Optimisation sous contraintes

Université Paris Cité 2022-2023
M1 MMA

Camille Pouchol

1 Généralités **2**
1.1 Contexte 2
1.2 Différentiabilité 4
1.3 Convexité 7
1.4 Problèmes bien posés et unicité 10

2 Approche agnostique **12**
2.1 Conditions d'optimalité du premier ordre . . 12
2.2 Projection sur un convexe fermé 15
2.3 Algorithme du gradient projeté 18

3 Dualité de Lagrange **20**
3.1 Approche géométrique 20
3.2 Lagrangien, dualités faible et forte 24
3.3 Conditions de KKT, le retour 28
3.4 Algorithme d'Uzawa 31

4 Compléments d'algorithmique **34**
4.1 Analyse de l'algorithme du gradient 34
4.2 Taux optimal et accélération de Nesterov . . . 37

A Théorèmes de séparation **39**

1 Généralités

1.1 Contexte

1.1.1 Vocabulaire de base

La théorie de l'optimisation a pour objectif

$$\text{minimiser } f(x) \text{ sous la contrainte } x \in S,$$

où S est un ensemble donné et $f : S \to \mathbb{R}$.On notera ce problème sous la forme concise

$$\begin{aligned} \min. \quad & f(x) \\ \text{s.c.} \quad & x \in S, \end{aligned} \tag{P}$$

où "min." est une abréviation pour "minimiser", et "s.c." pour "sous la/les contrainte(s)".

Si au contraire il s'agit de maximiser une fonction f, on se ramènera au cas ci-dessus en minimisant $-f$. Notons qu'on rencontrera parfois le problème de minimisation sur S d'une fonction à valeurs dans $\mathbb{R}\cup\{+\infty\}$ (ou à valeurs dans $\mathbb{R}\cup\{-\infty\}$ si l'on fait face à un problème de maximisation). On peut alors considérer le problème tel quel, ou se débarrasser de $+\infty$ en considérant le problème de minimiser f sur $\widetilde{S}$, où $\widetilde{S} = \{x \in S, f(x) < +\infty\}$. Les deux problèmes sont bien sûr strictement équivalents.

Problème bien posé. Commençons par préciser ce qu'on entend par "résoudre" un tel problème, du moins dans ce cours. Il s'agit dans un premier temps d'établir que le problème est *bien posé*, c'est-à-dire successivement que

(i) $S \neq \emptyset$,
(ii) $\inf_{x\in S} f(x) > -\infty$,
(iii) $\arg\min_{x\in S} f(x) \neq \emptyset$.

On notera toujours

$$p^\star := \inf_{x\in S} f(x).$$

La première condition (i) peut paraître saugrenue : qui diable prendrait $S = \emptyset$? Pourtant, la mention explicite à (i) est faite à dessein pour deux raisons principales. Tout d'abord, il n'est pas toujours clair qu'un ensemble S à l'expression complexe ne soit pas vide : il faut le vérifier. Deuxièmement, il arrive parfois qu'on pose un problème d'optimisation de la forme (P) non parce qu'on est intéressé par l'optimum, mais parce qu'on cherche à démontrer que $S \neq \emptyset$! Un point $x \in S$ est appelé point *admissible* pour le problème d'optimisation.

Ces différentes étapes sont parfaitement définies mathématiquement.

Objectifs. Si le problème est bien posé, la théorie de l'optimisation s'attelle dans un second temps à

(i) déterminer *le minimum*

$$p^\star = \min_{x\in S} f(x),$$

(ii) déterminer l'ensemble des *minimiseurs* (ou des points *optimaux*)

$$\{x \in S,\ f(x) = p^\star\} = \arg\min_{x\in S} f(x).$$

Ces deux étapes sont, elles, informelles : on cherche à obtenir des informations sur le réel $p^\star$ et l'ensemble des minimiseurs. Théoriquement, ce sont ce qu'on appelle des *conditions d'optimalité*, qui peuvent être nécessaires ou suffisantes. Numériquement, il s'agit de proposer des algorithmes générant une suite (x_k) telle que la suite réelle $(f(x_k))$ converge vers $p^\star$, et si possible, telle que la suite (x_k) converge vers un minimiseur.

Poursuivons en précisant la place occupée par ce cours au sein de la théorie de l'optimisation.

- *Optimisation continue.* Tous les problèmes abordés seront des problèmes d'optimisation *continue* (par opposition à des problèmes d'optimisation *discrète*). Plus précisément, l'ensemble S sera toujours un sous-ensemble de $\mathbb{R}^n$ pour un certain $n \in \mathbb{N}^*$. Par ailleurs, nombre de résultats établis dans le cours se généralisent à la situation où S est un sous-ensemble d'un espace de Hilbert. Pour ne pas alourdir l'exposition, on se contentera de travailler dans $\mathbb{R}^n$.
- *Optimisation sous contraintes.* Ce cours est plus particulièrement consacré à l'optimisation *sous contraintes*, c'est-à-dire à la situation où S est un sous-ensemble **strict** de l'espace ambiant $\mathbb{R}^n$. Néanmoins, de nombreux résultats s'appliqueront également au cas où le problème n'est pas contraint.
- *Optimisation convexe.* L'essentiel du cours se focalise sur les fonctions convexes (auquel cas l'ensemble S sera noté C et supposé convexe), mais le cas de fonctions et ensembles non convexes sera aussi abordé çà-et-là.
- *Optimisation différentiable.* Bien que cela ne figure pas dans son titre, les fonctions considérées dans ce cours seront toutes au moins *différentiables* (au sens de Gâteaux dans le pire des cas) sur un ouvert de $\mathbb{R}^n$ contenant S. Pour l'optimisation non différentiable (mais convexe), il vous faudra attendre votre cours d'optimisation de M2.

1.1.2 Contraintes

Dans ce cours, on abordera dans un premier temps la situation de contraintes écrites de manière générale et abstraite "$x \in S$", auquel cas on sera amené à considérer des propriétés géométriques et topologiques de S parmi la liste ci-dessous, relativement exhaustive :

- S est non vide,
- S est fermé,
- S est borné,
- $S = C$ est convexe.

Une partie entière de ce cours se spécialisera au cas où, f étant définie sur tout $\mathbb{R}^n$, S s'écrit sous la forme suivante :

$$S = \Big\{ x \in \Omega, \;\; f_i(x) \leq 0,\, i \in \{1, \dots, m\} \text{ et } h_j(x) = 0,\, j \in \{1, \dots, p\} \Big\},$$

les contraintes étant données par les fonctions $f_i : \Omega \to \mathbb{R}$, $i \in \{1, \dots, m\}$, $h_j : \Omega \to \mathbb{R}$, $j \in \{1, \dots, p\}$. Le problème d'optimisation s'écrira alors

$$\begin{array}{rll} \min. & f(x) & \\ \text{s.c.} & f_i(x) \leq 0, & i = 1, \dots, m \\ & h_j(x) = 0, & j = 1, \dots, p, \end{array} \tag{1.1}$$

où par convention on choisit d'écrire les contraintes inégalités avec $\leq$ (une contrainte d'inégalité $u(x) \geq 0$ étant équivalente à $-u(x) \leq 0$).

Bien entendu, tout ensemble S ne s'écrit pas nécessairement sous forme de contraintes d'égalité et d'inégalité, mais un des savoir-faire de l'optimisation consiste à connaître quels ensembles peuvent l'être, et surtout, de quelle manière il est bon de le faire.

En effet, l'écriture sous cette forme n'est pas unique, loin de là. Par exemple, on peut remplacer les p contraintes égalités par les $2p$ contraintes inégalités $h_j(x) \leq 0$ et $-h_j(x) \leq 0$ pour $j \in \{1, \dots, p\}$. À l'inverse, il est possible de ne garder que des contraintes égalités, voire même de résumer la contrainte sous la forme $h(x) = 0$ pour une seule fonction h (voir un exercice de TD).

Bien que ces différentes écritures soient équivalentes au sens où elle ne change pas le minimum ni l'ensemble des minimiseurs, elles ne se valent pas. Il sera en effet crucial d'impliquer des fonctions qui soient si possible

- différentiables sur un ouvert $\Omega \subset \mathbb{R}^n$,

- convexes (voire affines pour des contraintes d'égalité),
- indépendantes.

L'*indépendance* est à ce stade évoquée sans définition précise, mais elle fera l'objet de discussions dans le cadre de la dualité de Lagrange. Le choix des fonctions a un impact considérable sur

- l'applicabilité de résultats théoriques, que ce soit pour assurer que le problème est bien posé (et a, éventuellement, une unique position) ou obtenir des informations sur les minimiseurs,
- l'application fructueuse des divers algorithmes que nous introduirons.

1.2 Différentiabilité

Dans toute cette section, on considère $f : \Omega \to X$, avec Ω ouvert de $\mathbb{R}^n$, et X un $\mathbb{R}$-espace vectoriel normé de dimension finie. On rappelle que $L(\mathbb{R}^n, X)$ fait référence aux applications linéaires (donc continues car on est en dimension finie) de $\mathbb{R}^n$ dans X, et que dans le cas où $X = \mathbb{R}$, on note aussi $(\mathbb{R}^n)^* = L(\mathbb{R}^n, \mathbb{R})$ le dual de $\mathbb{R}^n$.

Enfin, toutes les notions de limite et de régularité afférente sont à entendre pour une norme quelconque puisque toutes les normes y sont équivalentes. Quoi qu'il en soit, la notation $\|\cdot\|$ fera toujours référence à la norme euclidienne.

1.2.1 Différentiabilité au sens de Gâteaux et Fréchet

Définition 1.1. On dit que $f : \Omega \to X$ est Gâteaux-différentiable (abrévié G-différentiable) en $x \in \Omega$ s'il existe une application linéaire (continue) $A \in L(\mathbb{R}^n, X)$ telle que

$$\forall h \in \mathbb{R}^n, \quad \frac{f(x+th) - f(x)}{t} \underset{t\to 0}{\longrightarrow} A(h).$$

Dans ce cas, A est unique, on la note $d^G f(x)$, et on dit que f est G-différentiable sur Ω si elle est G-différentable en tout x de Ω.

En particulier, une fonction G-différentiable est telle que, pour chaque $x \in \Omega$ et $h \in \mathbb{R}^n$, l'application $t \mapsto f(x+th)$ est dérivable en 0, de dérivée égale à $d^G f(x)$.

Attention, la notion de différentiabilité au sens de Gâteaux est assez faible, puisqu'on peut construire des fonctions G-différentiables en un point sans même qu'elles y soient continues. Cette notion faible se trouve néanmoins être assez naturelle en optimisation, puisqu'elle suffit pour écrire certains développements limités d'ordre 1.

Rappelons tout de même la notion plus forte, et plus standard, celle de la différentiabilité au sens de Fréchet.

Définition 1.2. On dit que $f : \Omega \to X$ est Fréchet-différentiable (abrévié F-différentiable) en $x \in \Omega$ s'il existe une application linéaire (continue) $B \in L(\mathbb{R}^n, X)$ telle que

$$f(x+h) = f(x) + B(h) + o(h).$$

Dans ce cas, B est unique, on la note $d^F f(x)$, et on dit que f est F-différentiable sur Ω si elle est F-différentiable en tout x de Ω.

Proposition 1.3

Si f est F-différentiable en $x \in \Omega$, elle est G-différentiable en x et $d^G f(x) = d^F f(x)$.

On note alors indifféremment $df(x)$ la différentielle de f en $x \in \Omega$.

Démonstration : Il suffit d'écrire la F-différentiabilité en th et faire tendre t vers 0. ■

Dans le cas particulier où $X = \mathbb{R}$ et si f est G-différentiable en $x \in \Omega$, $d^G f(x) \in (\mathbb{R}^n)^* = L(\mathbb{R}^n, \mathbb{R})$. Alors, par le théorème de représentation de Riesz [1] dans le cas de la dimension finie, il existe un unique vecteur $p \in \mathbb{R}^n$ tel que

$$\forall h \in \mathbb{R}^n, \quad d^G f(x)(h) = \langle p, h \rangle .$$

On note alors $\nabla f(x) := p$, appelé *vecteur gradient* de f en x. C'est donc a fortiori le cas si f est F-différentiable en $x \in \Omega$. En résumé, si f est G-différentiable en x et comme annoncé plus haut, on a le droit à un développement limite d'ordre 1 au sens où pour tout $h \in \mathbb{R}^n$,

$$f(x + th) = f(x) + t \langle \nabla f(x), h \rangle + o(t).$$

Réciproquement, si on peut écrire le développement limité $f(x + th) = f(x) + t \langle p, h \rangle + o(t)$, c'est que f est G-différentiable en x et $\nabla f(x) = p$.

Utilisation pratique. La G-différentiabilité est non seulement pertinente en optimisation (on verra qu'elle est suffisante pour établir diverses conditions du premier ordre), mais surtout bien pratique pour le calcul de gradients, même pour des fonctions qui sont manifestement très régulières. En effet, la G-différentiabilité consiste à prendre des limites de fonctions de la variable réelle $t \in \mathbb{R}$ et non de la variable vectorielle $h \in \mathbb{R}^n$ comme dans le cas de la F-différentiabilité.

Par exemple, justifions que, $A \in \mathcal{M}_n(\mathbb{R})$ étant donnée, la fonction $f : \mathbb{R}^n \to \mathbb{R}$ donnée par $f(x) = \frac{1}{2} \langle Ax, x \rangle$ (qui est clairement de classe C^∞ sur $\mathbb{R}^n$) a pour gradient $\nabla f(x) = \frac{1}{2}(A + A^T)x$. S'il avait fallu l'établir pour la F-différentiabilité, on aurait écrit

$$f(x + h) = f(x) + \left\langle \frac{1}{2}(A + A^T)x, h \right\rangle + \frac{1}{2} \langle Ah, h \rangle ,$$

puis justifié par un calcul supplémentaire que le terme tout à droite est bien un $o(h)$. En revanche, la G-différentiabilité mène au calcul immédiat, pour $h \in \mathbb{R}^n$ fixé :

$$f(x + th) = f(x) + t \left\langle \frac{1}{2}(A + A^T)x, h \right\rangle + \frac{1}{2} t^2 \langle Ah, h \rangle = f(x) + t \left\langle \frac{1}{2}(A + A^T)x, h \right\rangle + o(t),$$

le terme $\frac{1}{2} \langle Ah, h \rangle$ étant une brave constante.

1.2.2 Fonctions de classe C^1 et C^2

Définition 1.4. On dit qu'une fonction $f : \Omega \to X$ F-différentiable sur Ω est de classe C^1 sur Ω si l'application $x \in \Omega \mapsto d^F f(x) \in L(\mathbb{R}^n, X)$ est continue sur Ω.

Toujours par le théorème de représentation de Riesz, la continuité de $x \mapsto d^G f(x)$ en tant qu'application de Ω dans $L(\mathbb{R}^n, \mathbb{R})$ est équivalente à la continuité de $x \mapsto \nabla f(x)$ en tant qu'application de Ω dans $\mathbb{R}^n$.

Théorème 1.5

On suppose que $f : \Omega \to X$ est G-différentiable sur Ω et que l'application $x \in \Omega \mapsto d^G f(x) \in L(\mathbb{R}^n, X)$ est continue sur Ω. Alors f est F-différentiable (et donc de classe C^1) sur Ω.

1. L'appel au théorème de représentation de Riesz a tout de la pédanterie : dans le cas de $\mathbb{R}^n$, celui-ci est une évidence puisque, une forme linéaire (continue) $\varphi : \mathbb{R}^n \to \mathbb{R}$ étant donnée, on peut s'écrire $\varphi(x) = \sum_{i=1}^n x_i \varphi(e_i) = \langle a, x \rangle$ avec $a = (\varphi(e_1), \ldots, \varphi(e_n))$. Néanmoins, y faire référence permet d'instiller l'idée que tout ce qu'on raconte se généralise tel quel au cas où l'espace ambiant est un espace de Hilbert et pas seulement $\mathbb{R}^n$.

En particulier, une fonction de classe C^1 est automatiquement continue. Ce résultat a un intérêt pratique immédiat : pour montrer qu'une fonction est de classe C^1, il n'est pas nécessaire de démontrer qu'elle est Fréchet-différentiable, mais qu'elle est Gâteaux-différentiable, puis de montrer que la différentielle alors obtenue est continue. Or, comme on l'a vu, la Gâteaux-différentiabilité est bien plus simple à manipuler.

Démonstration : Soit $x \in \Omega$. Il nous faut montrer que $f(x+h) - f(x) - d^G f(x)(h) = o(h)$: on se donne donc $\varepsilon > 0$, et on cherche $\delta > 0$ tel que, si $\|h\| \leq \delta$, alors $\|f(x+h) - f(x) - d^G f(x)(h)\| \leq \varepsilon \|h\|$. Pour $h \in \mathbb{R}^n$ fixé et $t \in [0,1]$ notons $\gamma(t) = f(x+th)$; γ est une fonction dérivable sur $[0,1]$ puisque f est G-différentiable sur Ω. De plus, sa dérivée est donnée par $t \mapsto d^G f(x+th)(h)$. Or, comme on suppose que $x \in \Omega \mapsto d^G f(x) \in L(\mathbb{R}^n, X)$ est continue sur Ω, on trouve que f est même de classe C^1 sur $[0,1]$. Par conséquent, on a droit à la formule $\gamma(1) - \gamma(0) = \int_0^1 \gamma'(t)\,dt$, ce qui permet d'écrire

$$f(x+h) - f(x) - d^G f(x)(h) = \int_0^1 d^G f(x+th)(h)\,dt - d^G f(x)(h) = \int_0^1 (d^G f(x+th) - d^G f(x))(h)\,dt.$$

Par continuité de $d^G f$ en x, on peut alors se donner $\delta > 0$ tel que $\|x-y\| \leq \delta \implies \|d^G f(x) - d^G f(y)\| \leq \varepsilon$. Ainsi, si $\|h\| \leq \delta$, on trouve

$$\|f(x+h) - f(x) - d^G f(x)(h)\| \leq \int_0^1 \|d^G f(x+th) - d^G f(x)\|\|h\|\,dt \leq \varepsilon \|h\|.$$

■

Savoir qu'une fonction est de classe C^1 nous sera parfois bien utile, par exemple pour assurer la validité de formules telles que

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + t(y-x)), y - x \rangle\,dt.$$

Abordons le sujet épineux des fonctions plusieurs fois différentiables. Pour éviter une trop longue discussion, on se place dans le cas réel et directement dans la situation des fonctions de classe C^2.

