Opt}(D) = \max_\lambda \left\{ -1^T \lambda : \; W + \mathrm{Diag}(\lambda) \succeq 0 \right\}$$

Nous avons $\mathrm{Opt}(D) \leq \mathrm{Opt}(P)$.

Applications statistiques : régression linéaire

On suppose que les observations (b_i, a_i) sont liées par un modèle de régression linéaire :

$$b_i = a_i^T x_* + \xi_i, \quad i = 1, ..., m$$

ici

- $x_* \in \mathbb{R}^n$ est le paramètre vectoriel inconnu
- $\xi_i \in \mathbb{R}$ sont des bruits de mesure i.i.d., avec la densité p_ξ
- en écriture vectorielle, $b = Ax_* + \xi$, où A est la matrice avec des lignes a_i^T, $i = 1, ..., m$.

Estimateur de maximum de vraisemblance : on prend comme estimation de x_* une solution optimale de

$$\max_x \left\{ \ell(x) = \sum_{i=1}^{m} \log p_\xi(b_i - a_i^T x) \right\}$$

Exemples

- Loi normale $\mathcal{N}(0,\sigma^2)$: $p_\xi(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{z^2}{2\sigma^2}}$,

$$\ell(x) = -\frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^m (a_i^T x - b_i)^2,$$

et l'estimateur de ML est celui de *moindres carrés.*

- Loi de Laplace $\mathcal{L}(\tau)$: $p_\xi(z) = \frac{1}{2\tau} e^{-\frac{|z|}{\tau}}$,

$$\ell(x) = -m\log(2\tau) - \frac{1}{\tau}\sum_{i=1}^m |a_i^T x - b_i|,$$

et l'estimateur de ML *minimise la norme* ℓ_1 *des résidus.*

- Loi uniforme $U[-\tau,\tau]$: $p_\xi(z) = \frac{1}{2\tau} 1_{|z|\le\tau}$,

$$\ell(x) = \begin{cases} -m\log(2\tau) & \text{si } |a_i^T x - b_i| \le \tau,\ i = 1,...,m \\ -\infty & \text{sinon} \end{cases}$$

Pour trouver l'estimateur de ML on doit *trouver* x *qui satisfait* $|a_i^T x - b_i| \le \tau$, $i = 1,...,m$.

Problème des moindres carrés contraints

• Dans le cas du bruit normal, on cherche l'estimateur de x_* *qui satisfait la contrainte* $Cx = d$*.* On suppose que la matrice hessienne $A^T A$ *est inversible.* On doit résoudre le problème

$$\min_x \left\{ \tfrac{1}{2}(Ax-b)^T(Ax-b) : \; Cx = d \right\} \tag{P}$$

On remarque que le problème dual de (P) s'écrit

$$\max_\lambda -\tfrac{1}{2}(A^T b + C^T\lambda)^T (A^T A)^{-1}(A^T b + C^T\lambda) + \lambda^T d + b^T b, \tag{D}$$

et la solution optimale *(unique)* de (D) peut être calculée explicitement :

$$\lambda = (C(A^T A)^{-1}C^T)^{-1}(d - C(A^T A)^{-1}Ab).$$

Cela donne *l'estimateur de moindres carrés sous contraintes*

$$\begin{aligned} \widehat{x}_{CLS} &= (A^T A)^{-1}(A^T b + C^T\lambda) \\ &= \underbrace{(A^T A)^{-1}\Big(A^T b}_{\widehat{x}_{LS}} + \underbrace{C^T(C(A^T A)^{-1}C^T)^{-1}(d - C(A^T A)^{-1}Ab)}_{\text{correction de contrainte}}\Big) \end{aligned}$$

