Master 2 Informatique - Parcours IAAA
UE Mathématiques pour l'Intelligence Artificielle

Gradient, Minimum, Convexité et Optimisation

Loïc Neyrat, Yannis Formery

Table des matières

1 Calcul de gradient

Rappel : Le gradient est une application caractérisant la pente d'une fonction en chaque de son ensemble de définition.

Exercice (calcul de gradients). Soient $x \in \mathbf{R}^n$ un vecteur de dimension n et $A \in \mathcal{M}_n(\mathbf{R})$ une matrice réelle de taille n par n. Soient deux fonctions f et g définies par :

$$f : \left| \begin{array}{lll} \mathbf{R}^n & \to & \mathbf{R} \\ x & \mapsto & \|x\|_2 \end{array} \right.$$

$$g : \left| \begin{array}{lll} \mathbf{R}^n & \to & \mathbf{R} \\ x & \mapsto & x^T A x \end{array} \right.$$

Calculer les gradients ∇f et ∇g.

Correction :

— Nous pouvons réécrire la norme 2 de x de la manière suivante :

$$\|x\|_2 = \sqrt{x^T x}$$

Ainsi, f peut être vu comme la composition de deux fonctions f_1 et f_2 (i.e. $f = f_1 \circ f_2$) telles que :

$$f_1 : \left| \begin{array}{lll} \mathbf{R}^+ & \to & \mathbf{R} \\ x & \mapsto & \sqrt{x} \end{array} \right.$$

$$f_2 : \left| \begin{array}{lll} \mathbf{R}^n & \to & \mathbf{R} \\ x & \mapsto & x^T x \end{array} \right.$$

Donc, par application de la Chain Rule, et sachant que f_1 est dérivable sur $\mathbf{R}^+$, le gradient ∇f de f est définie par :

$$\nabla f : \left| \begin{array}{lll} \mathbf{R}^n & \to & \mathbf{R^n} \\ x & \mapsto & \frac{\partial f_1}{\partial f_2} \nabla f_2 \end{array} \right.$$

Calculons la dérivée partielle de f_2 par rapport à une composante x_k de x :

$$\begin{aligned} \frac{\partial}{\partial x_k} f_2 &= \frac{\partial}{\partial x_k} x^T x \\ &= \frac{\partial}{\partial x_k} \sum_{i=1}^{n} x_i^2 \\ &= \frac{\partial}{\partial x_k} x_k^2 \\ &= 2x_k \end{aligned}$$

Enfin, par définition du gradient, et en sachant que $\frac{\partial f_1}{\partial x^T x} = \frac{\partial \sqrt{x^T x}}{\partial x^T x} = \frac{1}{2\sqrt{x^T x}}$, on obtient :

$$\begin{aligned}
\nabla f &= \left[\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\right]^T \\
&= \left[\frac{\partial f_1}{\partial f_2}\frac{\partial}{\partial x_1}f_2, \dots, \frac{\partial f_1}{\partial f_2}\frac{\partial}{\partial x_n}f_2\right]^T \\
&= \frac{\partial f_1}{\partial f_2}\left[\frac{\partial}{\partial x_1}f_2, \dots, \frac{\partial}{\partial x_n}f_2\right]^T \\
&= \frac{\partial f_1}{\partial f_2}\left[\frac{\partial}{\partial x_1}f_2, \dots, \frac{\partial}{\partial x_n}f_2\right]^T \\
&= \frac{1}{2\sqrt{x^T x}}\left[2x_1, \dots, 2x_1\right]^T \\
&= \frac{x}{\|x\|}
\end{aligned}$$

— Le gradient de g est définie comme suit :

$$\nabla g : \left| \begin{array}{lcl} \mathbf{R}^n & \to & \mathbf{R}^n \\ x & \mapsto & \nabla x^T A x = \left[\frac{\partial}{\partial x_1}x^T A x, \dots, \frac{\partial}{\partial x_n}x^T A x\right]^T \end{array} \right. \tag{1}$$