Définition 1.6. On dit qu'une fonction $f : \Omega \to \mathbb{R}$ est de classe C^2 sur Ω si f F-différentiable sur Ω est telle que l'application $x \in \Omega \mapsto d^F f(x) \in L(\mathbb{R}^n, X)$ est de classe C^1 sur Ω. On note alors $d^2 f := d(df)$.

Dans ce cas, pour chaque $x \in \Omega$, l'application $d^2 f(x) \in L(\mathbb{R}^n, L(\mathbb{R}^n, \mathbb{R}))$ s'identifie à une application bilinéaire (continue) via l'écriture $d^2 f(x)(h)(k) = d^2(f)(x)(h,k)$: on garde la même notation $d^2 f(x)$ pour l'élément $L(\mathbb{R}^n, L(\mathbb{R}^n, \mathbb{R}))$ et son équivalent bilinéaire (continu). Il se trouve que cette application est toujours symétrique dans le cadre des fonctions de classe C^2.

Proposition 1.7

Soit $f : \Omega \to \mathbb{R}$ de classe C^2 sur Ω. Alors il existe une unique application $\nabla^2 f : \Omega \to L(\mathbb{R}^n, \mathbb{R}^n)$ appelée *Hessienne* de f telle que

$$\forall x \in \Omega,\ \forall h, k \in \mathbb{R}^n, \quad d^2 f(x)(h,k) = \left\langle \nabla^2 f(x)(h), k \right\rangle.$$

De plus, pour tout $x \in \Omega$, $\nabla f^2(x)$ est symétrique, *i.e.*,

$$\forall h, k \in \mathbb{R}^n, \quad \left\langle \nabla^2 f(x)(h), k \right\rangle = \left\langle h, \nabla^2 f(x)(k) \right\rangle.$$

On vérifie en fait que la Hessienne n'est autre que la différentielle de l'application gradient, *i.e.*, de l'application $x \in \Omega \mapsto \nabla f(x) \in \mathbb{R}^n$. Sa matrice dans la base canonique de $\mathbb{R}^n$ est donnée par $(\frac{\partial^2 f}{\partial x_i \partial x_j}(x))_{1 \leq i,j \leq n}$.

Démonstration : La première partie est encore une conséquence du théorème de représentation de Riesz. Seule la symétrie n'est pas évidente ; elle procède du théorème de Schwarz selon lequel $\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \frac{\partial^2 f}{\partial x_j \partial x_i}(x)$ pour tout $x \in \Omega$ et tous $i, j \in \{1, \ldots, n\}$ (voir vos cours de calcul différentiel de Licence). ■

1.3 Convexité

Dans cette section, on se donne $C \subset \mathbb{R}^n$ convexe, c'est-à-dire tel que pour tous $x, y \in C$ et $t \in [0,1]$, $(1-t)x + ty \in C$.

1.3.1 Les trois notions

Définition 1.8. On dit que $f : C \to \mathbb{R} \cup \{+\infty\}$ est

- *convexe* sur C si

$$\forall x, y \in C, \forall t \in [0,1], \quad f((1-t)x + ty) \leq (1-t)f(x) + tf(y).$$

- *strictement convexe* sur C si

$$\forall x \neq y \in C, \forall t \in]0,1[, \quad f((1-t)x + ty) < (1-t)f(x) + tf(y).$$

- *α-fortement convexe* sur C si

$$\forall x, y \in C, \forall t \in [0,1], \quad f((1-t)x + ty) \leq (1-t)f(x) + tf(y) - \alpha \frac{t(1-t)}{2}\|x-y\|^2.$$

Une fonction est dite fortement convexe sur C s'il existe $\alpha > 0$ tel que f soit α-fortement convexe sur C. Par exemple, si C est un ensemble convexe, la fonction qui vaut 0 sur C et $+\infty$ est convexe sur $\mathbb{R}^n$.

On a bien sûr la série d'implications

$$\alpha\text{-forte convexité} \quad \Longrightarrow \quad \text{stricte convexité} \quad \Longrightarrow \quad \text{convexité.}$$

Enfin, il est bon de noter que

$$f \text{ est } \alpha\text{-fortement convexe} \quad \Longleftrightarrow \quad x \mapsto f(x) - \frac{\alpha}{2}\|x\|^2 \text{ est convexe.}$$

1.3.2 Convexité et régularité

Avant de caractériser la convexité pour des fonctions régulières, établissons que la continuité est automatique pour les fonctions convexes (en dimension finie). On utilisera la notion $\text{int}(A)$ pour faire référence à l'intérieur dans $\mathbb{R}^n$ d'un ensemble $A \subset \mathbb{R}^n$.

Théorème 1.9

Soit $f : C \to \mathbb{R} \cup \{+\infty\}$ convexe. Alors f est continue sur $\text{int}(\text{dom}(f))$, où $\text{dom}(f) := \{x \in C,\ f(x) < +\infty\}$.

Démonstration : Soit $z \in \text{int(dom(f))}$. On procède eu trois temps, en montrant successivement que f est majorée, puis bornée et enfin continue (en fait même lipschitzienne) au voisinage de z. Commençons par la caractère majoré. Tout d'abord, par définition de z, il existe $\alpha > 0$ tel que $\overline{B}_1(z, \alpha) \subset \text{dom}(f)$. On prend ici la boule pour la norme $\|\cdot\|_1$ car cela simplifie un peu la preuve du petit résultat suivant : on peut écrire

$$\overline{B}_1(z, \alpha) \subset \text{conv}(\{y_j,\ j \in J\})$$

où J est fini et les y_j tous dans $\overline{B}_1(z, \alpha)$. [2] En effet, on peut (par exemple) à cet effet choisir les $z \pm e_i$, $i \in \{1, \ldots, n\}$.

On se donne désormais $M := \max\{f(y_j),\ j \in J\}$, qui est une quantité finie. Pour $x \in \overline{B}_1(z, \alpha)$, on peut d'après le résultat ci-dessus trouver des réels $t_j \in [0,1]$, $j \in J$ sommant à 1 tels que $x = \sum_{j \in J} t_j y_j$. Par convexité, on trouve alors

$$f(x) \leq \sum_{j \in J} t_j f(y_j) \leq M,$$

2. Pour $A \subset \mathbb{R}^n$, on rappelle que $\text{conv}(A)$ fait référence à l'enveloppe convexe de A, c'est-à-dire à l'ensemble des combinaisons convexes d'éléments de A.

c'est-à-dire que f est majorée sur $\overline{B}_1(z,\alpha)$. On en déduit immédiatement qu'elle est minorée sur cet ensemble : si on se donne $x \in \overline{B}_1(z,\alpha)$, il suffit de construire le point diamétralement opposé $u := 2z - x$ qui satisfait donc $f(u) \leq M$. Par convexité

$$f(z) = f\Big(\frac{x+u}{2}\Big) \leq \frac{1}{2}f(x) + \frac{1}{2}f(u) \leq \frac{1}{2}f(x) + M \implies f(x) \geq 2f(z) - M.$$

Ainsi, f est bornée sur $\overline{B}_1(z,\alpha)$: on peut fixer $C > 0$ tel que $|f| \leq C$ sur cet ensemble.

Montrons enfin le caractère lispchitzien, et ce sur $\overline{B}_1(z, \frac{\alpha}{2})$. On se donne x, y dans cette dernière boule, et on pose $u = x + (\frac{1}{t} - 1)(x - y)$ où $t \in]0, 1]$ est à choisir. Cela correspond encore à $x = (1-t)y + tu$, si bien que par convexité $f(x) - f(y) = f((1-t)y + tu) - f(y) \leq t(f(u) - f(y))$. Pour avoir $u \in \overline{B}_1(z,\alpha)$, on calcule

$$\|u - z\|_1 \leq \|x - z\|_1 + \Big(\frac{1}{t} - 1\Big)\|x - y\|_1 \leq \frac{\alpha}{2} + \Big(\frac{1}{t} - 1\Big)\|x - y\|_1,$$

et on voit qu'on peut fixer t à la valeur assurant que $(\frac{1}{t}-1)\|x-y\|_1 = \frac{\alpha}{2}$, ce qui correspond au choix $t = \frac{2\|x-y\|_1}{\alpha+2\|x-y\|_1}$. On trouve donc finalement

$$f(x) - f(y) \leq t(f(u) - f(y)) \leq 2Ct \leq \frac{4C}{\alpha}\|x - y\|_1.$$

L'échange des rôles de x et y donne le résultat. ■

En particulier, une fonction convexe de $\mathbb{R}^n$ dans $\mathbb{R}$ (qui ne prend donc pas la valeur $+\infty$) est continue sur tout $\mathbb{R}^n$. Dans le cas général, on ne peut pas affirmer que f est continue sur tout son domaine $\mathrm{dom}(f)$: la référence à l'intérieur est nécessaire. On le voit par exemple avec la fonction $f : \mathbb{R} \to \mathbb{R}$ qui vaut 0 sur $[0, 1]$ et $+\infty$ en dehors. Celle-ci est convexe sur $\mathbb{R}$, comme on l'a déjà vu, et elle est continue sur $]0, 1[$ (l'intérieur de son domaine), mais pas en 0 ni en 1.

Dans le cadre convexe (et en dimension finie), il n'y a pas lieu de faire de distinction entre la différentiabilité au sens de Gâteaux et celle au sens de Fréchet.

Théorème 1.10

Soit $f : \Omega \to \mathbb{R} \cup \{+\infty\}$ avec Ω ouvert contenant C. On suppose que f est convexe sur C. Pour $x \in \mathrm{int(dom(f))}$, f est G-différentiable en x si et seulement si f est F-différentiable en x.

Démonstration : Soit $x \in \mathrm{int(dom(f))}$. Par convexité, on sait par le résultat suivant que f est lipschitzienne au voisinage de x. Supposons désormais par l'absurde que f soit G-différentiable en x, mais pas F-différentiable. C'est donc qu'il existe une suite $(h_k) \in \mathbb{R}^n$ tendant vers 0 telle que la suite

$$\varepsilon_k := \frac{f(x + h_k) - f(x) - \langle \nabla f(x), h_k \rangle}{\|h_k\|},$$

satisfasse $|\varepsilon_k| \geq \varepsilon$ pour tout $k \in \mathbb{N}$, pour un certain $\varepsilon > 0$. Écrivons alors $h_k = t_k \frac{h_k}{\|h_k\|} =: t_k v_k$ (c'est donc que $t_k = \|h_k\|$), et la suite (v_k) est alors bornée. Puisqu'on est en dimension finie, on peut donc en extraire une sous-suite convergeant vers un certain $v \in \mathbb{R}^n$ via une extractrice φ, ce qui permet d'écrire

$$\begin{aligned}\varepsilon_{\varphi(k)} &= \left(\frac{f(x + t_{\varphi(k)}v_{\varphi(k)}) - f(x)}{t_{\varphi(k)}} - \left\langle \nabla f(x), v_{\varphi(k)} \right\rangle\right) \\ &= \left(\frac{f(x + t_{\varphi(k)}v) - f(x)}{t_{\varphi(k)}} - \langle \nabla f(x), v \rangle\right) + \frac{f(x + t_{\varphi(k)}v_{\varphi(k)}) - f(x + t_{\varphi(k)}v)}{t_{\varphi(k)}} + \left\langle \nabla f(x), v - v_{\varphi(k)} \right\rangle.\end{aligned}$$

Le premier terme tend vers 0 par Gâteaux-différentiabilité de f en x, le second aussi par le caractère Lipschitz de f, et le troisième tend lui aussi vers 0. C'est donc que la suite $\varepsilon_{\varphi(k)}$ tend vers 0, en contradiction avec la minoration $|\varepsilon_{\varphi(k)}| \geq \varepsilon$. ■

1.3.3 Conditions des premier et second ordres

La définition même de la convexité est difficile à manipuler. Si la fonction est régulière, on dispose heureusement de critère bien pratiques.

Proposition 1.11 (*Conditions du premier ordre*)

Soit $f : \Omega \to \mathbb{R} \cup \{+\infty\}$ G-différentiable sur Ω, ouvert contenant C. Alors f est

- *convexe* sur C si et seulement si

$$\forall x, y \in C, \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

si et seulement si

$$\forall x, y \in C, \quad \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0.$$

- *strictement convexe* sur C si et seulement si

$$\forall x \neq y \in C, \quad f(y) > f(x) + \langle \nabla f(x), y - x \rangle$$

si et seulement si

$$\forall x \neq y \in C, \quad \langle \nabla f(y) - \nabla f(x), y - x \rangle > 0.$$

- α-*fortement convexe* sur C si et seulement si

$$\forall x, y \in C, \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|^2$$

si et seulement si

$$\forall x, y \in C, \quad \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \alpha \|y - x\|^2.$$

Démonstration : Voir votre cours du premier semestre. ■

Proposition 1.12 (*Conditions du second ordre*)

Soit $f : \Omega \to \mathbb{R} \cup \{+\infty\}$ de classe C^2 sur Ω, ouvert contenant C. Alors f est

- *convexe* sur C si et seulement si

$$\forall x, y \in C, \quad \left\langle \nabla^2 f(x)(y - x), y - x \right\rangle \geq 0.$$

- *strictement convexe* sur C si

$$\forall x \neq y \in C, \quad \left\langle \nabla^2 f(x)(y - x), y - x \right\rangle > 0.$$

- α-*fortement convexe* sur C si et seulement si

$$\forall x, y \in C, \quad \left\langle \nabla^2 f(x)(y - x), y - x \right\rangle \geq \alpha \|y - x\|^2.$$

Attention au "si" (et non "si et seulement si") qui s'est caché dans le cas de la stricte convexité. Lorsque $C = \mathbb{R}^n$ (ou un ouvert quelconque de $\mathbb{R}^n$), notez que toutes les conditions s'écrivent de manière équivalente après remplacement de $y - x$ par un vecteur $h \in \mathbb{R}^n$ quelconque. On dispose ainsi du critère suivant, spectral, à partir des matrices symétriques $\nabla^2 f(x)$, $x \in \mathbb{R}^n$: f est

- convexe sur $\mathbb{R}^n$ si et seulement si, pour tout $x \in \mathbb{R}^n$, $\nabla^2 f(x)$ est positive (*i.e.*, les valeurs propres de $\nabla^2 f(x)$ sont positives)

- strictement convexe sur $\mathbb{R}^n$ si pour tout $x \in \mathbb{R}^n$, $\nabla^2 f(x)$ est définie positive (*i.e.*, les valeurs propres de $\nabla^2 f(x)$ sont strictement positives).
- α-fortement convexe sur $\mathbb{R}^n$ si et seulement si pour tout $x \in \mathbb{R}^n$, les valeurs propres de $\nabla^2 f(x)$ excèdent α.

Démonstration : Voir votre cours du premier semestre. ■

1.4 Problèmes bien posés et unicité

Les théorèmes assurant qu'un problème est bien posé reposent sur deux familles d'hypothèses : celles portant sur S, et celles portant sur f. Concernant S ou f, on verra apparaître

- des conditions d'ordre géométrique, comme la convexité.
- des conditions d'ordre topologique, comme par exemple le fait que S soit fermé ou que f soit continue.

Notons que dans le cas où S s'écrit via des contraintes d'inégalité et d'égalité sous la forme

$$S = \Big\{ x \in \mathbb{R}^n, \;\; f_i(x) \leq 0,\, i \in \{1, \dots, m\} \text{ et } h_j(x) = 0,\, j \in \{1, \dots, p\} \Big\},$$

on obtient un ensemble convexe si les fonctions f_i sont convexes sur $\mathbb{R}^n$ et les fonctions h_j affines sur $\mathbb{R}^n$, et fermé si toutes les fonctions f_i et h_j sont continues sur $\mathbb{R}^n$.

1.4.1 Problèmes bien posés : résultats généraux

On appellera *suite minimisante* une suite $(x_n) \in S^{\mathbb{N}}$ telle que

$$f(x_n) \underset{n \to +\infty}{\longrightarrow} \inf_{x \in S} f(x),$$

Une telle suite existe toujours par caractérisation séquentielle de l'infimum, y compris lorsque ce dernier vaut $-\infty$.

Dans le cadre de l'optimisation sous contraintes, la définition de la coercivité ne change pas.

Définition 1.13. On dit que $f : S \to \mathbb{R} \cup \{+\infty\}$ est coercive si

$$\forall A \in \mathbb{R},\; \exists R > 0, \quad x \in S \text{ et } \|x\| \geq R \implies f(x) > A.$$

Cette notion n'a bien sûr un intérêt que si S n'est pas borné : autrement dit, toute fonction sur S borné est coercive. Il est bon de connaître la caractérisation séquentielle de la coercivité,

$$f \text{ est coercive sur } S \quad \iff \quad \forall (x_n) \in S^{\mathbb{N}},\; \|x_n\| \to +\infty \implies f(x_n) \to +\infty.$$

Attention, la coercivité dépend de l'ensemble sur laquelle on la considère : la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par $f(x, y) = x^2$ n'est pas coercive sur $\mathbb{R}^2$, mais elle l'est sur $\{(x, y) \in \mathbb{R}^2,\, |y| \leq 1\}$.

Théorème 1.14

On suppose que
- S est fermé non vide,
- $f : S \to \mathbb{R} \cup \{+\infty\}$ est continue et coercive,

alors le problème (P) est bien posé.

Démonstration : Voir votre cours d'optimisation du premier semestre ou votre cours d'analyse fonctionnelle. ■

Lorsque S est borné, la coercivité est superflue dans le théorème ci-dessus (puisque cette condition devient vide). Pour insister sur cette situation, on rappelle le résultat associé.

Corollaire 1.15

On suppose que
- S est fermé borné non vide,
- $f : S \to \mathbb{R} \cup \{+\infty\}$ est continue,

alors le problème (P) est bien posé.

1.4.2 Cas convexe

Rappelons la définition d'un minimiseur local (dans le contexte de l'optimisation sous contraintes). Lorsque la distinction est nécessaire, un minimiseur pour (P) est appelé *minimiseur global*.