• *Régression de "'ridge"* consiste à imposer la contrainte $\|x\|_2 \leq r$ sur l'estimateur de moindres carrés :

$$\min_x \left\{(Ax-b)^T(Ax-b) : \ \|x\|_2 \leq r\right\} \qquad (C)$$

ou encore, considérer un estimateur pénalisé, la solution de

$$\min_x (Ax-b)^T(Ax-b) + \kappa x^T x \qquad (R)$$

L'estimateur pénalisé – la solution de (R) – s'écrit explicitement :

$$\widehat{x}_R = (A^TA + \kappa I)^{-1}A^Tb.$$

Par ailleurs, on remarque que la fonction de Lagrange du problème (C) s'écrit

$$L(x,\lambda) = (Ax-b)^T(Ax-b) + \lambda(x^Tx - r), \ \ x \in \mathbb{R}^n, \ \lambda \geq 0,$$

avec

$$\nabla_x L(x,\lambda) = \tfrac{1}{2}A^T(Ax-b) + \lambda x^T x, \quad x(\lambda) = (A^TA + \lambda I)^{-1}A^Tb$$

$$\min_x \left\{(Ax-b)^T(Ax-b) : \ \|x\|_2 \le r\right\} \qquad (C)$$

Maintenant il y a deux cas :

— soit $\lambda_* = 0$ (la contrainte correspondante n'est pas "active"), et

$$x_0 = (A^TA)^{-1}A^Tb \text{ satisfait } \|x_0\|_2 \le r.$$

Dans ce cas l'estimateur contraint $\widehat{x} = x_0$ coincide avec celui de moindres carrés ordinaires :

$$\widehat{x}_{LS} = (A^TA)^{-1}A^Tb$$

— soit $\|x_0\|_2 > r$, et on doit choisir $\lambda_* > 0$ tel que $\|x(\lambda_*)\|_2 = r$, avec l'estimateur contraint

$$\widehat{x}_C = (A^TA + \lambda_* I)^{-1}A^Tb$$

Estimateur de lasso [Hastie, Tibshirani, 1996]

$$\widehat{x}^{\text{lasso}} \in \operatorname*{Argmin}_{x} \left\{ \sum_{i=1}^{n} (b_i - a_i^T x)^2 \quad \text{s. c.} \quad \sum_{j=1}^{p} |x_j| \leq t \right\}$$

ou encore,

$$\widehat{x}^{\text{lasso}} \in \operatorname*{Argmin}_{x} \left\{ \sum_{i=1}^{n} (b_i - a_i^T x)^2 + \lambda \sum_{j=1}^{p} |x_j| \right\}$$

- par rapport au ridge : la pénalité $\|x\|_2^2 = \sum_{j=1}^{p} x_j^2$ est remplacé par $\|x\|_1 = \sum_{j=1}^{p} |x_j|$
- estimateur $\widehat{x}^{\text{lasso}}$ est non-linéaire
- quand $t \to \infty$ ($\lambda \to 0$) $\widehat{x}^{\text{lasso}} \to \widehat{x}^{\text{ls}}$, l'estimateur des moindres carrés ordinaires
- si $t \to 0$ (ou $\lambda \to \infty$), alors $\widehat{x}^{\text{lasso}} \to 0$, mais petite valeur de t (ou grande valeur de λ) cause *certains* des coefficients être exactement zéro
- Lasso est une (sorte de) procédure de sélection “continue” de support de $\widehat{x}$

Ridge, lasso, et sélection du support

- Régression ridge

$$\widehat{x}^{\text{ridge}} = \underset{x}{\operatorname{argmin}} \|b_i - Ax\|_2^2 + \lambda\|x\|_2^2$$

- Lasso

$$\widehat{x}^{\text{lasso}} \in \underset{x}{\operatorname{Argmin}} \|b - Ax\|_2^2 + \lambda\|x\|_1$$

- Sélection du meilleur support

$$\widehat{x}^{\text{sparse}} \in \operatorname{Argmin} \|b - Ax\|_2^2 + \lambda \underbrace{\sum_{j=1}^{p} I\{x_j \neq 0\}}_{\text{“norme” } \|x\|_0}$$

Ridge et (surtout) lasso sont deux alternatives “à coup numérique raisonnable” à la procedure difficile numériquement de sélection du meilleur sous-ensemble de prédicteurs.