Développons la dérivée partielle de g par rapport à une composante x_k de x :

$$\begin{aligned}
\frac{\partial}{\partial x_k}f_2 &= \frac{\partial}{\partial x_k}x^T A x \\
&= \frac{\partial}{\partial x_k}\sum_{i=1}^{n}\sum_{j=1}^{n} x_i a_{i,j} x_j \\
&= \frac{\partial}{\partial x_k}\sum_{i=1}^{n} x_i a_{i,k} x_k + x_k a_{k,i} x_i \\
&= \frac{\partial}{\partial x_k}\sum_{i=1}^{n} x_i a_{i,k} x_k + x_k a_{k,i} x_i \quad \text{(Si A est symétrique, } a_{i,k} = a_{k,i}\text{)} \\
&= \sum_{i=1}^{n} a_{i,k} x_i + a_{k,i} x_i \\
&= (A^T x)_k + (Ax)_k \\
&= ((A^T + A)x)_k
\end{aligned}$$

Donc par (1) :

$$\begin{aligned}
\nabla g &= \left[\frac{\partial}{\partial x_1}x^T A x, \dots, \frac{\partial}{\partial x_n}x^T A x\right]^T \\
&= \left[((A^T + A)x)_1, \dots, ((A^T + A)x)_n\right]^T \\
&= (A^T + A)x
\end{aligned}$$

2 Condition d'optimalité à l'ordre 1

Rappel sur les ouverts : Soit U un ensemble de $\mathbb{R}^n$. On dit que U est un ouvert de $\mathbb{R}^n$ si et seulement si $\forall u \in U, \exists\, r > 0$ tel que $\{x \in \mathbb{R}^n \mid \|x - u\|_2 < r\} \subset U$

Théorème 1 (Conditions d'optimalité nécessaires à l'ordre 1) *Si $x^\star$ est un minimum local et f est de classe $\mathcal{C}^1$ (dérivable une fois) sur un ouvert $U \in \mathbb{R}^n$ contenant $x^\star$, alors $\nabla f(x^\star) = 0$*

Puisque tout maximum global est un maximum local, il est utile de noter que le théorème 1 peut s'appliquer dans le cas d'un maximum global. Attention cependant, car la réciproque n'est pas vraie : si $\nabla f(x) = 0$ pour un x, selon ne veut pas forcément dire que f atteint un minimum local en x. En effet, x peut être un point selle ou plus généralement tout autre point où f admet une tangente constante.

Exemple d'ensemble ouvert :
Soit f une fonction définie par :

$$f : \left| \begin{array}{ccc} [0;\pi] & \to & \mathbf{R} \\ x & \mapsto & \sin(x) \end{array} \right.$$

dont on suppose d'être de classe $\mathcal{C}^1$ et d'admettre un minimum global sur $[0;\pi]$.
f est définie sur l'ensemble fermé $[0;\pi]$. Restreignons f à l'ouvert $]0;\pi[$ et notons $f_{]0;\pi[}$ cette restriction. Par disjonction de cas :

1. Si $x^\star$ est un minimum global de $f_{]0;\pi[}$, alors $x^\star$ est un minimum local se $f_{]0;\pi[}$ et $\nabla f(x^\star) = 0$. Soit $x \in]0;\pi[$ tel que $\nabla f(x) = \cos(x) = 0$. Alors $x = \frac{\pi}{2}$. Donc $x^\star = \frac{\pi}{2}$ et $\sin(x^\star) = 1$.
2. Si $x^\star$ est un minimum global de $f_{]0;\pi[}$, et $x^\star \notin]0;\pi[$, alors $x^\star = 0$ ou $x^\star = \pi$ donc $f(x^\star) = 0$. Or $\sin(\frac{\pi}{2}) = 1 > \sin(\pi) = \sin(0) = 0$ donc $\{0, \pi\}$ sont des minimas globaux de f.

3 Analyse : matrice hessienne

Soit f une fonction numérique

$$f : U \subset \mathbf{R}^n \to \mathbf{R}$$

et

$$x = [x_1, \ldots, x_n]^T \in \mathbf{R}^n$$

Si toutes les dérivées partielles seconde de f existent, la matrice hessienne de f, notée $\nabla^2 f$ ou $H(f)$, est définie par :

$$\nabla^2 f = H(f) = \begin{bmatrix} \frac{\partial}{\partial^2 x_1} f & \frac{\partial}{\partial x_1 \partial x_2} f & \cdots & \frac{\partial}{\partial x_1 \partial x_n} f \\ \frac{\partial}{\partial^2 x_2} f & \frac{\partial}{\partial^2 x_2} f & \cdots & \frac{\partial}{\partial x_2 \partial x_n} f \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_n \partial x_1} f & \frac{\partial}{\partial x_n \partial x_2} f & \cdots & \frac{\partial}{\partial^2 x_n} f \end{bmatrix}$$