Définition 1.16. On dit que $x^\star$ est un *minimiseur local* de $f : S \to \mathbb{R} \cup \{+\infty\}$ sur S si

$$\exists r > 0,\ \forall x \in S \cap \overline{B}(x^\star, r),\ f(x) \geq f(x^\star).$$

On rappelle quelques résultats généraux dans cette direction.

Proposition 1.17

On suppose que C et $f : C \to \mathbb{R} \cup \{+\infty\}$ sont convexes. Alors tout minimiseur local est global. Si de plus f est strictement convexe, il ne peut y avoir plus d'un minimiseur.

Démonstration : Voir votre cours d'optimisation du premier semestre. ■

Grâce à ce qui précède, on peut énoncer un résultat relativement général pour les fonctions fortement convexes. Celui-repose sur le lemme suivant :

Lemme 1.18

Soient C convexe et $f : C \to \mathbb{R} \cup \{+\infty\}$ qui soit à la fois (G-)différentiable sur un ouvert contenant C, et fortement convexe sur C. Alors f est coercive sur C.

Démonstration : En effet, on choisit $x_0 \in C$ quelconque et on utilise la caractérisation du premier ordre de la α-forte convexité pour un certain $\alpha > 0$, selon laquelle

$$\forall x \in C, \quad f(x) \geq f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{\alpha}{2}\|x - x_0\|^2 = \frac{\alpha}{2}\|x\|^2 + \langle \nabla f(x_0) - \alpha x_0, x \rangle + f(x_0) - \langle \nabla f(x_0), x_0 \rangle + \frac{\alpha}{2}\|x_0\|^2.$$

soit encore par l'inégalité de Cauchy-Schwarz

$$\forall x \in C, \quad f(x) \geq \frac{\alpha}{2}\|x\|^2 - \|\nabla f(x_0) - \alpha x_0\|\|x\| + \left(f(x_0) - \langle \nabla f(x_0), x_0 \rangle + \frac{\alpha}{2}\|x_0\|^2\right)$$

Le second membre est un trinôme en $\|x\|$, de terme dominant $\frac{\alpha}{2}\|x\|^2$. Par conséquent, il tend vers $+\infty$ quand $\|x\|$ tend vers $+\infty$, et donc $f(x)$ aussi. ■

La preuve montre qu'il suffit en fait que la fonction f soit (G-)différentiable en un point.

2 Approche agnostique

2.1 Conditions d'optimalité du premier ordre

À partir de maintenant et jusqu'à la fin de cette section, on suppose que la fonction f à minimiser est la restriction d'une fonction différentiable. En d'autres termes, on dispose d'une fonction f définie sur Ω, ouvert contenant S (en pratique, il arrive souvent que cet ouvert soit tout $\mathbb{R}^n$), et on cherche à minimiser f sur S. Ainsi, lorsqu'on fait référence à la différentiabilité de f en un point $x \in S$, on fait référence à celle de la fonction f définie sur Ω en ce point.

2.1.1 Défilé de cônes

Pour obtenir des conditions d'optimalité du premier ordre portant sur un minimiseur local, la philosophie est toujours la même : on perturbe localement la fonction autour du point. Avec des contraintes, il faut s'assurer que la perturbation en question ne nous fait pas sortir de l'ensemble admissible S. Il nous faut donc discuter des perturbations possibles, envisagées comme directions le long desquelles on perturbe le minimiseur. L'ensemble de telles directions aura naturellement la structure de cône, ce qui justifie un détour par le joli monde qu'ils forment.

Définition 2.1. On appelle cône tout ensemble P tel que

$$\forall p \in P,\ \forall \lambda > 0,\quad \lambda p \in P.$$

Attention, un cône ne contient pas nécessairement 0.

Exemples. $\mathbb{R}^n$, $\emptyset$ et plus généralement tout espace vectoriel sont des cônes. Pour $u \in \mathbb{R}^n$, $\{\lambda u,\ \lambda \geq 0\}$ est un cône de $\mathbb{R}^n$. Enfin, $P_1 := \{(x,y) \in \mathbb{R}^2,\ x > 0, y > 0\}$, $P_2 := \{(x,y) \in \mathbb{R}^2,\ x \geq 0\}$ ou $P_3 := \{(x,y) \in \mathbb{R}^2,\ |y| \geq x, x \geq 0\}$ sont des cônes de $\mathbb{R}^2$.

Un cône donné vient toujours avec un ami, son cône dual.

Définition 2.2. Si P est un cône, on définit son cône dual noté P^* et défini par

$$P^* = \{x \in \mathbb{R}^n,\ \langle x, y\rangle \geq 0,\ \forall y \in P\}.$$

Plus P est gros, plus son cône dual est petit, c'est-à-dire que si P_1 et P_2 sont deux cônes satisfaisant $P_1 \subset P_2$, alors $P_2^* \subset P_1^*$.

Exemples. Utilisant les notations précédentes, le cône dual de P_1 est $P_1^* = \{(x,y) \in \mathbb{R}^2,\ x \geq 0, y \geq 0\}$, et P_2 et P_3 ont même cône dual $P_2^* = P_3^* = \{\lambda(1,0),\ \lambda \geq 0\}$.

Bien qu'un cône puisse être non convexe (l'union de deux demi-droites par exemple) ou ne pas être fermé voire même ouvert (comme par exemple $(\mathbb{R}_+^*)^2$ dans $\mathbb{R}^2$), son dual est toujours convexe fermé.

Proposition 2.3

Le cône dual est toujours non vide, fermé, convexe et contient 0.

Démonstration : Le caractère non vide est immédiat, puisqu'on a toujours $0 \in P^*$. La convexité se démontre à la main, et le caractère fermé s'obtient aisément par caractérisation séquentielle : si (x_n) est une suite de P^* convergeant vers $x \in \mathbb{R}^n$, la continuité du produit scalaire permet de passer à la limite dans l'inégalité

$$\forall n \in \mathbb{N},\quad \langle x_n, y\rangle \geq 0,$$

pour chaque $y \in P$ fixé. ■

Comme on cherche à se donner des directions possibles le long desquelles perturber un optimum local $x^\star$, on veut se donner la notion de direction la plus large possible. On verra que cela donne des informations sur le gradient $\nabla f(x^\star)$ et son appartenance à un cône dual : plus le cône de départ sera grand, plus le cône dual sera petit et plus cela identifiera le gradient.

La notion la plus naturelle est la suivante.

Définition 2.4. On dit que $p \in \mathbb{R}^n$ est une *direction admissible* à S au point $x \in \mathbb{R}^n$ si

$$\exists \varepsilon > 0,\ \forall t \in]0, \varepsilon],\ x + tp \in S.$$

Une direction p est bien sûr admissible à S en x si et seulement si αp est admissible pour un certain (et donc tout) $\alpha > 0$: l'ensemble des directions admissibles est un cône.

Bien que tombant sous le sens, cette définition est trop restrictive : dans de nombreuses situations dignes d'intérêt (en particulier dans le cadre de contraintes d'égalité), seul $\{0\}$ est une direction admissible. Par exemple, on est confronté à cet écueil avec $S := \{(x_1, x_2) \in \mathbb{R}^2,\ x_1^2 + x_2^2 = 1\}$ en n'importe quel point $x \in S$.

On y préfère alors une notion plus large (mais plus complexe), celle de *cône tangent*.

Définition 2.5. On dit que $p \in \mathbb{R}^n$ est *tangent* à S en un point $x \in \mathbb{R}^n$ s'il existe une suite $(p_k) \in \mathbb{R}^n$ tendant vers p, une suite (t_k) strictement positive tendant vers 0 telles que

$$\forall k \in \mathbb{N}, \quad x + t_k p_k \in S.$$

Proposition 2.6

L'ensemble des vecteurs tangents à S en $x \in \mathbb{R}^n$ est un cône non vide appelé *cône tangent à S en x* et noté $T_S(x)$. De plus, il vérifie les propriétés suivantes :

- il contient les directions admissibles à S en x,
- $T_S(x) = \mathbb{R}^n$ si $x \in \mathrm{int}(S)$,
- $T_S(x) = \emptyset$ si $x \notin \overline{S}$,
- il est fermé.

Démonstration : Commençons par justifier qu'il s'agit d'un cône. Si $p \in T_S(x)$ et $\lambda > 0$, on se donne (t_k) et (p_k) associées à p : on montre que $\lambda p \in T_S(x)$ en posant pour $k \in \mathbb{N}$, $\widetilde{p}_k = \lambda p_k$, $\widetilde{t}_k = \frac{1}{\lambda} t_k$, deux suites qui tendent bien vers λp et 0 respectivement, et de plus $x + \widetilde{t}_k \widetilde{p}_k = x + t_k p_k \in S$ pour tout $k \in \mathbb{N}$.

Montrons que $T_S(x)$ contient les directions admissibles. En effet, si p est une direction admissible, on se donne $\varepsilon > 0$ tel que $x + tp \in S$ pour tout $t \in]0, \varepsilon]$, et on pose $p_k = p$ et $t_k = \frac{\varepsilon}{k}$ pour tout $k \in \mathbb{N}^*$, deux choix qui montrent que $p \in T_S(x)$.

En particulier, si $x \in \mathrm{int}(S)$, on se donne $r > 0$ tel que $\overline{B}(0, r) \subset S$, et on vérifie aisément que tout $p \in \mathbb{R}^n$ est une direction admissible à S en x (via le choix $\varepsilon = \frac{r}{\|p\|}$ si $p \neq 0$). Ainsi, $T_S(x)$ contient $\mathbb{R}^n$ et lui est donc égal.

Reste à justifier que $T_S(x)$ est fermé. On se donne $(p^{(n)})$ une suite de $T_S(x)$, convergeant vers $p \in \mathbb{R}^n$. Par définition, on peut pour chaque $n \in \mathbb{N}$ fixé trouver une suite $(p_k^{(n)})$ d'éléments de $\mathbb{R}^n$ convergeant vers $p^{(n)}$ et $(t_k^{(n)})$ une suite d'éléments de $\mathbb{R}_+^*$ convergeant vers 0 telles que

$$\forall k \in \mathbb{N}, \quad x + t_k^{(n)} p_k^{(n)} \in S.$$

Admettons un instant le lemme de topologie suivant : si une double suite $(u_k^{(n)})$ indexée par $k, n \in \mathbb{N}$ (d'un espace vectoriel normé quelconque) est telle que

- $(u_k^{(n)})_{k \in \mathbb{N}}$ converge vers $u^{(n)}$ pour chaque $n \in \mathbb{N}$,
- $(u^{(n)})_{n \in \mathbb{N}}$ converge vers u,

alors il existe une extraction φ telle que la suite $(u_{\varphi(k)}^{(k)})_{k \in \mathbb{N}}$ converge vers u.

On applique alors ce résultat aux deux doubles suites $(p_k^{(n)})$ et $(t_k^{(n)})$, obtenant deux suites $(p_{\varphi(k)}^{(k)})_{k\in\mathbb{N}}$ et $(t_{\varphi(k)}^{(k)})_{k\in\mathbb{N}}$ qui convergent respectivement vers p et 0. Comme par ailleurs,

$$\forall k \in \mathbb{N}, \quad x + t_{\varphi(k)}^{(k)} p_{\varphi(k)}^{(k)} \in S,$$

cela montre que $p \in T_S(x)$.

Pour ce qui est de la preuve du lemme, on construit φ par récurrence. Comme $(u_k^{(0)})_{k\in\mathbb{N}}$ converge vers $u^{(0)}$, on trouve un indice appelé $\varphi(0)$ tel que $\|u_{\varphi(0)}^{(0)} - u^{(0)}\| \leq \frac{1}{0+1}$. Comme $(u_k^{(1)})_{k>\varphi(0)}$ converge vers $u^{(1)}$, on trouve ensuite $\varphi(1) > \varphi(0)$ tel que $\|u_{\varphi(1)}^{(1)} - u^{(1)}\| \leq \frac{1}{1+1}$. Une fois $\varphi(0), \ldots, \varphi(k_0)$ construits, on utilise que $(u_k^{(k_0+1)})_{k>\varphi(k_0)}$ converge vers $u^{(k_0+1)}$ et se donne ensuite $\varphi(k_0+1) > \varphi(k_0)$ tel que $\|u_{\varphi(k_0+1)}^{(k_0+1)} - u^{(k_0+1)}\| \leq \frac{1}{k_0+2}$. Finalement, la suite $(u_k^{(\varphi(k))})_{k\in\mathbb{N}}$ est construite telle que $\|u_{\varphi(k)}^{(k)} - u^{(k)}\| \leq \frac{1}{k+1}$. Il suffit alors d'écrire

$$\forall k \in \mathbb{N}, \quad \|u_{\varphi(k)}^{(k)} - u\| \leq \|u_{\varphi(k)}^{(k)} - u^{(k)}\| + \|u^{(k)} - u\| \leq \frac{1}{k+1} + \|u^{(k)} - u\|,$$

les deux termes à droite tendant vers 0 quand $k \to +\infty$. ■

Par exemple, le cône tangent au cercle $S = \{(x_1, x_2) \in \mathbb{R}^2, x_1^2 + x_2^2 = 1\}$ en un point de la forme $(\cos(\theta), \sin(\theta)) \in S$ avec $\theta \in \mathbb{R}$, est donné par la droite engendrée par le vecteur $(-\sin(\theta), \cos(\theta))$.

Définition 2.7. On appelle *cône normal* à S au point $x \in \mathbb{R}^n$ l'ensemble $-(T_S(x))^*$, noté $N_S(x)$, c'est-à-dire

$$N_S(x) = \{d \in \mathbb{R}^n, \ \langle d, p\rangle \leq 0, \ \forall p \in T_S(x)\}.$$

D'après ce qui précède, cet ensemble est toujours convexe fermé (non vide).

2.1.2 Conditions du premier ordre

Proposition 2.8

Soit $f : \Omega \to \mathbb{R} \cup \{+\infty\}$ telle que f admet un minimiseur local en $x^\star$ sur S, en lequel f est Gâteaux-differentiable. Alors

- pour tout vecteur p tangent à S en $x^\star$ (*i.e.*, $p \in T_S(x^\star)$), $\langle \nabla f(x^\star), p\rangle \geq 0$, c'est-à-dire

$$\nabla f(x^\star) \in (T_S(x^\star))^* \quad \Longleftrightarrow \quad -\nabla f(x^\star) \in N_S(x^\star).$$

- si $S = C$ est de plus convexe, alors

$$\forall x \in C, \quad \langle \nabla f(x^\star), x^\star - x\rangle \leq 0.$$

On retrouve donc en particulier la condition nécessaire du premier ordre dans le cas non contraint : si $S = \mathbb{R}^n$, toute direction p est admissible (y compris $p = -\nabla f(x^\star)$) et on trouve alors $\|\nabla f(x^\star)\|^2 = 0$ et donc $\nabla f(x^\star) = 0$. Cet argument fonctionne plus généralement dès que $x \in \text{int}(S)$: un minimiseur local $x^\star$ doit alors nécessairement satisfaire $\nabla f(x^\star) = 0$.

Démonstration : Appliquant la différentiabilité supposée en $x^\star$, le long d'un vecteur p tangent à S en $x^\star$, on trouve

$$f(x^\star + tp) = f(x^\star) + t\,\langle \nabla f(x^\star), p\rangle + o(t),$$

On se donne désormais des suites (p_k) et (t_k) associées à p, et comme (t_k) tend vers 0, on peut écrire

$$\langle \nabla f(x^\star), p_k\rangle = \frac{f(x^\star + t_k p_k) - f(x^\star)}{t_k} + o(1),$$

où le $o(1)$ est à entendre avec $k \to +\infty$.

Pour k assez grand, on peut utiliser le fait que $x^\star$ est un minimiseur local de f sur S, soit $f(x^\star + t_k p_k) \geq f(x^\star)$, ce qui donne à la limite $k \to +\infty$

$$\langle \nabla f(x^\star), p \rangle = \lim_{k \to +\infty} \langle \nabla f(x^\star), p_k \rangle \geq 0.$$

Lorsque $S = C$ est convexe, il suffit de remarquer que, pour $x \in C$ quelconque (fixé), $p = x - x^\star$ est une direction admissible au point $x^\star$, et donc un vecteur tangent à S en $x^\star$. ■

Bien que nous n'en ayons pas besoin ici, il y a en fait équivalence pour $S = C$ convexe entre $-\nabla f(x^\star) \in N_C(x^\star)$ et le fait qu'on ait $\langle \nabla f(x^\star), x^\star - x \rangle \leq 0$ pour tout $x \in C$. En effet, comme le montre un exercice de TD, si $z \in C$, $T_C(z)$ coïncide avec l'adhérence de $\{t(y - z),\ y \in C,\ t > 0\}$.

Proposition 2.9

On suppose que $S = C$ est convexe, et que $f : \Omega \to \mathbb{R}$ est convexe et G-différentiable sur C. Alors $x^\star$ est un minimiseur global du problème (P) si et seulement si la relation

$$\nabla f(x^\star) \in (T_C(x^\star))^* \quad \Longleftrightarrow \quad -\nabla f(x^\star) \in N_C(x^\star). \tag{2.1}$$

est vérifiée.

Dans le cas non contraint où $C = \mathbb{R}^n$, $N_C(x^\star) = \{0\}$ et on retrouve alors le caractère nécessaire et suffisant de la condition du premier ordre $\nabla f(x^\star) = 0$ pour que $x^\star$ soit minimiseur global, dans le cas convexe.

Démonstration : Le sens direct est une conséquence immédiate des résultats qui précèdent. Le sens réciproque n'est guère plus difficile : par la caractérisation du premier ordre de la convexité, si $x^\star$ vérifie (2.1), on peut écrire

$$\forall x \in C, \quad f(x) \geq f(x^\star) + \langle \nabla f(x^\star), x - x^\star \rangle \geq f(x^\star),$$

d'où le résultat. ■

Dans la pratique, il est difficile de calculer $T_S(x)$. Le chapitre 3 donnera des critères suffisants, appelés conditions de *qualification de contraintes*, qui permettent d'exprimer $T_S(x)$ simplement lorsque S s'écrit par contraintes d'inégalité et d'égalité.

2.2 Projection sur un convexe fermé

2.2.1 Théorème de projection

Commençons par quelques rappels sur la projection sur un convexe fermé.

