
Analyse Convexe et Optimisation

Guillaume Garrigos

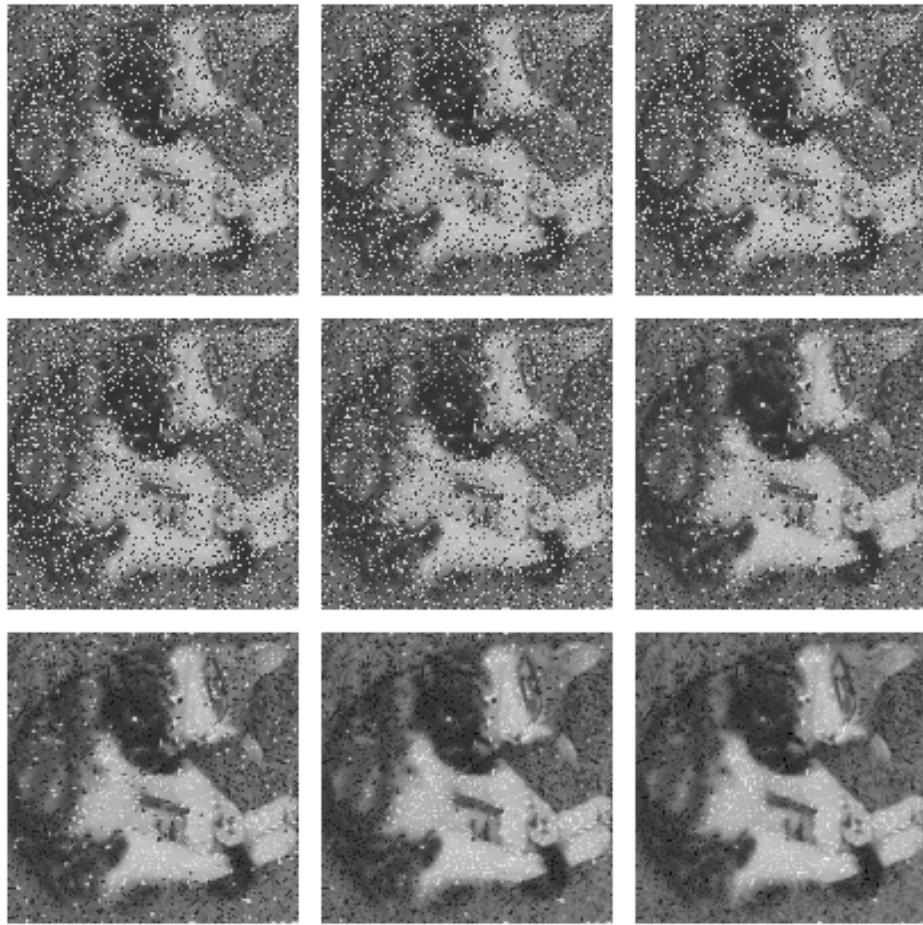


Table des matières

Avant-propos	5
Introduction: pourquoi s'intéresser à l'optimisation non lisse?	7
I Ensembles convexes	15
I.I Convexité	15
I.I.1 Définitions et calcul	15
I.I.2 Projection	20
I.II Approximation d'un convexe	25
I.II.1 Cônes	25
I.II.2 Cône polaire	28
I.II.3 Cônes tangent et normal à un convexe	30
II Analyse convexe non lisse	35
II.I Fonctions convexes s.c.i. propres	35
II.I.1 Fonctions à valeurs réelles étendues	35
II.I.2 Fonctions semi-continues inférieurement	37
II.I.3 Fonctions coercives et existence de minimiseurs	40
II.I.4 Fonctions convexes	41
II.I.5 Fonctions fortement convexes	48
II.II Sous-différentiel d'une fonction convexe	49
II.II.1 Sous-différentiel	49
II.II.2 Calcul sous-différentiel	53
II.II.3 Conditions d'optimalité	58
II.III Conjuguée de Fenchel	61
II.III.1 Définitions et calcul de la conjuguée	61
II.III.2 Propriétés duales de la conjuguée	63
II.III.3 Interlude: preuve de résultats jusque-là admis	66
II.III.4 Dualité de Fenchel-Rockafellar	68
III Algorithmes d'éclatement pour l'optimisation convexe	75
III.I Algorithmes élémentaires	75
III.I.1 Algorithme du Gradient	75

III.I.2 Algorithme Proximal	77
III.I.3 Calcul proximal	78
III.II Algorithmes d'éclatement	82
III.II.1 Éclatement simple : Algorithme du Gradient Proximal	83
III.II.2 Éclatement total : Algorithme de Davis-Yin	84
III.II.3 Éclatement composite total : Algorithme de Yan	89
A Annexe: Quelques éléments de modélisation mathématique	95
A.I Le transport optimal	95
A.II La classification	95
A.III Le traitement du signal et de l'image	104
B Annexe: Pour aller plus loin	105
B.I Résultats avancés sur les polyèdres	105
B.I.1 Cônes et lemme de Farkas	105
B.I.2 Théorème de Weyl: cônes finis = cônes polyédraux	107
B.I.3 Théorème de Motzkin et de Minkowski sur les polyèdres	109
B.I.4 Fonctions polyédrales	110
B.II Résultats avancés sur la conjuguée	117
B.II.1 Inf-convolution	118
B.II.2 Inf-composition	121
B.II.3 Dualité entre sous-différentiel et dérivée directionnelle	123
B.II.4 Dualité entre fonctions lisses et fortement convexes	130
B.III Résultats avancés sur les algorithmes	135
B.III.1 Éclatement Total	135
B.III.2 Preuve de convergence de Davis-Yin	138
B.III.3 Méthodes Lagrangiennes	145
B.III.4 Méthodes Lagrangiennes alternées	149
C Annexe: Encore quelques preuves	157
C.I Preuves alternatives et directes de certains résultats principaux	157
C.II Preuves de petits résultats laissés en exercice	164

Avant-propos

Note pour les étudiant·e·s. Ce document est presque terminé. Il est très possible que de nombreuses coquilles et erreurs soient encore présentes ; le cas échéant merci de bien vouloir me les signaler par mail.

Contenu du cours. L'analyse convexe est un sujet vaste, et le temps alloué à ce cours est réduit (24h de CM), j'ai donc effectué des choix concernant le contenu de ce cours et sa présentation. Tout d'abord, je me suis fixé quelques objectifs principaux:

- 1) Parler d'analyse *non-lisse*. Les problèmes faisant intervenir des fonctions non différentiables sont nombreux et il est je pense essentiel de savoir comment les traiter.
- 2) Parler d'*algorithmes*. Les problèmes d'optimisation sont partout, et je souhaite que mon audience sorte de ce cours en étant capables de résoudre la plupart des problèmes d'optimisation convexe.

En ce qui concerne le point 1) j'essaye d'aller le plus vite possible droit au but: notion de sous-différentiel, conditions d'optimalité, et dualité. L'objectif est qu'en sortie de ce cours les étudiant·e·s soient capables de faire du *calcul*: calculer le cône normal à une contrainte ; le sous-différentiel, la conjuguée et l'opérateur proximal de fonctions usuelles ; écrire des conditions d'optimalité ; écrire un problème dual. Certains des résultats principaux nécessitent parfois des preuves un peu laborieuses, j'ai donc fait le choix de déferrer leur preuve à des exercices avancés de TD. Je préfère prendre le temps de parler du théorème et de ses applications que de passer 1h à faire une preuve au tableau. Cela concerne notamment le Théorème de KKT non lisse et la dualité lisse / forte convexité.

En ce qui concerne le point 2) il y a des choix à faire, car la liste des algorithmes d'optimisation est sans fin. Plutôt que de faire un tour d'horizon étendu j'ai préféré me concentrer sur une famille d'algorithmes: les algorithmes d'éclatement. Ils permettent de résoudre n'importe quel problème d'optimisation qui font intervenir des sommes de fonctions lisses et/ou non-lisses, éventuellement composées avec des opérateurs linéaires, et éventuellement sous contraintes. Nous ne discuterons pas de l'efficacité de ces méthodes, ni de comment trouver des méthodes plus efficaces. Ce cours fera donc l'impasse sur les méthodes de Newton et de point intérieur (bien que l'on verra le concept de fonction barrière en TD). Nous ne verrons pas non plus la méthode du simplexe pour l'optimisation linéaire (trop spécifique à mon gout). Les méthodes d'accélération de Nesterov, ainsi

que les algorithmes pour les problèmes d'optimisation stochastique sont également hors-programme, et font l'objet d'un cours dédié *Optimization for Machine Learning* proposé aux M2 MIDS et M2MO [7].

Références. Le cours est en grande partie basé sur le livre *Convex Optimization in Normed Spaces* par Juan Peypouquet [12]. Il est en anglais mais je pense assez accessible et bien écrit. Pour compléter vos lectures :

- Le cours d'optimisation de L3 [6].
- *Optimisation et analyse convexe : Exercices et problèmes corrigés avec rappels de cours*, par Jean-Baptiste Hiriart-Urruty [8].
- *Convex Analysis and Nonlinear Optimization: Theory and Examples*, par Borwein and Lewis [4].
- *Convex optimization*, par Boyd and Vandenberghe [5].

Pour celleux qui veulent aller plus loin (gare au hors-piste!):

- *Infinite Dimensional Analysis - A Hitchhiker's Guide*, par Aliprantis and Border [2].
- *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, par Bauschke and Combettes [3].

Remerciements. Je remercie ici les collègues et étudiants qui m'ont fait part de retour sur le manuscrit, ou de coquilles: Olivier Bokanowski, Jules Herrmann.

Introduction: Pourquoi s'intéresser à l'optimisation non lisse ?