DDe manière plus intuitive, la matrice hessienne est une application décrivant une forme quadratique en chaque point de l'ensemble de définition d'une fonction (par exemple, si c'est une parabole,...). **Un exemple :** Posons $M \in \mathcal{M}_n(\mathbf{R})$ et

$$f : \left| \begin{array}{ccc} \mathbf{R}^n & \to & \mathbf{R} \\ x & \mapsto & x^T M x \end{array} \right.$$

En nous servant des résultats obtenus lors de la résolution de l'exercice de calculs de gradients, on a :

$$\nabla f = (M^T + M)x$$

Alors,

$$\nabla^2 f = H(f) = \begin{bmatrix} \frac{\partial}{\partial^2 x_1} f & \frac{\partial}{\partial x_1 \partial x_2} f & \cdots & \frac{\partial}{\partial x_1 \partial x_n} f \\ \frac{\partial}{\partial^2 x_2} f & \frac{\partial}{\partial^2 x_2} f & \cdots & \frac{\partial}{\partial x_2 \partial x_n} f \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_n \partial x_1} f & \frac{\partial}{\partial x_n \partial x_2} f & \cdots & \frac{\partial}{\partial^2 x_n} f \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1}(M^T + M)x \\ \frac{\partial}{\partial x_2}(M^T + M)x \\ \vdots \\ \frac{\partial}{\partial x_n}(M^T + M)x \end{bmatrix} = M^T + M$$

4 Conditions d'optimalité à l'ordre 2

Les deux théorèmes suivants explicitent les conditions nécessaires suffisantes d'optimalité à l'ordre 2.

Théorème 2 (Conditions nécessaires d'optimalité à l'ordre 2) *Si $x^\star$ est un minimum local et f est de classe $\mathcal{C}^2$ sur un ouvert $U \subset \mathbf{R}^n$ contenant x, alors $\nabla f(x^\star) = 0$ et $\nabla^2 f(x^\star)$ est semi-définie positive.*

Théorème 3 (Conditions suffisantes d'optimalité à l'ordre 2) *Si f est de classe $\mathcal{C}^2$ sur un ouvert $U \subset \mathbf{R}^n$ contenant x, si $\nabla f(x^\star) = 0$ et $\nabla^2 f(x^\star)$ est définie positive, alors $x^\star$ est un minimum local.*

Attention : contrairement à l'ordre 1, les théorèmes précédents ne s'appliquent qu'aux minimas locaux.
De plus, pour qualifier un maximum local, il est nécessaire que $-H$ soit semi-définie positive.

5 Existence du minimum

Voici deux théorèmes importants permettant de montrer l'existence d'un ou de plusieurs minimas globaux d'une fonction sous certaines conditions.

Théorème 4 (Théorème de Weierstrass) *Si f est définie et continue sur un sous-ensemble fermé et borné de $\mathbf{R}^n$, alors f admet un minimum global.*

Théorème 5 *Si f est définie et continue sur $\mathbf{R}^n$ et coercive (i.e. $f(x) \to +\infty$ quand $\|x\| \to +\infty$, alors f admet un minimum global sur tout sous-ensemble fermé de $\mathbf{R}^n$.*

Note : une fonction est dite *coercive* si elle tend vers $+\infty$ lorsque x s'éloigne en norme de l'origine. De façon plus formelle, f est coercive si $\|x\| \to +\infty \Rightarrow f(x) \to +\infty$

6 Unicité du minimum

Théorème 6 (Unicité du minimum) *Si f est une fonction strictement convexe définie sur un ensemble convexe $X \subset \mathbb{R}^n$, alors f admet au plus un minimum global.*

7 Convexité

Théorème 7 (Convexité d'un ensemble) *L'ensemble X est convexe si et seulement si $\forall\, x, y \in X$, le segment $[x, y] = \{tx + (1-t)y \mid t \in [0,1]\}$ est inclus dans X.*

Théorème 8 (Convexité d'une fonction à valeurs réelles) *Une fonction $f : X \subset \mathbb{R}^n \to R$ est convexe si et seulement si X est convexe et $\forall\, x, y \in X$ et $\forall\, t \in [0,1]$,*

$$f(tx + (1t)y) \leq tf(x) + (1-t)f(y)$$

Si l'inégalité est stricte pour $t \in]0,1[$, la fonction est dite strictement convexe.

Ces deux théorèmes, associés au théorème de l'unicité du minimum, permettent notamment de prouver que si une fonction f définie sur un ensemble convexe $X \subset \mathbf{R}^n$ est strictement convexe, alors f admet au plus un minimum global.

