Optimisation et apprentissage statistique

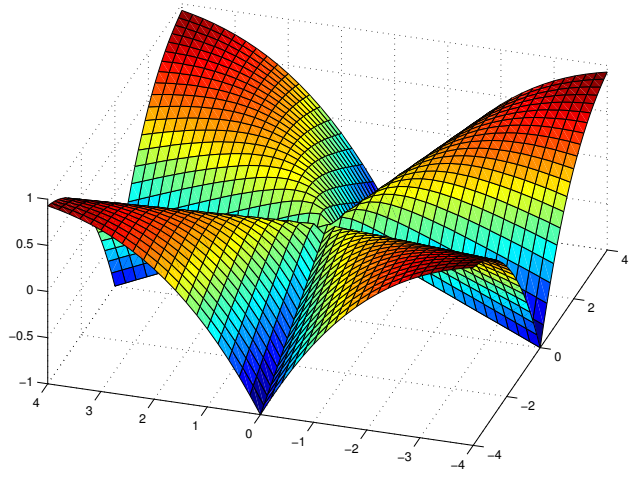
Une introduction à l'optimisation

Stéphane Canu
asi.insa-rouen.fr/enseignants/~scanu

Ecole d'été BasMatI 2015, Porquerolles

June 4, 2015

Plan



Unconstrained minimization problem

$$\begin{array}{rccc} J: & \mathbb{R}^p & \longrightarrow & \mathbb{R} \\ & \mathbf{x} & \longmapsto & J(\mathbf{x}) \end{array}$$

Unconstrained minimization

$$\min_{\mathbf{x}\in\mathbb{R}^p} J(\mathbf{x})$$

Example (Lasso)

$$\min_{\mathbf{x}\in\mathbb{R}^p} \underbrace{\frac{1}{2}\|A\mathbf{x}-\mathbf{y}\|^2 + \lambda \sum_{j=1}^{p}(|x_j|)}_{J(\mathbf{x})}$$

The program:

$$\min_{\mathbf{x}\in\mathbb{R}^p} \; J(\mathbf{x})$$

The program:

- existence of optimal solutions,
- characterization of optimal solutions,
- algorithms for computing optimal solutions: how to find a descent direction (if it exists)

The concept of derivative is fundamental

Existence of a solution

Definition (local minima)

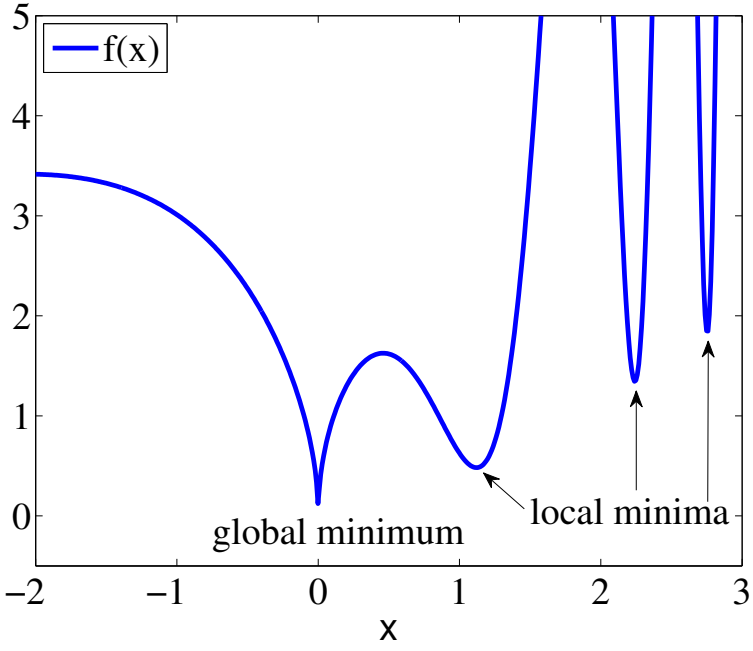
$\mathbf{x}^\star$ is a local minima for J if there exists $\varepsilon > 0$ such that

$$J(\mathbf{x}^\star) \leq J(\mathbf{x}) \qquad \forall \mathbf{x} \text{ with } \|\mathbf{x} - \mathbf{x}^\star\| \leq \varepsilon$$

Definition (global minimum)

$\mathbf{x}^\star$ is a global minimum for f if

$$J(\mathbf{x}^\star) \leq J(\mathbf{x}) \qquad \forall \mathbf{x} \in \mathbb{R}^p$$



The question of the existence of a global minima to this unconstrained minimization problem require some definition.

Definition (l.s.c.)

a real valued function f is lower semi-continuous (l.s.c.) at some point $\mathbf{x}$ if for every sequence $\{\mathbf{x}_k, k \in \mathbb{N}\}$ that converges to $\mathbf{x}$

$$J(\mathbf{x}) \leq \liminf_{k \to \infty} J(\mathbf{x}_k).$$

The 0/1 loss function, the counting function and the rank function are l.s.c.

Definition (coercive)

a real valued function J is coercive if for every sequence $\{\mathbf{x}_k, k \in \mathbb{N}\}$ such that $\lim_{k \to \infty} \|\mathbf{x}_k\| = \infty$

$$\lim_{k \to \infty} J(\mathbf{x}_k) = \infty.$$

Proposition

Weierstrass' theorem: existence of a solution (sufficient condition). If J is l.s.c. and coercive or admits a non empty bounded level set, then there exists a global minimizer for J.

Gradient

The gradient is a generalization of the usual concept of derivative of a function in one dimension to a function in several dimensions.

$$\begin{array}{rccc} J: & \mathbb{R}^p & \longrightarrow & \mathbb{R} \\ & \mathbf{x} & \longmapsto & J(\mathbf{x}) \end{array}$$

Assume J is differentiable in the sense that all its partial derivatives $\frac{\partial J}{\partial w_i}$ exists.

Definition (gradient)

The gradient $\nabla J(\mathbf{x})$ of a function J at point $\mathbf{x}$ is the vector whose components are the partial derivatives of J

Example

The least square

$$J_1(\mathbf{x}) = \|A\mathbf{x} - y\|^2 \qquad \nabla J_1(x) = 2A^t(A\mathbf{x} - y)$$

Ensemble de niveau

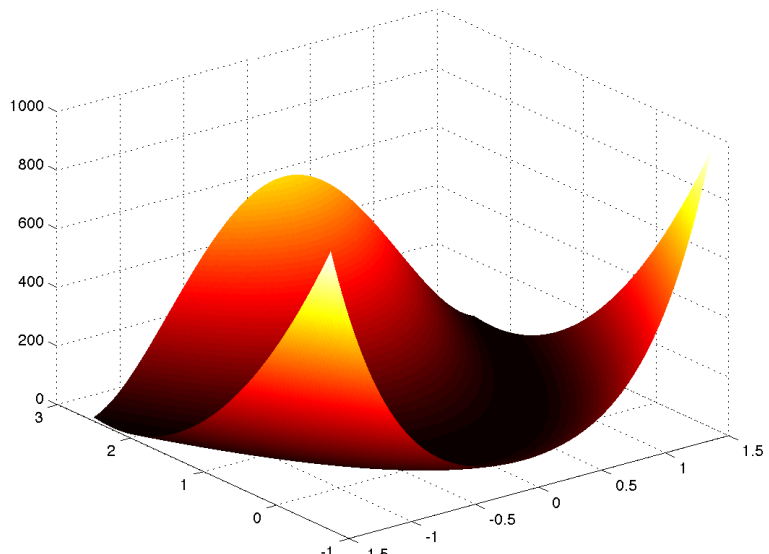
Definition (Ensemble de niveau (*level set*, iso contour, courbe de niveau...))

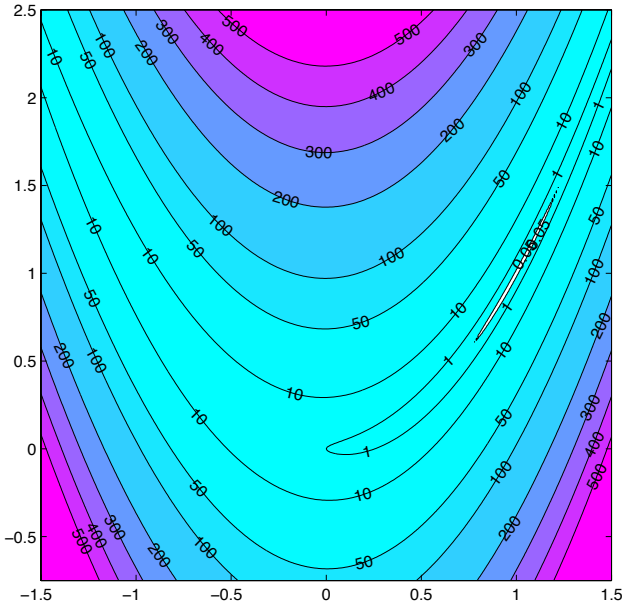
On appelle ensemble de niveau de J de $\mathbb{R}^n$ à valeur dans $\mathbb{R}$ (p=1) pour le niveau c l'ensemble :

$$N_c = \{\mathbf{x} \in \mathbb{R}^n \mid J(\mathbf{x}) = c\}$$

Exemple :

$$J_r(x, y) = (1 - x)^2 + 105(y - x^2)^2$$





Théorème

Theorem (Gradient)

S'il existe, le gradient de J au point $\mathbf{x}_0$ est perpendiculaire à l'ensemble de niveau $N_{J(\mathbf{x}_0)}$. Donc

$$\nabla J(\mathbf{x}_0)^\top \mathbf{v} = 0$$

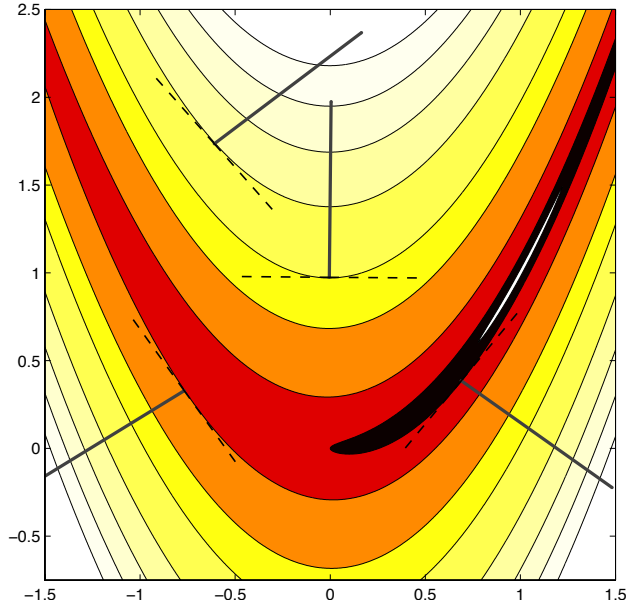
pour tous les vecteurs v tangents à J au point $\mathbf{x}_0$.

Exemples :

$$J_\ell(x, y) = ax + by$$
$$\nabla J_\ell(x, y) = (a, b)^\top$$

$$J_c(x, y, z) = x^2 + 2y^2 + z^2$$
$$\nabla J_c(x, y) = (2x, 4y, 2z)^\top$$

$$J_r(x, y) = (1 - x)^2 + 105(y - x^2)^2$$



dérivée directionnelle

Definition (dérivée directionnelle)

On appelle dérivée directionnelle de J au point $\mathbf{x}$ et dans la direction $\mathbf{d} \in \mathbb{R}^n$ la limite :

$$D_{\mathbf{x}} J(\mathbf{x}, \mathbf{d}) = \lim_{\varepsilon \to 0} \frac{J(\mathbf{x} + \varepsilon \mathbf{d}) - J(\mathbf{x})}{\varepsilon}$$

si elle existe

Exemples de dérivées directionnelles

soit J une fonction de $\mathbb{R}^n$ à valeur dans $\mathbb{R}^p$

$$J_1(\mathbf{x}) = \mathbf{b}^\top \mathbf{x} \qquad p = 1 \qquad D_{\mathbf{x}} J(\mathbf{x}, \mathbf{d}) = \mathbf{b}^\top \mathbf{d}$$

$$J_2(\mathbf{x}) = A\mathbf{x} - b \qquad p = n \qquad D_{\mathbf{x}} J(\mathbf{x}, \mathbf{d}) = A\mathbf{d}$$

$$J_3(f) = \|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i| \qquad p = 1 \qquad D_{\mathbf{x}} J(\mathbf{x}, \mathbf{d}) = \text{signe}(\mathbf{x})^\top \mathbf{d}$$

$$\ldots \qquad J_7(f) = \int_0^1 f(t)^2 dt \qquad n = p = \infty \qquad D_{\mathbf{x}} J(\mathbf{x}, \mathbf{d}) = 2 \int_0^1 f(t) g(t) dt$$

une manière commode de calculer cette dérivée est d'utiliser la définition suivante : $\varphi(\varepsilon) = J(\mathbf{x} + \varepsilon \mathbf{d})$

Theorem (Calcul pratique de la dérivée)

soit J une fonction de $\mathbb{R}^n$ à valeur dans $\mathbb{R}$:

$$D_{\mathbf{x}} J(\mathbf{x}, \mathbf{d}) = \left. \frac{dJ(\mathbf{x} + \varepsilon \mathbf{d})}{d\varepsilon} \right|_{\varepsilon=0} = \varphi'(0)$$

Démonstration :

$$\begin{aligned} \varphi'(0) &= \lim_{\epsilon \to 0} \frac{\varphi(\epsilon) - \varphi(0)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{J(\mathbf{x} + \epsilon \mathbf{d}) - J(\mathbf{x})}{\epsilon} \quad = D_{\mathbf{x}} J(\mathbf{x}, \mathbf{d}) \end{aligned}$$

une manière commode de calculer cette dérivée est d'utiliser la définition suivante : $\varphi(\varepsilon) = J(\mathbf{x} + \varepsilon\mathbf{d})$

Theorem (Calcul pratique de la dérivée)

soit J une fonction de $\mathbb{R}^n$ à valeur dans $\mathbb{R}$:

$$D_{\mathbf{x}}J(\mathbf{x},\mathbf{d}) = \left.\frac{dJ(\mathbf{x}+\varepsilon\mathbf{d})}{d\varepsilon}\right|_{\varepsilon=0} = \varphi'(0)$$

Démonstration :

$$\begin{aligned}\varphi'(0) &= \lim_{\epsilon\to 0}\frac{\varphi(\epsilon)-\varphi(0)}{\epsilon} \\ &= \lim_{\epsilon\to 0}\frac{J(\mathbf{x}+\epsilon\mathbf{d})-J(\mathbf{x})}{\epsilon} \quad = D_{\mathbf{x}}J(\mathbf{x},\mathbf{d})\end{aligned}$$

recette :