Théorème 2.10 (*projection sur un convexe fermé*)

Soient C un convexe fermé non vide de $\mathbb{R}^n$ et $x \in \mathbb{R}^n$. Il existe un unique minimiseur au problème

$$\begin{aligned} \min. \quad & \|x - y\| \\ \text{s.c.} \quad & y \in C, \end{aligned}$$

appelé projection de x sur C, et noté $P_C(x)$. De plus $P_C(x)$ est **caractérisé** par les conditions

(i) $P_C(x) \in C$,

(ii) $\forall y \in C,\ \langle x - P_C(x), y - P_C(x) \rangle \leq 0$,

et l'application $P_C : \mathbb{R}^n \to \mathbb{R}^n$ est 1-lipschitzienne.

Enfin, dans le cas où $C = F$ est un sous-espace vectoriel de $\mathbb{R}^n$, on a $\mathbb{R}^n = F \oplus F^\perp$ et, pour $x \in \mathbb{R}^n$, la projection P_F, appelée *projection orthogonale*, est linéaire, et $P_F(x)$ est caractérisé par les conditions

(i) $P_F(x) \in F$,

(ii) $x - P_F(x) \in F^\perp$.

En pratique, pour montrer qu'un candidat $z \in C$ est le projeté de x sur C, on utilise souvent la caractérisation (ii). Il peut aussi arriver qu'on se contente de justifier que $\|x - z\| \leq \|x - y\|$ pour tout y dans C, ce qui bien sûr caractérise aussi le projeté. Enfin, on trouve sporadiquement des situations où certaines techniques d'optimisation de ce cours conduisent à une expression ou une caractérisation équivalente plus simple.

Notons enfin le lemme évident (mais souvent pratique) suivant.

Lemme 2.11

Si C est convexe fermé non vide de $\mathbb{R}^n$, et $x \notin C$, alors $P_C(x) \in \partial C$.

Démonstration : On note $z = P_C(x)$, et on suppose par l'absurde que $z \in \text{int}(C)$. Alors, pour tout $t > 0$ suffisamment petit, $z + t(x - z) \in C$. Pourtant, pour de tels $t < 1$, on a

$$\|z + t(x - z) - x\| = (1 - t)\|x - z\| < \|x - z\|,$$

ce qui est en contradiction avec l'optimalité de z. ■

2.2.2 Projections usuelles

Voici une série de projections élémentaires à connaître.

- dans $\mathbb{R}$, $C = \mathbb{R}_+$, alors $P_C(x) = x_+ = \max(x, 0)$,
- dans $\mathbb{R}$, $C = [a, b]$, alors $P_C(x) = \begin{cases} a \text{ si } x < a \\ x \text{ si } a \leq x \leq b \\ b \text{ si } x > b \end{cases} = \max(a, \min(x, b))$,
- dans $\mathbb{R}^n$, $C = \mathbb{R}^n_+$, alors $P_C(x) = ((x_i)_+)_{1 \leq i \leq n}$,
- dans $\mathbb{R}^n$, $C = \prod_{i=1}^n [a_i, b_i]$, alors $P_C(x) = \left(P_{[a_i,b_i]}(x_i)\right)_{1 \leq i \leq n}$,

Plus généralement, si $C = C_1 \times C_2 \subset \mathbb{R}^n$ avec $C_1 \subset \mathbb{R}^p$, $C_2 \subset \mathbb{R}^q$, $p + q = n$, où C_1 et C_2 sont des convexes fermés non vides (de $\mathbb{R}^p$ et $\mathbb{R}^q$ respectivement), alors

$$P_C(x) = P_C(x_1, \ldots, x_n) = \left(P_{C_1}(x_1, \ldots, x_p), P_{C_2}(x_{p+1}, \ldots, x_n)\right).$$

- si $C = F$ est un sous-espace vectoriel de $\mathbb{R}^n$, alors en notant $(e_i)_{1 \leq i \leq r}$ une base orthonormée,

$$P_F(x) = \sum_{i=1}^r \langle x, e_i \rangle \, e_i,$$

- si F est un hyperplan de $\mathbb{R}^n$, en notant u un vecteur orthogonal à F (de sorte que $F = \{u\}^\perp = \{x \in \mathbb{R}^n, \langle u, x \rangle = 0\}$),

$$P_F(x) = x - \frac{\langle x, u \rangle}{\|u\|^2} u,$$

- si $C = \overline{B}(0, 1)$ (où la boule est bien sûr à entendre au sens de la norme euclidienne),

$$P_C(x) = \begin{cases} x & \text{si } \|x\| \leq 1 \\ \frac{x}{\|x\|} & \text{si } \|x\| > 1 \end{cases} = \frac{1}{\max(1, \|x\|)} x.$$

Proposition 2.12 (*double projection*)

Soit C un convexe fermé non vide. Si $C \subset F \subset \mathbb{R}^n$, où F est un sous-espace vectoriel de $\mathbb{R}^n$, alors

$$P_C = \widetilde{P}_C \circ P_F,$$

où $\widetilde{P}_C : F \to C$ est la projection sur C (définie sur F) et P_F la projection (orthogonale) sur F (définie sur $\mathbb{R}^n$).

Démonstration : Pour $x \in \mathbb{R}^n$, notons $z_F = P_F(x)$ et $z_C = P_C(x)$. On a

$$\forall y \in C, \quad \|x - z_C\|^2 \leq \|x - y\|^2,$$

c'est-à-dire, compte tenu du fait que $x - z_F$ est orthogonal à tout élément de F ,

$$\forall y \in C, \quad \|x - z_F\|^2 + \|z_F - z_C\|^2 \leq \|x - z_F\|^2 + \|z_F - y\|^2,$$

donc

$$\forall y \in C, \quad \|z_F - z_C\|^2 \leq \|z_F - y\|^2,$$

autrement dit

$$z_C = P_C(z_F) = \widetilde{P}_C(z_F).$$

■

2.2.3 Projection avec contrainte d'inégalité

Proposition 2.13

Soit C est un convexe fermé non vide de $\mathbb{R}^n$ pouvant s'écrire sous la forme

$$C = \{x \in \Omega,\ g(x) \leq 0\},$$

avec g G-différentiable sur Ω ouvert de $\mathbb{R}^n$. Alors pour tout $x \notin C$, le point $z = P_C(x)$ vérifie au moins l'une des deux conditions suivantes :

(i) $\nabla g(z) = 0$,

(ii) $\exists \lambda > 0,\ x = z + \lambda \nabla g(z)$.

Lorsqu'elle est possible, l'écriture d'un convexe donné sous la forme ci-dessus n'identifie pas g de manière unique : si g convient, alors g^3 aussi. Ce dernier choix, néanmoins, est à éviter : il nous fait systématiquement tomber dans le cas le moins informatif (i) dès que $x \notin C$. En effet, pour $x \notin C$ on a $P_C(x) \in \partial C$, et comme $C = \{g \leq 0\}$, l'inclusion claire $\partial C \subset \{g = 0\}$ (valable dès que g est continue) donne alors $g(z) = 0$ pour $z = P_C(x)$ avec $x \notin C$. En particulier, $\nabla g^3(z) = 3g^2(z)\nabla g(z) = 0$.

Démonstration : Si la condition (i) n'est pas vérifiée, alors $\nabla g(z) \neq 0$. Notant alors $F = \mathrm{Vect}(\nabla g(z))$, on dispose de la décomposition $\mathbb{R}^n = F \oplus F^\perp$ qui permet d'écrire $x - z$ sous la forme

$$x - z = \lambda \nabla g(z) + v,$$

où $\lambda \in \mathbb{R}$ et $\langle v, \nabla g(z) \rangle = 0$.

Montrons tout d'abord que $\lambda \geq 0$. Si $x = z$, on a $\lambda = 0$ (et $v = 0$). Dans le cas contraire, pour tout $t \in]0, 1[$ le point $u = z + t(x - z)$ est strictement plus proche de x que z, donc nécessairement $g(u) > 0$. Comme par ailleurs, $-g(z) \geq 0$, on trouve

$$\forall t \in]0, 1], \quad \frac{g(z + t(x - z)) - g(z)}{t} > 0,$$

et en passant à la limite quand $t \to 0$ il vient

$$\langle \nabla g(z), x - z \rangle \geq 0,$$

soit $\lambda \|\nabla g(z)\|^2 \geq 0$, ce qui prouve bien que $\lambda \geq 0$.

Montrons maintenant que $v = 0$. Considérons, pour $t > 0$ et $\varepsilon > 0$ le point $u = z + t(v - \varepsilon \nabla g(z))$. La fonction g étant G-différentiable, on a

$$\begin{aligned} g(u) - g(z) &= t \langle v - \varepsilon \nabla g(z), \nabla g(z) \rangle + o(t) \\ &= t \left(-\varepsilon \|\nabla g(z)\|^2 + o(1) \right). \end{aligned}$$

Pour $t > 0$ assez petit, le membre de droite est négatif donc $g(u) \leq 0$, soit $u \in C$ et par conséquent $\langle u - z, x - z \rangle \leq 0$. Après simplification par t, on obtient donc

$$\langle v - \varepsilon \nabla g(z), \lambda \nabla g(z) + v \rangle \leq 0,$$

ou encore

$$\|v\|^2 \leq \lambda\varepsilon \|\nabla g(z)\|^2.$$

En faisant alors tendre ε vers 0, on obtient donc bien que $v = 0$, d'où le résultat annoncé.

Pour finir, $\lambda > 0$ car sinon on aurait $x = z$, soit $x \in C$, ce qui est exclu par hypothèse. ■

2.3 Algorithme du gradient projeté

2.3.1 Définition

L'intuition sous-jacente à l'algorithme du gradient projeté provient alors d'une réécriture un peu alambiquée de la condition nécessaire et suffisante d'optimalité

$$\begin{aligned} x^\star \text{ est minimiseur global} &\iff \forall x \in C, \quad \langle \nabla f(x^\star), x - x^\star \rangle \geq 0 \\ &\iff P_C\big(x^\star - \nabla f(x^\star)\big) = x^\star \\ &\iff \exists \lambda > 0,\ P_C\big(x^\star - \lambda \nabla f(x^\star)\big) = x^\star. \end{aligned}$$

Autrement dit, $x^\star$ est minimiseur global si et seulement si, partant de $x^\star$, une descente selon le gradient de f en $x^\star$ suivie d'une projection sur C soit une manière originale de faire du sur-place.

Définition 2.14 (Algorithme du gradient projeté)**.** Soient C un convexe fermé non vide de $\mathbb{R}^n$ et $f : \Omega \to \mathbb{R} \cup \{+\infty\}$ G-différentiable sur Ω, ouvert contenant C. On appelle *algorithme du gradient projeté*, l'algorithme à deux paramètres $x_0 \in \mathbb{R}^n$ et $\lambda > 0$, donné par la suite définie par

$$\begin{cases} x_0 = x_0, \\ x_{k+1} = P_C\big(x_k - \lambda \nabla f(x_k)\big). \end{cases} \tag{2.2}$$

Cet algorithme est en général conceptuel : le calcul de $P_C(x)$ pour un x donné nécessite la résolution d'un problème d'optimisation quadratique, à savoir

$$\begin{aligned} \min. \quad & \|x - y\|^2 \\ \text{s.c.} \quad & y \in C. \end{aligned}$$

Ainsi, cet algorithme n'a d'intérêt que si l'on peut explicitement calculer P_C, ou à défaut si l'on peut efficacement résoudre le problème d'optimisation ci-dessus.

Lorsque $C = \mathbb{R}^n$, on note que l'on retrouve l'algorithme de descente de gradient usuel, puisque $P_C = \mathrm{Id}$.

2.3.2 Analyse de convergence

On peut néanmoins étudier la convergence de l'algorithme sous cette forme abstraite.

Proposition 2.15

On se donne f et C satisfaisant les hypothèses précisées dans la définition de l'algorithme

(i) f est α-fortement convexe sur C,

(ii) ∇f est L-lipschitzienne sur C.

Alors le problème (P) admet un unique minimiseur $x^\star \in C$ et pour tous $x_0 \in \mathbb{R}^n$ et $\lambda \in \left]0, \frac{2\alpha}{L^2}\right[$,

l'algorithme du gradient projeté converge linéairement vers $x^\star$, avec convergence linéaire de paramètre $q(\lambda) := \sqrt{1 - 2\alpha\lambda + L^2\lambda^2}$, via

$$\forall k \in \mathbb{N}, \quad \|x_k - x^\star\| \leq q^k(\lambda)\|x_0 - x^\star\|.$$

Rappelons que, sous ces hypothèses, on a nécessairement $\alpha \leq L$.

Démonstration : L'ensemble C est fermé non vide, et f y est coercive (car G-différentiable et fortement convexe, cf le lemme (1.18)). Sans hypothèse supplémentaire, il n'est pas garanti qu'elle soit continue sur C ce qui permettrait de conclure que le problème est bien posé et admet un unique minimiseur. On va voir par un argument direct que le problème admet un unique minimiseur $x^\star \in C$. La remarque faite plus haut montre que $x^\star$ est minimiseur global si et seulement s'il s'agit de l'unique point fixe de l'application

$$\Phi : x \mapsto P_C(x - \lambda\nabla f(x)),$$

qui correspond exactement à l'application itérée par l'algorithme au sens où $x_{k+1} = \Phi(x_k)$. On va maintenant chercher à montrer que cette application est en fait contractante pour λ assez petit, ce qui permettra de conclure à la convergence linéaire grâce au théorème du point fixe appliqué dans le complet $\mathbb{R}^n$. Pour $z, z' \in \mathbb{R}^n$ quelconques, on calcule en utilisant le caractère 1-lipschitzien de P_C puis les deux hypothèses du théorème :

$$\begin{aligned}
\|\Phi(z) - \Phi(z')\|^2 &= \|P_C(z - \lambda\nabla f(z)) - P_C(z' - \lambda\nabla f(z'))\|^2 \\
&\leq \|z - \lambda\nabla f(z) - z' + \lambda\nabla f(z')\|^2 = \|z - z'\|^2 - 2\lambda\langle z - z', \nabla f(z) - \nabla f(z')\rangle + \lambda^2\|\nabla f(z) - \nabla f(z')\|^2 \\
&\leq \|z - z'\|^2 - 2\lambda\alpha\|z - z'\|^2 + \lambda^2 L^2\|z - z'\|^2 \\
&\leq (1 - 2\alpha\lambda + L^2\lambda^2)\|z - z'\|^2
\end{aligned}$$

Ainsi, la fonction Φ est $q(\lambda)$-lipschitzienne, avec

$$q(\lambda)^2 := (1 - 2\alpha\lambda + L^2\lambda^2).$$

On vérifie sans difficulté que $q(\lambda) < 1$ si et seulement si $\lambda \in]0, \frac{2\alpha}{L^2}[$. Lorsque λ est choisi ainsi, le théorème du point fixe donne l'existence d'un unique minimiseur $x^\star$ ainsi que la convergence linéaire de (x_k) vers $x^\star$ avec l'estimation voulue $\|x_k - x^\star\| \leq q^k(\lambda)\|x_0 - x^\star\|$. ■

3 Dualité de Lagrange

Dans cette section, on se consacre au cas où S s'écrit par le biais de contraintes d'inégalité et d'égalité, c'est-à-dire

$$\begin{aligned} \text{min.} \quad & f(x) \\ \text{s.c.} \quad & f_i(x) \leq 0, \quad i = 1, \dots, m \\ & h_j(x) = 0, \quad j = 1, \dots, p, \end{aligned} \tag{3.1}$$

qui correspond à l'ensemble de contraintes

$$S := \Big\{ x \in \Omega, \;\; f_i(x) \leq 0,\, i \in \{1, \dots, m\} \text{ et } h_j(x) = 0,\, j \in \{1, \dots, p\} \Big\},$$

où $\Omega \subset \mathbb{R}^n$ est un ouvert de $\mathbb{R}^n$ (on aura très souvent $\Omega = \mathbb{R}^n$).

De nombreux problèmes s'écrivent sous cette forme qui, on le rappelle, est tout sauf unique en terme de choix des fonctions.

Définition 3.1. On dira que le problème (3.1) est convexe si Ω est convexe et si
- les fonctions f et f_i, $i = 1, \dots, m$ sont convexes sur Ω
- les fonctions h_j, $j = 1, \dots, p$ sont affines.

On résumera alors les contraintes d'égalité $h_j(x) = 0$, $j = 1, \dots, p$ sous forme matricielle concise $Ax = b$ avec $A \in \mathcal{M}_{p,n}(\mathbb{R})$ et $b \in \mathbb{R}^p$.

Dans le cas où le problème (3.1) est convexe, notons que l'ensemble S est alors manifestement convexe.

Régularité. Enfin, on fait l'hypothèse que les fonctions f, f_i, $i = 1, \dots, m$, h_j, $j = 1, \dots, p$ sont de classe C^1 sur Ω. On mentionnera çà-et-là les hypothèses minimales nécessaires : pour la plupart des énoncés, la différentiabilité au sens de Gâteaux est en réalité suffisante.

3.1 Approche géométrique

3.1.1 Cône linéarisant

Comme on l'a vu, la condition nécessaire du premier ordre pour l'optimalité locale est donnée par

$$\nabla f(x) \in (T_S(x))^* \quad \Longleftrightarrow \quad -\nabla f(x) \in N_S(x).$$

Dans le cadre de contraintes d'inégalité et d'égalité, tout l'enjeu est de caractériser $T_S(x)$ de manière à ce que son cône dual ait une expression propice aux calculs. L'idée centrale va être de linéariser les contraintes au voisinage du point x.

Définition 3.2. Soit $x \in S$. On appelle *contraintes actives* en x celles qui s'annulent parmi les contraintes d'inégalité, et on note

$$I(x) := \{ i \in \{1, \dots, m\},\, f_i(x) = 0 \},$$

les indices correspondants.

Définition 3.3. Pour $x \in S$, on appelle *cône linéarisant* à S en x l'ensemble des $p \in \mathbb{R}^n$ tels que

$$\forall i \in I(x),\; \langle \nabla f_i(x), p \rangle \leq 0, \quad \forall j \in \{1, \dots p\},\; \langle \nabla h_j(x), p \rangle = 0.$$

On note $T_S^l(x)$ cet ensemble.