Ridge et LASSO dans un cas particulier

Soit $n = m$ et $A = I$, une matrice identité, c.-à-d.

$$b_i = x_i + \xi_i,\ i = 1, ..., n.$$

- Estimateur de moindres carrés : $\widehat{x}^{\mathsf{ls}} = \operatorname{argmin}_x \sum_{i=1}^n (b_i - x_i)^2$,

$$\widehat{x}_i^{\mathsf{ls}} = b_i, \quad i = 1, \ldots, n.$$

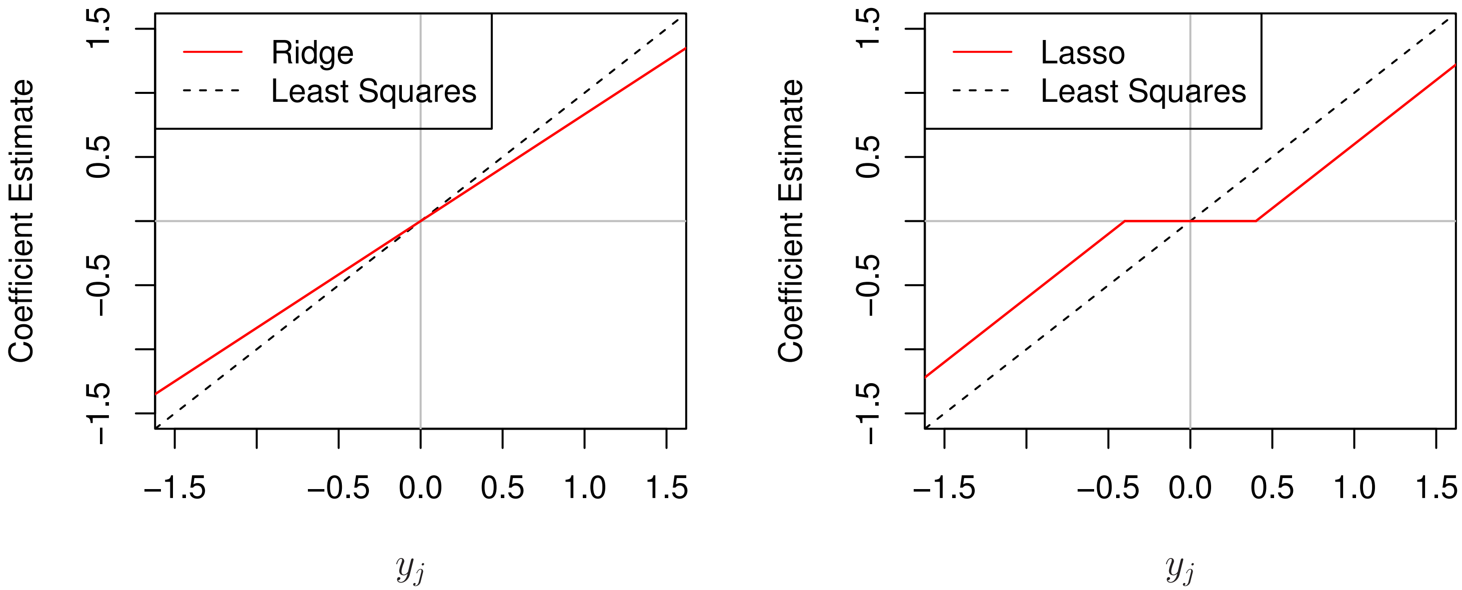
- Régression ridge : $\widehat{x}^{\mathsf{ridge}} = \operatorname{argmin}_x \sum_{i=1}^n (b_i - x_i)^2 + \lambda \sum_{i=1}^n x_i^2$,

$$\widehat{x}_i^{\mathsf{ridge}} = \frac{b_i}{1+\lambda}, \quad i = 1, \ldots, n.$$

- Lasso : $\widehat{x}^{\mathsf{lasso}} \in \operatorname{Argmin}_x \sum_{i=1}^n (b_i - x_i)^2 + \lambda \sum_{i=1}^n |x_i|$,

$$\widehat{x}_i^{\mathsf{lasso}} = \begin{cases} b_i - \lambda/2, & b_i > \lambda/2, \\ b_i + \lambda/2, & b_i < -\lambda/2, \\ 0, & |b_i| \leq \lambda/2 \end{cases} \quad i = 1, \ldots, n.$$