1. calculer $J(\mathbf{x} + \varepsilon\mathbf{d})$
2. calculer la dérivée par rapport à ε
3. prendre la valeur de la dérivée en $\varepsilon = 0$

Exemples de dérivées directionnelles

Exemples

$$J_4(\mathbf{x}) = \|\mathbf{x}\|^2 \qquad D_{\mathbf{x}}J(\mathbf{x},\mathbf{d}) = 2\mathbf{x}^\top\mathbf{d}$$

$$J_5(\mathbf{x}) = \|A\mathbf{x}-b\|^2 \qquad D_{\mathbf{x}}J(\mathbf{x},\mathbf{d}) = 2(A\mathbf{x}-b)^\top A\mathbf{d}$$

$$J_6(\mathbf{x}) = \|\mathbf{x}\|_2 \qquad D_{\mathbf{x}}J(\mathbf{x},\mathbf{d}) = \frac{\mathbf{x}^\top\mathbf{d}}{\|\mathbf{x}\|}$$

$$\begin{aligned}\varphi_4(\varepsilon) &= \|\mathbf{x}+\varepsilon\mathbf{d}\|^2 \\ &= \|\mathbf{x}\|^2 + 2\varepsilon\mathbf{x}^\top\mathbf{d} + \|\varepsilon\mathbf{d}\|^2 \\ &= \underbrace{a}_{\|\mathbf{x}\|^2} + \underbrace{b}_{2\mathbf{x}^\top\mathbf{d}}\varepsilon + \underbrace{c}_{\|\mathbf{d}\|^2}\varepsilon^2\end{aligned}$$

$$\varphi_4'(\varepsilon) = b + 2c\varepsilon \quad \Rightarrow \quad \varphi_4'(0) = b \quad \Leftrightarrow \quad D_{\mathbf{x}}J_4(\mathbf{x},\mathbf{d}) = 2\mathbf{x}^\top\mathbf{d}$$

Dérivée directionnelle et gradient

Theorem (p=1)

Si, pour $\mathbf{x}$ *fixé, l'application* $\mathbf{d} \to D_{\mathbf{x}}J(\mathbf{x},\mathbf{d})$ *est linéaire continue,*

$$D_{\mathbf{x}}J(\mathbf{x},\mathbf{d}) = \nabla J_{\mathbf{x}}(\mathbf{x})^{\top}\mathbf{d}$$

Calcul pratique du gradient :

1. calculer $J(\mathbf{x}+\varepsilon\mathbf{d})$
2. calculer la dérivée par rapport à ε
3. prendre la valeur de la dérivée en $\varepsilon = 0$
4. écrire $\varphi'(0)$ sous la forme $\nabla J_{\mathbf{x}}(\mathbf{x})^{\top}\mathbf{d}$ et identifier le gradient

Exemple de calcul de dérivée directionnelle

$$
\begin{aligned}
J(\mathbf{x}) &= \sum_{i=1}^{n} y_i A_i^\top \mathbf{x} - \sum_{i=1}^{n} \log\left(1 + \exp^{A_i^\top \mathbf{x}}\right) \\
&= \mathbf{y}^\top A\mathbf{x} - \mathbf{e}^\top \log\left(\mathbb{1} + \exp^{A\mathbf{x}}\right) \\
J(\mathbf{x} + \varepsilon\mathbf{d}) &= \mathbf{y}^\top A(\mathbf{x} + \varepsilon\mathbf{d}) - \mathbf{e}^\top \log\left(\mathbb{1} + \exp^{A(\mathbf{x}+\varepsilon\mathbf{d})}\right) \\
&= \mathbf{y}^\top A\mathbf{x} + \varepsilon\mathbf{y}^\top A\mathbf{d} - \mathbf{e}^\top \log\left(\mathbb{1} + \exp^{A\mathbf{x}} \exp^{\varepsilon A\mathbf{d}}\right)
\end{aligned}
$$

pour $\mathbf{x}$ et $\mathbf{d}$ fixés :

$$
\begin{aligned}
\varphi(\varepsilon) &= \mathbf{y}^\top A\mathbf{x} + \varepsilon\mathbf{y}^\top A\mathbf{d} - \mathbf{e}^\top \log\left(\mathbb{1} + \exp^{A\mathbf{x}} \exp^{\varepsilon A\mathbf{d}}\right) \\
&= \mathbf{y}^\top A\mathbf{x} + \varepsilon\mathbf{y}^\top A\mathbf{d} - \sum_{i=1}^{n} \log\left(1 + \exp^{A_i^\top \mathbf{x}} \exp^{\varepsilon A_i^\top \mathbf{d}}\right) \\
\varphi'(\varepsilon) &= \mathbf{y}^\top A\mathbf{d} - \sum_{i=1}^{n} \frac{\exp^{A_i^\top \mathbf{x}} \exp^{\varepsilon A_i^\top \mathbf{d}} A_i^\top \mathbf{d}}{1 + \exp^{A_i^\top \mathbf{x}} \exp^{\varepsilon A_i^\top \mathbf{d}}} \\
\varphi'(0) &= \mathbf{y}^\top A\mathbf{d} - \sum_{i=1}^{n} \frac{\exp^{A_i^\top \mathbf{x}} A_i^\top \mathbf{d}}{1 + \exp^{A_i^\top \mathbf{x}}} \\
&= \mathbf{y}^\top A\mathbf{d} - \frac{\exp^{A\mathbf{x}}}{\mathbb{1}+\exp^{A\mathbf{x}}}^\top A\mathbf{d} \quad = \underbrace{\left(A^\top \mathbf{y} - A^\top \frac{\exp^{A\mathbf{x}}}{\mathbb{1} + \exp^{A\mathbf{x}}}\right)}_{\nabla_{\mathbf{x}} J(\mathbf{x})}{}^\top \mathbf{d} \\
\nabla_{\mathbf{x}} J(\mathbf{x}) &= A^\top(\mathbf{y} - \mathbf{p}) \qquad \text{avec} \quad p_i = \frac{\exp^{A_i^\top \mathbf{x}}}{1+\exp^{A_i^\top \mathbf{x}}}
\end{aligned}
$$

Règles de calcul pour la dérivée

c'est un opérateur linéaire :

$$\nabla(J_1 + \alpha J_2)_{\mathbf{x}}(\mathbf{x}) = \nabla_{\mathbf{x}} J_1(\mathbf{x}) + \alpha \nabla_{\mathbf{x}} J_2(\mathbf{x})$$

Combinaison d'applications :
exemples

- $J_\alpha(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} - \mathbf{x}^\top \mathbf{b}$
- $J_\beta(\mathbf{x}) = \mathbb{1}^\top \log(A\mathbf{x})$
- $J_\gamma(\mathbf{x}) = f(\mathbf{a}^\top \mathbf{x})$

$$\nabla_{\mathbf{x}} J_\gamma(\mathbf{x}) = f'(\mathbf{a}^\top \mathbf{x})\mathbf{a}$$

Développement limité au premier ordre

Propriété : développement au premier ordre

si $\|\mathbf{d}\| = 1$ et $\varepsilon > 0$,

$$J(\mathbf{x} + \varepsilon\mathbf{d}) = J(\mathbf{x}) + \varepsilon \underbrace{\nabla_x J(x)^\top \mathbf{d}}_{DJ_{\mathbf{x}}(\mathbf{x},\mathbf{d})} + o(\varepsilon)$$

Application dans le cas p=1 :
si on cherche une direction de descente $\mathbf{d}$ (qui fait diminuer J)

$$J(\mathbf{x} + \varepsilon\mathbf{d}) < J(\mathbf{x})$$

On peut choisir $\mathbf{d} = -\nabla_x J(x)$ (ce n'est pas la seule solution)

Démonstration

$$J(\mathbf{x} + \varepsilon\mathbf{d}) = J(\mathbf{x} - \varepsilon\nabla_x J(\mathbf{x})) = J(\mathbf{x}) - \varepsilon\nabla_x J(x)^\top \nabla_x J(x) + o(\varepsilon) = J(\mathbf{x}) - \underbrace{\varepsilon\|\nabla_x J(x)\|^2}_{>0} + o(\varepsilon)$$

Dérivée seconde

Definition (Matrice Hessienne)

C'est la matrice des dérivée seconde de l fonctionnelle J, si elle existe :

$$H_J(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 J(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 J(\mathbf{x})}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 J(\mathbf{x})}{\partial x_j \partial x_1} & \cdots & \frac{\partial^2 J(\mathbf{x})}{\partial x_n \partial x_1} \\ \frac{\partial J(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 J(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 J(\mathbf{x})}{\partial x_j \partial x_2} & \cdots & \frac{\partial^2 J(\mathbf{x})}{\partial x_n \partial x_2} \\ \vdots & \ddots & & & & \\ \frac{\partial^2 J(\mathbf{x})}{\partial x_i \partial x_1} & \cdots & \cdots & \frac{\partial^2 J(\mathbf{x})}{\partial x_j \partial x_i} & \cdots & \frac{\partial^2 J(\mathbf{x})}{\partial x_n \partial x_i} \\ \vdots & \ddots & & & & \\ \frac{\partial^2 J(\mathbf{x})}{\partial x_n \partial x_1} & \cdots & \cdots & \frac{\partial^2 J(\mathbf{x})}{\partial x_j \partial x_n} & \cdots & \frac{\partial^2 J(\mathbf{x})}{\partial x_n \partial x_n} \end{pmatrix}$$

Exemples de dérivée seconde

Exemples

$J_2(\mathbf{x}) = A\mathbf{x} - b$	$\nabla_{\mathbf{x}} J(\mathbf{x}) = A$	$H_{\mathbf{x}}(\mathbf{x}) = 0$
$J_3(\mathbf{x}) = \|\mathbf{x}\|_1$	$\nabla_{\mathbf{x}} J(\mathbf{x}) = \text{sign}(\mathbf{x})$	$H_{\mathbf{x}}(\mathbf{x}) = 0$
$J_4(\mathbf{x}) = \|\mathbf{x}\|^2$	$\nabla_{\mathbf{x}} J(\mathbf{x}) = 2\mathbf{x}$	$H_{\mathbf{x}}(\mathbf{x}) = 2I$
$J_5(\mathbf{x}) = \|A\mathbf{x} - b\|^2$	$\nabla_{\mathbf{x}} J(\mathbf{x}) = 2A^\top(A\mathbf{x} - b)$	$H_{\mathbf{x}}(\mathbf{x}) = 2A^\top A$
$J_6(\mathbf{x}) = \|\mathbf{x}\|_2$	$\nabla_{\mathbf{x}} J(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|}$	$H_{\mathbf{x}}(\mathbf{x}) = \frac{1}{\|\mathbf{x}\|^2} M(\mathbf{x})$
$J_8(f) = \int_0^1 f(t)^2 dt$	$DJ_f(f) = f$	$H_f(f) = I$

Développement au second ordre

Soit $J : \mathbb{R}^n \longrightarrow \mathbb{R}$

Propriété : développement de Taylor au second ordre

$$J(\mathbf{x}+\mathbf{d}) = J(\mathbf{x}) + \nabla_{\mathbf{x}} J(\mathbf{x})\mathbf{d} + \frac{1}{2}\mathbf{d}^\top H_{\mathbf{x}}(\mathbf{x})\mathbf{d} + o(\|\mathbf{d}\|^2)$$

Application : si au point $\mathbf{x}$ on a $\nabla_{\mathbf{x}} J(\mathbf{x}) = 0$ et $H_{\mathbf{x}} J(\mathbf{x})$ définie positive, alors $\mathbf{x}$ est un minimum local de J.

ce résultat peut être généralisé

Exemple

$$J(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} - \mathbf{b}^\top \mathbf{x}$$

$$\nabla_{\mathbf{x}} J(\mathbf{x}) = A\mathbf{x} - \mathbf{b} \qquad\qquad H_{\mathbf{x}}(\mathbf{x}) = A$$

Si A est définie positive la solution du problème d'optimisation est $x^\star = A^{-1}\mathbf{b}$ c'est un minimum global

Si A n'est pas définie positive...

Sous-gradient

Definition

Sous-gradient On appelle sous gradient de J au point $\mathbf{x}_0$ tout vecteur $\mathbf{g} \in \mathbb{R}^n$ vérifiant

$$\forall \mathbf{x} \in \mathcal{V}(x_0), \qquad J(\mathbf{x}) \geq J(\mathbf{x}_0) + \mathbf{g}^\top (\mathbf{x} - \mathbf{x}_0)$$

Definition

sous-différentielle On appelle sous différentielle de J au point $\mathbf{x}_0$ l'ensemble de tous les sous-gradient de J au point $\mathbf{x}_0$. On le note $\partial J(\mathbf{x}_0)$

exemple

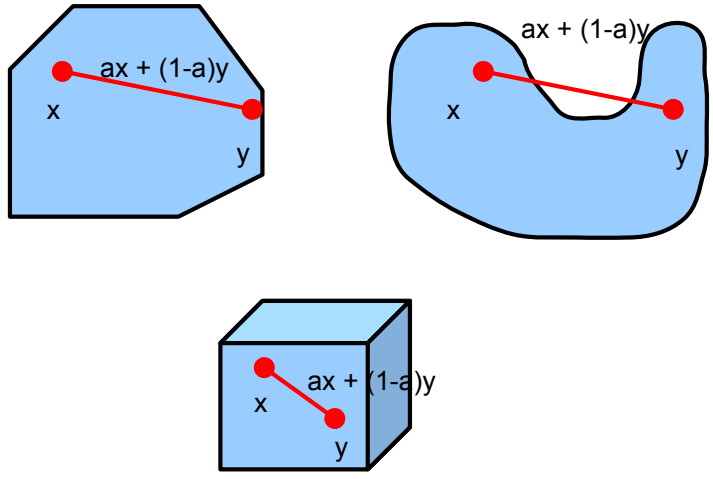
$$J_5(\mathbf{x}) = \sum_{i=1}^{d} |x_i| \qquad J_6(\mathbf{x}) = \sum_{i=1}^{d} |x_i| \mathbb{1}_{x_i<0}$$

Ensemble convexe

Definition (ensemble convexe)

un ensemble Ω est dit convexe si, pour tout $\mathbf{x}, \mathbf{y} \in \Omega$ on a

$$\forall a \in]0,1[, \quad a\mathbf{x} + (1-a)\mathbf{y} \in \Omega$$



exemples : convexes ou non convexes ?