Un autre théorème utile est le suivant :

Théorème 9 *Si f est une fonction convexe définie sur un ensemble convexe $X \subset \mathbb{R}^n$, alors tout minimum local de f est aussi un minimum global de f. Si X est ouvert, alors $\nabla f(x^\star) = 0$ et $x^\star$ est un minimum global de f.*

Voici une liste de quelques familles de fonctions convexes à savoir reconnaître. **Astuces pour reconnaître qu'une fonction est convexe :**

— les fonctions linéaires sont convexes.
— les combinaisons linéaires positives (à poids positifs) de fonctions convexes sont convexes.
— le maximum (point par point) d'un ensemble de fonctions convexes est convexe.

Enfin, pour reconnaître une fonction convexe, on peut utiliser le théorème suivant :

Théorème 10 *Si f est de classe $\mathcal{C}^2$, alors f est convexe si et seulement si $\nabla^2(f)$ est semi-définie positive sur l'intérieur de X. Si $\nabla^2(f)$ est définie positive sur l'intérieur de X, f est strictement convexe.*

Exercice d'application :
Déterminons la convexité ou non-convexité de la fonction f définie par :

$$f : \left| \begin{array}{lcl} \mathbf{R}^2 & \to & \mathbf{R} \\ x, y & \mapsto & xy + x^2 + 2y + 1 \end{array} \right.$$

Calculons son gradient et sa matrice hessienne :

$$\nabla f(x,y) = \begin{bmatrix} \frac{\partial}{\partial x} f(x,y) \\ \frac{\partial}{\partial y} f(x,y) \end{bmatrix} = \begin{bmatrix} y + 2x \\ x + 2 \end{bmatrix}$$

$$\nabla^2 f = H(f) = \begin{bmatrix} \frac{\partial}{\partial^2 x} f & \frac{\partial}{\partial x \partial y} f \\ \frac{\partial}{\partial y \partial x x} f & \frac{\partial^2}{\partial y} f \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix}$$

Son polynôme caractéristique est donné par :

$$\chi = \det(\nabla^2 f) = |\nabla^2 f - XI| = (X - 1 - \sqrt{2})(X - 1 + \sqrt{2})$$

Par conséquent, toutes les valeurs propres de la matrice hessienne ne sont pas toutes positives (par exemple, $1 - \sqrt{2} < 0$ est une racine de χ). Donc f n'est pas convexe.
De façon générale, pour montrer qu'une n'est pas convexe dans les conditions de l'exercice précédent, il faut soit montrer que la matrice hessienne est (semi)-définie positive, ou alors montrer qu'il existe $(u, v) \in \mathbf{R}^2$ tels que :

$$\begin{bmatrix} u & v \end{bmatrix} \nabla^2 f(x,y) \begin{bmatrix} u \\ v \end{bmatrix} < 0$$

Une dernière approche se base sur la formule définissant le déterminant d'une matrice A comme $\det(A) = \prod_i \lambda_i$, $\lambda_i \in sp(A)$ avec $sp(A)$ le spectre de A. Elle consiste à montrer que $\det(\nabla^2 f(x,y)) < 0$, c'est-à-dire qu'il existe une valeur propre λ de $\nabla^2 f(x,y)$ strictement négative, et donc que $\nabla^2 f(x,y)$ n'est pas semi-définie positive.

8 Application 1 : Méthode des moindres aux carrés

Soit

$$C = \|Ax - y\|_2$$

avec $A \in \mathcal{M}_{m,n}(\mathbf{R})$, $x \in \mathbf{R}^n$ et $y \in \mathbf{R}^m$. Montrer que C admet un minimum global $x^\star$ et donner une expression de ce minimum.
Le problème posé revient à résoudre le problème d'optimisation :

$$x^\star = \min_{x \in \mathbf{R}^n} C$$

L'application d'une fonction croissante ne change pas l'antécédent du minimum par C. Or la fonction carrée est croissante sur $\mathbf{R}^+$. Donc minimiser C revient à minimiser C^2.
On pose

$$C^2 : \min_{x \in \mathbf{R}^n} \|Ax - y\|_2^2$$

Calculons le gradient C^2.