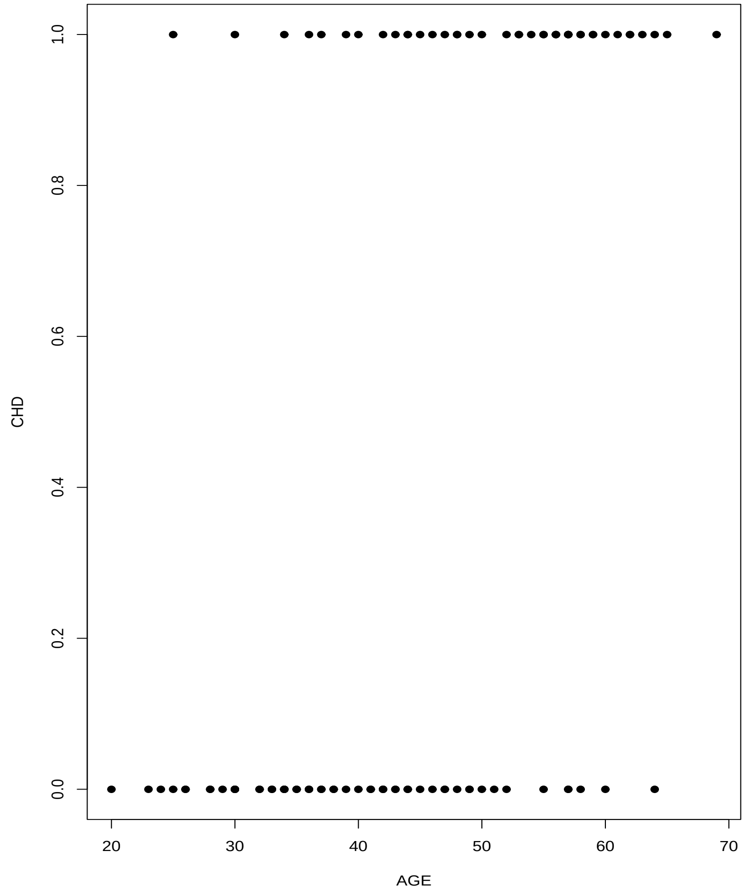
- $\widehat{x}^{\text{ridge}} = \text{ponderation}(b_i)$
- $\widehat{x}_i^{\text{lasso}} = \text{seuillage}(b_i)$



Modèle de régression logistique

Exemple Données de l'état de maladie coronarienne (CHD) et d'age : 100 sujets

réponse η - absence ou presence (0/1) de la CHD, prédicteur ζ - age.



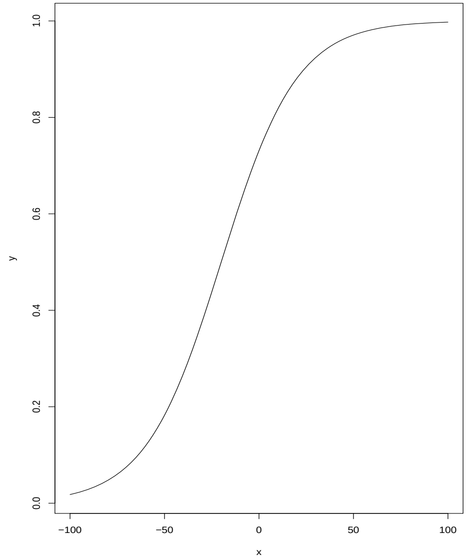
— Régression linéaire n'est pas appropriée :

$$E(\eta|\zeta = a) = P(\eta = 1|\zeta = a) = x_0 + x_1 a$$

doit être dans $[0, 1]$, pour tout a.

— L'idee est de modéliser la relation entre $p(a) = P(\eta = 1|\zeta = a)$ et a en utilisant la *fonction de réponse logistique* :

$$p(a) = \frac{e^{x_0 + x_1 a}}{1 + e^{x_0 + x_1 a}} \quad \Leftrightarrow \quad \mathrm{logit}\{p(a)\} := \log \frac{p(a)}{1 - p(a)} = x_0 + x_1 a.$$



Fonction de réponse logistique

Interprétation

— Il s'agit d'un cas spécial d'un *modèle linéaire généralisé* (GLM) avec la fonction de lien logit :

$$g(E(\eta|\zeta = a)) = x_0 + x_1 a, \quad g(z) = \log\frac{z}{1-z}, \quad 0 \le z < 1.$$

— Pourquoi logit ? Pour un a fixé, evidence, ou échelle des chances $\frac{p(a)}{1-p(a)}$ est naturellement logarithmique : d'habitude, on compte les chances comme '10 contre 1', ou '2 contre 1'.

$p(a) = 0.75 \Rightarrow$ *chances* d'avoir la CHD à l'age a sont 3 contre 1.