- $\{\mathbf{x} \,|\|\mathbf{x}\|^2 < k\}$
- $\{\mathbf{x} \,|A\mathbf{x} \leq \mathbf{b}\}$
- $\{\mathbf{x} \,|\sum_j \sqrt{|x_j|} < k\}$
- union de 2 convexes disjoints

Fonction convexe

Definition (fonction convexe)

une fonction J est dite convexe si, pour tout $\mathbf{x}, \mathbf{y} \in \Omega$ on a

$$\forall \alpha \in]0,1[, \quad J(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha J(\mathbf{x}) + (1-\alpha) J(\mathbf{y})$$

si J est une fonction convexe, la fonction $-J$ est dite concave.

exemple

$$J_1(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} \qquad J_2(\mathbf{x}) = \tfrac{1}{2} \mathbf{x}^\top A \mathbf{x}$$
$$J_3(\mathbf{x}) = \sum_{i=1}^{d} \exp^{-x_i} \qquad J_4(\mathbf{x}) = -\log \mathbf{a}^\top \mathbf{x}$$

On vérifiera que toute somme de fonctions convexe est convexe.

Theorem

si x_0 est une solution locale d'un problème d'optimisation convexe, alors c'est aussi la solution globale.

Conclusions

- Minimiser c'est annuler un gradient
- pour calculer le gradient : la recette (ou un programme)
- confortable dans le cas convexe (unicité de la solution)
- (petits) problèmes dans le cas non différentiable

Unconstrained optimization (Fermat's rule)

For J convex & differentiable, global minima

$$\mathbf{x}^\star = \arg\min_{\mathbf{x}} J(\mathbf{x}) \quad \Longleftrightarrow \quad \nabla J(\mathbf{x}^\star) = 0$$

For J convex & nondifferentiable, global minima

$$\mathbf{x}^\star = \arg\min_{\mathbf{x}} J(\mathbf{x}) \quad \Longleftrightarrow \quad 0 \in \partial J(\mathbf{x}^\star)$$

For J non convex & nondifferentiable, local minima, necessary condition

$$\mathbf{x}^\star = \arg\min_{\mathbf{x} \in V(\mathbf{x}^\star)} J(\mathbf{x}) \quad \Longrightarrow \quad 0 \in \partial_c J(\mathbf{x}^\star)$$

or equivalently (sometimes referred as Oresme's rule)

$$\mathbf{x}^\star = \arg\min_{\mathbf{x} \in V(\mathbf{x}^\star)} J(\mathbf{x}) \quad \Longrightarrow \quad D_c J(\mathbf{x}^\star, \mathbf{x}) \geq 0, \quad \forall \mathbf{x} \in V(\mathbf{x}^\star)$$

Fermat's rule and critical points

Table : Fermat's rule and critical (or stationary) points

	differentiable	non differentiable
cvx	$\mathbf{x}^\star = \arg\min_{\mathbf{x}} J(\mathbf{x}) \Leftrightarrow \nabla J(\mathbf{x}^\star) = 0$	$\mathbf{x}^\star = \arg\min_{\mathbf{x}} J(\mathbf{x}) \Leftrightarrow 0 \in \partial J(\mathbf{x}^\star)$
non cvx	$\mathbf{x}^\star = \arg\min_{\mathbf{x} \in V(\mathbf{x_0})} J(\mathbf{x}) \Rightarrow \nabla J(\mathbf{x}^\star) = 0$	$\mathbf{x}^\star = \arg\min_{\mathbf{x} \in V(\mathbf{x_0})} J(\mathbf{x}) \Rightarrow 0 \in \partial_c J(\mathbf{x}^\star)$

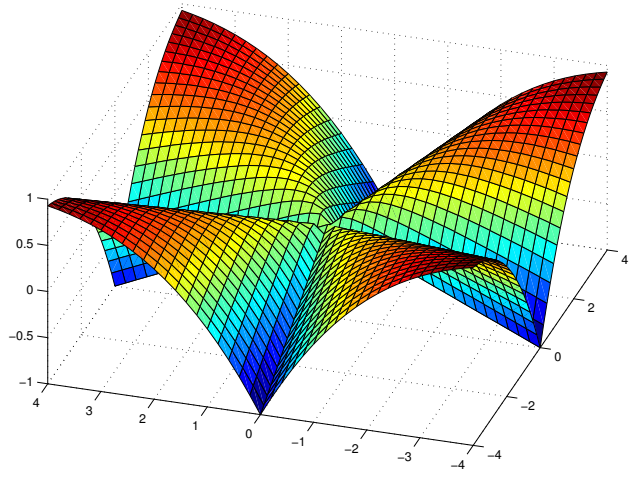
Definition (Clarke critical (or stationary) point)

Let f be a locally Lipschitz continuous function at $\mathbf{x}$. A point $\mathbf{x}^\star$ is called a (Clarke) critical point if

$$0 \in \partial_c J(\mathbf{x})$$

Local minima are critical points but the converse is not always true.

Plan



Plan

problème : $\min_{\mathbf{x} \in \mathbb{R}^n} J(\mathbf{x})$

Solution : trouver $\mathbf{x}^* \in \mathbb{R}^n$ tel que $\nabla J(\mathbf{x}^*) = 0$

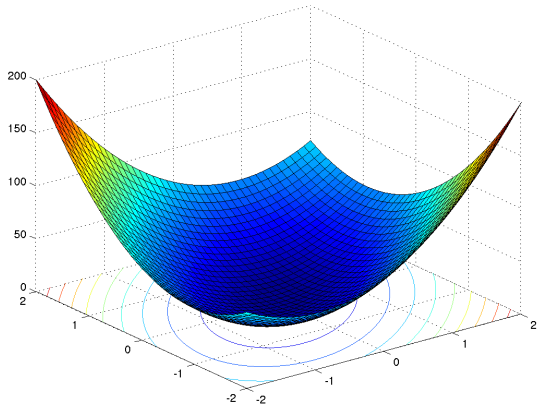
n équations à n inconnues ... c'est une condition nécessaire et suffisante pour J convexe !

Exemples de problème d'optimisation sans contr.

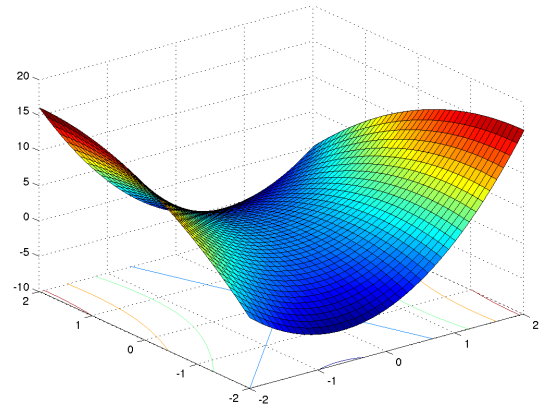
- linéaire : $\min_{\mathbf{x}\in\mathbb{R}^n} J_\ell(\mathbf{x}) = \mathbf{b}^\top \mathbf{x}$ $\quad \rightarrow$ pas de solution
- quadratique convexe (moindres carrés)

$$\min_{\mathbf{x}\in\mathbb{R}^n} J_q(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} - \mathbf{b}^\top \mathbf{x} \qquad A \text{ définie positive}$$

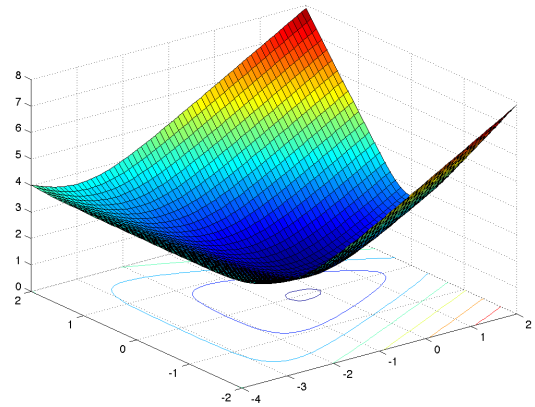
- quadratique non convexe
- non quadratique convexe : $\exp^{x_1+3x_2-1/10} + \exp^{x_1-x_2-1/10} + \exp^{-x_1-1/10}$



quadratique convexe



quadratique non convexe



non quadratique convexe

Illustration 2d

$$J(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} - \mathbf{b}^\top \mathbf{x}$$

lignes d'iso cout : $\{\mathbf{x} \in \mathbb{R}^2 \mid J(\mathbf{x}) = \text{ Cte } \}$

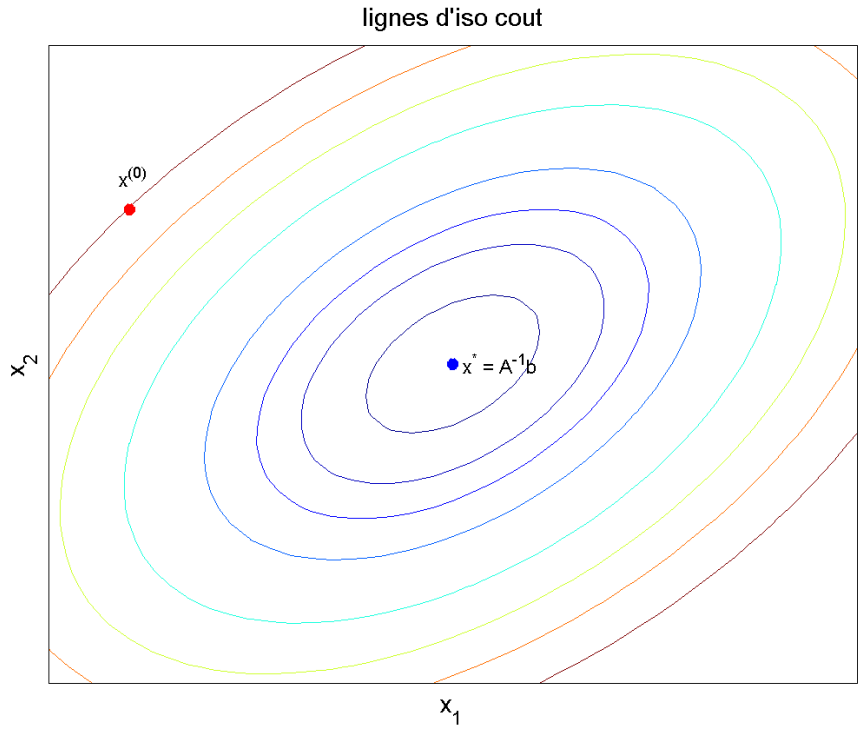
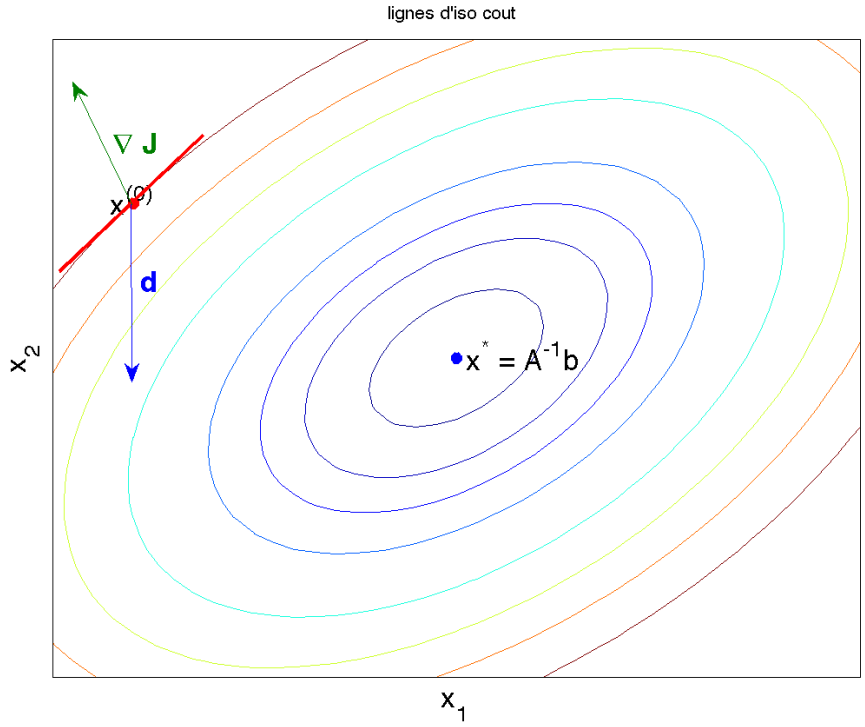


Illustration 2d

$$J(\mathbf{x}) = \frac{1}{2}\,\mathbf{x}^\top A\mathbf{x} - \mathbf{b}^\top \mathbf{x}$$

lignes d'iso cout : $\{\mathbf{x} \in \mathbb{R}^2 \mid J(\mathbf{x}) = \text{ Cte } \}$



une direction de descente $\mathbf{d}$ doit vérifier : $\mathbf{d}^\top \underbrace{(A\mathbf{x} - \mathbf{b})}_{\nabla J(\mathbf{x})} < 0$

deux approches itératives classiques

problème : $\min_{\mathbf{x}\in\mathbb{R}^n} J(\mathbf{x})$

Solution : partir d'un point $\mathbf{x}^0 \in \mathbb{R}^n$
et construire une suite $x^{(1)}, x^{(2)}, \ldots x^{(k)}, \ldots k \in \mathbb{N}$
qui converge vers la solution $\mathbf{x}^*$ telle que $\nabla J(\mathbf{x}^*) = 0$

deux approches itératives classiques

- recherche linéaire (*line search*) $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \rho_k \mathbf{d}^{(k)}$