Notons que cet ensemble est bien un cône, et qu'il est toujours convexe fermé. Par ailleurs, il est bel et bien lié au cône tangent.

Lemme 3.4

Pour tout $x \in S$, on a

$$T_S(x) \subset T_S^l(x).$$

Démonstration : Soit $p \in T_S(x)$. On trouve donc (t_k) positive tendant vers 0, (p_k) tendant vers p, telles que $x + t_k p_k \in S$ pour tout $k \in \mathbb{N}$. On a donc pour tout $k \in \mathbb{N}$,

$$\forall i \in \{1, \dots, m\},\ f_i(x + t_k p_k) \leq 0, \quad \forall j \in \{1, \dots, p\},\ h_j(x + t_k p_k) = 0.$$

On fixe $i \in I(x)$, de sorte que $f_i(x) = 0$. Un développement limité fournit

$$f_i(x + t_k p_k) = t_k \langle \nabla f_i(x), p_k \rangle + o(t_k) \leq 0.$$

On divise par t_k puis fait tendre k vers $+\infty$ pour obtenir le résultat. L'idée est la même pour montrer que $\langle \nabla h_j(x), p \rangle = 0$ pour tout $j \in \{1, \dots, p\}$. ■

De plus, contrairement à son pendant non linéarisé, on peut calculer de manière très explicite le cône dual de $T_S^l(x)$.

Proposition 3.5

Le cône dual $(T_S^l(x))^*$ est donné par l'ensemble des $q \in \mathbb{R}^n$ tels qu'il existe $\lambda = (\lambda_i)_{i \in I(x)} \in \mathbb{R}_+^{|I(x)|}$, $\nu \in \mathbb{R}^p$ tels que

$$-q = \sum_{i \in I(x)} \lambda_i \nabla f_i(x) + \sum_{j=1}^{q} \nu_j \nabla h_j(x).$$

Démonstration : Pour montrer ce résultat, il vaut mieux l'abstraire et justifier que, étant donnés l vecteurs $v_1, \dots, v_l$, et p vecteurs $w_1, \dots, w_p$ de $\mathbb{R}^n$, le cône

$$P := \{p \in \mathbb{R}^n,\ \forall i \in \{1, \dots l\}, \langle v_i, p \rangle \leq 0,\ \forall j \in \{1, \dots, p\}, \langle w_j, p \rangle = 0\}$$

est de dual donné par

$$P^* = \left\{ -\sum_{i=1}^{l} \lambda_i v_i - \sum_{j=1}^{p} \nu_j w_j,\ \lambda \in \mathbb{R}_+^l,\ \nu \in \mathbb{R}^p \right\}.$$

On note momentanément V cet ensemble. Montrons le sens évident. Soit $q \in V$ qui s'écrit donc $q = -\sum_{i=1}^{l} \lambda_i v_i - \sum_{j=1}^{p} \nu_j w_j$, $\lambda \in \mathbb{R}_+^r$, $\nu \in \mathbb{R}^p$. Un calcul immédiat montre que $\langle p, q \rangle \geq 0$ pour $p \in P$, ce qui montre que $q \in P^*$.

Le sens réciproque est plus difficile, et repose sur le théorème de séparation stricte A.4. Soit $q \in P^*$, dont on suppose par l'absurde qu'il n'est pas dans V. Or V est un convexe fermé non vide ; on peut donc séparer V et $\{q\}$ strictement, c'est-à-dire qu'on trouve $u \in \mathbb{R}^n$ non nul tel que

$$\langle u, q \rangle < \inf_{z \in V} \langle u, z \rangle .$$

Comme $0 \in V$, l'infimum est négatif est $\langle u, q \rangle < 0$. En fait, l'infimum vaut alors exactement 0 du fait que V est un cône. Si ce n'était pas le cas, c'est-à-dire si l'infimum était strictement négatif, on trouverait $z_0 \in V$ tel que $\langle u, z_0 \rangle < 0$. Mais alors tz_0 est dans V pour tout $t > 0$, d'où $\langle u, q \rangle < \langle u, tz_0 \rangle$, ce qui après division par t et passage à la limite $t \to +\infty$, donnerait $\langle u, z_0 \rangle \geq 0$.

On va désormais montrer qu'on a nécessairement $u \in P$, ce qui conduire à $\langle u, q \rangle \geq 0$ puisque $q \in P^*$, et donc à une contradiction. Pour $j \in \{1, \dots, p\}$, on commence par choisir $z = w_j$ puis $z = w_j$ qui sont bien des vecteurs de V, ce qui donne $\langle u, w_j \rangle \geq 0$ et $\langle u, -w_j \rangle \leq 0$ et donc $\langle u, w_j \rangle = 0$. Pour $i \in \{1, \dots, l\}$, on choisit ensuite $z = -v_i \in V$, obtenant $\langle u, -v_i \rangle \geq 0$, et donc $\langle u, v_i \rangle \leq 0$. Finalement, $u \in P$ ce qu'il fallait démontrer. ■

La délivrance viendrait donc de l'inclusion réciproque à $T_S(x) \subset T_S^l(x)$. Voici deux contre-exemples montrant qu'il n'y a pas égalité en général. Si $S = \{(x_1, x_2) \in \mathbb{R}^2,\ 0 \leq x_2 \leq x_1^3\}$, on constate que, pour $x = (0,0)$, $T_S(x) = \mathbb{R}_+ \times \{0\}$, et $T_S^l(x) = \mathbb{R} \times \{0\}$. De même, le cas $S = \{(x_1, x_2) \in \mathbb{R}^2,\ x_1^2 \leq x_2^2\}$ conduit pour $x = (0,0)$ à $T_S(x) = S$ et $T_S^l(x) = \mathbb{R}^2$. Notons que $T_S^l(x)$ est toujours convexe alors que $T_S(x)$ ne l'est pas systématiquement ; le dernier contre-exemple est une situation de ce type.

3.1.2 Qualification des contraintes et conditions de KKT

Vue l'importance de l'égalité $T_S(x) = T_S^l(x)$, on donne à un nom à cette situation.

Définition 3.6. Soit $x \in S$. On dit que les contraintes sont *qualifiées* en $x \in S$ si $T_S(x) = T_S^l(x)$.

Attention, cette propriété n'est pas intrinsèque à l'ensemble S, mais dépend des fonctions utilisées pour le décrire.

Définition 3.7. On dit que $x \in \Omega$ satisfait les *conditions de Karush-Kuhn-Tucker (KKT)* s'il existe $(\lambda, \nu) \in \mathbb{R}^m \times \mathbb{R}^p$ tels que

- *Admissibilité primale* :
$$\forall i \in \{1, \dots, m\},\ f_i(x) \leq 0, \quad \text{et} \quad \forall j \in \{1, \dots, p\},\ h_j(x) = 0.$$
- *Admissibilité duale* :
$$\forall i \in \{1, \dots, m\},\ \lambda_i \geq 0$$
- *Condition de complémentarité* :
$$\forall i \in \{1, \dots, m\},\ \lambda_i\, f_i(x) = 0,$$
- *Condition de stationnarité* :
$$\nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + \sum_{j=1}^{p} \nu_j \nabla h_j(x) = 0.$$

Les vecteurs $(\lambda, \nu) \in \mathbb{R}_+^m \times \mathbb{R}^p$ sont appelés *multiplicateurs de Lagrange* associés à x.

Notons que la première condition signifie simplement que x est admissible, *i.e.*, $x \in S$. La condition de complémentarité, quant à elle, s'interprète ainsi : si pour $i \in \{1, \dots, m\}$, $\lambda_i > 0$, alors $f_i(x) = 0$ (*i.e.*, $i \in I(x)$, la contrainte est active). Réciproquement, si la contrainte n'est pas active, alors nécessairement $\lambda_i = 0$.

Proposition 3.8

On suppose que $x^\star \in S$ est un minimiseur local pour le problème (3.1). Si les contraintes sont qualifiées en $x^\star$, alors $x^\star$ satisfait les conditions de KKT.

Démonstration : Soit $x^\star$ un minimiseur local. Comme $x^\star \in S$, on a automatiquement la première condition.

Le point étant un minimiseur local, on a en outre $\nabla f(x^\star) \in (T_S(x^\star))^* = (T_S^l(x^\star))^*$, la dernière égalité étant l'hypothèse des contraintes qualifiées. Ainsi, il existe $\widetilde{\lambda} = (\widetilde{\lambda}_i)_{i \in I(x^\star)} \in \mathbb{R}_+^{|I(x^\star)|}$, $\nu \in \mathbb{R}^p$ tels que

$$-\nabla f(x^\star) = \sum_{i \in I(x)} \widetilde{\lambda}_i \nabla f_i(x^\star) + \sum_{j=1}^{q} \nu_j \nabla h_j(x^\star).$$

On considère alors le vecteur $\lambda \in \mathbb{R}^m$ obtenu en complétant le vecteur $\widetilde{\lambda}$ par des zéros. On a donc bien la deuxième condition puisque $\lambda \geq 0$, la troisième puisque $\lambda_i^\star\, f_i(x^\star) = 0$ si $i \in I(x^\star)$ (auquel cas $f_i(x^\star) = 0$) et si $i \notin I(x^\star)$

(auquel cas $\lambda_i = 0$ par construction). Enfin, la quatrième condition

$$\nabla f(x^\star) + \sum_{i=1}^{m} \lambda_i^\star \nabla f_i(x^\star) + \sum_{j=1}^{p} \nu_j^\star \nabla h_j(x^\star) = 0.$$

est vérifiée. ■

Dans ce cours, on considèrera la condition de qualification suivante.

Définition 3.9. On dit que $x \in S$ est un *point régulier* par rapport aux contraintes si les gradients

$$\nabla f_i(x),\ i \in I(x), \quad \nabla h_j(x),\ j \in \{1, \dots, p\}$$

sont linéairement indépendants.

Par exemple, on vérifie que dans le cas de $S = \{(x_1, x_2) \in \mathbb{R}^2,\ 0 \leq x_2 \leq x_1^3\}$, $(0,0)$ n'est pas régulier par rapport aux contraintes (ce qu'on savait déjà puisqu'on a vu que $T_S(x) \neq T_S^l(x)$).

Théorème 3.10

Soit $x \in S$. Si x est un point régulier par rapport aux contraintes, alors les contraintes sont qualifiées en x.

Démonstration : Le résultat est admis. Il s'appuie sur le théorème des fonctions implicites et requiert donc la régularité C^1 (au moins localement). ■

Mettant bout à bout tous les résultats obtenus, on a obtenu ce qui suit.

Corollaire 3.11

On suppose que $x^\star \in S$ est un minimiseur local pour le problème (3.1). Si $x^\star$ est régulier pour les contraintes, alors $x^\star$ satisfait les conditions de KKT.

3.1.3 Pratique des conditions de KKT

En pratique, on utilise ce résultat comme on le fait pour toute condition suffisante : on commence par étudier l'ensemble des points réguliers de l'ensemble S pour l'ensemble des contraintes d'inégalité et d'égalité que l'on s'est donné pour le décrire. On écrit alors les conditions de KKT pour ces points, ce qui sélectionne un ensemble de points très restreint, surtout lorsque l'on travaille en petite dimension. On compare enfin la valeur prise par la fonction f en ces points et en les points non réguliers pour voir ceux qui l'emportent parmi ces candidats.

Considérons un exemple pour illustrer la méthode, donné par le problème

$$\begin{aligned} \max. \quad & x_1 + x_2 \\ \text{s.c.} \quad & x_1^2 + x_2^2 \leq 4 \\ & (x_1 - 1)^2 + x_2^2 \geq 1. \end{aligned}$$

Mise sous forme canonique. Sa forme canonique est

$$\begin{aligned} \min. \quad & -x_1 - x_2 \\ \text{s.c.} \quad & x_1^2 + x_2^2 - 4 \leq 0 \\ & 1 - (x_1 - 1)^2 - x_2^2 \leq 0, \end{aligned}$$

Caractère bien posé. Le problème n'est pas convexe, mais l'ensemble est fermé borné et la fonction objectif continue : le problème est bien posé. On note

$$\forall (x_1,x_2),\quad f(x_1,x_2)=-x_1-x_2,\ f_1(x_1,x_2)=x_1^2+x_2^2-4,\ f_2(x_1,x_2)=1-(x_1-1)^2-x_2^2.$$

Recherche des points réguliers. Les fonctions impliquées sont de classe C^1, et un calcul direct donne

$$\forall (x_1,x_2)\in\mathbb{R}^2,\quad \nabla f(x_1,x_2)=(-1,-1),\ \nabla f_1(x_1,x_2)=(2x_1,2x_2),\ \nabla f_2(x_1,x_2)=(-2(x_1-1),-2x_2).$$

On vérifie sans difficulté qu'il n'y a qu'un seul point où les deux contraintes sont actives, j'ai nommé $(2,0)$. En ce point, les gradients ∇f_1 et ∇f_2 sont liés, donc ce point n'est pas régulier. Pour le reste, lorsque seule la première ou la seconde contrainte est active, on vérifie que la gradient correspondant ne s'annule pas : tous les points de S sont réguliers pour les contraintes, à l'exception de $(2,0)$.

Conditions de KKT. Soit $x=(x_1,x_2)\in S$ un minimiseur local, dont on suppose qu'il est régulier, *i.e.*, $x\neq(2,0)$. Il satisfait donc les conditions de KKT : il existe λ_1,λ_2 positifs tels que

$$\lambda_1 f_1(x_1,x_2)=0,\ \lambda_2 f_2(x_1x_2)=0$$

et

$$\nabla f(x_1,x_2)+\lambda_1\nabla f_1(x_1,x_2)+\lambda_2\nabla f_1(x_1,x_2)=0.$$

On distingue alors plusieurs cas selon la nullité des λ_i (ou de manière équivalente, selon les contraintes actives en x).

Premier cas : $\lambda_1>0$, $\lambda_2>0$. Ce cas de figure ne se présente pas puisque cela impose $f_1(x_1,x_2)=f_2(x_1,x_2)=0$ et donc $x=(2,0)$, point non régulier qu'on a exclu.

Deuxième cas : $\lambda_1=0$, $\lambda_2=0$. Ce cas correspond à un minimum local à l'intérieur de S, soit $\nabla f(x)=0$, condition qui n'est jamais vérifiée puisque le gradient de f ne s'annule pas.

Troisième cas : $\lambda_1=0$, $\lambda_2>0$. On a alors $f_2(x)=0$, soit $(x_1-1)^2+x_2^2=1$. La condition de stationnarité donne en outre

$$(-1,-1)+\lambda_2(-2(x_1-1),-2x_2)=0,$$

et ce système en (λ_2,x_1,x_2) se résout de manière unique en $x=(1-\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2})$, associé à $\lambda_1=0$, $\lambda_2=\frac{\sqrt{2}}{2}$.

Quatrième cas : $\lambda_1>0$, $\lambda_2=0$. On a alors $f_1(x)=0$, soit $x_1^2+x_2^2=4$. La condition de stationnarité donne en outre

$$(-1,-1)+\lambda_1(2x_1,2x_2)=0,$$

et ce système en (λ_1,x_1,x_2) se résout de manière unique en $x=(\sqrt{2},\sqrt{2})$, associé à $\lambda_1=\frac{1}{2\sqrt{2}}$, $\lambda_2=0$.

Conclusion. On a donc au final trois points candidats pour le titre de minimiseur global : $x=(1-\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2})$, $x=(\sqrt{2},\sqrt{2})$ et le point non régulier $x=(2,0)$. Évaluant f en chacun de ces points, on trouve que $x=(\sqrt{2},\sqrt{2})$ est l'unique minimiseur local, qui est donc global.

3.2 Lagrangien, dualités faible et forte

Venons-en à une autre manière d'introduire les conditions de KKT, par dualité (de Lagrange). Cette approche est complémentaire à la précédente.

3.2.1 Pénalisation et Lagrangien

La philosophie de nombreux problèmes qui font appel à la *dualité* consiste à remplacer le problème initial, contraint, par un problème qui ne l'est plus : on bénéficie alors de toute l'outillerie de l'optimisation sans contraintes.

La manière la plus brutale et directe de raisonner ainsi consiste, pour un problème générique contraint

$$\begin{aligned} \min. \quad & f(x) \\ \text{s.c.} \quad & x \in S, \end{aligned} \tag{3.2}$$

à considérer le problème non contraint totalement équivalent

$$\min. \quad f(x) + \delta_S(x),$$

où la fonction δ_S vaut 0 sur S et $+\infty$ ailleurs. Le défaut principal de cette réécriture réside dans la perte de régularité qu'elle engendre : la fonction $f + \delta_S$ est terriblement irrégulière. Nous n'explorerons pas cette voie dans ce cours, qui est consacré aux problèmes différentiables.

Pourtant, sachez que cette approche est au centre de nombreuses approches théoriques et numériques en optimisation (au moins dans le cas convexe), via la théorie de *l'analyse convexe*. C'est ce point de vue extrêmement élégant et fécond que vous adopterez en grande partie lors de votre cours de M2.

A contrario, nous étudierons ici une approche plus douce, sous la forme

$$\min. \quad f(x) + p(x), \tag{3.3}$$

où p pénalise le fait que x ne soit pas dans S, mais de manière bien moins extrême qu'en prenant la valeur $+\infty$. Voici la manière dont nous pénaliserons les contraintes, c'est l'approche par *dualité de Lagrange*.

Définition 3.12. On appelle Lagrangien associé au problème (3.1) la fonction $L : \Omega \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ définie par

$$\forall (x, \lambda, \nu) \in \Omega \times \mathbb{R}^m \times \mathbb{R}^p, \quad L(x, \lambda, \nu) = f(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j h_j(x).$$

Définition 3.13. On appelle fonction duale associée au problème (3.1) la fonction $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R} \cup \{-\infty\}$ définie par

$$\forall (\lambda, \nu) \in \mathbb{R}^m \times \mathbb{R}^p, \quad g(\lambda, \nu) = \inf_{x \in \Omega} L(x, \lambda, \nu).$$

Pour un couple (λ, ν) donnés, il se peut que l'infimum soit $-\infty$, valeur que peut donc prendre la fonction g.

Le premier résultat à propos de la dualité de Lagrange est évident mais pourtant crucial.