$a = 0 \Rightarrow$

$$\log\frac{p(0)}{1-p(0)} = x_0 \quad \Leftrightarrow \quad \frac{p(0)}{1-p(0)} = e^{x_0}.$$

Ainsi e^{x_0} peut être interprété comme niveau de référence

(surtout si zéro est dans la plage des données de la variable prédictive) ζ.

En augmentant a de 1, on multiplie les chances par e^{x_1}. Si $x_1 > 0$ alors $e^{x_1} > 1$ et les chances augmentent ; si $x_1 < 0$ alors les chances diminuent.

Fonction de vraisemblance

- Modèle et données : $\{(\eta_i, a_i),\ i = 1, \ldots, n\}$, $\eta_i \in \{0, 1\}$, i.i.d.

$$\pi_i = \pi(a_i) = P(\eta_i = 1|a_i) = E(\eta_i|a_i) = \frac{e^{x_0+x_1a_i}}{1+e^{x_0+x_1a_i}}, \quad i = 1, \ldots, n.$$

- Vraisemblance et $\log$–vraisemblance à maximiser par rapport à (x_0, x_1) :

$$\begin{aligned} L(x_0, x_1; D_n) &= \prod_{i=1}^{n} \pi_i^{\eta_i}(1-\pi_i)^{1-\eta_i} \\ &= \prod_{i=1}^{n} \left(\frac{e^{x_0+x_1a_i}}{1+e^{x_0+x_1a_i}}\right)^{\eta_i}\left(\frac{1}{1+e^{x_0+x_1a_i}}\right)^{1-\eta_i} \\ &= \prod_{i=1}^{n} \frac{e^{(x_0+x_1a_i)\eta_i}}{1+e^{x_0+x_1a_i}} \\ \log\{L(x_0, x_1; D_n)\} &= \sum_{i=1}^{n} \eta_i(x_0+x_1a_i) - \sum_{i=1}^{n} \log\Big\{1+e^{x_0+x_1a_i}\Big\}. \end{aligned}$$

Pas de solution analytique, mais une solution numérique comme solution d'un problème d'optimisation

$$\min_{x_0,x_1} \log\{L(x_0, x_1; D_n)\}$$

Plus généralement on considère le problème de classification, dans lequel on observe les couples (a_i, η_i), ou $a_i \in \mathbb{R}^n$ et $\eta_i \in \{0, 1\}$.

• On admet que *les étiquettes* (labels) η_i sont des realisations des v.a. indépendantes de loi de Bernoulli $B(p_i)$ de paramètre p_i qui depend de $a_i \in \mathbb{R}^n$ (lien logistique) :

$$p_i = \mathrm{Prob}\{\eta_i = 1\} = \frac{\exp(a_i^T x)}{1 + \exp(a_i^T x)}$$

où x est le paramètre à estimer à partir des observations.

• *Fonction log-vraisemblance* (on admet que $y_1 = ... = y_k = 1$ et $y_{k+1} = ... = y_m = 0$)

$$\begin{aligned}\ell(u, v) &= \log \left(\prod_{i=1}^{k} \frac{\exp(a_i^T x)}{1 + \exp(a_i^T x)} \prod_{i=k+1}^{m} \frac{1}{1 + \exp(a_i^T x)} \right) \\ &= \sum_{i=1}^{k} a_i^T x - \sum_{i=1}^{m} \log(1 + \exp(a_i^T x))\end{aligned}$$