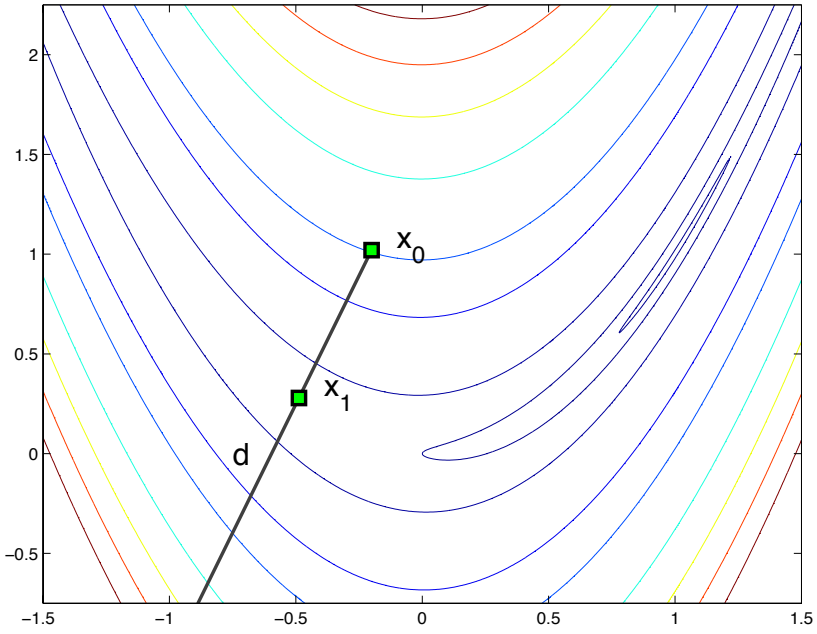
  $\min_{\rho\in\mathbb{R}} J(\mathbf{x}^{(k)} + \rho\mathbf{d}^{(k)})$ $\mathbf{d}^{(k)}$ étant une direction de descente

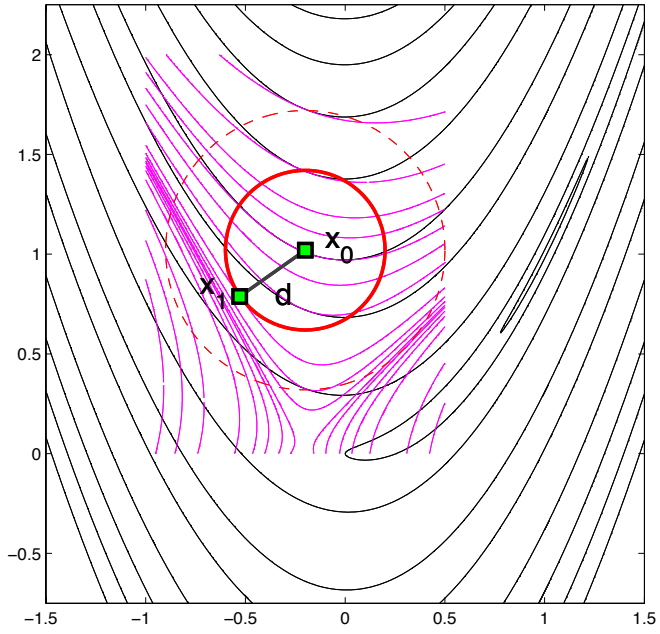
- région de confiance (*trust region*) $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{d}^{(k)}$

  $\min_{\mathbf{d}\in R(\mathbf{x})} M(\mathbf{x}+\mathbf{d})$ $R(\mathbf{x})$ étant une région de confiance autour de $\mathbf{x}$

avec le modèle M : $M(\mathbf{x}+\mathbf{d}) = J(\mathbf{x}) + \mathbf{d}^\top \nabla J(\mathbf{x}) + \frac{1}{2}\mathbf{d}^\top B(\mathbf{x})\mathbf{d}$ et $R(x) = \{\mathbf{x}+\mathbf{p} \mid \|\mathbf{p}\| \leq \Delta\}$

Recherche linéaire vs. Région de confiance

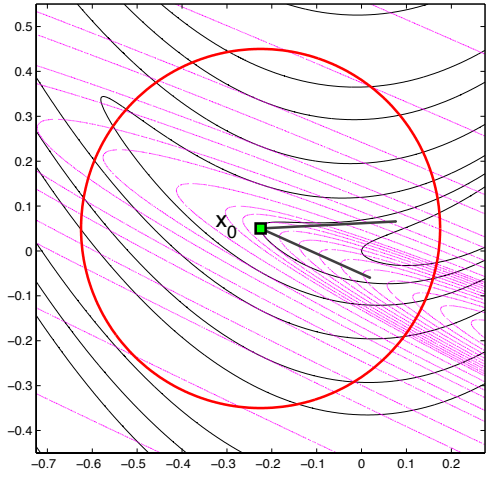




$\mathbf{d} \leftarrow$ trouve une direction

$\rho \leftarrow$calcul le pas$(\mathbf{x}_0, \mathbf{d}, J)$

$\mathbf{x}_1 \leftarrow \mathbf{x}_0 + \rho\mathbf{d}$



définir une région R

$R = \{\mathbf{d} \mid \|\mathbf{d}\|^2 < \delta\}$

définir un modèle M

$M(\mathbf{x}_0 + \mathbf{d}) =$

. $\quad M(\mathbf{x}_0) + \mathbf{d}^\top \nabla J + \frac{1}{2}\mathbf{d}^\top B\mathbf{d}$

$\mathbf{d} \leftarrow \arg\min_{\mathbf{d} \in R} M(\mathbf{x}_0 + \mathbf{d})$

$\mathbf{x}_1 \leftarrow \mathbf{x}_0 + \mathbf{d}$

Méthodes de recherche linéaire

$$\mathbf{x}^{(\text{new})} = \mathbf{x}^{(\text{old})} + \rho\, \mathbf{d}, \quad \mathbf{x}, \mathbf{d} \in \mathbb{R}^d$$

- initialisation $\mathbf{x}^{(0)}$
- direction $\mathbf{d} \in \mathbb{R}^n$
- pas $\rho \in \mathbb{R}$

Fonction $\mathbf{x} \leftarrow \texttt{min}(J, \mathbf{x})$
While (on a pas convergé) **do**
  $\mathbf{d} \leftarrow$ trouve une direction
  $\rho \leftarrow$calcul le pas$(\mathbf{x}, \mathbf{d}, J)$
  $\mathbf{x} \leftarrow \mathbf{x} + \rho\mathbf{d}$
done

- suite minimisante : $\mathbf{x}^{(k)}, \rho_k, \mathbf{d}^{(k)}, \quad k \in \mathbb{N}$
  convergente : $\mathbf{x}^{(k)} \longrightarrow \mathbf{x}^*$ et $\nabla J(\mathbf{x}^{(k)}) \longrightarrow 0$
- arrêt (on a convergé) :
  - minimum atteint : $\|\nabla J\| \leq \epsilon$,
  - le cout ne diminue plus

$$|J(\mathbf{x}^{(\text{new})}) - J(\mathbf{x}^{(\text{old})})| \leq \varepsilon$$

  - ou on dépasse un nombre d'itération maximum fixé : $k \geq k_{\text{max}}$

Direction de descente

Definition (Direction de descente)

d est une direction de descente s'il existe $\rho_0 > 0$ tel que $\forall \rho \leq \rho_0$

$$J(\mathbf{x}^{(\text{new})}) < J(\mathbf{x}^{(\text{old})})$$

Theorem

Toute direction $\mathbf{d}$ *telle que* $\nabla J^\top \mathbf{d} < 0$ *est une direction de descente*

Elément de démonstration : $J(\mathbf{x} + \rho\mathbf{d}) = J(\mathbf{x}) + \rho \nabla J^\top \mathbf{d} + o(\rho^2)$

Les enjeux : le compromis

- précision (descendre un bon coup)
- efficacité (temps de calcul)

Direction de descente du gradient et autres

Theorem (Descente de gradient)

L'opposé du gradient est une direction de descente : $\mathbf{d} = -\nabla J(\mathbf{x})$

Elément de démonstration : pour $\mathbf{d} = -\nabla J(x)$ et ρ suffisamment petit

$$J(x - \rho \nabla J) = J(x) - \rho \|\nabla J\|^2 + \rho^2/2 \underbrace{\nabla J^\top H \nabla J}_{H \text{ définie positive}} < J(x)$$

Principaux choix de direction de descente

Gradient $\mathbf{d} = -\nabla J$ $\rightarrow$ peu précis mais efficace $\mathcal{O}(n)$

Grad. conjugué $\mathbf{d}^{(k)} = -\nabla J + \beta_k \mathbf{d}^{(k-1)}$ $\rightarrow \mathcal{O}(n)$

Quasi Newton $\mathbf{d} = -B \nabla J$ (de nombreuses variantes) $\rightarrow \mathcal{O}(n^2)$

- $B = (diag(H))^{-1}$
- DFP (Davidson Fletcher Powell) : $B^{k+1} = B^k + \varphi(B^k, \nabla J, \mathbf{x})$
- BFGS (Broyden Fletcher Goldfarb Shanno)

Newton $\mathbf{d} = -H^{-1} \nabla J$ $\rightarrow$ plus précis mais couteux $\mathcal{O}(n^3)$

- si ∇J est le vecteur gradient en $\mathbf{x}$ et H la matrice hessienne

Propriété : développement de Taylor au second ordre

$$J(\mathbf{x}+\mathbf{d}) = J(\mathbf{x}) + \nabla_{\mathbf{x}} J(\mathbf{x})^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top H(\mathbf{x})\mathbf{d} + o(\|\mathbf{d}\|^2)$$

- que l'on cherche à minimiser par rapport à $\mathbf{d}$. si on a

$$G(\mathbf{d}) = J(\mathbf{x}) + \nabla_{\mathbf{x}} J(\mathbf{x})^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top H(\mathbf{x})\mathbf{d} + o(\|\mathbf{d}\|^2)$$

$$\nabla_{\mathbf{d}} G_{\mathbf{d}}(\mathbf{d}) = 0 \qquad \Leftrightarrow \qquad \mathbf{d} = -H(\mathbf{x})^{-1}\nabla_{\mathbf{x}} J(\mathbf{x})$$

théorème

Si la matrice hessienne est définie positive (et régulière), la direction de descente de la méthode de Newton est donnée par

$$\mathbf{d} = -H(\mathbf{x})^{-1}\nabla_{\mathbf{x}} J(\mathbf{x})$$

La Méthode de Newton

Fonction $\mathbf{x} \leftarrow \texttt{Newton}(J, \mathbf{x}, k)$
While (on a pas convergé) **do**
  $(\mathbf{g}, H) \leftarrow$ grad_et_hessienne$(J, \mathbf{x})$
  $\mathbf{d} \leftarrow -H^{-1}\mathbf{g}$
  $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{d}$
done

on peut aussi rechercher un pas optimal

$$\mathbf{x}^{(\text{new})} = \mathbf{x}^{(\text{old})} + \rho\, \mathbf{d}$$

La Méthode de Newton et région de confiance

Le problème : construire une suite $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{d}$ solution de

$$\begin{cases} \min\limits_{\mathbf{d}\in\mathbb{R}^n} & J(\mathbf{x}^{(k)}) + \nabla_{\mathbf{x}} J(^{(k)})\mathbf{d} + \frac{1}{2}\mathbf{d}^\top H(\mathbf{x}^{(k)})\mathbf{d} \\ \text{avec} & \|\mathbf{d}\|^2 \leq \Delta_k \end{cases}$$

La solution : $\Delta_k \Leftrightarrow \lambda_k$

$$\mathbf{d} = -(H(\mathbf{x}^{(k)}) + \lambda_k I)^{-1} \nabla J(^{\mathbf{x}(k)}) \qquad \text{et} \qquad \mathbf{d}^\top \mathbf{d} = \Delta_k$$

Fonction $\mathbf{x} \leftarrow$ `Newton_reg_conf`$(J, \mathbf{x}, k)$
While (on a pas convergé) **do**
  $\lambda \leftarrow$
  $(\mathbf{g}, H) \leftarrow$ grad_et_hessienne$(J, \mathbf{x})$
  $\mathbf{d} \leftarrow -(H + \lambda I)^{-1}\mathbf{g}$
  $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{d}$
done

pré conditionnement de la matrice hessienne

conclusion

$\mathcal{O}(n)$ Gradient

- pas fixe
- pas optimal
- heuristique (règle d'Armijo)

$\mathcal{O}(n^2)$ Gradient conjugué

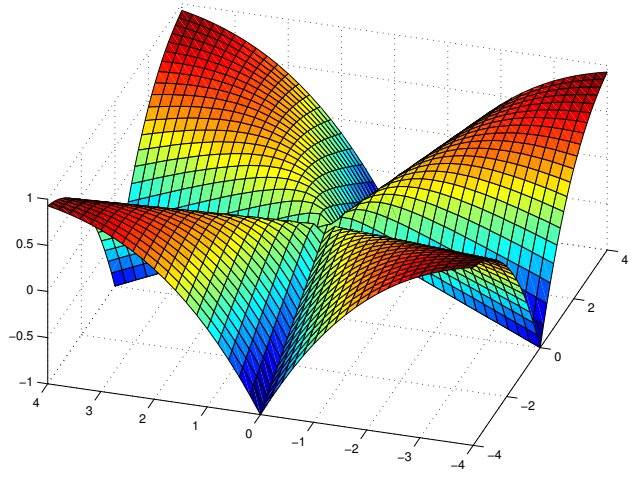
- quasi newton
- approximation $d = H^{-1}\nabla J$
- Levenberg Marquart

$\mathcal{O}(n^3)$ Newton

méthode	*direction de descente*	*temps de calcul*	*convergence*
gradient	$d = -\nabla J$	$\mathcal{O}(n)$	lineaire
quasi Newton	$d = -B\nabla J$	$\mathcal{O}(n^2)$	super linéaire
Newton	$d = -H^{-1}\nabla J$	$\mathcal{O}(n^3)$	quadratic

le temps de calcul du pas optimal peut aussi varier

Plan



Linear SVM: the problem

Linear SVM are the solution of the following problem (called primal)

Let $\{(\mathbf{x}_i, y_i);\ i = 1 : n\}$ be a set of labelled data with $\mathbf{x}_i \in \mathbb{R}^d, y_i \in \{1, -1\}$.
A support vector machine (SVM) is a linear classifier associated with the following decision function: $D(x) = \text{sign}\left(\mathrm{w}^\top \mathbf{x} + b\right)$ where $\mathrm{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ a given thought the solution of the following problem:

$$\begin{cases} \min\limits_{\mathrm{w},b} & \frac{1}{2}\|\mathrm{w}\|^2 = \frac{1}{2}\mathrm{w}^\top \mathrm{w} \\ \text{with} & y_i(\mathrm{w}^\top \mathbf{x}_i + b) \geq 1 \qquad\qquad i = 1, n \end{cases}$$