La théorie de l'optimisation consiste à étudier des problèmes de la forme

$$\underset{x \in \mathbb{R}^N}{\text{minimiser}} \ f(x)$$

où f est une fonction dépendant de paramètres $x = (x_1, \dots, x_N)$ à choisir. Lorsqu'on s'intéresse à problème d'optimisation, on se pose en général trois types de questions:

- Est-ce que mon problème est bien posé ? Y a-t-il existence, unicité d'une solution à mon problème ?
- Est-ce que les solutions vérifient des propriétés qui vont m'aider à les calculer ? Par exemple le fait que le gradient s'annule
- Est-ce que je peux calculer une solution approchée à l'aide d'un algorithme ?

Si $f : \mathbb{R}^N \rightarrow \mathbb{R}$ est une fonction bien définie et différentiable, alors le calcul différentiel standard permet de répondre à toutes ces questions (voir le cours d'optimisation de L3 [6]). Comme vous le savez, la notion centrale du calcul différentiel est celle de *differentielle*, qui pour notre fonction à valeurs réelles est équivalente à celle de *gradient*. Notamment, le théorème de Fermat nous dit que le gradient s'annule en les solutions, ce qui va nous pousser à résoudre l'équation $\nabla f(x) = 0$. De plus, l'algorithme du gradient

$$x_{n+1} = x_n - \lambda \nabla f(x_n)$$

est un algorithme très simple à implémenter avec de bonnes garanties théoriques de convergence vers une solution lorsque $n \rightarrow +\infty$.

Mais ceci n'est pas réaliste.

La plupart des problèmes issus de la vie réelle se modélisent comme des problèmes d'optimisation pour lesquels:

- 1) la fonction f n'est **pas définie** partout;
- 2) il y a une **contrainte** à respecter ;
- 3) la fonction f n'est **pas différentiable** partout.

L'objectif de ce cours est d'apprendre à gérer ces difficultés, afin d'apporter une réponse dans ce cadre général aux trois questions que l'on s'est posées (problème bien posé? conditions d'optimalité? algorithmes?). Nous allons traiter chacune de ces difficultés de la façon suivante:

- 1) Si une fonction f n'est pas définie en un point x , on lui attribuera d'office une valeur *infinie*: $f(x) = +\infty$. Ceci nous permettra de travailler avec des fonctions qui sont virtuellement définies partout, et on verra que la contrepartie est minime car travailler avec des valeurs $+\infty$ n'est pas tant problématique.
- 2) Si on a une contrainte, afin d'éviter de travailler avec deux objets différents (fonction vs. contrainte) on va transformer la contrainte en ensemble. Pour cela on utilisera la notion d'*indicatrice* d'un ensemble, qui vaut 0 sur l'ensemble et $+\infty$ en dehors. Ceci nous permettra virtuellement de retirer toutes les contraintes, et de se focaliser la minimisation sans contraintes de fonctions $f : \mathbb{R}^N \rightarrow \mathbb{R} \cup \{+\infty\}$.
- 3) Si f n'est pas différentiable, on introduira la notion de *sous-gradient* qui généralise la notion de gradient à toutes les fonctions, même celles non différentiables ou à valeurs infinies. Ceci nous permettra d'établir des conditions d'optimalités (un sous-gradient va s'annuler) et de proposer des algorithmes (méthodes basées sur le sous-gradient).

Dans le reste de cette introduction, je propose quelques exemples de problèmes pratiques qui se modélisent comme des problèmes d'optimisation possédant ces propriétés. Par exemple les exemples .1 et .2 où la difficulté du problème se situe dans les contraintes. L'exemple .3 présente une famille de problèmes où les fonctions ne sont pas différentiables, et parfois à valeurs infinies. Cette liste ne se veut pas exhaustive, et n'est que le reflet des problèmes qui intéressent leur auteur.

Exemple .1 (Le problème du transport optimal). Le problème du transport optimal consiste à trouver comment transporter, de la façon la plus efficace/économique possible, un objet d'un point A vers un point B . Ou, plus exactement, de nombreux objets depuis tout un tas de points de départ A_i vers des points d'arrivée B_i (cf. Figure 1). Introduit à l'origine par Monge pour résoudre un problème de déplacement de tas de sable, ce problème permet de nos jours de répondre à des questions sur le « déplacement » d'objets plus abstraits, comme des images (cf. Figure 2). Ce problème peut être modélisé comme un problème d'optimisation, et plus précisément comme un problème d'optimisation linéaire de la forme

$$\underset{x \in \mathbb{R}^N}{\text{minimiser}} \langle c, x \rangle \text{ sous la contrainte que } \langle a_i, x \rangle \leq b_i,$$

où $c \in \mathbb{R}^N$, $a_1, \dots, a_M \in \mathbb{R}^N$ et $b_1, \dots, b_M \in \mathbb{R}$ dépendent du problème. Pour plus de détails sur cette modélisation, vous pouvez lire [cet](#) et [cet](#) article, dont je me suis inspiré ici.

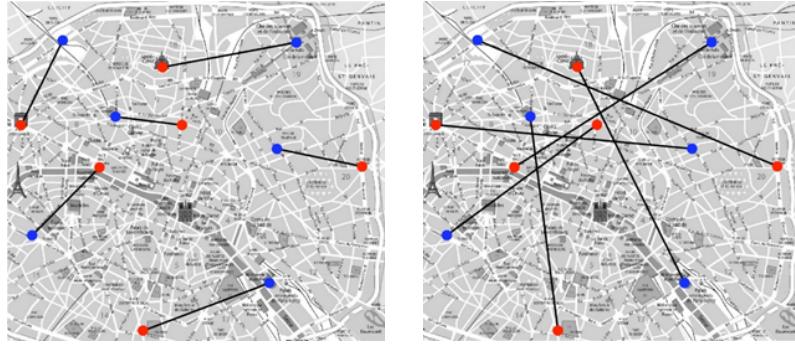


FIGURE 1 – Si chaque point bleu doit aller sur un point rouge, lequel doit aller où pour minimiser la somme des trajets à vol d'oiseau ? Et surtout: comment répondre à cette question sans avoir à tester les $n!$ combinaisons ?



FIGURE 2 – Application du Transport optimal: Une fois calculé un chemin optimal entre deux images (ici aux extrémités) on peut trouver au milieu de ce chemin une image (ici au centre) qui combine la forme d'une image avec le style de l'autre. Tout l'art ici consiste à définir correctement ce que « optimal » veut dire, qui est un problème *beaucoup* plus difficile que résoudre le problème de transport en lui-même. Extrait de l'article [Style transfer by relaxed optimal transport and self-similarity](#) par Kolkin et al., 2019 [10].

Exemple .2 (Problème de classification). On suppose que l'on dispose d'un certain type de données, et on veut être capable de les **classer** en deux groupes. Ce type de problème peut être très facile à réaliser pour un humain, mais toute la question est de savoir comment automatiser cette prise de décision pour l'implémenter sur une machine.

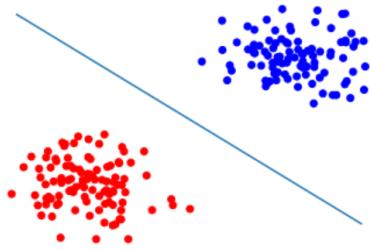


FIGURE 3 – Classifier deux groupes de points dans \mathbb{R}^2 , relativement facile.

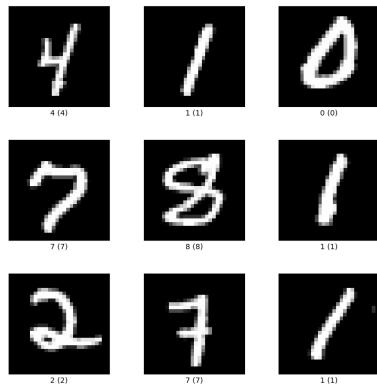


FIGURE 4 – Classifier des nombres écrits à la main, difficulté moyenne. Issu du jeu de données [MNIST](#), utilisé abondamment pour tester les réseaux de neurones.

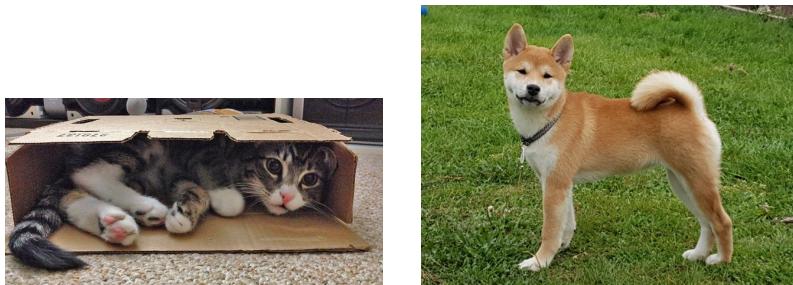


FIGURE 5 – Classifier des photos dans \mathbb{R}^N , $N > 10^6$, en deux catégories (chat/chien), très difficile.



FIGURE 6 – Classifier des visages humains, très très difficile.

Une modélisation possible de ce problème consiste en le transformer en un problème d'optimisation convexe, ayant la forme suivante:

$$\underset{x \in \mathbb{R}^N}{\text{minimiser}} \|Px\|^2 \text{ sous la contrainte que } \langle a_i, x \rangle \leq b_i,$$

où $P \in \mathcal{M}_N(\mathbb{R})$, $a_1, \dots, a_M \in \mathbb{R}^N$ et $b_1, \dots, b_M \in \mathbb{R}$ sont construites à partir des données à classer. Dans ce contexte, ce problème est communément appelé **Machine à vecteur de support** (Support Vector Machine, ou SVM). Pour plus d'informations sur ce problème, consulter la section [A.II](#) dédiée à la modélisation de ce problème, et le TP associé.