est *concave en x.*

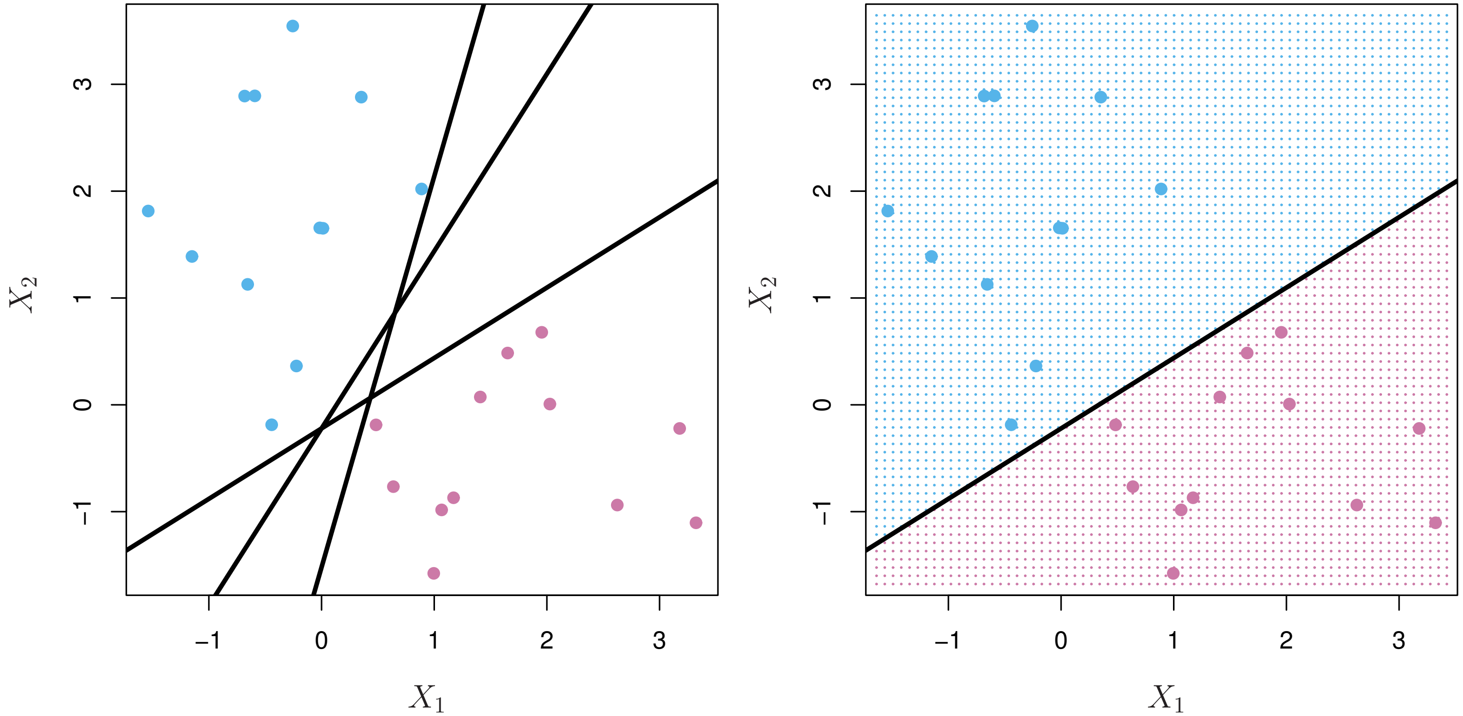
Machine à vecteur de support

On considère un *problème de classification (binaire)* avec les données (a_i, ℓ_i), $i = 1, \ldots, m$, où $a_i \in \mathbb{R}^n$ et $\ell_i \in \{-1, 1\}$.

• On dit que l'échantillon admet une séparation linéaire si il existe un *hyperplan de séparation* $f(a) := a^T u + v = 0$ tel que

$$v + a_i^T u \geq 0 \text{ si } \ell_i = 1, \quad \text{et} \quad v + a_i^T u < 0 \text{ si } \ell_i = -1.$$

Si $f(a) = 0$ est un plan de séparation alors un classifieur "naturel" est $\mathrm{sign}\{f(a)\}$.



Remarque : un plan de separation satisfait $\ell_i(v + u^T a_i) \geq 0$, $\forall i$.

Classifieur à marge maximale

— Si l'ensemble de données admet une séparation linéaire, il est naturel de chercher l'hyperplan de séparation *à marge maximale*, c.-à-d., *l'hyperplan de séparation le plus éloigné des observations*.

— *Problème d'optimisation :*

$$\begin{aligned} & \min_{u,v} \left\{ \tfrac{1}{2} u^T u : \ \ell_i(v + u^T a_i) \geq 1, \ i = 1, \ldots, m \right\} \\ = \ & \min_{u,v} \left\{ \tfrac{1}{2} u^T u : \ \Lambda(\mathbf{1}v + Au) \geq \mathbf{1} \right\} \end{aligned} \qquad (P_0)$$

$\|u\|_2^{-1}$ étant la marge de séparation, $\Lambda = \mathrm{Diag}(\ell_i)$, et A la matrice avec les lignes a_i^T.