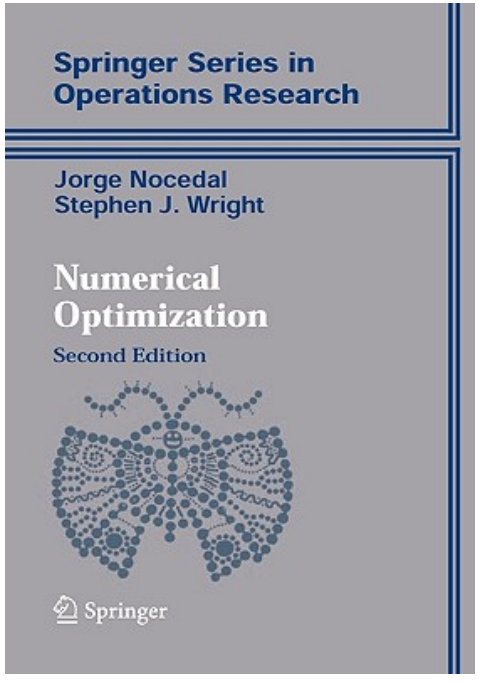
This is a quadratic program (QP):

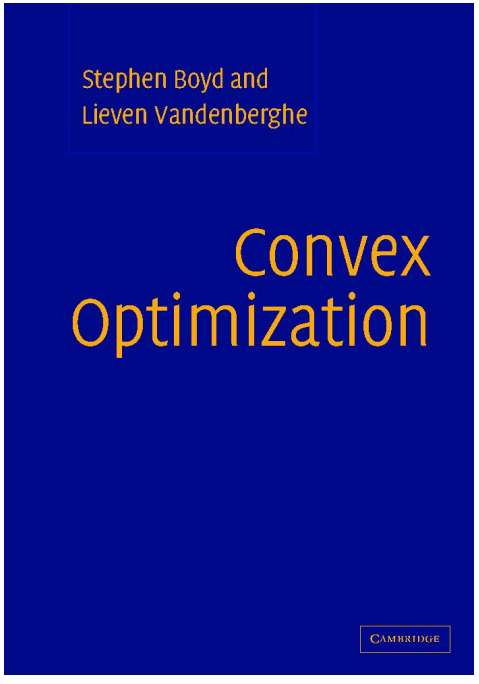
$$\begin{cases} \min\limits_{\mathbf{z}} & \frac{1}{2}\mathbf{z}^\top A\mathbf{z} - \mathbf{d}^\top \mathbf{z} \\ \text{with} & B\mathbf{z} \leq \mathbf{e} \end{cases}$$

$\mathbf{z} = (\mathrm{w}, b)^\top$, $\mathbf{d} = (0, \ldots, 0)^\top$, $A = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$, $B = -[\text{diag}(\mathbf{y})X, \mathbf{y}]$ et $\mathbf{e} = -(1, \ldots, 1)^\top$

Road map

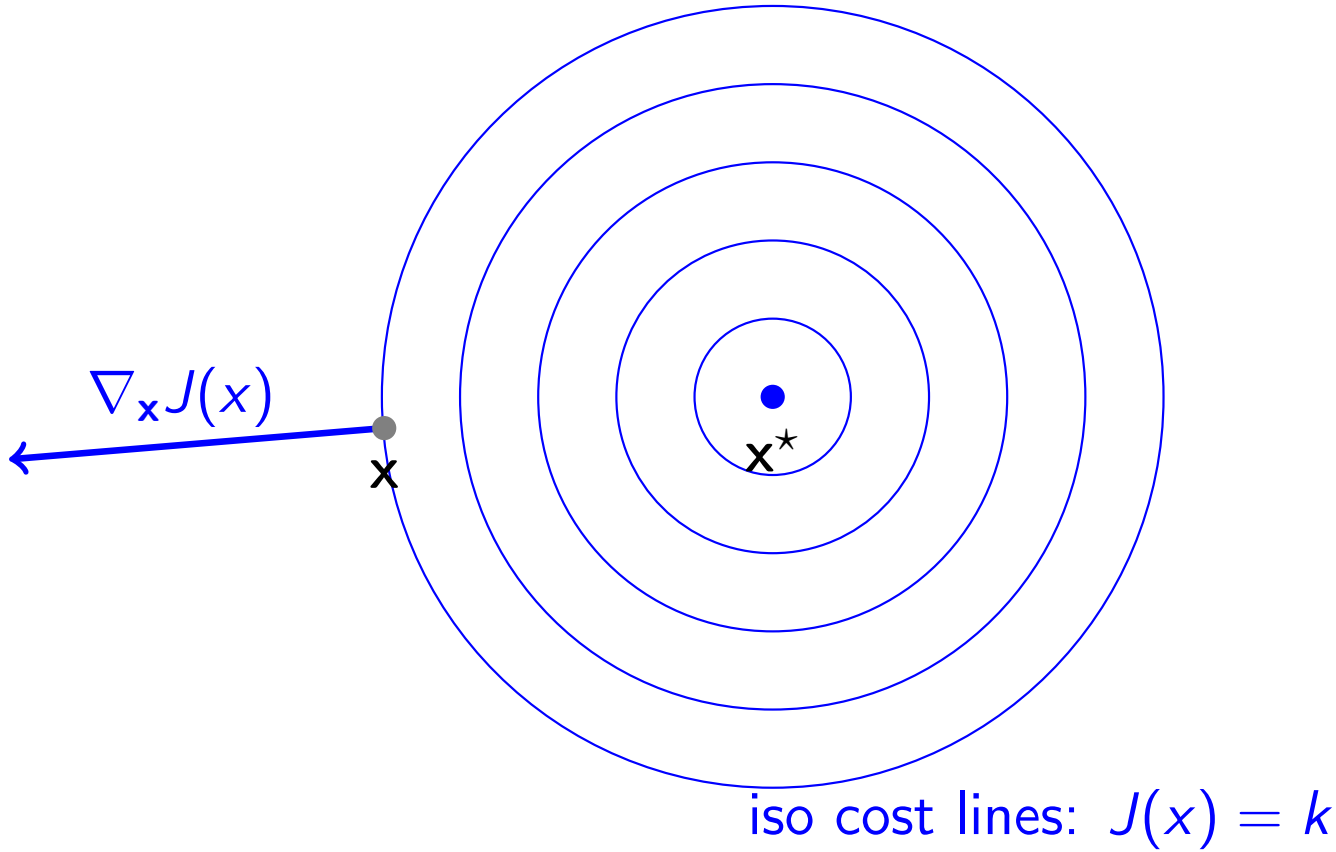
1. Optimisation
2. Introduction au gradient et autres dérivées
3. Algorithmes pour l'optimisation sans contraintes
4. Optimisation avec contraintes





A simple example (to begin with)

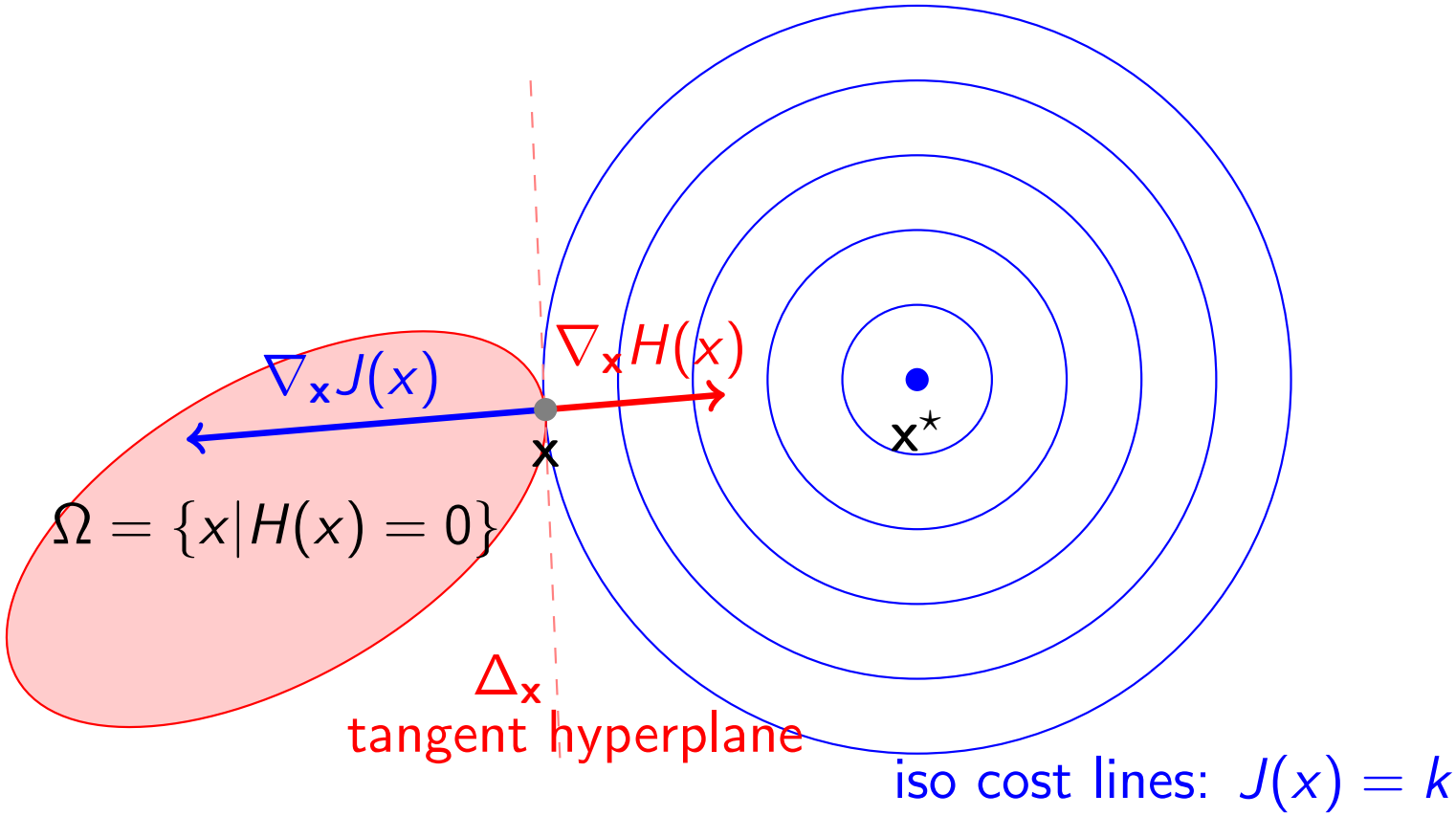
$$\begin{cases} \min\limits_{x_1,x_2} & J(\mathbf{x}) = (x_1 - a)^2 + (x_2 - b)^2 \\ \text{with} \end{cases}$$



iso cost lines: $J(x) = k$

A simple example (to begin with)

$$\begin{cases} \min\limits_{x_1,x_2} & J(\mathbf{x}) = (x_1 - a)^2 + (x_2 - b)^2 \\ \text{with} & H(\mathbf{x}) = \alpha(x_1 - c)^2 + \beta(x_2 - d)^2 + \gamma x_1 x_2 - 1 \end{cases}$$



$$\nabla_{\mathbf{x}} H(x) = \lambda \, \nabla_{\mathbf{x}} J(x)$$

The only one equality constraint case

$$\begin{cases} \min\limits_{\mathbf{x}} & J(\mathbf{x}) \\ \text{with} & H(\mathbf{x}) = 0 \end{cases} \qquad \begin{aligned} J(\mathbf{x}+\varepsilon\mathbf{d}) &\approx J(\mathbf{x}) + \varepsilon\nabla_{\mathbf{x}} J(\mathbf{x})^\top \mathbf{d} \\ H(\mathbf{x}+\varepsilon\mathbf{d}) &\approx H(\mathbf{x}) + \varepsilon\nabla_{\mathbf{x}} H(\mathbf{x})^\top \mathbf{d} \end{aligned}$$

Loss J : $\mathbf{d}$ is a descent direction if it exists $\varepsilon_0 \in \mathbb{R}$ such that $\forall \varepsilon \in \mathbb{R},\ 0 < \varepsilon \leq \varepsilon_0$

$$J(\mathbf{x}+\varepsilon\mathbf{d}) < J(\mathbf{x}) \quad \Rightarrow \quad \nabla_{\mathbf{x}} J(\mathbf{x})^\top \mathbf{d} < 0$$

constraint H : $\mathbf{d}$ is a feasible descent direction if it exists $\varepsilon_0 \in \mathbb{R}$ such that $\forall \varepsilon \in \mathbb{R},\ 0 < \varepsilon \leq \varepsilon_0$

$$H(\mathbf{x}+\varepsilon\mathbf{d}) = 0 \quad \Rightarrow \quad \nabla_{\mathbf{x}} H(\mathbf{x})^\top \mathbf{d} = 0$$

If at $x^\star$, vectors $\nabla_{\mathbf{x}} J(\mathbf{x}^\star)$ and $\nabla_{\mathbf{x}} H(\mathbf{x}^\star)$ are collinear there is no feasible descent direction $\mathbf{d}$. Therefore, $x^\star$ is a local solution of the problem.

Lagrange multipliers

Assume J and functions H_i are continuously differentials (and independent)

$$\mathcal{P} = \begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^n} & J(\mathbf{x}) \\ \text{avec} & H_1(\mathbf{x}) = 0 \\ \text{et} & H_2(\mathbf{x}) = 0 \\ & \dots \\ & H_p(\mathbf{x}) = 0 \end{cases}$$

Lagrange multipliers

Assume J and functions H_i are continuously differentials (and independent)

$$\mathcal{P} = \begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^n} & J(\mathbf{x}) & \\ \text{avec} & H_1(\mathbf{x}) = 0 & \lambda_1 \\ \text{et} & H_2(\mathbf{x}) = 0 & \lambda_2 \\ & \dots & \\ & H_p(\mathbf{x}) = 0 & \lambda_p \end{cases}$$

each constraint is associated with λ_i : the Lagrange multiplier.

Lagrange multipliers

Assume J and functions H_i are continuously differentials (and independent)

$$\mathcal{P} = \begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^n} & J(\mathbf{x}) & \\ \text{avec} & H_1(\mathbf{x}) = 0 & \lambda_1 \\ \text{et} & H_2(\mathbf{x}) = 0 & \lambda_2 \\ & \dots & \\ & H_p(\mathbf{x}) = 0 & \lambda_p \end{cases}$$

each constraint is associated with λ_i : the Lagrange multiplier.