Proposition 3.14

La fonction g est concave sur $\mathbb{R}^m \times \mathbb{R}^p$.

Démonstration : Pour $x \in \Omega$ fixé, la fonction $(\lambda, \nu) \mapsto L(x, \lambda, \nu)$ est affine donc concave. Ainsi, g l'est aussi comme infimum de fonctions concaves. ■

Pour véritablement prendre en compte notre aversion à ce qu'on ait $f_i(x) > 0$, il est temps de nous restreindre à $\lambda \geq 0$, ce qui permet de comparer la fonction g à la fonction f. En effet, pour $x \in S$, $(\lambda, \nu) \in \mathbb{R}^m_+ \times \mathbb{R}^p$, on trouve $L(x, \lambda, \nu) \leq f(x)$, ce dont on tire

$$\forall x \in S, \ \forall (\lambda, \nu) \in \mathbb{R}^m_+ \times \mathbb{R}^p, \quad f(x) \geq g(\lambda, \nu),$$

3.2.2 Problème dual

Notons que cela peut encore s'écrire

$$\forall(\lambda,\nu)\in\mathbb{R}_+^m\times\mathbb{R}^p,\quad p^\star\geq g(\lambda,\nu)$$

et que ce résultat signifie, en d'autres termes, que $g(\lambda,\nu)$ est une borne inférieure à l'infimum $p^\star$ pour n'importe quel choix de $(\lambda,\nu)\in\mathbb{R}_+^m\times\mathbb{R}^p$.

Par conséquent, il est naturel de considérer la meilleure telle borne, c'est-à-dire le problème d'optimisation sous contrainte suivant.

Définition 3.15. On appelle *problème dual* du problème d'optimisation d'origine le problème

$$\begin{aligned}\max.\quad & g(\lambda,\nu)\\ \text{s.c.}\quad & \lambda\geq 0.\end{aligned}\tag{D}$$

On note $d^\star$ le supremum associé à ce problème.

Le problème d'optimisation d'origine est alors appelé *problème primal*. Une variable $x^\star\in S$ minimiseur du problème primal est souvent appelée *variable primale optimale*, alors qu'une variable $(\lambda^\star,\nu^\star)\in\mathbb{R}_+^m\times\mathbb{R}^p$ maximiseur du problème dual est appelée *variable duale optimale*.

Notons qu'il s'agit d'un problème convexe au sens de la terminologie de cette section (modulo les changements de signe appropriés), puisque la fonction g est concave, les contraintes d'inégalité sont convexes (elles sont même affines, *i.e.*, $-\lambda\leq 0$), et il n'y a pas de contraintes d'égalité.

Exemple détaillé. Donnons-nous un premier exemple instructif

$$\begin{aligned}\min.\quad & \langle c,x\rangle\\ \text{s.c.}\quad & x\geq 0,\\ & Ax=b,\end{aligned}$$

où $c\in\mathbb{R}^n$, $A\in\mathcal{M}_{p,n}(\mathbb{R})$, $b\in\mathbb{R}^p$. On trouve alors

$$\forall(x,\lambda,\nu)\in\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^p,\quad L(x,\lambda,\nu)=-\langle b,\nu\rangle+\left\langle c-\lambda+A^T\nu,x\right\rangle,$$

et donc pour $(\lambda,\nu)\in\mathbb{R}_+^m\times\mathbb{R}^p$, $g(\lambda,\nu)=-\infty$ si $c-\lambda-A^T\nu\neq 0$, et $-\langle b,\nu\rangle$ sinon. Une manière compacte d'écrire le problème dual est alors

$$\begin{aligned}\max.\quad & -\langle b,\nu\rangle\\ \text{s.c.}\quad & \lambda\geq 0,\\ & c-\lambda+A^T\nu=0,\end{aligned}$$

ou encore

$$\begin{aligned}\max.\quad & -\langle b,\nu\rangle\\ \text{s.c.}\quad & A^T\nu+c\geq 0,\end{aligned}$$

Ces différents problèmes équivalents (au sens où ils s'obtiennent les uns à partir des autres de manière immédiate) ne constituent pas stricto sensu le problème dual (D), mais on se permet abusivement d'y référer comme tel.

Les formulations équivalentes jouent un rôle important en ce qui concerne la régularité du problème dual : dans sa forme originelle, on a très souvent $g(\lambda,\nu)=-\infty$ sur une portion conséquente de l'ensemble $\mathbb{R}_+^m\times\mathbb{R}^p$, c'est-à-dire que le problème dual est donc a priori très irrégulier. À l'inverse, le fait d'encoder les situations par le biais de contraintes peut permettre d'exhiber une fonction très régulière, comme le montre l'exemple ci-dessus.

3.2.3 Dualités faible et forte

Rappelons le résultat évident $f(x) \geq g(\lambda, \nu)$ pour tout $x \in S$, $(\lambda, \nu) \in \mathbb{R}^m_+ \times \mathbb{R}^p$ de manière équivalente, mais compacte.

Proposition 3.16 (*dualité faible*)

Il y a *dualité faible*, c'est-à-dire

$$p^\star \geq d^\star.$$

Comme le but originel est de calculer $p^\star$, on se demande naturellement si par hasard on aurait égalité.

Définition 3.17 (dualité forte)**.** On dit qu'il y a *dualité forte* si $p^\star = d^\star$.

Lorsqu'il y a dualité forte, on peut donc de manière alternative résoudre un autre problème d'optimisation pour essayer de déterminer $p^\star$. Ce problème alternatif, dit *dual*, est toujours convexe. On est donc passé d'un problème d'optimisation en dimension n avec m contraintes d'inégalité et p contraintes d'égalité, à un problème d'optimisation convexe en dimension $m + p$ avec m contraintes d'inégalité (affines).

3.2.4 Interprétation par point-selle

De manière tout à fait générale, on peut discuter des dualités faible et forte en exploitant le résultat

$$\inf_{u \in U} \sup_{v \in V} \varphi(u, v) \geq \sup_{v \in V} \inf_{u \in U} \varphi(u, v), \tag{3.4}$$

valable pour n'importe quels ensembles U, V, et fonction $\varphi : U \times V \to \mathbb{R}$.

En effet, on se convainc facilement qu'on obtient la dualité faible via

$$\begin{aligned} p^\star = \inf_{x \in S} f(x) &= \inf_{x \in S} \sup_{(\lambda,\nu) \in \mathbb{R}^m_+ \times \mathbb{R}^p} L(x, \lambda, \nu) && (3.5)\\ &\geq \sup_{(\lambda,\nu) \in \mathbb{R}^m_+ \times \mathbb{R}^p} \inf_{x \in S} L(x, \lambda, \nu) && (3.6)\\ &\geq \sup_{(\lambda,\nu) \in \mathbb{R}^m_+ \times \mathbb{R}^p} \inf_{x \in \Omega} L(x, \lambda, \nu) && (3.7)\\ &= \sup_{(\lambda,\nu) \in \mathbb{R}^m_+ \times \mathbb{R}^p} g(\lambda, \nu) = d^\star, && (3.8) \end{aligned}$$

appliquant la propriété (3.4) au cas $U = S$, $V = \mathbb{R}^m_+ \times \mathbb{R}^p$ et $\varphi = L$ pour la première inégalité, et passant de l'infimum sur S à celui sur le plus grand ensemble Ω dans le second.

Continuous avec le cadre abstrait général, remarquant que le cas d'égalité dans (3.4) s'obtient dès qu'il existe un point-selle à la fonction φ au sens suivant.

Définition 3.18. On dit que $(u^\star, v^\star)$ est un point-selle de φ sur $U \times V$ si

$$\forall (u, v) \in U \times V, \quad \varphi(u^\star, v) \leq \varphi(u^\star, v^\star) \leq \varphi(u, v^\star).$$

Cela signifie que $u^\star$ minimise la fonction $u \mapsto \varphi(u, v^\star)$ sur U, et que $v^\star$ maximise la fonction $v \mapsto \varphi(u^\star, v)$ sur V.

Lemme 3.19

Il y a équivalence entre

(i) $(u^\star, v^\star)$ est point-selle de la fonction φ sur $U \times V$,

(ii) il y a égalité dans (3.4), *i.e.*,

$$\inf_{u\in U}\sup_{v\in V}\varphi(u,v)=\sup_{v\in V}\inf_{u\in U}\varphi(u,v),$$

et de plus

$$u^\star \text{ minimise } u\mapsto\sup_{v\in V}\varphi(u,v) \text{ sur } U,\ v^\star \text{ maximise } v\mapsto\inf_{u\in U}\varphi(u,v) \text{ sur } V.$$

Démonstration : Montrons le sens direct. On a en effet d'une part

$$\varphi(u^\star,v^\star)=\inf_{u\in U}\varphi(u,v^\star)\le\sup_{v\in V}\inf_{u\in U}\varphi(u,v),$$

et d'autre part

$$\varphi(u^\star,v^\star)=\sup_{v\in V}\varphi(u^\star,v)\ge\inf_{u\in U}\sup_{v\in V}\varphi(u,v).$$

L'inégalité dans (3.4) montre que toutes ces quantités sont égales, d'où l'égalité voulue, mais aussi la deuxième assertion puisque l'on dispose alors des égalités $\sup_{v\in V}\varphi(u^\star,v)=\inf_{u\in U}\sup_{v\in V}\varphi(u,v)$ et $\sup_{v\in V}\inf_{u\in U}\varphi(u,v)=\inf_{u\in U}\varphi(u,v^\star)$.

Réciproquement, l'égalité dans (3.4) et l'hypothèse donnent de concert

$$\sup_{v\in V}\varphi(u^\star,v)=\inf_{u\in U}\sup_{v\in V}\varphi(u,v)=\sup_{v\in V}\inf_{u\in U}\varphi(u,v)=\inf_{u\in U}\varphi(u,v^\star).$$

Or

$$\varphi(u^\star,v^\star)\le\sup_{v\in V}\varphi(u^\star,v)=\inf_{u\in U}\varphi(u,v^\star)\le\varphi(u^\star,v^\star),$$

ce qui montre que toutes ces quantités sont égales et donc le résultat. ■

Interprétons ces résultats dans le contexte qui nous intéresse. Il faut prendre garde au fait qu'il y a une inégalité en plus de (3.4) qui est impliquée dans la série d'inégalités (3.5), celle-ci devant aussi être une égalité pour avoir dualité forte. [3]

Corollaire 3.20

Il y a équivalence entre

(i) il existe un point-selle $(x^\star,(\lambda^\star,\nu^\star))\in S\times(\mathbb{R}^m_+\times\mathbb{R}^p)$ au Lagrangien sur $\Omega\times(\mathbb{R}^m_+\times\mathbb{R}^p)$,

(ii) il y a dualité forte, et de plus $x^\star\in S$ est une variable primale optimale, $(\lambda^\star,\nu^\star)\in\mathbb{R}^m_+\times\mathbb{R}^p$ une variable duale optimale.

3.3 Conditions de KKT, le retour

3.3.1 Dualité forte et KKT

Il n'y a plus qu'à détailler un peu ce que signifie être point-selle du Lagrangien pour retrouver sur les conditions de KKT.

Proposition 3.21

On suppose qu'il y a dualité forte $p^\star=d^\star$ et que $p^\star$ et $d^\star$ sont atteints par $x^\star$ et $(\lambda^\star,\nu^\star)$, respectivement. Alors $x^\star$ satisfait les conditions de KKT, de multiplicateurs de Lagrange associés $(\lambda^\star,\nu^\star)$.

3. Le résultat utilisé est donc en fait une généralisation immédiate du précédent, étant donnée une fonction $\varphi:U_r\times V\to\mathbb{R}$ avec $U\subset U_r$ (cf S et Ω dans le cas de la dualité de Lagrange) pour laquelle on a toujours $\inf_{u\in U}\sup_{v\in V}\varphi(u,v)\ge\sup_{v\in V}\inf_{u\in U_r}\varphi(u,v)$.

Démonstration : La condition de point-selle s'écrit

$$x^\star \in S \text{ minimise } x \mapsto L(x,\lambda^\star,\nu^\star) \text{ sur } \Omega, \qquad (\lambda^\star,\nu^\star) \in \mathbb{R}^m_+ \times \mathbb{R}^p \text{ maximise } (\lambda,\nu) \mapsto L(x^\star,\lambda,\nu) \text{ sur } \mathbb{R}^m_+ \times \mathbb{R}^p,$$

et les deux premières conditions de KKT sont déjà vérifiées.

Comme Ω est ouvert, la première condition implique que le gradient de la fonction $x \mapsto L(x,\lambda^\star,\nu^\star)$, s'annule en $x^\star$, c'est-à-dire précisément la condition de stationnarité

$$\nabla f(x^\star) + \sum_{i=1}^m \lambda_i \nabla f_i(x^\star) + \sum_{j=1}^p \nu_j \nabla h_j(x^\star) = 0.$$

Quant à la seconde condition, elle équivaut au fait que $(\lambda^\star,\nu^\star) \in \mathbb{R}^m_+ \times \mathbb{R}^p$ maximise $\sum_{i=1}^m \lambda_i f_i(x^\star)$. Cette dernière quantité apparaît comme la somme de quantités toutes négatives et ne peut donc excéder 0, valeur qu'elle atteint si et seulement si chaque terme est nul, ce qui correspond à $\lambda_i^\star f_i(x^\star) = 0$ pour tout $i \in \{1,\dots,m\}$: revoilà la condition de complémentarité. ■

3.3.2 Suffisance et nécessité des conditions de KKT

Commençons par regarder la suffisance des conditions ; vous devez commencer à être habitués au fait que la convexité est une condition qui mène typiquement à cet état de fait.

Proposition 3.22

On suppose que le problème (3.1) est convexe. Si $x^\star \in \Omega$ satisfait les conditions de KKT de multiplicateurs de Lagrange associés $(\lambda^\star,\nu^\star)$, alors $x^\star$ est primal optimal (et de plus, $(\lambda^\star,\nu^\star)$ est dual optimal et il y a dualité forte).

Démonstration : Grâce à l'interprétation par point-selle fournie par le corollaire 3.20, la preuve est immédiate. En effet, la condition du premier ordre, qui est nécessaire pour exprimer la minimisation de $x \mapsto L(x,\lambda^\star,\nu^\star)$ sur Ω par $x^\star \in S$, devient suffisante. Ainsi, $(x^\star,(\lambda^\star,\nu^\star)) \in S \times (\mathbb{R}^m_+ \times \mathbb{R}^p)$ est un point-selle de L sur $\Omega \times (\mathbb{R}^m_+ \times \mathbb{R}^p)$. ■

Pour ce qui est de la nécessité, la proposition (**??**) laisse un goût d'inachevé puisqu'il faut supposer la dualité forte. L'idéal est de disposer de conditions suffisantes qui l'assurent.

On considère dans ce cours une des plus célèbres et des plus usitées d'entre elles, la condition *de Slater*, qui a le mérite d'être très faible. À l'instar de la notion de point régulier, celle-ci dépend des contraintes utilisées pour définir S ; elle n'est pas intrinsèque à S.

Définition 3.23. Lorsque que le problème (3.1) est convexe, on dit qu'il satisfait les *conditions de Slater* s'il existe $x \in \Omega$ tel que

$$\forall i \in \{1,\dots,m\}, \quad f_i(x) < 0 \qquad \text{et} \qquad Ax = b.$$

Théorème 3.24

On suppose que problème (3.1) est convexe et que les conditions de Slater sont satisfaites. Alors il y a dualité forte et l'optimum $d^\star$ du problème dual est atteint s'il est fini.

Corollaire 3.25

On suppose que le problème (3.1) est convexe et satisfait les conditions de Slater. Si $x^\star \in \Omega$ est un minimiseur de (3.1), alors il satisfait les conditions de KKT (et le multiplicateur de Lagrange associé $(\lambda^\star,\nu^\star)$ est dual optimal, avec dualité forte).

Démonstration du corollaire : D'après le théorème, il y a dualité forte et comme on suppose avoir un minimiseur du problème (3.1), c'est que $p^\star$ est fini et donc $d^\star = p^\star$ aussi, et celui-ci donc atteint. Si on note alors $(\lambda^\star,\nu^\star)$

une variable duale optimale, la proposition 3.21 montre que $x^\star$ satisfait les conditions de KKT de multiplicateur de Lagrange associé $(\lambda^\star, \nu^\star)$. ■

Démonstration du théorème : On peut montrer que les conditions de Slater sont des conditions de qualification de contraintes, c'est-à-dire que sous les conditions de Slater, tout point $x \in S$ satisfait $T_S(x) = T_S^l(x)$. On utilise ici une approche cousine, que l'on présente dans le cas plus simple où il n'y a pas de contraintes d'égalité.

Si $p^\star = -\infty$, $d^\star = -\infty$ par dualité faible, donc $p^\star = d^\star$. Comme la condition de Slater donne l'existence de $\overline{x} \in \Omega$ tel que $f_i(\overline{x}) < 0$ pour tout $i \in \{1, \dots, m\}$, on dispose d'un point admissible si bien que $p^\star \neq +\infty$. Il nous reste donc à traiter le cas $p^\star$ fini.

On note

$$C := \{(u, v) \in \mathbb{R}^m \times \mathbb{R},\ \exists x \in \Omega,\ f(x) \leq v \text{ et } \forall i \in \{1, \dots, m\},\ f_i(x) \leq u_i\},$$

qui est convexe, et non vide puisqu'il contient tous les points $(f_1(x), \dots, f_m(x), f(x))$ pour $x \in \Omega$. On note aussi qu'il satisfait la propriété suivante : si $(u, v) \in C$, $(u', v') \in A$ pour tous $u' \geq u$, $v' \geq v$. En fait, on peut obtenir C à partir de

$$B = \{(f_1(x), \dots, f_m(x), f(x)),\ x \in \Omega\},$$

l'ensemble C se construisant en y ajoutant tous les points qui y sont "supérieurs", c'est-à-dire que $C := B + \mathbb{R}_+^{m+1}$.

Montrons que $(0, p^\star) \in \mathbb{R}^m \times \mathbb{R}$ n'est pas dans l'intérieur de C. Sinon, $(0, p^\star - \varepsilon)$ serait dans C pour ε assez petit, ce qui fournirait l'existence de x admissible tel que $f(x) \leq p^\star - \varepsilon$, contredisant la définition de $p^\star$.