$$\begin{aligned}
\nabla C^2 &= \nabla \|Ax - y\|_2^2 \\
&= \nabla (Ax - y)^T (Ax - y) \\
&= \nabla (x^T A^T A x - x^T A^T y - y^T A x + y^T y) \\
&= \nabla (x^T A^T A x) - \nabla (x^T A^T y) - \nabla (y^T A x) + \nabla (y^T y) \\
&= \nabla (x^T A^T A x) - \nabla (x^T A^T y) - \nabla (y^T A x) && \text{Car } \nabla y^T y = 0\text{, puisque ce terme ne dépend pas de } x \\
&= \left[A^T A + (A^T A)^T\right] x - 2A^T y && \text{Car } \nabla y^T A x \in \mathbf{R} \text{ et } (y^T A x = x^T A^T y)^T\text{, donc } y^T A x = x^T A^T y \\
&= 2A^T A x - 2A^T y \\
&= 2A^T (Ax - y)
\end{aligned}$$

De même, on peut calculer la hessienne de C^2 :

$$\begin{aligned}
\nabla^2 C^2 &= \nabla \nabla C^2 \\
&= \nabla 2A^T (Ax - y) \\
&= \nabla 2A^T A x - 2A^T y \\
&= 2A^T A
\end{aligned}$$

Or on sait que les valeurs propres de $A^T A$ sont les carrés des valeurs singulières de A. Donc $\nabla^2 C^2$ est semi-définie positive, nous permettant de conclure que C^2 est convexe. Il restera alors à montrer que C^2 admet un minimum global. Une fois l'existence d'un tel minimum démontrée, on a, si $A^T A$ est inversible :

$$\begin{aligned}
&\nabla C^2 = 0 \\
\Leftrightarrow\ & 2A^T (Ax^\star - y) = 0 \\
\Leftrightarrow\ & A^T (Ax^\star - y) = 0 \\
\Leftrightarrow\ & A^T A x^\star - A^T y = 0 \\
\Leftrightarrow\ & A^T A x^\star = A^T y \\
\Leftrightarrow\ & x^\star = (A^T A)^{-1} A^T y
\end{aligned}$$

9 Matrice pseudo-inverse

Une vision non-formelle et imparfaite d'une matrice pseudo-inverse est la généralisation de la notion de matrice inverse aux matrices rectangulaires.

Soit A une matrice réelle de taille m par n et de rang r. Soit $A = U\Sigma V^T$ sa décomposition en valeurs singulières avec, en notant σ_i la i^e valeur propre de A :

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 & \dots & \dots & 0 \\ 0 & \sigma_2 & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \sigma_r & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & 0 & 0 & \dots & 0 & 0 \end{bmatrix}$$

Alors on définit le conjugué de Σ comme :

$$\Sigma^\dagger = \begin{bmatrix} 1/\sigma_1 & 0 & \dots & 0 & \dots & \dots & 0 \\ 0 & 1/\sigma_2 & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & 1/\sigma_r & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & 0 & 0 & \dots & 0 & 0 \end{bmatrix}$$

et la pseudo-inverse de A comme :

$$A^\dagger = V\Sigma^\dagger$$

Ainsi, on a que $AA^\dagger = U\Sigma\Sigma^\dagger U^T = \Sigma\Sigma^\dagger$

$$AA^\dagger = U\Sigma\Sigma^\dagger U^T = \Sigma\Sigma^\dagger = \begin{bmatrix} I_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

Avec I_r la matrice identité de taille r.

Pour plus de détails, voir Moore-Penrose Pseudo-inverse

10 Descente de gradient

'Backtracking' avec la règle d'Armijo

Input: $x_0 \in \mathbf{R}^n, f : \mathbf{R}^n \to \mathbf{R}$ de classe $\mathcal{C}^1, s > 0, 0 < \beta < 1, 0 < \sigma < 1$.

$$\text{Iteration : } x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$
$$\alpha_k = \beta^{m_k} s$$

m_k plus petit entier positif tel que

$$f(x_k) - f(x_{k+1}) \geq \sigma\beta^{m_k} s \nabla f(x_k)^T \nabla f(x_k)$$

15

'Backtracking' avec la règle d'Armijo

Input: $x_0 \in \mathbf{R}^n, f : \mathbf{R}^n \to \mathbf{R}$ de classe $\mathcal{C}^1, s > 0, 0 < \beta < 1, 0 < \sigma < 1$.

$$\text{Iteration :}\ \ x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

$$\alpha_k = \beta^{m_k} s$$

m_k plus petit entier positif tel que

$$f(x_k) - f(x_{k+1}) \geq \sigma \beta^{m_k} s \nabla f(x_k)^T \nabla f(x_k)$$

Convergence

Soit (x_k) la suite des points générés par l'algorithme de descente de gradient avec pas choisi par la règle d'Armijo. Alors, tout point limite (i.e. valeur d'adhérence) de (x_k) est un point stationnaire.