Exemple .3 (Traitement de l'image). Les problèmes en traitement de l'image sont nombreux et incluent notamment les problèmes de défloutage (retirer le flou qu'à subi une image) et de débruitage (retirer un bruit qui se traduit par des perturbations sur les pixels). On encode un image comme un vecteur $x \in \mathbb{R}^N$ dont les coefficients correspondent aux couleurs de chaque pixel, et les problèmes sus-mentionnés peuvent se modéliser comme des problèmes d'optimisation de la forme

$$\underset{x \in \mathbb{R}^N}{\text{minimiser}} R(x) + D(Ax; b),$$

où $R : \mathbb{R}^N \rightarrow \mathbb{R}$ est une fonction dite de régularisation, $A \in \mathcal{M}_N(\mathbb{R})$ est une matrice qui va par exemple encoder l'opération de floutage, $b \in \mathbb{R}^N$ correspond souvent à l'image dégradée, et $D : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$ est une fonction de pénalisation qui peut se voir comme une distance (bien que ce ne soit jamais une distance à strictement parler). L'idée derrière ce type de modélisation est assez simple: d'un côté on veut que x soit proche de l'image dégradée (donc que $D(Ax; b)$ soit petit) tandis que de l'autre on veut que x soit « de bonne qualité », ou « naturelle » (cela va s'obtenir en prenant $R(x)$ petit). Il existe toute une littérature explorant quelle type de fonction R choisir pour l'image soit de bonne qualité, et quel type de distance D prendre en fonction du bruit sur les données. En voici les exemples les plus standards:

- $R(x) = \|Wx\|_1$ où l'on note $\|\cdot\|_1$ la norme ℓ^1 , et W est une matrice appelée la transformée d'*ondelettes*. Il est connu que la transformée d'ondelette d'une image « naturelle » est un vecteur dont les coefficients sont pour la plupart égaux à zéro; et il est également connu que minimiser la norme ℓ^1 met à zéro de nombreux coefficients.

- $R(x) = \|Dx\|_1$ où D est une matrice de différences finies (on dit que R est la *variation totale* de x). Au vu de ce qui précède, minimiser $R(x)$ va faire en sorte que Dx ait beaucoup de coefficients à zéro. Or avoir des différences égales à zéro équivaut à ce que l'image soit localement constante: c'est une propriété désirable des images naturelles.
- $D(y - b) = \frac{1}{2}\|y - b\|^2$ qui est la pénalisation *quadratique* standard. On la retrouve en statistiques dans les problèmes de moindre carrés, mais aussi en traitement d'image, notamment lorsque les données sont corrompues par du bruit gaussien. Par exemple le modèle ROF combine une pénalisation quadratique avec la régularisation par variation totale.
- $D(y - b) = \|y - b\|_1$ qui est une pénalisation *robuste*, typiquement utilisée en présence de données aberrantes, car on sait qu'elle est plus robuste que la pénalisation quadratique.
- $D(y - b) = KL(y; b)$ où la pénalisation de Kullback-Liebler est définie par

$$KL(y; b) = \begin{cases} \sum_{i=1}^N b_i \log \left(\frac{b_i}{y_i} \right) - b_i + y_i & \text{si } b_i, y_i > 0, \\ +\infty & \text{sinon.} \end{cases}$$

Cette pénalisation est particulièrement appropriée en présence de bruit de Poisson, que l'on retrouve typiquement en imagerie astronomique.

La figure 7 présente une photo de trou noir, qui a été sélectionnée en minimisant un modèle combinant essentiellement les régularisations et pénalisations sus-mentionnées. La figure 8 illustre un problème de reconstruction d'image en utilisant des modèles combinant les ingrédients précédemment cités.

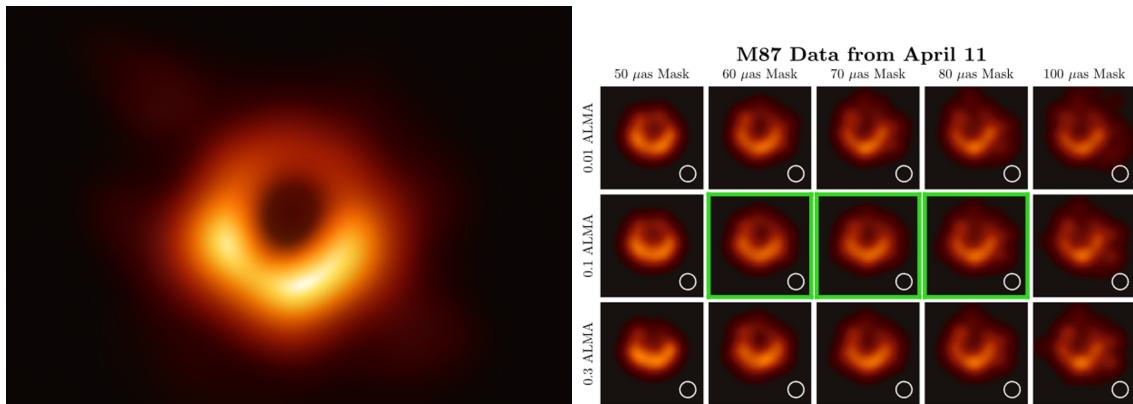


FIGURE 7 – Gauche: reconstruction d'une photo de trou noir, sélectionnée parmi plusieurs candidates (droite) en minimisant un modèle [1].

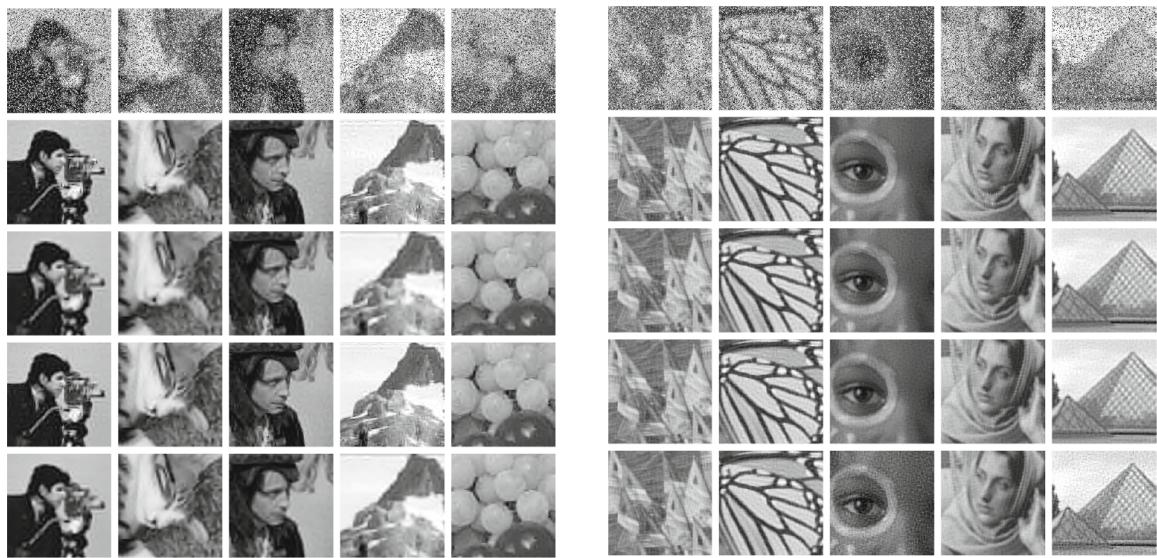


FIGURE 8 – Reconstruction d'image. Première ligne: image floutée et bruitée. Lignes suivantes: reconstruction en utilisant divers modèles.

Chapitre I

Ensembles convexes

Notations. On notera $[n] = \{1, \dots, n\}$. L'ordre partiel sur \mathbb{R}^N se définit par $x \leq y \Leftrightarrow x_i \leq y_i$ pour tout $i \in [N]$. On définira également l'ordre partiel strict $x < y \Leftrightarrow x_i < y_i$ pour tout $i \in [N]$. On notera $[Ax = b]$ au lieu de $\{x \in \mathbb{R}^N \mid Ax = b\}$. De même, on notera $[Ax \leq b]$ au lieu de $\{x \in \mathbb{R}^N \mid Ax \leq b\}$.

I.I Convexité

I.I.1 Définitions et calcul

Définition I.1 (Ensemble convexe). On dit que $C \subset \mathbb{R}^N$ est **CONVEXE** si

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

Exemple I.2 (Ensembles convexes). Voici quelques exemples d'ensembles convexes que l'on retrouvera tout au long de ce cours:

- F un sous-espace vectoriel. Par exemple, les **droites** $\mathbb{R}a$, où $a \in \mathbb{R}^N$.
- \mathcal{F} un sous-espace affine. Par exemple, les **hyperplans** $[\langle a, x \rangle = b]$ où $a \in \mathbb{R}^N$, $b \in \mathbb{R}$, ou encore les solutions de systèmes linéaires $[Ax = b]$.
- Une **demi-droite** $\mathbb{R}_{+}a$, où $a \in \mathbb{R}^N$.
- Un **demi-espace** (affine) $[\langle a, x \rangle \leq b]$ où $a \in \mathbb{R}^N$, $b \in \mathbb{R}$. On parlera de **demi-espace linéaire** lorsque $b = 0$.
- Une **boule** (fermée) $\mathbb{B}(a, \delta)$, où $a \in \mathbb{R}^N$, $\delta > 0$.
- Le **simplexe unité** $\Delta^N := \{\lambda \in \mathbb{R}^N \mid \lambda_i \geq 0 \text{ et } \sum_{i=1}^N \lambda_i = 1\}$.
- L'**orthant positif** $\mathbb{R}_+^N = \{x \in \mathbb{R}^N \mid x_i \geq 0\}$.

Définition I.3 (Combinaison convexe). Une **COMBINAISON CONVEXE** de vecteurs $x_1, \dots, x_p \in \mathbb{R}^N$ où $p \geq 1$ est un vecteur de la forme $\sum_{i=1}^p \lambda_i x_i$, où $(\lambda_i)_{i=1}^p \in \Delta^p$.