Theorem (First order optimality conditions)

for $\mathbf{x}^\star$ being a local minima of $\mathcal{P}$, it is necessary that:

$$\nabla_x J(\mathbf{x}^\star) + \sum_{i=1}^{p} \lambda_i \nabla_x H_i(\mathbf{x}^\star) = 0 \qquad \textit{and} \qquad H_i(\mathbf{x}^\star) = 0, \quad i = 1, p$$

Un exemple où ça marche

$$\begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^3} & J(\mathbf{x}) = -x_1x_2 - x_1x_3 - x_2x_3 \\ \text{avec} & H(\mathbf{x}) = x_1 + x_2 + x_3 - 3 = 0 \end{cases}$$

Un exemple où ça marche

$$\begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^3} & J(\mathbf{x}) = -x_1x_2 - x_1x_3 - x_2x_3 \\ \text{avec} & H(\mathbf{x}) = x_1 + x_2 + x_3 - 3 = 0 \end{cases}$$

$$\nabla_x J(x) = -\begin{pmatrix} x_2 + x_3 \\ x_1 + x_3 \\ x_1 + x_2 \end{pmatrix} \qquad \nabla_x H(x) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Les conditions d'optimalité sont

$$\begin{cases} & -x_2 & -x_3 & +\lambda & = 0 \\ -x_1 & & -x_3 & +\lambda & = 0 \\ -x_1 & -x_2 & & +\lambda & = 0 \\ x_1 & +x_2 & +x_3 & & = 3 \end{cases}$$

la résolution du système $(A\backslash b)$ donne :

$$x_1 = x_2 = x_3 = 1 \text{ et } \lambda = 2$$

Un autre exemple où ça marche

$$\begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^n} & J(\mathbf{x}) = -\mathbf{x}^\top A\mathbf{x} - \mathbf{x}^\top \mathbf{b} \\ \text{avec} & C\mathbf{x} = \mathbf{d} \end{cases}$$

$$\begin{aligned} \nabla_{\mathbf{x}} J &= A\mathbf{x} - \mathbf{b} \\ \nabla_{\mathbf{x}} H &= C^\top \end{aligned}$$

$$\begin{aligned} \nabla_x J(x) + \sum_{i=1}^{p} \lambda_i \nabla_x H_i(\mathbf{x}) = 0 \quad &\Rightarrow \quad A\mathbf{x} + C^\top \lambda = \mathbf{b} \\ H(\mathbf{x}) = 0 \quad &\Rightarrow \quad C\mathbf{x} = \mathbf{d} \end{aligned}$$

...et on résoud le système linéaire.

Attention si en plus on cherche $\mathbf{x} \geq 0$ ça devient plus compliqué

Un exemple où ça ne marche pas

$$\begin{cases} \min\limits_{x_1,x_2} & x_1 + x_2 \\ \text{avec} & (x_1+1)^2 + x_2^2 = 1 \\ \text{et} & (x_1-2)^2 + x_2^2 = 4 \end{cases}$$

le minimum est $(0,0)$ l'unique solution réalisable ! Dans ce cas, il n'existe pas de multiplicateurs de Lagrange

lagrangien

Une fonction bien pratique :

définition : lagrangien

On appelle lagrangien du problème $\mathcal{P}$ la fonction L définie par :

$$L(\mathbf{x}, \lambda) = J(x) + \sum_{i=1}^{p} \lambda_i H_i(\mathbf{x})$$

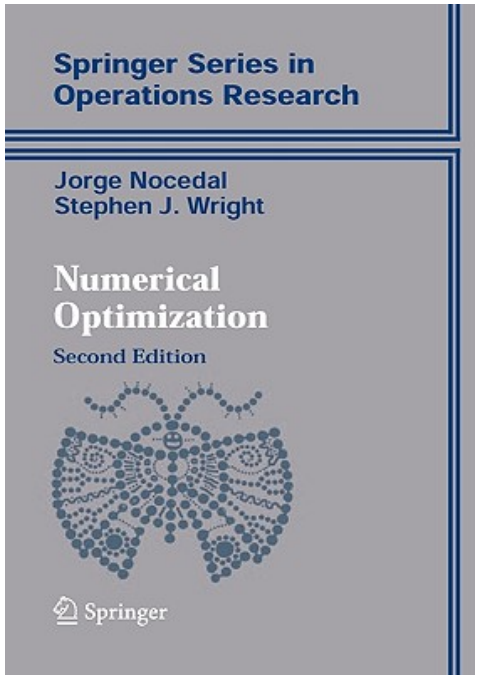
Grâce au lagrangien on retrouve les conditions d'optimalité :

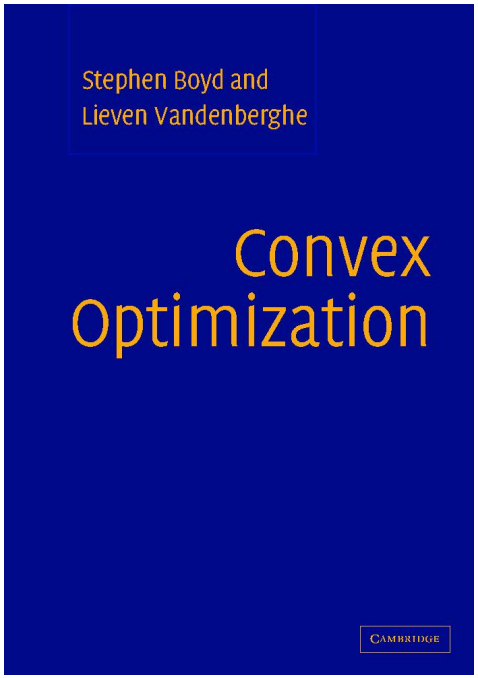
$$\begin{cases} \nabla_x L(\mathbf{x}, \lambda) = 0 & \Rightarrow \quad \nabla_x J(\mathbf{x}) + \sum_{i=1}^{p} \lambda_i \nabla_x H_i(\mathbf{x}) = 0 \\ \nabla_{\lambda_i} L(\mathbf{x}, \lambda) = 0 & \Rightarrow \quad H_i(\mathbf{x}) = 0 \end{cases}$$

Interprétation graphique : optimisation multicritères.

Plan

1. Optimisation
2. Introduction au gradient et autres dérivées
3. Algorithmes pour l'optimisation sans contraintes
4. Optimisation avec contraintes





The only one inequality constraint case

$$\begin{cases} \min\limits_{\mathbf{x}} & J(\mathbf{x}) & J(\mathbf{x}+\varepsilon\mathbf{d}) \approx J(\mathbf{x}) + \varepsilon\nabla_{\mathbf{x}}J(\mathbf{x})^{\top}\mathbf{d} \\ \text{with} & G(\mathbf{x}) \leq 0 & G(\mathbf{x}+\varepsilon\mathbf{d}) \approx G(\mathbf{x}) + \varepsilon\nabla_{\mathbf{x}}G(\mathbf{x})^{\top}\mathbf{d} \end{cases}$$

cost J : $\mathbf{d}$ is a descent direction if it exists $\varepsilon_0 \in \mathbb{R}$ such that $\forall \varepsilon \in \mathbb{R},\ 0 < \varepsilon \leq \varepsilon_0$

$$J(\mathbf{x}+\varepsilon\mathbf{d}) < J(\mathbf{x}) \quad \Rightarrow \quad \nabla_{\mathbf{x}}J(\mathbf{x})^{\top}\mathbf{d} < 0$$

constraint G : $\mathbf{d}$ is a feasible descent direction if it exists $\varepsilon_0 \in \mathbb{R}$ such that $\forall \varepsilon \in \mathbb{R},\ 0 < \varepsilon \leq \varepsilon_0$

$$G(\mathbf{x}+\varepsilon\mathbf{d}) \leq 0 \quad \Rightarrow \quad \begin{array}{ll} G(\mathbf{x}) < 0 : & \text{no limit here on } \mathbf{d} \\ G(\mathbf{x}) = 0 : & \nabla_{\mathbf{x}}G(\mathbf{x})^{\top}\mathbf{d} \leq 0 \end{array}$$

Two possibilities

If $x^\star$ lies at the limit of the feasible domain ($G(\mathbf{x}^\star) = 0$) and if vectors $\nabla_{\mathbf{x}}J(\mathbf{x}^\star)$ and $\nabla_{\mathbf{x}}G(\mathbf{x}^\star)$ are collinear and in opposite directions, there is no feasible descent direction $\mathbf{d}$ at that point. Therefore, $x^\star$ is a local solution of the problem... Or if $\nabla_{\mathbf{x}}J(\mathbf{x}^\star) = 0$

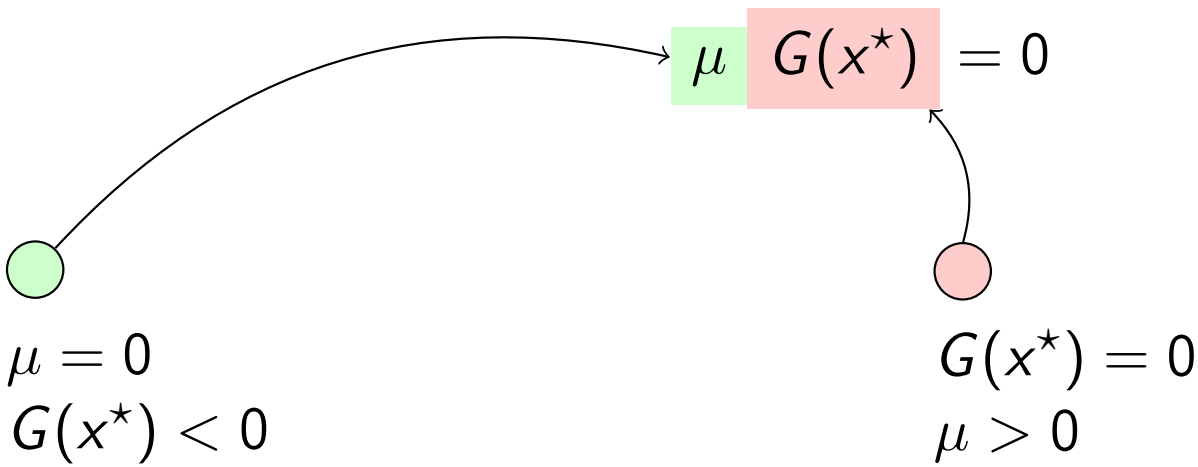
Two possibilities for optimality

$$\nabla_{\mathbf{x}} J(\mathbf{x}^\star) = -\mu \, \nabla_{\mathbf{x}} G(\mathbf{x}^\star) \qquad \text{and} \qquad \mu > 0; G(\mathbf{x}^\star) = 0$$

or

$$\nabla_{\mathbf{x}} J(\mathbf{x}^\star) = 0 \qquad \text{and} \qquad \mu = 0; G(\mathbf{x}^\star) < 0$$

This alternative is summarized in the so called complementarity condition:



First order optimality condition (1)

$$\text{problem } \mathcal{P} = \begin{cases} \min\limits_{\mathbf{x}\in\mathbf{R}^{n}} & J(\mathbf{x}) \\ \text{with} & h_j(x) = 0 \quad j = 1, \ldots, p \\ \text{and} & g_i(x) \leq 0 \; i = 1, \ldots, q \end{cases}$$

Definition: Karush, Kuhn and Tucker (KKT) conditions

$$\text{stationarity} \quad \nabla J(x^\star) + \sum_{j=1}^{p} \lambda_j \nabla h_j(x^\star) + \sum_{i=1}^{q} \mu_i \nabla g_i(x^\star) = 0$$

$$\begin{aligned} \text{primal admissibility} \quad & h_j(x^\star) = 0 && j = 1, \ldots, p \\ & g_i(x^\star) \leq 0 && i = 1, \ldots, q \\ \text{dual admissibility} \quad & \mu_i \geq 0 && i = 1, \ldots, q \\ \text{complementarity} \quad & \mu_i g_i(x^\star) = 0 && i = 1, \ldots, q \end{aligned}$$

λ_j and μ_i are called the Lagrange multipliers of problem $\mathcal{P}$

First order optimality condition (2)

Theorem (12.1 Nocedal & Wright pp 321)

If a vector $x^\star$ is a stationary point of problem $\mathcal{P}$
Then there exists[a] Lagrange multipliers such that $\left(x^\star, \{\lambda_j\}_{j=1:p}, \{\mu_i\}_{i=1:q}\right)$ fulfill KKT conditions

[a] under some conditions *e.g.* linear independence constraint qualification

If the problem is convex, then a stationary point is the solution of the problem

A quadratic program (QP) is convex when...

$$(QP) \quad \begin{cases} \min\limits_{\mathbf{z}} & \frac{1}{2}\mathbf{z}^\top A\mathbf{z} - \mathbf{d}^\top \mathbf{z} \\ \text{with} & B\mathbf{z} \leq \mathbf{e} \end{cases}$$

...when matrix A is positive definite

Exemple

$$\mathcal{P} = \left\{ \begin{array}{ll} \min\limits_{\mathbf{x}\in\mathbb{R}^2} & J(\mathbf{x}) = x_1^2 + x_2^2 \\ \text{avec} & 2x_1 + x_2 \leq -4 \end{array} \right.$$

Exemple

$$\mathcal{P} = \begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^2} & J(\mathbf{x}) = x_1^2 + x_2^2 \\ \text{avec} & 2x_1 + x_2 \leq -4 \end{cases}$$

$$\begin{aligned} \text{stationarité} \quad & 2x_1 + 2\mu = 0 \\ & 2x_2 + \mu = 0 \\ \text{admissibilité primal} \quad & 2x_1 + x_2 + 4 \leq 0 \\ \text{admissibilité duale} \quad & \mu \geq 0 \\ \text{complementarité} \quad & \mu(2x_1 + x_2 + 4) = 0 \end{aligned}$$

$$x_1 = -\frac{8}{5},\ x_2 = -\frac{4}{5},\ \mu = \frac{8}{5},$$

KKT condition - Lagrangian (3)

$$\text{problem } \mathcal{P} = \begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^n} & J(\mathbf{x}) \\ \text{with} & h_j(x) = 0 \quad j = 1, \ldots, p \\ \text{and} & g_i(x) \leq 0 \; i = 1, \ldots, q \end{cases}$$

Definition: Lagrangian

The lagrangian of problem $\mathcal{P}$ is the following function:

$$\mathcal{L}(\mathbf{x}, \lambda, \mu) = J(x) + \sum_{j=1}^{p} \lambda_j h_j(x) + \sum_{i=1}^{q} \mu_i g_i(x)$$

The importance of being a lagrangian

- the stationarity condition can be written: $\nabla \mathcal{L}(\mathbf{x}^\star, \lambda, \mu) = 0$
- the lagrangian saddle point $\max\limits_{\lambda,\mu} \min\limits_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda, \mu)$

Primal variables: x and dual variables λ, μ (the Lagrange multipliers)

Duality – definitions (1)