De plus, si x^* est le seul point stationnaire de f dans un ensemble ouvert, il existe un ensemble ouvert S contenant x^* tel que si il existe k_0 tel que $x_{k_0} \in S$, alors $x_k \in S$ pour tout $k \geq k_0$ et $x_k \to x^*$.

Vitesse de convergence ?

11 Conditions nécessaires d'optimalité : intuition

Soit f définie sur $\mathbf{R}^n$ une fonction à minimiser sous les contraintes c_i. Le problème d'optimisation sous contraintes s'exprime de la manière suivante :

$$\min_{x \in \mathbf{R}^n} f(x) \text{ subject to } \begin{cases} c_i(x) = 0, & i \in \mathcal{E} \\ c_i(x) \geq 0, & i \in \mathcal{I} \end{cases}$$

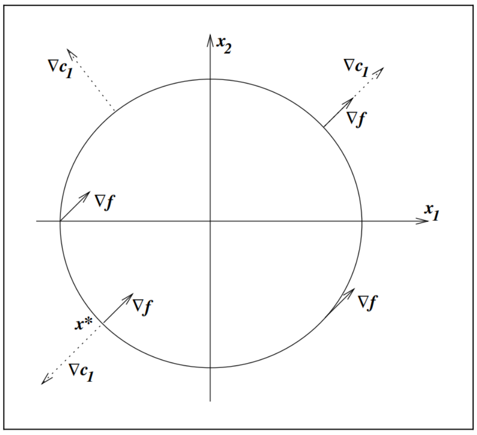
Théorème 11 *$x^\star$ est une solution locale du problème si et seulement si $x^\star \in \Omega$ et il existe un ensemble ouvert $U \subset \mathbf{R}^n$ contenant $x^\star$ tel que pour tout $x \in U \cap \Omega$, $f(x^\star) \leq f(x)$*

Exemple : Soit

$$f : \left| \begin{array}{lcl} \mathbf{R}^2 & \to & \mathbf{R} \\ x_1, x_2 & \mapsto & x_1^2 + x_2^2 - 2 \end{array} \right. \tag{2}$$

Définissons la contrainte $c_1 : x \mapsto x_1^2 + x_2^2 - 2$

Soit $A(x)$ *l'active set*, c'est-à-dire l'ensemble des indices des contraintes d'égalité et d'inégalité tels que $c_i(x) = 0$ pour un point x donné.



La solution se trouve au point où la tangente de f est orthogonale au gradient. Aussi, la normale à l'espace de contraintes est l'opposé du gradient.

12 Conditions d'optimalité nécessaires : Karush-Kuhn-Tucker

Les conditions de Karush-Kuhn-Tucker constituent une méthode de résolution de problèmes d'optimisation à contraintes d'inégalités, généralisant la méthode des multiplicateurs de Lagrange.
Considérons par exemple le problème suivant :

$$\mathcal{L}(x, \lambda) = f(x) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i c_i(x)$$

sous contraintes

$$\begin{aligned} \nabla_x \mathcal{L}(x^\star, \lambda^\star) &= 0, \\ c_i(x^\star)) &= 0, \quad \text{for all } i \in \mathcal{E} \\ c_i(x^\star)) &\geq 0, \quad \text{for all } i \in \mathcal{I} \\ \lambda_i^\star &\geq 0, \quad \text{for all } i \in \mathcal{I} \\ \lambda_i^\star c_i(x^\star) &= 0, \quad \text{for all } i \in \mathcal{E} \cup \mathcal{I} \end{aligned}$$

Comme on sait que $x^\star$ est la valeur de x en laquelle f atteint son minimum, on cherche à annuler les contraintes indiquées ci-dessus pour ce $x^\star$ fixé en trouvant les $\lambda_i^\star$ adéquats . D'où les $\lambda_i^\star c_i(x^\star) = 0, \text{ for all } i \in \mathcal{E} \cup \mathcal{I}$. Les coefficients de Lagrange $\lambda_i^\star$ sont toujours positifs et on retrouve les contraintes sur les $c_i(x^\star)$ imposées dans le problème d'optimisation défini ci-dessus.