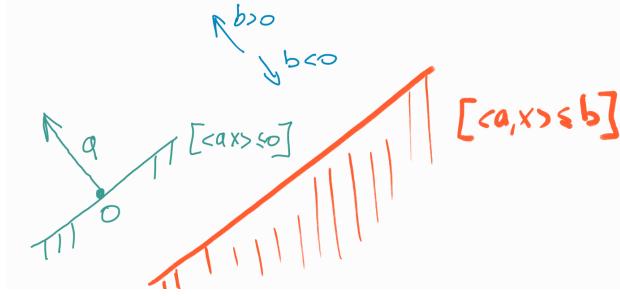


FIGURE I.1 – Un demi-espace

Proposition I.4 (Stabilité par combinaisons convexes). *Un ensemble $C \subset \mathbb{R}^N$ est convexe si et seulement si il est stable par combinaisons convexes.*

Démonstration. Si C est stable par combinaison convexe pour tout p , il l'est en particulier pour $p = 2$, ce qui fait de lui un convexe par définition. Supposons maintenant que C soit convexe et montrons qu'il est stable par combinaisons convexes. Le résultat est trivial si $p = 1$, et par définition de la convexité le résultat est vrai pour $p = 2$. Raisonnons par récurrence sur p , et supposons que C est stable par combinaisons convexes de taille $p \geq 2$. Considérons alors x_1, \dots, x_p, x_{p+1} ainsi que $(\lambda_i)_{i=1}^{p+1} \in \Delta^{p+1}$, et montrons que $x = \sum_{i=1}^{p+1} \lambda_i x_i \in C$. Pour ce faire, il suffit de poser $\mu = \sum_{i=1}^p \lambda_i$, tel que $\mu + \lambda_{p+1} = 1$. Si $\mu = 0$ alors la conclusion est triviale: $x = x_{p+1} \in C$. Si $\mu \neq 0$, on peut écrire

$$x = \lambda_{p+1} x_{p+1} + \sum_{i=1}^p \lambda_i x_i = (1 - \mu)x_{p+1} + \mu \left(\sum_{i=1}^p \frac{\lambda_i}{\mu} x_i \right).$$

Or on voit que $\sum_{i=1}^p \frac{\lambda_i}{\mu} = 1$ ce qui veut dire que $x' := \sum_{i=1}^p \frac{\lambda_i}{\mu} x_i \in C$ par récurrence. On conclut alors via la définition de la convexité que $x = (1 - \mu)x_{p+1} + \mu x' \in C$. ■

Proposition I.5 (Intersection de convexes). *Soient $(C_i)_{i \in I} \subset \mathbb{R}^N$ une famille de convexes. Alors $\bigcap_{i \in I} C_i$ est convexe.*

Démonstration. Immédiat. ■

Définition I.6 (Polyèdre). Un **POLYÈDRE** est un ensemble de la forme $[Ax \leq b]$, où $A \in \mathcal{M}_{m,n}(\mathbb{R})$ et $b \in \mathbb{R}^m$. Si on note a_i la i -ème ligne de A , alors on peut écrire

$$[Ax \leq b] = \{x \in \mathbb{R}^N \mid \langle a_i, x \rangle \leq b_i, \forall i \in [m]\} = \bigcap_{i=1}^m [\langle a_i, x \rangle \leq b_i].$$

Remarque I.7 (Intersections de demi-espaces). Par définition, un ensemble est un polyèdre si et seulement si il est une intersection *finie* de demi-espaces affines. Au vu de la propo-

sition précédente, les polyèdres sont clairement convexes. Nous verrons plus tard que les intersections *infinies* de demi-espaces affines sont exactement les ensembles convexes fermés (voir théorème I.19). Nous verrons également que les intersections infinies de demi-espaces *linéaires* sont exactement les cônes fermés (voir théorème I.39), et en particulier les intersections finies de demi-espaces linéaires sont les cônes polyédraux.

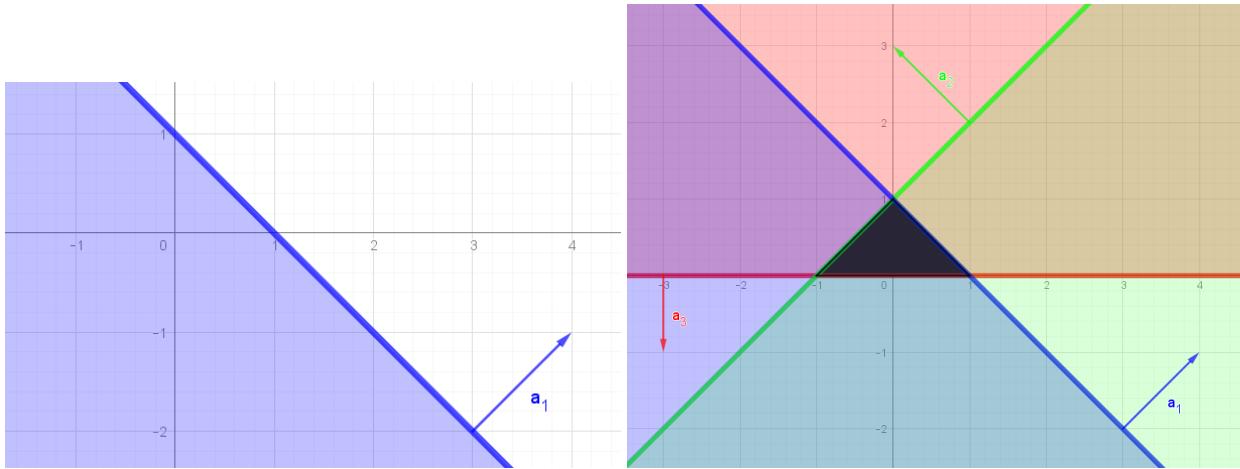


FIGURE I.2 – Gauche : En bleu, le demi-espace $\{z = (x, y) \in \mathbb{R}^2 \mid x + y \leq 1\}$, que l'on peut décrire comme $[\langle a_1, \cdot \rangle \leq b_1] = \{z = (x, y) \mid \langle a_1, z \rangle \leq b_1\}$ avec $a_1 = (1, 1)^\top$, $b_1 = 1$; En gras, l'hyperplan supporté par a_1 . Droite : Trois demi-espaces de la forme $[\langle a_i, \cdot \rangle \leq b_i]$ avec $a_1 = (1, 1)^\top$, $b_1 = 1$, $a_2 = (-1, 1)$, $b_2 = 1$ et $a_3 = (0, -1)$, $b_3 = 0$; et leur intersection, un triangle (en noir).

Exemple I.8 (Polyèdres). Faisons un peu de zoologie:

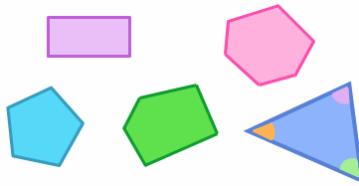


FIGURE I.3 – Quelques polyèdres bornés dans \mathbb{R}^2 . Ce sont des polygones convexes.

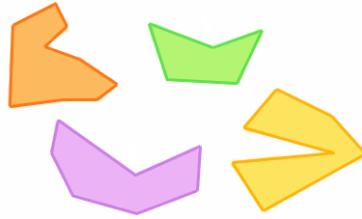


FIGURE I.4 – Ces polygones du plan ne sont pas des polyèdres.

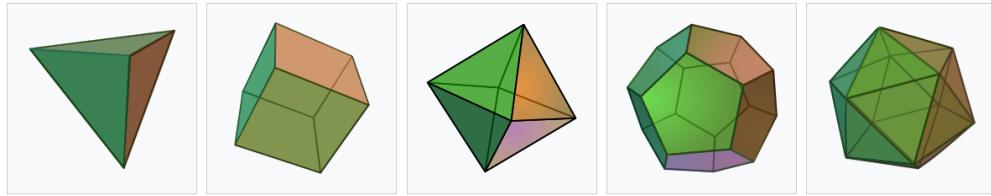


FIGURE I.5 – Cinq polyèdres bornés dans \mathbb{R}^3 (connus comme les cinq solides de Platon).

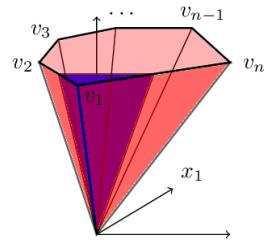


FIGURE I.6 – Un polyèdre de \mathbb{R}^3 qui est également un cône (non borné). Le cône a été tronqué afin de ne pas occuper un espace infini.

Proposition I.9 (Somme finie de convexes). Si $C, D \subset \mathbb{R}^N$ sont convexes alors $C + D$ est convexe. Plus généralement, toute somme finie de convexes est convexe.

Démonstration. Immédiat. ■

Exemple I.10 (S.e.a). Tout sous-espace affine \mathcal{F} peut s'écrire $F + a$, c'est-à-dire la somme d'un espace vectoriel F et d'un singleton a . Ces deux derniers étant convexes, on en déduit que \mathcal{F} est convexe.

Proposition I.11 (Produit cartésien de convexes). Si $C \subset \mathbb{R}^N$ et $D \subset \mathbb{R}^M$ sont convexes, alors $C \times D \subset \mathbb{R}^{N+M}$ est convexe.

Démonstration. Observer que $C \times D = (C \times \{0\}) + (\{0\} \times D)$, et que $C \times \{0\}$ est trivialement convexe. ■

Proposition I.12 (Convexité et topologie). Soit $C \subset \mathbb{R}^N$ convexe. Alors $\text{adh } C$ et $\text{int } C$ sont convexes.

Démonstration. Laissé en exercice, voir TD. ■

Définition I.13 (Enveloppe convexe). Soit $A \subset \mathbb{R}^N$. On définit son **ENVELOPPE CONVEXE** notée $\text{co } A$ comme l'ensemble des combinaisons convexes de A , c'est-à-dire

$$\text{co } A = \left\{ \sum_{i=1}^p \lambda_i a_i \mid p \geq 1, a_1, \dots, a_p \in A, (\lambda_i)_{i=1}^p \in \Delta^p \right\}.$$

Proposition I.14 (L'enveloppe convexe est convexe). Soit $A \subset \mathbb{R}^N$. Alors $\text{co } A$ est convexe.