Finalement $(0, p^\star) \notin \text{int}(C)$, et on montre aisément que $(0, p^\star) \in \overline{C}$ à l'aide d'une suite minimisante. On peut donc trouver par théorème de séparation A.3 un hyperplan séparant C et $(0, p^\star)$, c'est-à-dire qu'il existe $\lambda \in \mathbb{R}^m$, $\lambda_0 \in \mathbb{R}$ avec $(\lambda, \lambda_0) \neq 0$ tels que

$$\forall (u, v) \in C, \quad \langle (\lambda, \lambda_0), (u, v) \rangle_{\mathbb{R}^{m+1}} = \langle \lambda, u \rangle_{\mathbb{R}^m} + \lambda_0 v \geq \langle (\lambda, \lambda_0), (0, p^\star) \rangle_{\mathbb{R}^{m+1}} = \lambda_0 p^\star.$$

Cette inégalité fournit immédiatement le fait que $\lambda \geq 0$, $\lambda_0 \geq 0$. En effet, si ce n'était pas le cas et ayant fixé $(u, v) \in C$ quelconque, il suffirait d'augmenter la composante de ce vecteur associée à la composante strictement négative de (λ, λ_0), jusqu'à ce que l'inégalité ci-dessus soit violée.

Supposons momentanément avoir démontré que $\lambda_0 > 0$. Après division par λ_0 et notant à nouveau λ le vecteur λ/λ_0, on trouve

$$\inf_{(u,v) \in C} \sum_{i=1}^m \lambda_i u_i + v = \inf_{(u,v) \in C} \langle \lambda, u \rangle + v = p^\star.$$

Or, x parcourant Ω, les points $(f_1(x), \dots, f_m(x), f(x))$ sont dans A et on trouve donc en particulier

$$g(\lambda) = \inf_{x \in \Omega} \Big(\sum_{i=1}^m \lambda_i f_i(x) + f(x) \Big) \geq \inf_{(u,v) \in C} \Big(\sum_{i=1}^m \lambda_i u_i + v \Big) \geq p^\star.$$

C'est donc que $d^\star \geq g(\lambda) \geq p^\star$: la dualité forte est établie, et il existe bien un maximiseur dual puisqu'on a alors $g(\lambda) = d^\star$.

Finissons par justifier que $\lambda_0 > 0$ en raisonnant par l'absurde : si $\lambda_0 = 0$, on a

$$\inf_{(u,v) \in C} \langle \lambda, u \rangle = 0.$$

Or utilisant le point $u = (f_1(\overline{x}), \dots, f_m(\overline{x}))$, $v = f(\overline{x})$ où $\overline{x}$ est donné par la condition de Slater, on a

$$\langle \lambda, u \rangle = \sum_{i=1}^m \lambda_i f_i(\overline{x}) < 0.$$

puisque la nullité de cette dernière somme imposerait $\lambda = 0$, ce qui est impossible puisqu'on a supposé que $\lambda_0 = 0$, et que $(\lambda, \lambda_0) \neq 0$. C'est une contradiction. ■

3.4 Algorithme d'Uzawa

3.4.1 Dualité de Lagrange et algorithmique

La première idée qui vient à l'esprit lorsque l'on a accès au problème dual, c'est de résoudre le problème dual plutôt que le primal. Le dual est a priori plus séduisant, lui qui a l'avantage de toujours être convexe, et avec des contraintes très simples, contrairement au problème primal. La contrainte de positivité sur la variable λ signifie même que l'on peut envisager un algorithme de gradient projeté, la projection se faisant sur le convexe fermé $\mathbb{R}^m_+ \times \mathbb{R}^p$, qui s'écrit en effet explicitement.

Cette approche peut en effet être mise en place dans de diverses situations, mais elle a plusieurs défauts et limites, que voici.

- pour que la résolution du dual apporte des informations précises sur le primal, il faut qu'il y ait dualité forte ; or cela suppose en pratique au moins la convexité, la dualité forte étant une situation très rare en son absence.
- la résolution du dual suppose qu'on connaît la fonction duale g ; or $g(\lambda, \nu)$ s'obtient précisément en résolvant un problème de minimisation, celui du Lagrangien $x \mapsto L(x, \lambda, \nu)$. Toute approche directement écrite sur le dual suppose donc que l'on sait résoudre explicitement (ou très rapidement) ce problème, qui a au moins l'avantage d'être non contraint.
- même lorsqu'on a accès à la fonction g, on a vu que celle-ci peut valoir $-\infty$ sur de larges portions de $\mathbb{R}^m_+ \times \mathbb{R}^p$, ce qui empêche toute utilisation d'algorithme de gradient projeté... à moins que l'on intègre comme contrainte le fait que l'on ne considère que les $(\lambda, \nu) \in \mathbb{R}^m_+ \times \mathbb{R}^p$ en lesquels $g(\lambda, \nu) > -\infty$, mais au prix que la contrainte ne se prête alors plus nécessairement à un algorithme de gradient projeté.
- une fois le problème dual "résolu", on dispose d'une approximation de $d^\star$ (et donc de $p^\star = d^\star$ s'il y a dualité forte), ainsi que d'approximations de variables duales optimales. En l'état, une telle approche ne dit donc pas comment remonter à des (approximations) de variables primales optimales.

L'algorithme d'Uzawa ci-dessous souffre des mêmes défauts, à l'exception du dernier puisqu'il produit *à la fois* une suite de variables duales et une suite de variables primales.

3.4.2 Présentation de l'algorithme d'Uzawa

L'algorithme d'Uzawa cherche à résoudre le primal via la recherche de points-selles du Lagrangien. Pour alléger les notations, on note ici $F = (f_1, \dots, f_m)$, $h = (h_1, \dots h_p)$.

Définition 3.26 (Algorithme d'Uzawa)**.** On appelle *algorithme d'Uzawa*, l'algorithme de paramètres $\lambda_0 \in \mathbb{R}^m_+$, $\nu^0 \in \mathbb{R}^p$, $\rho > 0$ suivant, donné par les suites définies par les récurrences

$$\begin{cases} x_k \in \arg\min\limits_{x \in \Omega} L(x, \lambda_k, \nu_k), \\ \lambda_{k+1} = \max(0, \lambda_k + \rho\, F(x_k)) \\ \nu_{k+1} = \nu_k + \rho\, h(x_k). \end{cases} \tag{3.9}$$

Pour que cet algorithme ait un sens, il faut que, pour tout $(\lambda, \nu) \in \mathbb{R}^m_+ \times \mathbb{R}^p$, le problème non contraint (au sens où la minimisation a lieu sur tout Ω)

$$\min. \quad L(x, \lambda, \nu),$$

soit bien posé, ce qui est une véritable restriction comme on l'a déjà vu.

La deuxième partie de l'itération est bien une montée de gradient projeté, cherchant à résoudre le problème d'optimisation (contraint) de maximisation de $(\lambda, \nu) \mapsto L(x_k, \lambda, \nu)$ sur $\mathbb{R}^m_+ \times \mathbb{R}^p$. En effet, la seconde partie de l'algorithme peut s'écrire

$$(\lambda_{k+1}, \nu_{k+1}) = P_{\mathbb{R}^m_+ \times \mathbb{R}^p}\big((\lambda_k, \nu_k) + \rho\, \nabla_{\lambda,\nu} L(x_k, \lambda_k, \nu_k)\big).$$

Cet algorithme suppose que l'on sache résoudre explicitement (ou très efficacement) le problème de minimisation du Lagrangien. Dans le cas contraire, il existe d'autres algorithmes, mais moins efficaces.

Enfin, rien ne garantit que, pour k fixé, on ait $x_k \in S$: en pratique, arrêter l'algorithme à une itérée donnée peut tout à fait conduire à un point non admissible.

3.4.3 Convergence de l'algorithme d'Uzawa

Pour simplifier, on se place dans le cas où $\Omega = \mathbb{R}^n$. Comme on y est habitué, le pas ρ ne doit pas trop grand pour que l'algorithme converge, ainsi que le montre le théorème ci-dessous dans le cadre convexe.

Théorème 3.27

On suppose que le problème est convexe, et en outre que f est α-fortement convexe sur $\mathbb{R}^n$ et les fonctions f_i sont M-lipschitziennes sur $\mathbb{R}^n$, et on note

$$\rho_c := \frac{2\alpha}{M^2 + \|A\|_2^2}.$$

Enfin, on suppose que le Lagrangien possède un point-selle $(x^\star, (\lambda^\star, \nu^\star)) \in S \times (\mathbb{R}^m_+ \times \mathbb{R}^p)$. Alors l'algorithme d'Uzawa est bien défini, et pour tous $\lambda_0 \in \mathbb{R}^m_+$, $\nu_0 \in \mathbb{R}^p$, $\rho \in]0, \rho_c[$, la suite (x_k) associée converge vers $x^\star$.

L'hypothèse finale du théorème que constitue l'existence d'un point-selle n'est pas très explicite ; néanmoins, les hypothèses de convexité faites assurent que le problème est bien posé (cf la preuve ci-dessous), et pour qu'il existe un point-selle, il suffit alors par exemple que les conditions de Slater soient satisfaites d'après le corollaire 3.25.

Démonstration : Sous les hypothèses du théorème, on sait que le problème est bien posé et admet un unique minimiseur grâce au lemme 1.18 conjugué au théorème 1.14. Enfin, comme $(x^\star, (\lambda^\star, \nu^\star)) \in S \times (\mathbb{R}^m_+ \times \mathbb{R}^p)$ est un point-selle du Lagrangien, $x^\star \in S$ est une variable primale optimale d'après le corollaire 3.20, c'est donc l'unique minimiseur.

Enfin, la forte convexité (et sa régularité) de f signifie aussi que, pour (λ, ν) fixé, la fonction régulière $x \mapsto L(x, \lambda, \nu)$ est-elle aussi fortement convexe et régulière (comme somme d'une fonction convexe et de fonctions convexes), et donc admet un unique minimiseur sur $\mathbb{R}^n$. En d'autres termes, le problème sous-jacent au calcul de x_k est bien posé et s'écrit ici

$$\{x_k\} = \arg\min_{x \in \mathbb{R}^n} L(x, \lambda_k, \nu_k).$$

Pour simplifier, on va noter $\mu_k = (\lambda_k, \nu_k) \in \mathbb{R}^m_+ \times \mathbb{R}^p$ (et $\mu^\star = (\lambda^\star, \nu^\star)$), P la projection sur $\mathbb{R}^m_+ \times \mathbb{R}^p$ dans $\mathbb{R}^m \times \mathbb{R}^p$, et $u = (F, h) = (f_1, \dots, f_m, h_1, \dots, h_p)$, de sorte que les itérations de l'algorithme d'Uzawa s'écrivent

$$\begin{cases} \{x_k\} = \arg\min_{x \in \mathbb{R}^n} L(x, \mu_k), \\ \mu_{k+1} = P(\mu_k + \rho\, u(x_k)). \end{cases}$$

Commençons par montrer que

$$\mu^\star = P(\mu^\star + \rho\, u(x^\star)).$$

Pour ce faire, il suffit de justifier que $\lambda^\star = P_{\mathbb{R}^m_+}(\lambda^\star + \rho\, F(x^\star))$, c'est-à-dire encore

$$\forall i \in \{1, \dots m\}, \quad \lambda_i^\star = \max(0, \lambda_i^\star + \rho f_i(x^\star)).$$

Or on a vu que tout point-selle du Lagrangien vérifie les conditions de KKT, donc on vérifie aisément l'égalité en examinant les conditions de complémentarité par disjonction de cas ($\lambda_i^\star = 0$ ou $\lambda_i^\star > 0$).

Ainsi on peut écrire, à l'aide du caractère 1-lipschitzien de P :

$$\begin{aligned} \|\mu_{k+1} - \mu^\star\|^2 &= \|P(\mu_k + \rho u(x_k)) - P(\mu^\star + \rho u(x^\star))\|^2 \\ &\leq \|(\mu_k + \rho u(x_k)) - (\mu^\star + \rho u(x^\star))\|^2 = \|(\mu_k - \mu^\star) + \rho(u(x_k) - u(x^\star))\|^2 \\ &= \|\mu_k - \mu^\star\|^2 + \rho^2 \|u(x_k) - u(x^\star)\|^2 + 2\rho \left\langle \mu_k - \mu^\star, u(x_k) - u(x^\star) \right\rangle. \end{aligned}$$

Or $x^\star$ minimisant la fonction $x \mapsto L(x, \mu^\star) = f(x) + \langle \mu^\star, u(x) \rangle$, on a la condition nécessaire du premier ordre (voir l'exercice 8 du TD1)

$$\forall x \in \mathbb{R}^n, \quad \langle \nabla f(x^\star), x - x^\star \rangle + \langle \mu^\star, u(x) - u(x^\star) \rangle \geq 0.$$

De même, x_k minimisant $x \mapsto L(x, \mu_k) = f(x) + \langle \mu_k, u(x) \rangle$, on a

$$\forall x \in \mathbb{R}^n, \quad \langle \nabla f(x_k), x - x_k \rangle + \langle \mu_k, u(x) - u(x_k) \rangle \geq 0.$$

On applique la première inégalité à $x = x_k$, la seconde à $x = x^\star$, et on somme les deux inégalités pour obtenir

$$\langle \nabla f(x^\star) - \nabla f(x_k), x_k - x^\star \rangle + \langle \mu^\star - \mu_k, u(x_k) - u(x^\star) \rangle \geq 0,$$

Le caractère fortement convexe de f fournit alors

$$\langle \mu^\star - \mu_k, u(x^\star) - u(x_k) \rangle \leq - \langle \nabla f(x^\star) - \nabla f(x_k), x^\star - x_k \rangle \leq -\alpha \|x^\star - x_k\|^2$$

On obtient donc finalement

$$\begin{aligned} \|\mu_{k+1} - \mu^\star\|^2 &\leq \|\mu_k - \mu^\star\|^2 + \rho^2 \|u(x_k) - u(x^\star)\|^2 - 2\rho\alpha \|x^\star - x_k\|^2 \\ &\leq \|\mu_k - \mu^\star\|^2 + (\rho^2 L^2 - 2\rho\alpha) \|x^\star - x_k\|^2 \end{aligned}$$

où L est une constante de Lipschitz pour u. Donc si $\rho \in]0, \frac{\alpha}{2L^2}[$, alors $\rho^2 L^2 - 2\rho\alpha < 0$ et donc

$$\|\mu_{k+1} - \mu^\star\|^2 < \|\mu_k - \mu^\star\|^2$$

La suite (ε_k) définie pour $k \in \mathbb{N}$ par $\varepsilon_k := \|\mu_k - \mu^\star\|^2$ est donc une suite décroissante positive, elle converge. Comme $(\varepsilon_{k+1} - \varepsilon_k)$ tend alors vers 0, l'encadrement

$$\varepsilon_{k+1} - \varepsilon_k \leq (\rho^2 L^2 - 2\rho\alpha) \|x^\star - x_k\|^2 \leq 0$$

montre que (x_k) tend vers $x^\star$. ■

4 Compléments d'algorithmique

L'objectif de cette section est une discussion autour des limites et propriétés fines des algorithmes du premier ordre. Les résultats sont exposés dans le cadre convexe, différentiable et sans contraintes avec $\Omega = \mathbb{R}^n$.

4.1 Analyse de l'algorithme du gradient

4.1.1 Cadre et inégalités utiles

Le cadre général sera celui d'une fonction $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ qui

(i) convexe sur $\mathbb{R}^n$,
(ii) de classe C^1 sur $\mathbb{R}^n$,
(iii) de gradient L-Lipschitz sur $\mathbb{R}^n$.

Sous ces hypothèses on a toujours les deux lemmes suivants.

Lemme 4.1

Sous les hypothèses (ii) et (iii) :

$$\forall x, y \in \mathbb{R}^n, \quad f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2.$$

Démonstration : Pour $x, y \in C$ et $\lambda \in [0, 1]$, f étant de classe C^1, on peut écrire

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle \, dt$$

et donc en utilisant l'inégalité de Cauchy-Schwarz ainsi que le caractère lispchitzien de ∇f :

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x \rangle &= \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle \, dt \\ &\leq \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| \|x - y\| \\ &\leq L\|x - y\|^2 \int_0^1 t \, dt = \frac{1}{2} L \, \|x - y\|^2. \end{aligned}$$

■

Notons que l'inégalité ci-dessus montre que toute fonction qui satisfait (ii), (iii) et qui est en outre α-fortement convexe vérifie $\alpha \leq L$ puisqu'on a alors

$$\frac{\alpha}{2}\|x - y\|^2. \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2}\|x - y\|^2.$$

Lemme 4.2 (*Baillon-Haddad*)

Sous les hypothèses (i), (ii) et (iii) :

$$\forall x, y \in \mathbb{R}^n, \quad \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2,$$

Démonstration : On fait la preuve dans le cas C^2. On fixe $x, y \in \mathbb{R}^n$; le caractère C^2 permet d'écrire la formule avec reste intégral

$$\forall x, y \in \mathbb{R}^n, \quad \nabla f(y) - \nabla f(x) = \int_0^1 \nabla^2 f(x + t(y - x))(y - x) \, dt = \Big(\int_0^1 \nabla^2 f(x + t(y - x)) \, dt \Big)(y - x).$$

Discutons de l'application linéaire $\nabla^2 f(z)$ pour $z \in \mathbb{R}^n$. Comme f est convexe, ses valeurs propres sont positives. Comme ∇f est L-Lipschitz, ses valeurs propres sont plus petites que L. Il en va donc de même pour l'application linéaire

$$A := \int_0^1 \nabla^2 f(x + t(y-x))\, dt,$$

à propos de laquelle on vient de voir que $\nabla f(y) - \nabla f(x) = A(y-x)$. On peut alors conclure en écrivant

$$\begin{aligned}\|\nabla f(y) - \nabla f(x)\|^2 = \|A(y-x)\|^2 &= \left\langle AA^{1/2}(y-x), A^{1/2}(y-x)\right\rangle \\ &\leq L\left\langle A^{1/2}(y-x), A^{1/2}(y-x)\right\rangle = \langle A(y-x), y-x\rangle = \langle \nabla f(y) - \nabla f(x), y-x\rangle .\end{aligned}$$

■

4.1.2 Problème d'optimisation et algorithme

Le problème d'optimisation d'intérêt est non contraint, donné par

$$\text{min.} \quad f(x),$$

et on rappelle que l'algorithme de descente de gradient (à pas constant) est donné par

$$x_{k+1} = x_k - \lambda \nabla f(x_k), \quad k \in \mathbb{N},$$

où $\lambda > 0$, partant de $x_0 \in \mathbb{R}^n$.