Primal and (Lagrange) dual problems

$$\mathcal{P} = \begin{cases} \min\limits_{\mathbf{x}\in\mathbf{R}^n} & J(\mathbf{x}) \\ \text{with} & h_j(x) = 0 \quad j = 1, p \\ \text{and} & g_i(x) \leq 0 \quad i = 1, q \end{cases} \qquad \mathcal{D} = \begin{cases} \max\limits_{\lambda\in\mathbf{R}^p, \mu\in\mathbf{R}^q} & Q(\lambda, \mu) \\ \text{with} & \mu_j \geq 0 \quad j = 1, q \end{cases}$$

Dual objective function:

$$\begin{aligned} Q(\lambda, \mu) &= \inf_x \quad \mathcal{L}(\mathbf{x}, \lambda, \mu) \\ &= \inf_x \quad J(x) + \sum_{j=1}^{p} \lambda_j h_j(x) + \sum_{i=1}^{q} \mu_i g_i(x) \end{aligned}$$

Wolf dual problem

$$\mathcal{W} = \begin{cases} \max\limits_{\mathbf{x},\lambda\in\mathbf{R}^p, \mu\in\mathbf{R}^q} & \mathcal{L}(\mathbf{x}, \lambda, \mu) \\ \text{with} & \mu_j \geq 0 \quad j = 1, q \\ \text{and} & \nabla J(x) + \sum\limits_{j=1}^{p} \lambda_j \nabla h_j(x) + \sum\limits_{i=1}^{q} \mu_i \nabla g_i(x) = 0 \end{cases}$$

Duality – theorems (2)

Theorem (12.12, 12.13 and 12.14 Nocedal & Wright pp 346)

If f, g and h are convex and continuously differentiable[a], then the solution of the dual problem is the same as the solution of the primal

[a] under some conditions *e.g.* linear independence constraint qualification

$$\begin{aligned}(\lambda^\star, \mu^\star) &= \text{solution of problem } \mathcal{D} \\ \mathbf{x}^\star &= \arg\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda^\star, \mu^\star)\end{aligned}$$

$$\begin{aligned}Q(\lambda^\star, \mu^\star) = \arg\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda^\star, \mu^\star) &= \mathcal{L}(\mathbf{x}^\star, \lambda^\star, \mu^\star) \\ &= J(\mathbf{x}^\star) + \lambda^\star H(\mathbf{x}^\star) + \mu^\star G(\mathbf{x}^\star) = J(\mathbf{x}^\star)\end{aligned}$$

and for any feasible point $\mathbf{x}$

$$Q(\lambda, \mu) \leq J(\mathbf{x}) \qquad \rightarrow \qquad 0 \leq J(\mathbf{x}) - Q(\lambda, \mu)$$

The duality gap is the difference between the primal and dual cost functions

Road map

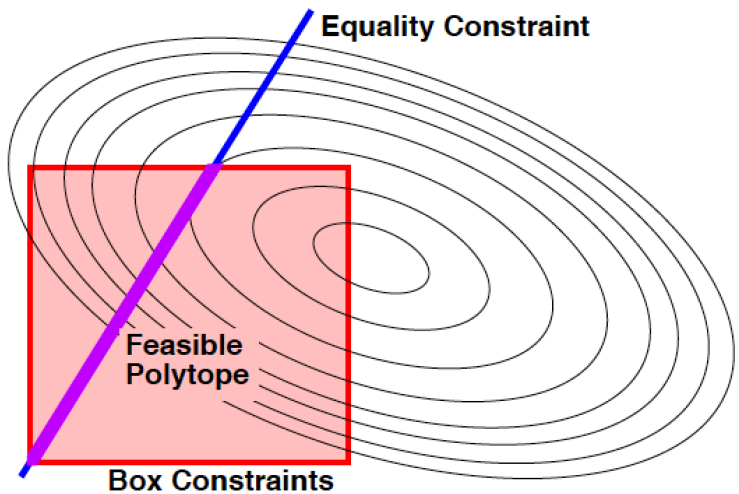


Figure from L. Bottou & C.J. Lin, Support vector machine solvers, in Large scale kernel machines, 2007.

Linear SVM dual formulation - The lagrangian

$$\begin{cases} \min\limits_{\mathrm{w},b} & \frac{1}{2}\|\mathrm{w}\|^2 \\ \text{with} & y_i(\mathrm{w}^\top \mathbf{x}_i + b) \geq 1 \qquad i = 1, n \end{cases}$$

Looking for the lagrangian saddle point $\max\limits_{\alpha} \min\limits_{\mathrm{w},b} \mathcal{L}(\mathrm{w}, b, \alpha)$ with so called lagrange multipliers $\alpha_i \geq 0$

$$\mathcal{L}(\mathrm{w}, b, \alpha) = \frac{1}{2}\|\mathrm{w}\|^2 - \sum_{i=1}^{n} \alpha_i \big(y_i(\mathrm{w}^\top \mathbf{x}_i + b) - 1\big)$$

α_i represents the influence of constraint thus the influence of the training example (x_i, y_i)

Stationarity conditions

$$\mathcal{L}(\mathrm{w}, b, \alpha) = \frac{1}{2}\|\mathrm{w}\|^2 - \sum_{i=1}^{n} \alpha_i \big(y_i(\mathrm{w}^\top \mathbf{x}_i + b) - 1\big)$$

Computing the gradients:

$$\begin{cases} \nabla_{\mathrm{w}} \mathcal{L}(\mathrm{w}, b, \alpha) & = \mathrm{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \\ \dfrac{\partial \mathcal{L}(\mathrm{w}, b, \alpha)}{\partial b} & = \sum_{i=1}^{n} \alpha_i \; y_i \end{cases}$$

we have the following optimality conditions

$$\begin{cases} \nabla_{\mathrm{w}} \mathcal{L}(\mathrm{w}, b, \alpha) = 0 \quad \Rightarrow \quad \mathrm{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \\ \dfrac{\partial \mathcal{L}(\mathrm{w}, b, \alpha)}{\partial b} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \alpha_i \; y_i = 0 \end{cases}$$

KKT conditions for SVM

$$\text{stationarity} \quad \mathrm{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} \alpha_i \, y_i = 0$$

$$\text{primal admissibility} \quad y_i(\mathrm{w}^\top \mathbf{x}_i + b) \geq 1 \qquad i = 1, \ldots, n$$

$$\text{dual admissibility} \quad \alpha_i \geq 0 \qquad i = 1, \ldots, n$$

$$\text{complementarity} \quad \alpha_i \Big(y_i(\mathrm{w}^\top \mathbf{x}_i + b) - 1 \Big) = 0 \qquad i = 1, \ldots, n$$

The complementary condition split the data into two sets

- $\mathcal{A}$ be the set of active constraints: usefull points

$$\mathcal{A} = \{ i \in [1, n] \;\big|\; y_i(\mathrm{w}^{*\top} \mathbf{x}_i + b^*) = 1 \}$$

- its complementary $\bar{\mathcal{A}}$ useless points

$$\text{if } i \notin \mathcal{A}, \alpha_i = 0$$

The KKT conditions for SVM

The same KKT but using matrix notations and the active set $\mathcal{A}$

stationarity	$\mathrm{w} - X^\top D_y \alpha = 0$
	$\alpha^\top y = 0$
primal admissibility	$D_y(Xw + b\mathbb{1}) \geq \mathbb{1}$
dual admissibility	$\alpha \geq 0$
complementarity	$D_y(X_{\mathcal{A}}\mathrm{w} + b\mathbb{1}_{\mathcal{A}}) = \mathbb{1}_{\mathcal{A}}$
	$\alpha_{\bar{\mathcal{A}}} = 0$

Knowing $\mathcal{A}$, the solution verifies the following linear system:

$$\begin{cases} \mathrm{w} & -X_{\mathcal{A}}^\top D_y \alpha_{\mathcal{A}} & & = 0 \\ -D_y X_{\mathcal{A}} \mathrm{w} & & -b\mathbf{y}_{\mathcal{A}} & = -\mathbf{e}_{\mathcal{A}} \\ & -\mathbf{y}_{\mathcal{A}}^\top \alpha_{\mathcal{A}} & & = 0 \end{cases}$$

with $D_y = \mathrm{diag}(\mathbf{y}_{\mathcal{A}})$, $\alpha_{\mathcal{A}} = \alpha(\mathcal{A})$, $\mathbf{y}_{\mathcal{A}} = \mathbf{y}(\mathcal{A})$ et $X_{\mathcal{A}} = X(X_{\mathcal{A}}; :)$.

The KKT conditions as a linear system

$$\left\{\begin{array}{lll} \mathrm{w} & -X_{\mathcal{A}}^{\top} D_y \alpha_{\mathcal{A}} & = 0 \\ -D_y X_{\mathcal{A}} \mathrm{w} & -b\mathbf{y}_{\mathcal{A}} & = -\mathbf{e}_{\mathcal{A}} \\ & -\mathbf{y}_{\mathcal{A}}^{\top} \alpha_{\mathcal{A}} & = 0 \end{array}\right.$$

with $D_y = \mathrm{diag}(\mathbf{y}_{\mathcal{A}})$, $\alpha_{\mathcal{A}} = \alpha(\mathcal{A})$, $\mathbf{y}_{\mathcal{A}} = \mathbf{y}(\mathcal{A})$ et $X_{\mathcal{A}} = X(X_{\mathcal{A}}; :)$.

$$\begin{array}{|c|c|c|} \hline I & -X_{\mathcal{A}}^{\top} D_y & 0 \\ \hline -D_y X_{\mathcal{A}} & 0 & -\mathbf{y}_{\mathcal{A}} \\ \hline 0 & -\mathbf{y}_{\mathcal{A}}^{\top} & 0 \\ \hline \end{array} \quad \begin{array}{|c|} \hline \mathrm{w} \\ \hline \alpha_{\mathcal{A}} \\ \hline b \\ \hline \end{array} \quad = \quad \begin{array}{|c|} \hline 0 \\ \hline -\mathbf{e}_{\mathcal{A}} \\ \hline 0 \\ \hline \end{array}$$

we can work on it to separate w from $(\alpha_{\mathcal{A}}, b)$

The SVM dual formulation

The SVM Wolfe dual

$$\begin{cases} \max\limits_{\mathrm{w},b,\alpha} & \frac{1}{2}\|\mathrm{w}\|^2 - \sum\limits_{i=1}^{n} \alpha_i \big(y_i(\mathrm{w}^\top \mathbf{x}_i + b) - 1\big) \\ \text{with} & \alpha_i \geq 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 1, \ldots, n \\ \text{and} & \mathrm{w} - \sum\limits_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 \text{ and } \sum\limits_{i=1}^{n} \alpha_i \, y_i = 0 \end{cases}$$

using the fact: $\mathrm{w} = \sum\limits_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$

The SVM Wolfe dual without w and b

$$\begin{cases} \max\limits_{\alpha} & -\frac{1}{2} \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} \alpha_j \alpha_i y_i y_j \mathbf{x}_j^\top \mathbf{x}_i + \sum\limits_{i=1}^{n} \alpha_i \\ \text{with} & \alpha_i \geq 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 1, \ldots, n \\ \text{and} & \sum\limits_{i=1}^{n} \alpha_i \, y_i = 0 \end{cases}$$

Linear SVM dual formulation

$$\mathcal{L}(\mathrm{w}, b, \alpha) = \frac{1}{2}\|\mathrm{w}\|^2 - \sum_{i=1}^{n} \alpha_i \big(y_i(\mathrm{w}^\top \mathbf{x}_i + b) - 1\big)$$

Optimality: $\mathrm{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \qquad \sum_{i=1}^{n} \alpha_i \, y_i = 0$

$$\begin{aligned}\mathcal{L}(\alpha) &= \tfrac{1}{2} \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_j \alpha_i y_i y_j \mathbf{x}_j^\top \mathbf{x}_i}_{\mathrm{w}^\top \mathrm{w}} - \sum_{i=1}^{n} \alpha_i y_i \underbrace{\sum_{j=1}^{n} \alpha_j y_j \mathbf{x}_j^\top}_{\mathrm{w}^\top} \mathbf{x}_i - b \underbrace{\sum_{i=1}^{n} \alpha_i y_i}_{=0} + \sum_{i=1}^{n} \alpha_i \\ &= -\frac{1}{2} \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_j \alpha_i y_i y_j \mathbf{x}_j^\top \mathbf{x}_i + \sum_{i=1}^{n} \alpha_i\end{aligned}$$

Dual linear SVM is also a quadratic program

$$\text{problem } \mathcal{D} \quad \begin{cases} \min\limits_{\alpha \in \mathbf{R}^n} & \frac{1}{2}\alpha^\top G \alpha - \mathbf{e}^\top \alpha \\ \text{with} & \mathbf{y}^\top \alpha = 0 \\ \text{and} & 0 \leq \alpha_i \qquad\qquad i = 1, n \end{cases}$$

with G a symmetric matrix $n \times n$ such that $G_{ij} = y_i y_j \mathbf{x}_j^\top \mathbf{x}_i$

SVM primal vs. dual

Primal

$$\begin{cases} \min\limits_{w \in \mathbb{R}^d, b \in \mathbb{R}} & \frac{1}{2}\|w\|^2 \\ \text{with} & y_i(w^\top x_i + b) \geq 1 \\ & i = 1, n \end{cases}$$

- $d + 1$ unknown
- n constraints
- classical QP
- perfect when $d << n$

Dual

$$\begin{cases} \min\limits_{\alpha \in \mathbb{R}^n} & \frac{1}{2}\alpha^\top G \alpha - \mathbf{e}^\top \alpha \\ \text{with} & \mathbf{y}^\top \alpha = 0 \\ \text{and} & 0 \leq \alpha_i \qquad\qquad i = 1, n \end{cases}$$

- n unknown
- G Gram matrix (pairwise influence matrix)
- n box constraints
- easy to solve
- to be used when $d > n$

SVM primal vs. dual

Primal

$$\begin{cases} \min\limits_{w \in \mathbb{R}^d, b \in \mathbb{R}} & \frac{1}{2}\|w\|^2 \\ \text{with} & y_i(w^\top x_i + b) \geq 1 \\ & i = 1, n \end{cases}$$

- $d + 1$ unknown
- n constraints
- classical QP
- perfect when $d << n$

Dual

$$\begin{cases} \min\limits_{\alpha \in \mathbb{R}^n} & \frac{1}{2}\alpha^\top G \alpha - \mathbf{e}^\top \alpha \\ \text{with} & \mathbf{y}^\top \alpha = 0 \\ \text{and} & 0 \leq \alpha_i \qquad\qquad i = 1, n \end{cases}$$