Démonstration. Notons $C := \text{co } A$ et montrons que C est convexe. Pour cela prenons $x, y \in C$, $t \in [0, 1]$ et montrons que $z = (1 - t)x + ty \in C$. Par définition nous avons $z = (1 - t) \sum_{i=1}^p \lambda_i a_i + t \sum_{j=1}^q \lambda'_j a'_j$ où $\lambda_i \in \Delta^p$ et $\lambda'_j \in \Delta^q$. Nous voyons donc que z est une combinaison d'éléments de A (les a_i et a'_j), avec des coefficients positifs dont la somme vaut $\sum_i (1 - t)\lambda_i + \sum_j t\lambda'_j = 1$. Donc z est bien une combinaison convexe de A . ■

Exemple I.15 (Polytopes). Un **polytope** est l'enveloppe convexe d'un nombre *fini* de points, c'est-à-dire un ensemble de la forme $\text{co}(a_1, \dots, a_p)$. On peut constater que les polytopes de \mathbb{R}^2 sont exactement les polygones convexes. Si $A \in \mathcal{M}_{M,N}(\mathbb{R})$ est une matrice dont les colonnes sont $A = [a_1, \dots, a_p]$, on notera simplement $\text{co}(A)$ au lieu de $\text{co}(a_1, \dots, a_p)$.

Exemple I.16 (Boule unité de la norme ℓ^1). Si on considère la boule unité de la norme ℓ^1

$$\mathbb{B}_1 := \{x \in \mathbb{R}^N \mid \|x\|_1 = \sum_{i=1}^N |x_i| \leq 1\},$$

alors on peut montrer que c'est un polytope. Plus précisément, c'est l'enveloppe convexe des vecteurs $\pm e_1, \dots, \pm e_N$, où les e_i sont les vecteurs de la base canonique.

Remarque I.17 (Polytopes = polyèdres bornés). Un Théorème dû à Minkowski¹ énonce que les polytopes sont exactement les polyèdres *bornés*. On se convainc assez facilement que c'est vrai en faisant des dessins dans \mathbb{R}^2 . Ceci dit le résultat est loin d'être trivial (voir Théorème B.10 en Annexe)! Si on nous dit que $C = [Ax \leq b]$ est borné, comment trouver des points c_1, \dots, c_p dont l'enveloppe convexe redonne C ? Il faut trouver les sommets du polyèdre, ce qui n'est pas chose aisée en pratique, bien qu'il existe des algorithmes² pour le faire.

¹H. Minkowski, *Allgemeine Lehrsätze über die konvexe Polyeder*, 1897.

²G. M. Ziegler, *Lectures on Polytopes*, 1995. Voir le Chapitre 1.

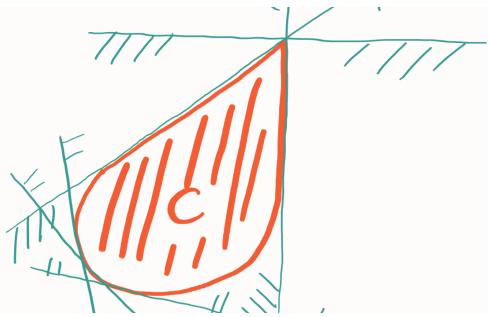


FIGURE I.7 – Un convexe vu comme intersection de demi-espaces

Proposition I.18 (Les polytopes sont fermés). Soient $a_1, \dots, a_p \in \mathbb{R}^N$. Alors $\text{co}(a_1, \dots, a_p)$ est fermé.

Démonstration. La preuve est laissée en exercice (voir TD). ■

Théorème I.19 (Caractérisation des convexes comme intersection de demi-espaces). Un ensemble $C \subset \mathbb{R}^N$ est convexe fermé si et seulement si il est l'intersection de demi-espaces.

Démonstration. Admis pour l'instant, on en verra une preuve dans le prochain chapitre, voir le corollaire II.108. L'étudiant·e curieu·x·se peut également consulter une preuve directe dans la section C.I.3 en annexe. ■

I.I.2 Projection

Définition I.20 (Projection). Soit $C \subset \mathbb{R}^N$ fermé convexe non vide, et $x \in \mathbb{R}^N$. On dit que $p \in \mathbb{R}^N$ est une **PROJECTION** de x sur C , si $p \in C$ et

$$(\forall c \in C) \quad \|p - x\| \leq \|c - x\|.$$

Si la projection existe et est unique, on notera $p = \text{proj}_C(x)$.

Remarque I.21 (Points fixes de la projection). Notons que si $x \in C$ alors forcément $\text{proj}_C(x) = x$.

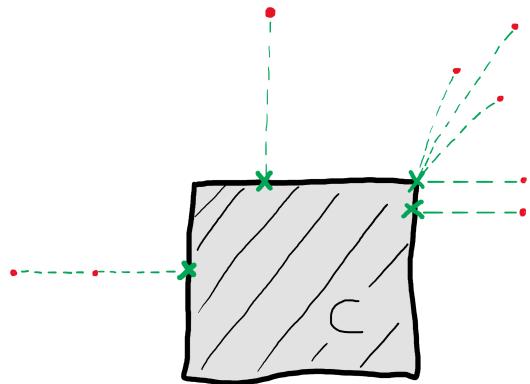


FIGURE I.8 – Diverses projections sur un carré. Des points différents (en rouge) peuvent se projeter sur le même point (en vert).

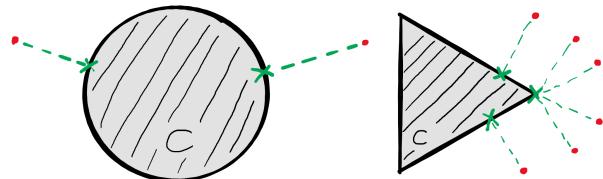


FIGURE I.9 – Encore quelques projections sur des convexes.

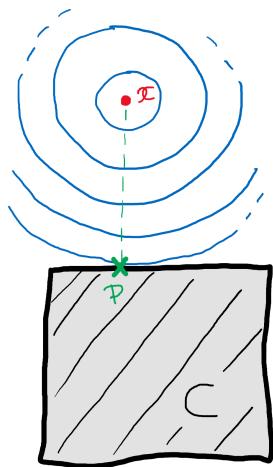


FIGURE I.10 – Un ensemble convexe C , un point x (en rouge) et sa projection $p = \text{proj}_C(x)$ sur C (en vert), qui est le point de C qui est le plus proche possible de x . Pour trouver cette projection on peut imaginer une boule centrée en x dont le rayon grossit jusqu'à toucher C : lorsque l'intersection entre cette boule et C est réduite à un point, alors ce point est exactement $\text{proj}_C(x)$.

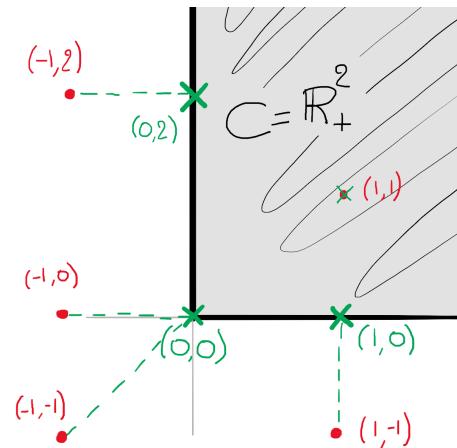


FIGURE I.11 – Divers points x (en rouge) et leurs projections $p = \text{proj}_C(x)$ (en vert) sur l'orthant positif $C = \mathbb{R}^2_+$. Dans ce cas la projection a pour effet de mettre tous les coefficients négatifs à zéro.

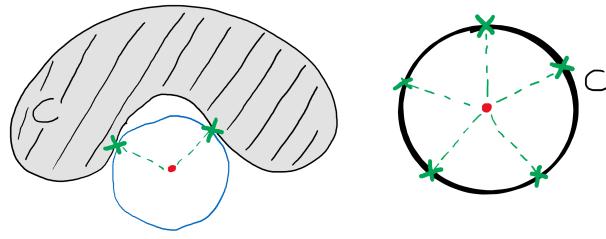


FIGURE I.12 – La projection n'est pas bien définie si C n'est pas convexe! Ici deux ensembles C non convexes, une patate et un cercle ($\text{cercle} \neq \text{disque}$) pour lesquels le point rouge peut trouver plus d'un point vert dans C qui minimise la distance.

Proposition I.22 (Caractérisation de la projection par les angles). Soient $C \subset \mathbb{R}^N$, $x \in \mathbb{R}^N$ et $p \in C$. Alors p est une projection de x sur C si et seulement si

$$(\forall c \in C) \quad \langle x - p, c - p \rangle \leq 0.$$

Démonstration. Soient $c \in C$, $t \in]0, 1]$ et $z_t = (1 - t)p + tc \in C$. Alors on peut écrire que

$$\begin{aligned} \|p - x\|^2 &\leq \|z_t - x\|^2 \\ \Leftrightarrow \|p - x\|^2 &\leq \|p - x + t(c - p)\|^2 \\ \Leftrightarrow 0 &\leq t^2\|c - p\|^2 - 2t\langle x - p, c - p \rangle \\ \Leftrightarrow 0 &\leq t\|c - p\|^2 - 2\langle x - p, c - p \rangle. \end{aligned}$$

Si p est une projection, alors on sait que cette inégalité est vérifiée, pour tout $t > 0$. En faisant tendre t vers zéro, on conclut en effet que $0 \geq \langle x - p, c - p \rangle$.