La question générale est la suivante : supposant que le problème est bien posé (de sorte que $p^\star \in \mathbb{R}$),

$$\text{a-t-on } f(x_k) \underset{k\to+\infty}{\longrightarrow} p^\star, \text{ et si oui, à quelle vitesse ?}$$

On se concentrera sur cette question et non celle (parfois tout aussi importante) de la convergence des itérées.

4.1.3 Le cas fortement convexe

On commence par le cas où f est α-fortement convexe sur $\mathbb{R}^n$, hypothèse sous laquelle le problème est bien posé : $p^\star \in \mathbb{R}$, et il existe (un unique) $x^\star$ tel que $f(x^\star) = p^\star$. On rappelle *l'inégalité de Polyak-Lojasiewicz* (PL).

Lemme 4.3

Pour f qui satisfait (i), (ii) et qui est de plus α-fortement convexe sur $\mathbb{R}^n$, on a

$$\forall x \in \mathbb{R}^n, \quad f(x) - p^\star \leq \frac{1}{2\alpha}\|\nabla f(x)\|^2.$$

Démonstration : Voir votre cours du premier semestre. ■

Proposition 4.4

Soit f qui satisfait (i), (ii) et (iii) et qui est de plus α-fortement convexe. On suppose que $\lambda < \frac{2}{L}$. Alors l'algorithme de descente de gradient converge linéairement avec

$$f(x_k) - p^\star \leq q(\lambda)^k (f(x_0) - p^\star),$$

où $q(\lambda) := 1 - 2\alpha\lambda + \alpha L \lambda^2$.

Cette convergence linéaire admet donc pour paramètre $q(\lambda)$, qui est optimisé pour $\lambda = \frac{1}{L}$, valeur pour laquelle on obtient le taux $1 - \frac{\alpha}{L}$. On peut en fait affiner ces estimations et établir une meilleure convergence linéaire, qui une fois optimisée pour $\lambda = \frac{1}{L}$, est contrôlée par le taux $\frac{L-\alpha}{L+\alpha}$.

Démonstration : Soit $k \in \mathbb{N}$. On commence par utiliser le premier lemme en les points x_k et $x_{k+1} = x_k - \lambda f(x_k)$, puis l'inégalité de PL

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), -\lambda \nabla f(x_k) \rangle + \frac{L}{2} \| - \lambda \nabla f(x_k)\|^2 = f(x_k) - \lambda \langle \nabla f(x_k), \nabla f(x_k) \rangle + \frac{L}{2}\lambda^2 \|\nabla f(x_k)\|^2 \\ &= f(x_k) - \big(\lambda - \frac{L}{2}\lambda^2\big)\|\nabla f(x_k)\|^2 \\ &\leq f(x_k) - 2\alpha\big(\lambda - \frac{L}{2}\lambda^2\big)(f(x_k) - p^\star). \end{aligned}$$

Finalement, on trouve

$$f(x_k) - p^\star \leq \big(1 - 2\alpha\lambda + \alpha L \lambda^2\big)^k (f(x^0) - p^\star).$$

■

4.1.4 Le cas convexe

On s'affranchit de l'hypothèse de forte convexité, mais on suppose a minima que le problème est bien posé, de sorte que $p^\star \in \mathbb{R}$, et qu'il existe $x^\star$ tel que $f(x^\star) = p^\star$. Autrement dit, on suppose que f admet un minimiseur. C'est le cas si f est coercive, par exemple.

Pour alléger les énoncés, on note donc Γ_L l'ensemble des fonctions $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ qui soient convexes, de classe C^1, de gradient L-Lipschitz, admettant au moins un minimiseur sur $\mathbb{R}^n$.

Lemme 4.5

Soit $f \in \Gamma_L$. Soit $x^\star$ un minimiseur quelconque de f. Alors, pour tout $\lambda \leq \frac{2}{L}$, la suite $(\|x_k - x^\star\|)$ décroît.

Démonstration : On note $T_\lambda : \mathbb{R}^n \to \mathbb{R}^n$ l'application itérée par l'algorithme du gradient, c'est-à-dire que $T_\lambda : x \mapsto x - \lambda \nabla f(x)$, application dont $x^\star$ est manifestement un point fixe. La propriété cruciale est le caractère 1-lipschitzien de cette application sous les hypothèses faites sur le pas : pour $x, y \in \mathbb{R}^n$,

$$\begin{aligned} \|T_\lambda(x) - T_\lambda(y)\|^2 &= \|(x-y) - \lambda(\nabla f(x) - \nabla f(y))\|^2 \\ &= \|x-y\|^2 - 2\lambda \langle x - y, \nabla f(x) - \nabla f(y) \rangle + \lambda^2 \|\nabla f(x) - \nabla f(y)\|^2 \\ &\leq \|x-y\|^2 - \lambda\Big(\frac{2}{L} - \lambda\Big)\|\nabla f(x) - \nabla f(y)\|^2, \\ &\leq \|x-y\|^2. \end{aligned}$$

où la première inégalité vient du lemme de Baillon-Haddad, la seconde du choix du pas. On écrit alors simplement l'inégalité ci-dessus pour $x = x_k$, $y = x^\star$ pour conclure à la décroissance voulue. ■

Proposition 4.6

Soit $f \in \Gamma_L$. Pour $\lambda < \frac{2}{L}$,

$$f(x_k) - p^\star \leq M(\lambda) \frac{\|x_0 - x^\star\|^2}{k+1},$$

où $M(\lambda) := \frac{2}{\lambda(2 - \lambda L)}$.

Notons que la constante $M(\lambda)$ est optimisée pour $\lambda = \frac{1}{L}$ et vaut alors $2L$.

Démonstration : Soit $k \in \mathbb{N}$. Comme on l'a vu au début de la preuve de convergence linéaire, on a

$$f(x_{k+1}) \leq f(x_k) - \big(\lambda - \frac{L}{2}\lambda^2\big)\|\nabla f(x_k)\|^2.$$

Il nous faut réussir à estimer le terme en gradient, sans l'inégalité de PL. On commence par écrire l'inégalité de convexité

$$f(x^\star) \geq f(x_k) + \langle \nabla f(x_k, x^\star - x_k \rangle .$$

Ainsi, supposant $x_k \neq x^\star$, car sinon l'inégalité demandée est une évidence, on trouve

$$\frac{f(x_k) - f(x^\star)}{\|x_0 - x^\star\|} \leq \frac{f(x_k) - f(x^\star)}{\|x_k - x^\star\|} \leq \|\nabla f(x_k)\|,$$

où l'on a utilisé le lemme qui précède, puis l'inégalité de Cauchy-Schwarz.

Notant $\alpha_k = f(x_k) - f(x^\star)$ pour $k \in \mathbb{N}$, cette suite positive vérifie ainsi

$$\alpha_{k+1} \leq \alpha_k - \frac{\lambda(2 - \lambda L)}{2\|x_0 - x^\star\|^2} \alpha_k^2.$$

La conclusion provient alors du lemme (de Grönwall discret) suivant : si (α_k) est une suite positive telle que $\alpha_{k+1} \leq \alpha_k - c^{-1}\alpha_k^2$ avec $c > 0$ une constante, alors $\alpha_k \leq \frac{c}{k+1}$. ■

4.2 Taux optimal et accélération de Nesterov

4.2.1 Rudiments sur les taux optimaux

On se concentre à nouveau sur le cas convexe, sans hypothèse de forte convexité, et plus précisément sur la classe de fonctions Γ_L. Une manière équivalente d'énoncer le résultat de convergence de la proposition 4.6 est de la forme suivante pour x_0, $k \in \mathbb{N}$ fixés et $\lambda < \frac{2}{L}$,

$$\forall f \in \Gamma_L, \quad f(x_k) - p^\star \leq M(\lambda) \frac{\|x_0 - x^\star\|^2}{k+1} \quad \iff \quad \sup_{f \in \Gamma_L} \frac{f(x_k) - p^\star}{\|x_0 - x^\star\|^2} \leq \frac{M(\lambda)}{k+1},$$

où $M(\lambda)$ est une constante qui dépend de λ.

Ainsi, uniformément sur le classe de fonctions Γ_L, l'algorithme de descente de gradient partant de x_0 et de pas $\lambda < \frac{2}{L}$ commet une erreur en $O(\frac{1}{k})$. Notons que "l'algorithme de descente de gradient" fait en fait référence à une famille d'algorithmes indexée par $\lambda > 0$ (une fois x_0 fixé), pour laquelle la constante $M(\lambda)$, n'excède jamais $2L$. Sur cette classe d'algorithmes, on a donc

$$\inf_{\lambda > 0} \sup_{f \in \Gamma_L} \frac{f(x_k) - p^\star}{\|x_0 - x^\star\|^2} \leq \frac{2L}{k+1},$$

Il existe en outre des fonctions pour lesquelles l'algorithme de descente de gradient ne fait pas mieux, voir un exercice de TD pour un résultat dans cette direction (en dimension $n = 1$).

La question est la suivante : peut-on trouver un algorithme qui fasse mieux que l'algorithme de descente de gradient ? Par *faire mieux*, on entend qui soit tel que la suite des erreurs commises le long des itérations converge plus vite vers 0 que $1/k$. Pour répondre à cette question de manière sensée, il faut préciser la famille d'algorithmes à laquelle on se restreint. Ici, on considère la famille des algorithmes qui reposent sur des évaluations du gradient de la fonction, à laquelle l'algorithme de descente de gradient appartient pour tout $\lambda > 0$.

Définition 4.7. Soit $x_0 \in \mathbb{R}^n$ fixé. On note $\mathcal{A}(x_0)$ l'ensemble des suites (x_k) de premier terme x_0, telles que de plus

$$\forall k \in \mathbb{N}^*, \quad x_k \in x_0 + \mathrm{Vect}(\nabla f(x_0), \ldots, \nabla f(x^{(k-1)})).$$

On peut alors montrer le résultat suivant :

Théorème 4.8

Soit $k \leq \frac{n-1}{2}$. Alors pour tout $x_0 \in \mathbb{R}^n$,

$$\inf_{(x_k)\in\mathcal{A}(x_0)} \sup_{f\in\Gamma_L} \frac{f(x_k)-p^\star}{\|x_0-x^\star\|^2} \geq \frac{3}{32}\frac{L}{(k+1)^2},$$

Démonstration : Voir l'exercice de TD consacré. ■

Ce résultat suppose que le nombre d'itérations reste petit devant la dimension, ce qui est le cas pour les problèmes typiques de l'apprentissage machine où la dimension n est trop grande pour qu'on puisse espérer itérer autant de fois. Pourvu que cette condition soit satisfaite, le résultat énonce qu'on ne peut faire mieux que $1/k^2$. Notons aussi que si l'on restreignait encore la classe de fonctions aux seules fonctions quadratiques, l'infimum ci-dessus vaudrait 0 pour $k \geq n$, puisque l'algorithme du gradient conjugué trouve dans ce cadre l'optimum en au plus n itérations.

4.2.2 Accélération de Nesterov

Est-ce que cette borne inférieure est trop grossière, c'est-à-dire qu'en vérité le membre de gauche est en $1/k$? Dans ce cas, l'algorithme de descente de gradient est optimal. La possibilité alternative, c'est que cette borne soit en fait précise et qu'il existe un algorithme de la famille $\mathcal{A}(x_0)$ qui atteigne le taux $1/k^2$. On va découvrir avec stupeur que c'est la deuxième situation qui se produit.

Définition 4.9 (Nesterov, 1983)**.** Soit $f : \mathbb{R}^n \to \mathbb{R}$ de classe C^1 sur $\mathbb{R}^n$. On appelle algorithme *de descente de gradient accélérée de Nesterov* l'algorithme de paramètres $x_{-1} = x_0 \in \mathbb{R}^n$, $\lambda > 0$, défini par la récurrence indexée par $k \in \mathbb{N}$

$$\begin{cases} y_k = x_k + \gamma_k(x_k - x_{k-1}), \\ x_{k+1} = y_k - \lambda\nabla f(y_k), \end{cases} \tag{4.1}$$

avec $\gamma_k = \frac{k-1}{k+2}$.

Pour $x_0 \in \mathbb{R}^n$ fixé, la suite produite est bien un élément de l'ensemble des algorithmes de gradient $\mathcal{A}(x_0)$.

Théorème 4.10

Soit $f \in \Gamma_L$. Pour $\lambda \leq \frac{1}{L}$,

$$f(x_k) - p^\star \leq M(\lambda)\frac{\|x_0-x^\star\|^2}{(k+1)^2},$$

avec $M(\lambda) = \frac{2}{\lambda}$.

Démonstration : Admise. ■

La convergence est donc optimisée pour $\lambda = \frac{1}{L}$, et on trouve alors $M(\lambda) = 2L$. En résumé, on a donc l'encadrement très précis suivant :

$$\frac{3}{32}\frac{L}{(k+1)^2} \leq \inf_{(x_k)\in\mathcal{A}(x_0)} \sup_{f\in\Gamma_L} \frac{f(x_k)-p^\star}{\|x_0-x^\star\|^2} \leq 2\frac{L}{(k+1)^2}.$$

A Théorèmes de séparation

Dans cette appendice, on s'intéresse à des théorèmes géométriques consistant à séparer deux ensembles A et B de $\mathbb{R}^n$ par un hyperplan. On rappelle qu'un hyperplan est un ensemble de la forme

$$\{x \in \mathbb{R}^n,\, \langle u, x\rangle = \alpha\}.$$

où $u \in \mathbb{R}^n \setminus \{0\}$ et $\alpha \in \mathbb{R}$.

A.1 Séparation et séparation stricte

Définition A.1. Soient $A, B \subset \mathbb{R}^n$. On dit qu'on peut

- *séparer* A et B s'il existe $u \in \mathbb{R}^n \setminus \{0\}$, $\alpha \in \mathbb{R}$, tels que

$$\forall x \in A,\ \forall y \in B, \quad \langle u, x\rangle \leq \alpha \leq \langle u, y\rangle\,.$$

- *séparer* A et B *strictement* s'il existe $u \in \mathbb{R}^n \setminus \{0\}$, $\alpha_1, \alpha_2 \in \mathbb{R}$, tels que

$$\forall x \in A,\ \forall y \in B, \quad \langle u, x\rangle \leq \alpha_1 < \alpha_2 \leq \langle u, y\rangle\,.$$

On dit alors dans le premier cas que l'hyperplan $\{x \in \mathbb{R}^n, \langle u, x\rangle = \alpha\}$ sépare A et B, et dans le second que, pour $\alpha \in]\alpha_1, \alpha_2[$, l'hyperplan $\{x \in \mathbb{R}^n, \langle u, x\rangle = \alpha\}$ sépare strictement A et B.

Notons que l'on peut s'affranchir de toute mention aux paramètres réels, puisqu'on peut séparer A et B si et seulement s'il existe $u \neq 0$ tel que

$$\sup_{x\in A} \langle u, x\rangle \leq \inf_{y\in B} \langle u, y\rangle\,,$$

et séparer strictement A et B si et seulement s'il existe $u \neq 0$ tel que

$$\sup_{x\in A} \langle u, x\rangle < \inf_{y\in B} \langle u, y\rangle\,.$$

Pourtant, on écrit la définition comme ci-dessus car il est souvent pratique de manipuler un hyperplan séparateur (ou séparateur strict) lorsqu'on fait appel à un résultat de séparation.

A.2 Séparation large

Théorème A.2

Soient C_1 et C_2 deux ensembles convexes non vides disjoints de $\mathbb{R}^n$. Alors on peut séparer C_1 et C_2.

Démonstration : À venir. ■

Ce résultat n'est valable qu'en dimension finie. Dans le cas d'un espace de Hilbert, on peut montrer que la conclusion reste vraie si l'on suppose que l'un des deux ensembles est ouvert.

Un corollaire simple (mais important) est donné par le résultat suivant, souvent appelé *théorème de l'hyperplan support*.

Corollaire A.3

Soient C un convexe non vide et $x_0 \in \partial C$. Alors il existe un hyperplan qui sépare C et $\{x_0\}$.

Démonstration : À venir. ■

A.3 Séparation stricte

Venons-en à un résultat de séparation stricte, qui lui s'applique tel quel en dimension infinie (tout en sachant que les compacts s'y font plus rares).

Théorème A.4

Soient C_1 et C_2 deux ensembles convexes fermés non vides disjoints de $\mathbb{R}^n$, dont l'un au moins est compact. Alors on peut séparer C_1 et C_2 strictement.

On se sert souvent de ce résultat pour strictement séparer un point d'un convexe fermé auquel il n'appartient pas.

Démonstration : L'idée est de séparer strictement l'ensemble $\{0\}$ de $C_1 - C_2$, c'est-à-dire trouver $u \neq 0$

$$\inf_{z \in C_1 - C_2} \langle u, z \rangle = \inf_{x \in C_1,\, y \in C_2} \langle u, x - y \rangle > \langle u, 0 \rangle = 0,$$

auquel cas on vérifie en effet que C_1 et C_2 sont bel et bien strictement séparés.

Tout d'abord, notons que $C_1 - C_2$ est convexe, non vide, et ne contient pas 0 puisque C_1 et C_2 sont disjoints. Par ailleurs, un petit résultat de topologie assure que c'est un fermé en tant que somme d'un fermé et d'un compact.

On peut donc projeter 0 sur $C_1 - C_2$, et on note alors u le projeté qui satisfait $u \neq 0$. Par caractérisation du projeté, on sait que $\langle 0 - u, z - u \rangle \leq 0$ pour tout $z \in C_1 - C_2$. Or cela se réécrit $\langle u, z \rangle \geq \|u\|^2$ d'où en effet

$$\inf_{z \in C_1 - C_2} \langle u, z \rangle \geq \|u\|^2 > 0.$$

■