- Le problème d'optimisation *(P_0) est convexe.*
- On appelle également ce classifieur *hard margin classifier* (classifieur à marge dure)

Une reformulation

- On écrit le problème dual de (P$_0$) (avec $\lambda \geq 0$) :

$$L(u, v; \lambda) = \tfrac{1}{2} u^T u - \lambda^T (\Lambda(\mathbf{1} v + A u) - \mathbf{1}),$$

avec

$$\begin{array}{lll} \nabla_u L(u, v; \lambda) = u - A^T \Lambda \lambda, & \Rightarrow & u(\lambda) = A^T \Lambda \lambda, \\ L'_v(u, v; \lambda) = \Lambda \mathbf{1} := \ell, & \Rightarrow & \ell^T \lambda = 0. \end{array}$$

$\Rightarrow$ problème dual de (D$_0$) :

$$\begin{aligned} & \max_\lambda \left\{ -\tfrac{1}{2} \lambda^T \Lambda A A^T \Lambda \lambda + \mathbf{1}^T \lambda : \ \ell^T \lambda = 0, \ \lambda \geq 0 \right\} \\ = \ & -\min_\lambda \left\{ \tfrac{1}{2} \textstyle\sum_{i,j=1}^m \lambda_i \lambda_j \ell_i \ell_j a_i^T a_j - \sum_{i=1}^m \lambda_i : \ \begin{array}{l} \sum_{i=1}^m \lambda_i \ell_i = 0, \\ \lambda_i \geq 0, \quad \forall i \end{array} \right\} \quad (D_0) \end{aligned}$$

Proposition *Soit $[u^*; v^*]$ une solution optimale de (P$_0$). Si λ^* est une solution optimale duale, alors*

$$u^* = A^T \Lambda \lambda^* = \sum_{i=1}^m \ell_i \lambda_i^* a_i,$$

et pour tout k tel que $\lambda_k > 0$

$$v^* = \ell_k - a_k^T u^* = \ell_k - \textstyle\sum_{i=1}^m \lambda_i^* \ell_i a_i^T a_k.$$

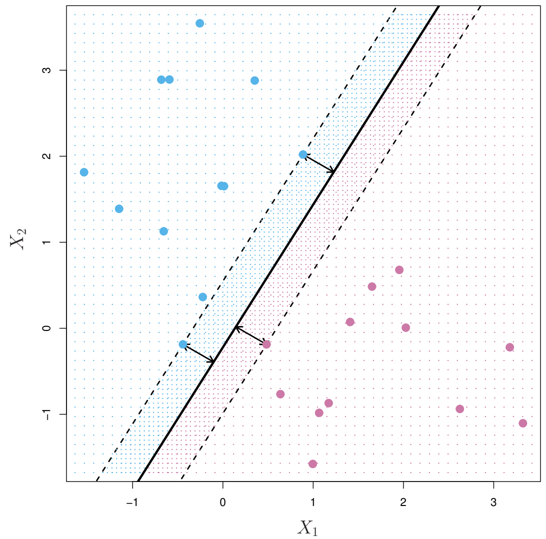
Remarques

— *solution duale creuse :* la condition de complémentarité implique que λ^* et (u^*, v^*) satisfont

$$\lambda_i^* \Big\{ \ell_i [a_i^T u^* + v^*] - 1 \Big\} = 0, \ \forall i = 1, \ldots, m.$$

Autrement dit, seuls les vecteurs a_i pour lesquels $a_i^T u^* + v^* = \ell_i$ correspondent à $\lambda_i^* > 0$, les autres λ_i^* sont nulls.

On appelle ces a_i *vecteurs de support (support vectors)*



— sensibilité – une seule observation peut modifier significativement la solution.

— *Et si l'hyperplan de separation n'existait pas ?*

Classifieur "à marge douce"

- *L'idée :* admettre des individus mal classés – imposer une *marge douce (soft margin)*.
- Problème d'optimisation

$$\begin{aligned}
& \min_{v,u,\epsilon}\left\{\tfrac{1}{2}u^Tu + C\textstyle\sum_{i=1}^m \epsilon_i : \begin{array}{l} \ell_i(v+a_i^Tu) \geq 1-\epsilon_i, \\ \epsilon_i \geq 0,\ i=1,\dots,m \end{array}\right\} \\
= \ & \min_{v,u,\epsilon}\left\{\tfrac{1}{2}u^Tu + C\mathbf{1}^T\epsilon :\ \Lambda(v\mathbf{1}+Au) \geq \mathbf{1}-\epsilon,\ \epsilon \geq 0\right\} \quad (P_1)
\end{aligned}$$

  où $\epsilon = [\epsilon_1; \dots; \epsilon_n]$ est vecteur des variables d'écart (slacks), et $C \geq 0$ est un paramètre d'ajustement.