Si cette propriété est vérifiée, alors l'inégalité ci-dessus est vraie, donc avec $t = 1$ nous voyons que $\|p - x\|^2 \leq \|c - x\|^2$. Ceci nous permet de conclure que p est une projection. ■

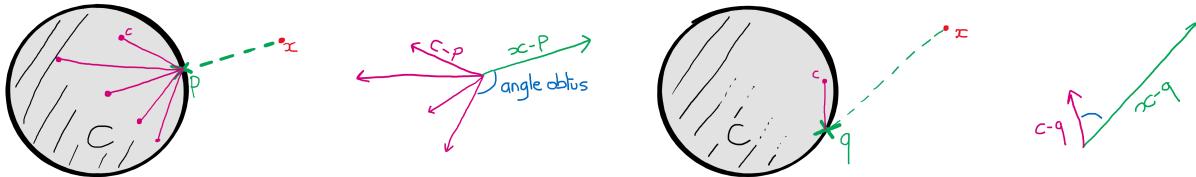
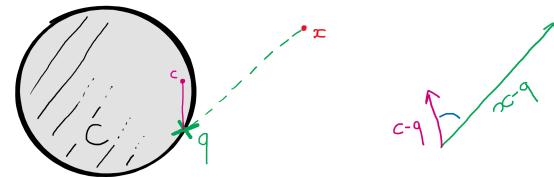


FIGURE I.13 – Caractérisation de la projection par les angles : on voit que si $p = \text{proj}_C(x)$, alors pour tout $c \in C$, le vecteur $c - p$ forme un angle obtus avec $x - p$.

La projection est toujours bien définie:

FIGURE I.14 – Caractérisation de la projection par les angles : on voit que si $q \neq \text{proj}_C(x)$, alors il existe un $c \in C$ tel que le vecteur $c - q$ forme un angle aigu avec $x - q$.



Théorème I.23 (de la Projection). Soit $C \subset \mathbb{R}^N$ fermé convexe non vide, et $x \in \mathbb{R}^N$. Alors la projection de x sur C existe et est unique.

Démonstration. Nous allons montrer existence et unicité.

- Existence : Définissons $d := \inf_{c \in C} \|c - x\|$. Par définition de l'inf, pour tout $n \in \mathbb{N}$ il doit exister un $c_n \in C$ tel que

$$d \leq \|c_n - d\| \leq d + \frac{1}{n}.$$

Clairement c_n est une suite bornée, donc elle admet une sous-suite c_{n_k} qui converge vers un certain \bar{c} . Puisque C est fermé on sait que $\bar{c} \in C$. De plus on peut passer à la limite dans l'inégalité précédente pour obtenir $d = \|\bar{c} - x\|$. Donc \bar{c} est une projection de x sur C .

- Unicité : Supposons qu'il existe deux projections $p, p' \in C$ de x sur C . On peut donc appliquer la caractérisation de la projection par les angles (proposition I.22) pour écrire

$$(\forall c \in C) \quad \langle x - p, c - p \rangle \leq 0 \quad \text{et} \quad \langle x - p', c - p' \rangle \leq 0.$$

En faisant la somme de ces deux inégalités, et en prenant $c = p$ ou p' , on obtient

$$\langle x - p, p' - p \rangle + \langle x - p', p - p' \rangle \leq 0.$$

Ceci est équivalent à $\|p' - p\|^2 \leq 0$, d'où l'unicité. ■

Exemple I.24 (Projection). On peut calculer explicitement la projection d'un vecteur x sur un ensemble C pour des ensembles simples. La plupart des exemples ci-dessous seront vus en TD.

- Si C est la boule unité $\mathbb{B}(0, 1)$, alors

$$\text{proj}_C(x) = \begin{cases} \frac{1}{\|x\|}x & \text{si } \|x\| \geq 1, \\ x & \text{si } \|x\| \leq 1. \end{cases}$$

- Si C est une boule quelconque $\mathbb{B}(a, r)$, alors

$$\text{proj}_C(x) = \begin{cases} \frac{r}{\|x-a\|}x + \left(1 - \frac{r}{\|x-a\|}\right)a & \text{si } \|x-a\| \geq r, \\ x & \text{si } \|x-a\| \leq r. \end{cases}$$

- Si C est l'orthant positif \mathbb{R}_+^N , alors $\text{proj}_C(x)$ consiste à mettre à zéro tous les coefficients négatifs de x , c'est-à-dire³ :

$$\text{proj}_C(x) = ((x_i)_+)^N_{i=1}.$$

³Ici $x_+ = \max\{x, 0\}$ désigne la partie positive de x .

- Si C est une droite $\mathbb{R}a$, alors $\text{proj}_C(x) = \frac{\langle a, x \rangle}{\|a\|^2} a$.
- Si C est une demi-droite $\mathbb{R}_+ a$, alors $\text{proj}_C(x) = \frac{(\langle a, x \rangle)_+}{\|a\|^2} a$.
- Si C est un hyperplan $[\langle a, x \rangle = b]$, alors $\text{proj}_C(x) = x - \frac{\langle a, x \rangle - b}{\|a\|^2} a$.
- Si C est un demi-espace $[\langle a, x \rangle \leq b]$, alors $\text{proj}_C(x) = x - \frac{(\langle a, x \rangle - b)_+}{\|a\|^2} a$.
- Si C est le simplexe Δ^N , il n'y a pas de forme close pour calculer sa projection, mais on dispose d'un algorithme qui la calcule en temps fini, avec une complexité de l'ordre de $O(N \log(N))$.

Proposition I.25 (Continuité de la projection). Soit $C \subset \mathbb{R}^N$ convexe fermé non vide. Alors $\text{proj}_C : \mathbb{R}^N \rightarrow \mathbb{R}^N$ est continue car elle est 1-Lipschitzienne:

$$(\forall x, y \in \mathbb{R}^N) \quad \|\text{proj}_C(y) - \text{proj}_C(x)\| \leq \|y - x\|.$$

Démonstration. Admis pour l'instant, nous en verrons une preuve dans le dernier chapitre, voir le corollaire B.64. On peut également en voir une preuve directe dans la section C.I.1 en annexe. ■

On termine cette section avec un théorème de séparation de Hahn-Banach. Il en existe plusieurs versions, avec des hypothèses plus ou moins fortes sur C . Ici on suppose C fermé, ce qui nous permet d'avoir une séparation forte. On réfère à la section C.I.2 en annexe pour des exemples de séparation faible sans condition de fermeture.

Théorème I.26 (de séparation forte de Hahn-Banach). Soit $C \subset \mathbb{R}^N$ convexe fermé non vide, tel que $x \notin C$. Alors il existe $\alpha \neq 0$ et $\varepsilon > 0$ tels que pour tout $c \in C$, $\langle \alpha, c \rangle \geq \varepsilon + \langle \alpha, x \rangle$.

Démonstration du théorème I.26. Puisque C est convexe fermé et non vide, nous pouvons considérer $p := \text{proj}_C(x)$, qui existe bien d'après le théorème I.23. D'après la caractérisation de la projection par les angles (proposition I.22), nous avons pour tout $c \in C$ que

$$\langle x - p, c - p \rangle \leq 0,$$

ce qui peut être écrit de manière équivalente en

$$\|p - x\|^2 + \langle p - x, x \rangle \leq \langle p - x, c \rangle.$$

Ainsi, en posant $\alpha = p - x$ et $\varepsilon := \|\alpha\|^2$, nous voyons que l'énoncé est vérifié. On pense à vérifier que $\varepsilon \neq 0$ et $\alpha \neq 0$, ce qui est vrai ici puisque $x \notin C$ implique que $p \neq x$. ■

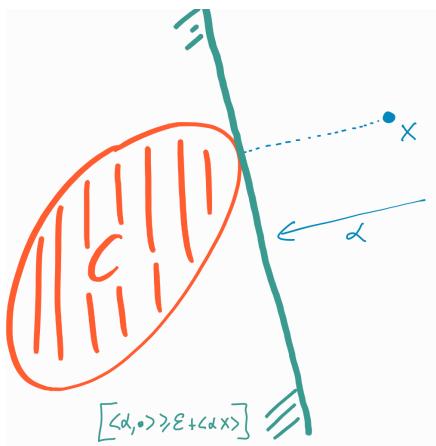


FIGURE I.15 – Séparation entre un convexe fermé et un point extérieur.

I.II Approximation d'un convexe

I.II.1 Cônes

Définition I.27 (Cône). On dit que $K \subset \mathbb{R}^N$ est un **CÔNE** si K est convexe et si il est positivement homogène, au sens où $\mathbb{R}_+K \subset K$.

Remarque I.28 (Observations). Notons qu'un cône non vide contient toujours 0 puisque $0K \subset K$. Notons également que si $x \in K$ alors $\mathbb{R}_+x \subset K$, autrement dit un cône est une réunion (possiblement infinie) de demi-droites.

Remarque I.29 (Cône non convexe?). Sachez que dans la littérature le mot *cône* peut être utilisé pour décrire un ensemble qui est simplement positivement homogène (sans hypothèse de convexité). Faites donc attention lorsque vous lisez un document de bien vérifier quelle est la définition de cône. Certains auteurs vont dans ce cas parler de cônes, puis de cônes convexes. Dans ce cours nous faisons le choix d'appeler cône un ensemble positivement homogène et convexe, notamment parce que nous ne manipulerons jamais de cône non-convexe.

Exemple I.30 (Cônes). Nous avons déjà rencontré les cônes suivants:

- la demi-droite \mathbb{R}_+a ,
- le sous-espace vectoriel F ,
- l'orthant \mathbb{R}_+^N ,
- les cônes polyédraux $[Ax \leq 0]$,
- le demi-espace linéaire $[\langle a, x \rangle \leq 0]$.

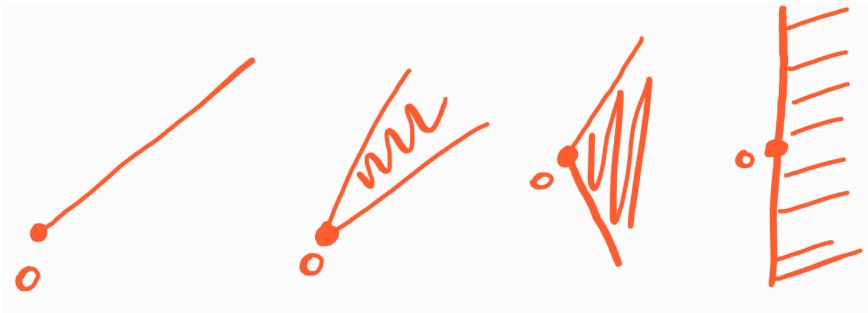


FIGURE I.16 – Cônes.