- n unknown
- G Gram matrix (pairwise influence matrix)
- n box constraints
- easy to solve
- to be used when $d > n$

$$f(x) = \sum_{j=1}^{d} w_j x_j + b = \sum_{i=1}^{n} \alpha_i \, y_i (x^\top x_i) + b$$

The bi dual (the dual of the dual)

$$\begin{cases} \min\limits_{\alpha \in \mathbf{R}^n} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top \alpha \\ \text{with} & \mathbf{y}^\top \alpha = 0 \\ \text{and} & 0 \leq \alpha_i \qquad\qquad i = 1, n \end{cases}$$

$$\begin{aligned} \mathcal{L}(\alpha, \lambda, \mu) &= \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top \alpha + \lambda\, \mathbf{y}^\top \alpha - \mu^\top \alpha \\ \nabla_\alpha \mathcal{L}(\alpha, \lambda, \mu) &= G\alpha - \mathbf{e} + \lambda\, \mathbf{y} - \mu \end{aligned}$$

The bidual

$$\begin{cases} \max\limits_{\alpha, \lambda, \mu} & -\frac{1}{2}\alpha^\top G\alpha \\ \text{with} & G\alpha - \mathbf{e} + \lambda\, \mathbf{y} - \mu = 0 \\ \text{and} & 0 \leq \mu \end{cases}$$

since $\|\mathrm{w}\|^2 = \frac{1}{2}\alpha^\top G\alpha$ and $DX\mathrm{w} = G\alpha$

$$\begin{cases} \max\limits_{\mathrm{w}, \lambda} & -\frac{1}{2}\|\mathrm{w}\|^2 \\ \text{with} & DX\mathrm{w} + \lambda\, \mathbf{y} \geq \mathbf{e} \end{cases}$$

by identification (possibly up to a sign)

$b = \lambda$ is the Lagrange multiplier of the equality constraint

Cold case: the least square problem

Linear model

$$y_i = \sum_{j=1}^{d} w_j x_{ij} + \varepsilon_i \quad , \qquad i = 1, n$$

n data and d variables; $d < n$

$$\min_{w} = \sum_{i=1}^{n} \left(\sum_{j=1}^{d} x_{ij} w_j - y_i \right)^2 = \|Xw - \mathbf{y}\|^2$$

Solution: $\widetilde{w} = (X^\top X)^{-1} X^\top \mathbf{y}$

$$f(\mathbf{x}) = \mathbf{x}^\top \underbrace{(X^\top X)^{-1} X^\top \mathbf{y}}_{\widetilde{w}}$$

What is the influence of each data point (matrix X lines) ?

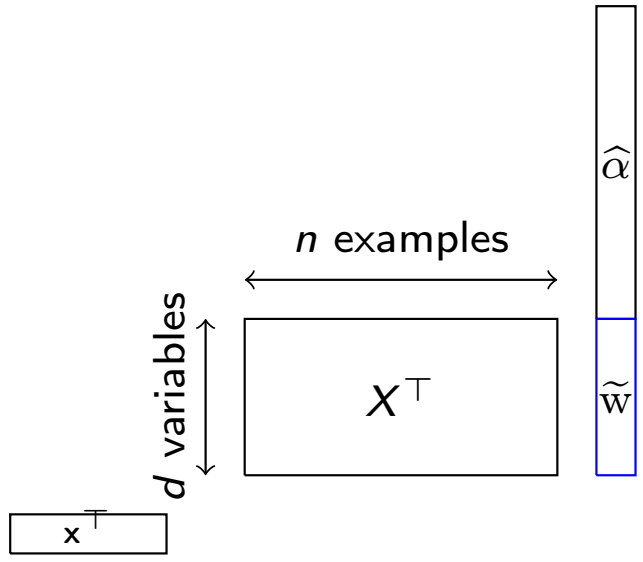
Shawe-Taylor & Cristianini's Book, 2004

data point influence (contribution)

for any new data point $\mathbf{x}$

$$f(\mathbf{x}) = \mathbf{x}^\top (X^\top X)(X^\top X)^{-1} \underbrace{(X^\top X)^{-1} X^\top \mathbf{y}}_{\widetilde{\mathrm{w}}}$$

$$= \mathbf{x}^\top X^\top \underbrace{X(X^\top X)^{-1}(X^\top X)^{-1} X^\top \mathbf{y}}_{\widehat{\alpha}}$$



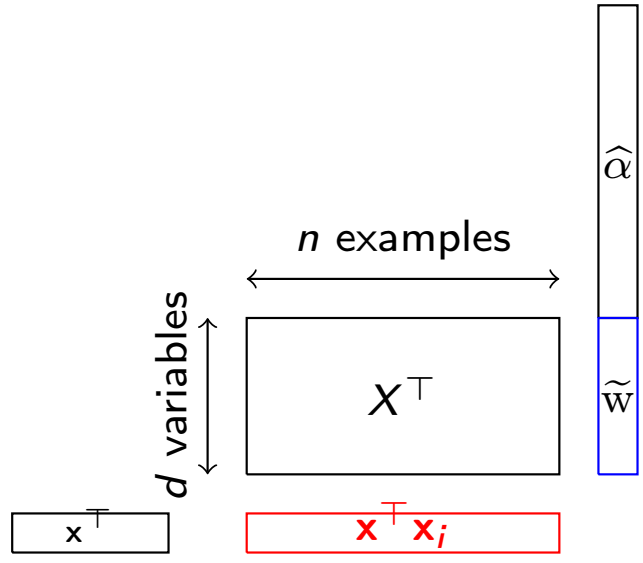
$$f(\mathbf{x}) = \sum_{j=1}^{d} \widetilde{\mathrm{w}}_j x_j$$

data point influence (contribution)

for any new data point $\mathbf{x}$

$$f(\mathbf{x}) = \mathbf{x}^\top (X^\top X)(X^\top X)^{-1} \underbrace{(X^\top X)^{-1} X^\top \mathbf{y}}_{\widetilde{\mathrm{w}}}$$

$$= \mathbf{x}^\top X^\top \underbrace{X(X^\top X)^{-1}(X^\top X)^{-1} X^\top \mathbf{y}}_{\widehat{\alpha}}$$



$$f(\mathbf{x}) = \sum_{j=1}^{d} \widetilde{\mathrm{w}}_j x_j = \sum_{i=1}^{n} \widehat{\alpha}_i \, (\mathbf{x}^\top \mathbf{x}_i)$$

from variables to examples

$$\underbrace{\widehat{\alpha} = X(X^\top X)^{-1} \widetilde{\mathrm{w}}}_{n \text{ examples}} \quad \text{et} \quad \underbrace{\widetilde{\mathrm{w}} = X^\top \widehat{\alpha}}_{d \text{ variables}}$$

what if $d \geq n$!

SVM primal vs. dual

Primal

$$\begin{cases} \min\limits_{w \in \mathbb{R}^d, b \in \mathbb{R}} & \frac{1}{2}\|w\|^2 \\ \text{with} & y_i(w^\top x_i + b) \geq 1 \\ & i = 1, n \end{cases}$$

- $d + 1$ unknown
- n constraints
- classical QP
- perfect when $d << n$

Dual

$$\begin{cases} \min\limits_{\alpha \in \mathbb{R}^n} & \frac{1}{2}\alpha^\top G \alpha - \mathbf{e}^\top \alpha \\ \text{with} & \mathbf{y}^\top \alpha = 0 \\ \text{and} & 0 \leq \alpha_i \qquad\qquad i = 1, n \end{cases}$$

- n unknown
- G Gram matrix (pairwise influence matrix)
- n box constraints
- easy to solve
- to be used when $d > n$

$$f(x) = \sum_{j=1}^{d} w_j x_j + b = \sum_{i=1}^{n} \alpha_i \, y_i (x^\top x_i) + b$$

Road map

1. Optimisation
2. Introduction au gradient et autres dérivées
3. Algorithmes pour l'optimisation sans contraintes
4. Optimisation avec contraintes

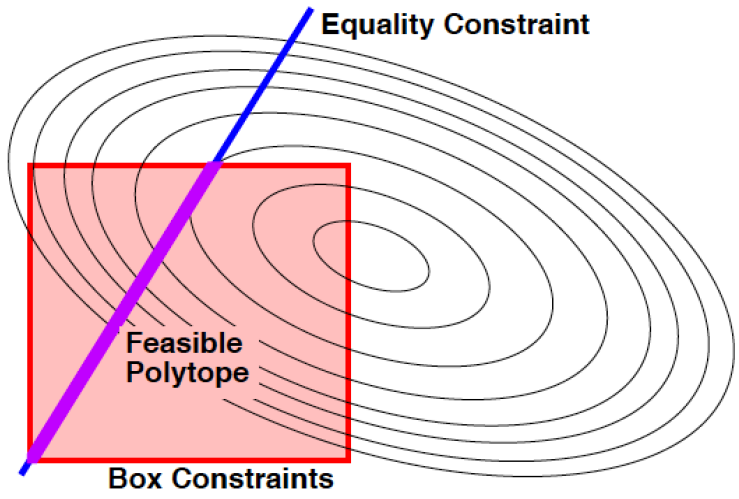
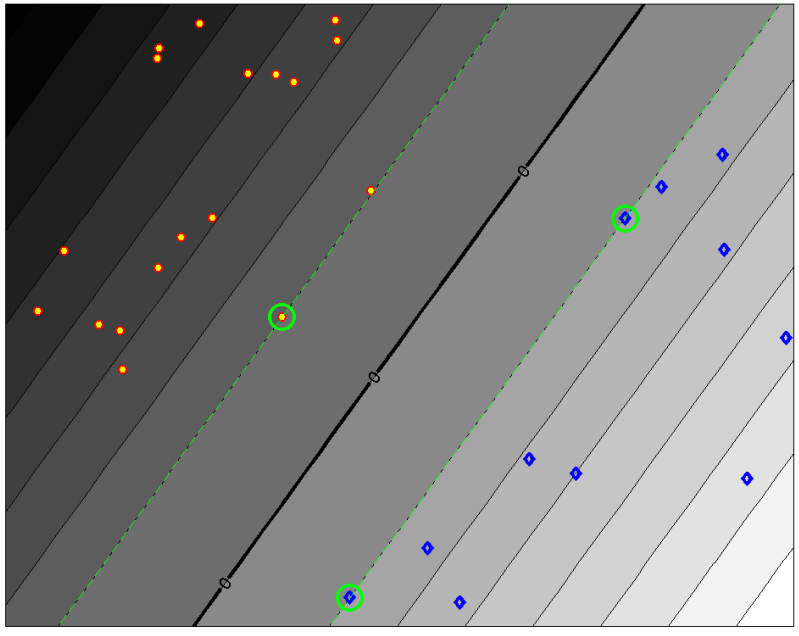


Figure from L. Bottou & C.J. Lin, Support vector machine solvers, in Large scale kernel machines, 2007.

Solving the dual (1)

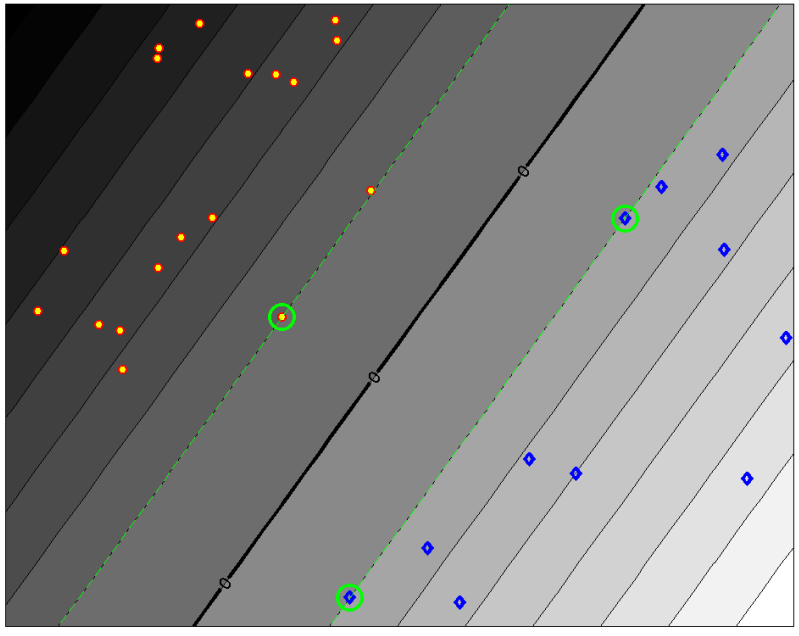


Data point influence

- $\alpha_i = 0$ this point is useless
- $\alpha_i \neq 0$ this point is said to be support

$$f(\mathbf{x}) = \sum_{j=1}^{d} w_j x_j + b = \sum_{i=1}^{n} \alpha_i \, y_i (\mathbf{x}^\top \mathbf{x}_i) + b$$

Solving the dual (1)



Data point influence

- $\alpha_i = 0$ this point is useless
- $\alpha_i \neq 0$ this point is said to be support

$$f(\mathbf{x}) = \sum_{j=1}^{d} w_j x_j + b = \sum_{i=1}^{3} \alpha_i \, y_i (\mathbf{x}^\top \mathbf{x}_i) + b$$

Decison border only depends on 3 points ($d + 1$)

Solving the dual (2)

Assume we know these 3 data points

$$\begin{cases} \min\limits_{\alpha \in \mathbb{R}^n} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top \alpha \\ \text{with} & \mathbf{y}^\top \alpha = 0 \\ \text{and} & 0 \leq \alpha_i \qquad\qquad i = 1, n \end{cases} \Longrightarrow \begin{cases} \min\limits_{\alpha \in \mathbb{R}^3} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top \alpha \\ \text{with} & \mathbf{y}^\top \alpha = 0 \end{cases}$$

$$L(\alpha, b) = \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top \alpha + b\, y^\top \alpha$$

solve the following linear system

$$\begin{cases} G\alpha + & b\,\mathbf{y} & = \mathbf{e} \\ \mathbf{y}^\top \alpha & & = 0 \end{cases}$$

```
U = chol(G);  % upper
a = U\ (U'\e);
c = U\ (U'\y);
b = (y'*a)\(y'*c)
alpha = U\ (U'\(e - b*y));
```

Conclusion: variables or data point?

- seeking for a universal learning algorithm
  - no model for $\mathbb{P}(\mathbf{x}, y)$
- the linear case: data is separable
  - the non separable case
- double objective: minimizing the error together with the regularity of the solution
  - multi objective optimisation
- dualiy : variable – example
  - use the primal when $d < n$ (in the liner case) or when matrix G is hard to compute
  - otherwise use the dual
- universality = nonlinearity
  - kernels