- *Variable d'écart (slack)* ϵ_i nous dit où se trouve la i-ème observation :
  - $\epsilon_i = 0$: i-ème observation est "de bon cote" de la marge
  - $\epsilon_i > 0$: i-ème observation viole la marge
  - $\epsilon_i > 1$: i-ème observation est "de mauvais coté" (mal classée).
- *Paramètre C à choisir* établit une pénalité pour la violation de la marge

Formulation duale

- On écrit le dual de (P_1) (avec $\lambda, \nu \geq 0$) :

$$L(u, v, \epsilon; \lambda, \nu) = \tfrac{1}{2}u^T u + C\mathbf{1}^T \epsilon - \lambda^T(\Lambda(\mathbf{1}v + Au) - \mathbf{1} + \epsilon) - \nu^T \epsilon,$$

avec

$$\begin{array}{lcl} \nabla_u L(u, v, \epsilon; \lambda) = u - A^T \Lambda \lambda & \Rightarrow & u(\lambda) = A^T \Lambda \lambda, \\ L'_v(u, v, \epsilon; \lambda) = \Lambda \mathbf{1} := \ell & \Rightarrow & \ell^T \lambda = 0, \\ \nabla_\epsilon L'_v(u, v, \epsilon; \lambda) = C\mathbf{1} - \lambda - \nu & \Rightarrow & \lambda + \nu = C\mathbf{1}. \end{array}$$

$\Rightarrow$ *problème dual :*

$$\min_\lambda \left\{ -\tfrac{1}{2}\lambda^T \Lambda A A^T \Lambda \lambda + \mathbf{1}^T \lambda : \ \ell^T \lambda = 0, \ \lambda, \nu \geq 0, \ \lambda + \nu = C\mathbf{1} \right\},$$

et, en éliminant ν, on arrive à

$$\min_\lambda \left\{ -\tfrac{1}{2}\lambda^T \Lambda A A^T \Lambda \lambda + \mathbf{1}^T \lambda : \ \ell^T \lambda = 0, \ 0 \leq \lambda \leq C\mathbf{1} \right\} \qquad (D_1)$$

$$= \ -\min_\lambda \left\{ \tfrac{1}{2} \sum_{i,j=1}^m \lambda_i \lambda_j \ell_i \ell_j a_i^T a_j - \sum_{i=1}^m \lambda_i : \begin{array}{l} \sum_{i=1}^m \lambda_i \ell_i = 0, \\ 0 \leq \lambda_i \leq C, \ \forall i \end{array} \right\}$$

— Avantage principal d'une fonction de pénalité linéaire est que les variables de slack disparaissent du problème dual ;

— Si λ^* est une solution optimale duale, alors la solution optimale primal u^* est donnée par $u^* = A^T \Lambda \lambda^* = \sum_{i=1}^n \ell_i \lambda_i^* a_i$, avec $\lambda_i^* > 0$ seulement pour les observations i t.q.

$$\ell_i(a_i^T u^* + v^*) = 1 - \epsilon_i \leq 1$$

Les a_i correspondants sont les *vecteurs de support* dans le cas d'un *classifieur à marge douce.*

Les solutions duales $0 < \lambda_i^* < C$ correspondent aux vecteurs de support a_i sur les "bords de la marge" (avec $\epsilon_i = 0$) ; si a_i viole la marge ($\epsilon_i > 0$), nous avons $\lambda_i^* = C$.

— *Le classifieur*

$$g(a) = \mathrm{sign}\{a^T u^* + v^*\} = \mathrm{sign}\Big\{ \sum_{i=1}^n \ell_i \lambda_i^* a_i^T a + v^* \Big\}.$$

ne nécessite pas de calcule explicite de u^*, seule les produits $a_i^T a$ sont utilisés $\Rightarrow$ on peut faire les calculs pour un n "très grand" (l'idee du "Kernel trick").