Proposition I.31 (Intersection de cônes). Soient $(K_i)_{i \in I} \subset \mathbb{R}^N$ une famille de cônes. Alors $\bigcap_{i \in I} K_i$ est un cône.

Démonstration. Nous savons déjà que l'intersection de convexes est convexe (proposition I.5). De plus il est immédiat de voir que l'intersection d'ensembles positivement homogène l'est aussi. ■

Proposition I.32 (Somme de cônes). Soient $K, K' \subset \mathbb{R}^N$ deux cônes. Alors $K + K'$ est un cône.

Démonstration. Immédiat, notamment avec la proposition I.9. ■

Proposition I.33 (Adhérence d'un cône). Si $K \subset \mathbb{R}^N$ est un cône, alors $\text{cl } K$ est un cône.

Démonstration. On sait que l'adhérence d'un convexe est convexe (I.12). Montrons donc que K est positivement homogène. Soit $x \in \text{cl } K$, $\lambda \geq 0$, et montrons que $\lambda x \in \text{cl } K$. Par définition $x = \lim_n x_n$ où $x_n \in K$, donc $\lambda x = \lim_n \lambda x_n$ où $\lambda x_n \in \mathbb{R}_+ K \subset K$. D'où le résultat. ■

Définition I.34 (Enveloppe conique). Soit $A \subset \mathbb{R}^N$. On définit son **ENVELOPPE CONIQUE** notée $\text{cone } A$ comme étant l'ensemble des combinaisons coniques de A , c'est-à-dire

$$\text{cone } A = \left\{ \sum_{i=1}^p \lambda_i a_i \mid p \geq 1, a_1, \dots, a_p \in A, (\lambda_i)_{i=1}^p \in \mathbb{R}_+^p \right\}.$$

Proposition I.35 (Les combinaisons coniques forment l'enveloppe conique). Soit $A \subset \mathbb{R}^N$. Alors $\text{cone } A$ est un cône, tel que

$$\text{cone } A = \mathbb{R}_+ \text{co } A = \text{co } \mathbb{R}_+ A.$$

Démonstration. Notons $K = \text{cone}(A)$.

Montrons que $K \subset \mathbb{R}_+ \text{co } A$: Si $x = \sum_i \lambda_i a_i$, alors $x = \frac{1}{\sum_i \lambda_i} \sum_i \frac{\lambda_i}{\sum_j \lambda_j} a_i \in \mathbb{R}_+ \text{co } A$.

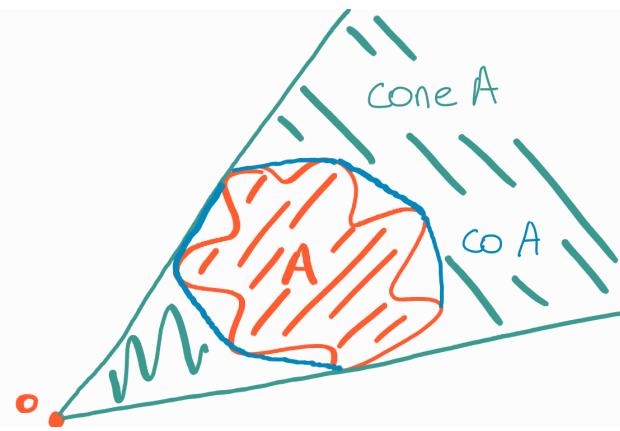


FIGURE I.17 – Enveloppe conique d'un ensemble.

Montrons que $\mathbb{R}_+ \text{co } A \subset \text{co } \mathbb{R}_+ A$: Si $x = \mu \sum_i \lambda_i a_i$ alors $x = \sum_i \lambda_i (\mu a_i) \in \text{co } \mathbb{R}_+ A$.

Montrons que $\text{co } \mathbb{R}_+ A \subset K$: Si $x = \sum_i \lambda_i (t_i a_i)$ alors $x = \sum_i (\lambda_i t_i) a_i \in K$.

On peut alors conclure que $\text{cone } A$ est convexe (car c'est une enveloppe convexe) et positivement homogène (car $\mathbb{R}_+ \text{cone } A = \mathbb{R}_+ \mathbb{R}_+ \text{co } A = \text{cone}(A)$). ■

Exemple I.36 (Cône à base finie). L'enveloppe conique d'une ensemble fini $\text{cone}(a_1, \dots, a_p)$ est parfois appelé un cône à base finie. Pour faire l'analogie avec les polytopes (qui sont l'enveloppe convexe d'un ensemble fini) on pourrait parler de cône polytopal. Si $A \in \mathcal{M}_{M,N}(\mathbb{R})$ est une matrice dont les colonnes sont $A = [a_1, \dots, a_p]$, on notera simplement $\text{cone}(A)$ au lieu de $\text{cone}(a_1, \dots, a_p)$.

Remarque I.37 (Cône à base finie = cône polyédral). Un Théorème dû à Weyl⁴ énonce que les cônes à base finie sont exactement les cônes polyédraux. Autrement dit, pour tout cône K de la forme $[Ax \leq 0]$, il existe k_1, \dots, k_p tels que $K = \text{cone}(k_1, \dots, k_p)$, et vice-versa (voir le Théorème B.7 en Annexe). De manière similaire à la remarque I.17, il n'est pas trivial de passer d'une forme à l'autre, bien que des algorithmes existent. Pour cette raison il reste utile de garder les deux noms (base finie vs. polyédral) pour indiquer de quelle présentation du cône on dispose.

Remarque I.38 (Polyèdre = polytope + cône polyédral). Un Théorème dû à Motzkin⁵ énonce que tout polyèdre peut s'écrire comme la somme d'un polytope et d'un cône polyédral (voir le Théorème B.9 en Annexe). Autrement dit, pour tout C de la forme $[Ax \leq b]$, il existe $c_1, \dots, c_p, d_1, \dots, d_q$ tels que $C = \text{co}(c_1, \dots, c_p) + \text{cone}(d_1, \dots, d_q)$ (et vice-versa).

⁴H. Weyl, *Elementare theorie der konvexen polyeder*, 1934.

⁵T. Motzkin, *Beiträge zur Theorie der linearen Ungleichungen*, Thesis, 1936.

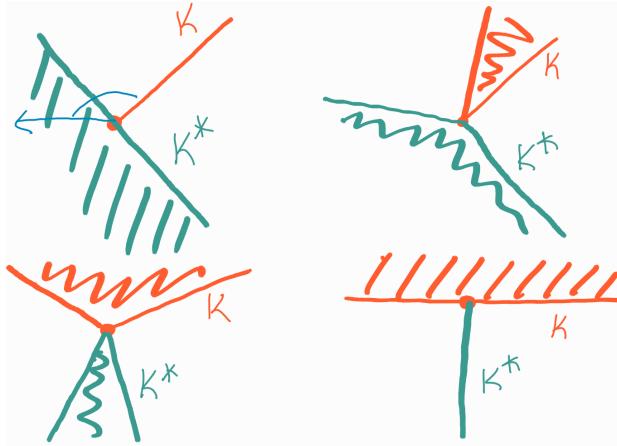


FIGURE I.18 – Cônes polaires.

Théorème I.39 (Caractérisation des cônes fermés par intersection). *Les cônes fermés sont exactement les intersections de demi-espaces linéaires.*

Démonstration. On admet ce résultat pour l'instant, on en verra une preuve dans le prochain chapitre, voir le corollaire II.109. L'étudiant·e curieu·x·se peut également consulter une preuve directe dans la section C.I.4 en annexe. ■

I.II.2 Cône polaire

Définition I.40 (Cône polaire). Soit $K \subset \mathbb{R}^N$ un cône. On définit son **POLAIRE** noté K^* par

$$K^* = \{x^* \in \mathbb{R}^N \mid (\forall x \in K) \langle x^*, x \rangle \leq 0\}.$$

Proposition I.41 (Cône polaire est un cône). *Si $K \subset \mathbb{R}^N$ est un cône, alors K^* est un cône fermé non vide.*

Démonstration. Il nous faut vérifier que K^* est convexe, positivement homogène, fermé et non vide.

K^* est non vide: cela vient du fait qu'il contient nécessairement 0 par définition.

K^* est convexe: Si $x^*, y^* \in K^*$ et $t \in [0, 1]$ alors pour tout $x \in K$ nous avons que

$$\langle (1-t)x^* + ty^*, x \rangle = (1-t)\langle x^*, x \rangle + t\langle y^*, x \rangle \leq 0.$$

K^* est p.h.: Si $x^* \in K^*$ et $\lambda \geq 0$, alors pour tout $x \in K$ nous avons que $\langle \lambda x^*, x \rangle \leq 0$.

K^* est fermé: Si $x_n^* \in K^*$ converge vers x^* , alors pour tout $x \in K$ nous avons que $\langle x^*, x \rangle = \lim_n \langle x_n^*, x \rangle \leq 0$. ■

Proposition I.42 (Cône polaire du produit cartésien). Soient $K_1 \subset \mathbb{R}^N$ et $K_2 \subset \mathbb{R}^M$ deux cônes. Alors $(K_1 \times K_2)^* = K_1^* \times K_2^*$.

Démonstration. On procède par double inclusion.

\subset : Soit $x^* = (x_1^*, x_2^*) \in (K_1 \times K_2)^*$ et montrons que $x^* \in K_1^* \times K_2^*$. Par symétrie, il suffit de montrer que $x_1^* \in K_1^*$. Prenons donc $x_1 \in K_1$ quelconque, et définissons $x = (x_1, 0) \in K_1 \times K_2$. On voit alors bien que $\langle x_1^*, x_1 \rangle = \langle x^*, x \rangle \leq 0$.

\supset : Soit $x^* = (x_1^*, x_2^*) \in K_1^* \times K_2^*$ et montrons que $x^* \in (K_1 \times K_2)^*$. Prenons donc $x = (x_1, x_2) \in K_1 \times K_2$ quelconque et montrons que $\langle x^*, x \rangle \leq 0$. Ce produit est égal à $\langle x_1^*, x_1 \rangle + \langle x_2^*, x_2 \rangle$ qui est bien négatif puisque $x_i^* \in K_i^*$ et $x_i \in K_i$. ■

Exemple I.43 (Cône polaire). On peut calculer le cône polaire de nombreux cônes élémentaires. Pour plus de détails, voir la feuille de TD.

- Si $K = F$ un s.e.v., alors K^* est son orthogonal F^\perp .
- Si $K = \mathbb{R}_+ a$, alors K^* est le demi-espace linéaire $[\langle a, x \rangle \leq 0]$.
- Si $K = [\langle a, x \rangle \leq 0]$, alors K^* est la demi-droite $\mathbb{R}_+ a$.
- Si $K = [0, +\infty[$ alors $K^* =]-\infty, 0]$.
- Si $K = \mathbb{R}_+^N$, alors K^* est l'orthant négatif \mathbb{R}_-^N .
- Si $K = \{0\}$ alors $K^* = \mathbb{R}^N$, et vice-versa.
- Si K est un cône polyédral, alors K^* est aussi un cône polyédral, voir proposition I.47.

Proposition I.44 (Le passage au polaire est décroissant). Si $K_1, K_2 \subset \mathbb{R}^N$ sont des cônes, alors $K_1 \subset K_2 \Rightarrow K_2^* \subset K_1^*$.

Démonstration. Soit $x_2^* \in K_2^*$, et montrons que $x_2^* \in K_1^*$. Pour cela prenons $x_1 \in K_1$ et vérifions que $\langle x_2^*, x_1 \rangle \leq 0$. Or ceci est vrai car $x_1 \in K_1 \subset K_2$ et $x_2^* \in K_2^*$. ■

Théorème I.45 (du cône bipolaire). Si $K \subset \mathbb{R}^N$ est un cône fermé, alors $K^{**} = K$.

Démonstration. On prouvera ce résultat dans le prochain chapitre, voir le corollaire II.110. L'étudiant·e curieu·x·se peut également consulter une preuve directe dans la section C.I.5 en annexe. ■

Remarque I.46 (Biorthogonal d'un s.e.v.). On retrouve ici un résultat bien connu d'algèbre linéaire, qui dit que si F est un s.e.v. alors $(F^\perp)^\perp = F$. Pour le voir, il suffit de se rappeler que F est un cône et que $F^* = F^\perp$.

Proposition I.47 (Polaire d'un cône polyédral - Lemme de Farkas). Soit $A \in \mathcal{M}_{M,N}(\mathbb{R})$ une matrice et notons a_1, \dots, a_M les lignes de A . Alors

$$[Ax \leq 0]^* = \text{cone}(a_1, \dots, a_M) \quad \text{and} \quad \text{cone}(a_1, \dots, a_M)^* = [Ax \leq 0].$$

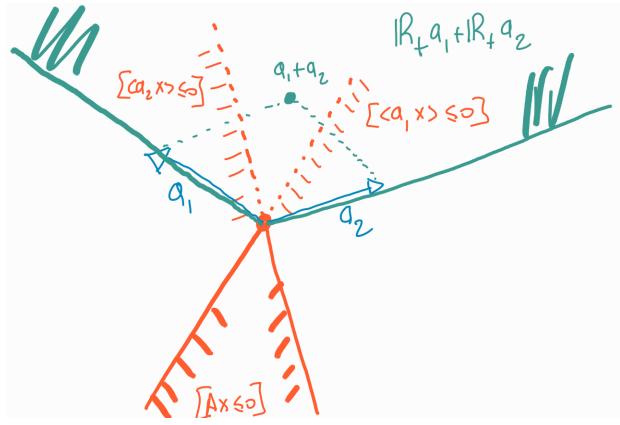


FIGURE I.19 – Polaire d'un cône polyédral.

Démonstration. La preuve est laissée en exercice, voir la feuille de TD correspondante. On peut aussi regarder en Annexe la Proposition B.3. ■

Remarque I.48 (Une généralisation d'un résultat bien connu). Le résultat de la proposition I.47, que l'on peut écrire sous forme matricielle comme

$$[Ax \leq 0]^* = \text{cone}(A^*) \quad \text{and} \quad \text{cone}(A)^* = [A^*x \leq 0], \quad (\text{I.1})$$

peut se voir comme une généralisation du résultat bien connu que

$$(\ker A)^\perp = \text{Im } A^* \quad \text{and} \quad (\text{Im } A)^\perp = \ker A^*. \quad (\text{I.2})$$

En effet, le noyau $\ker A$ peut aussi s'écrire $[Ax = 0]$ tandis que l'image $\text{Im } A^*$ peut s'écrire $\text{span}(a_1, \dots, a_p)$, c'est-à-dire que l'image d'une matrice est l'enveloppe linéaire de ses colonnes. Autrement dit (I.1) est aux inégalités ce que (I.2) est aux égalités. Une autre manière de voir le lien entre ces deux résultats est d'écrire

$$[Ax \leq 0] = A^{-1}\mathbb{R}_-, \quad [Ax = 0] = A^{-1}\{0\}, \quad \text{cone}(A^*) = A^*\mathbb{R}_+^N, \quad \text{Im } A^* = A^*\mathbb{R}^N.$$

On voit alors que (I.2) s'écrit $(A^{-1}\{0\})^* = A^*\mathbb{R}^N$ tandis que nous venons de prouver (I.1) qui s'écrit $(A^{-1}\mathbb{R}_-)^* = A^*\mathbb{R}_+^N$. Notez que pour chaque résultat il y a un couple de cônes polaires qui apparaissent : $\{0\}$ et \mathbb{R}^N pour (I.2), et \mathbb{R}_- et \mathbb{R}_+^N pour (I.1). Ce n'est pas une coïncidence ! Il est possible de montrer (c'est un exercice du TD) que pour tout cône fermé K on a l'égalité $(A^{-1}K)^* = \text{cl } A^*K^*$.

I.II.3 Cônes tangent et normal à un convexe

Dans cette section on considérera toujours (sauf mention contraire) que $C \subset \mathbb{R}^N$ est un ensemble convexe fermé non vide.

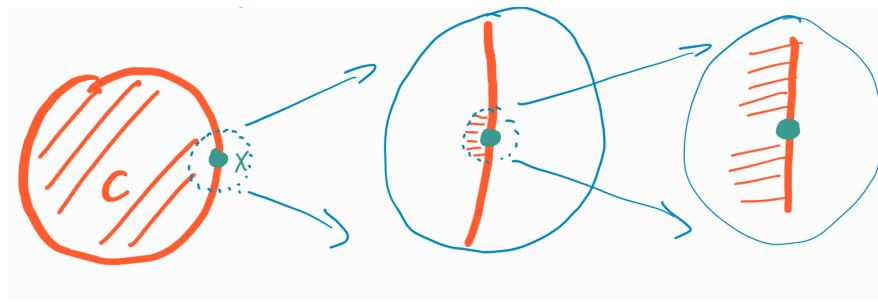


FIGURE I.20 – Le cône tangent à la boule unité est un demi-espace. On le voit en zoomant au voisinage du point.

Définition I.49 (Cônes tangent et normal). Soit $x \in C$. On définit le **CÔNE TANGENT** de C en x par

$$T_C(x) = \text{cl } \mathbb{R}_+(C - x).$$

On définit également le **CÔNE NORMAL** de C en x par $N_C(x) = T_C(x)^*$. Si $x \notin C$, on notera par convention que $T_C(x) = N_C(x) = \emptyset$.

Proposition I.50 (Cônes tangent et normal sont un cône). Pour tout $x \in C$, $T_C(x)$ et $N_C(x)$ sont des cônes fermés non vides. En particulier, $T_C(x) = N_C(x)^*$.

Démonstration. Le cas de $N_C(x)$ est trivial en temps que cône polaire (voir proposition I.41). En ce qui concerne $T_C(x)$: il est fermé par définition (c'est l'adhérence de $\mathbb{R}_+(C - x)$); c'est un cône car $C - x$ est convexe donc $\mathbb{R}_+(C - x) = \mathbb{R}_+ \text{co}(C - x) = \text{cone}(C - x)$ d'après la proposition I.35, et on sait que l'adhérence d'un cône est un cône (voir proposition I.33); enfin il est non vide car $x \in C$ implique que $0 \in T_C(x)$. Enfin, le fait que $T_C(x) = N_C(x)^*$ est une conséquence du théorème I.45 sur le cône bipolaire. ■

Proposition I.51 (Caractérisation du cône normal par les angles). Soit $x \in C$. Alors $x^* \in N_C(x)$ si et seulement si

$$(\forall c \in C) \quad \langle x^*, c - x \rangle \leq 0.$$

Démonstration. On procède par double implication.

\Rightarrow : On a par définition $C - x \subset T_C(x)$, donc pour tout $c - x \in C - x$ et pour $x^* \in T_C(x)^*$ on a bien un angle obtus entre les deux.

\Leftarrow : Soit x^* tel que $\langle x^*, c - x \rangle \leq 0$ pour tout $c \in C$, et montrons que $x^* \in N_C(x)$. Prenons donc $d \in T_C(x)$ quelconque, et montrons que $\langle x^*, d \rangle \leq 0$. Par définition de $T_C(x)$, on a $d = \lim d_n$ où $d_n = \lambda_n(c_n - x)$ avec $\lambda_n \geq 0$ et $c_n \in C$. Clairement on a $\langle x^*, d_n \rangle = \lambda_n \langle x^*, c_n - x \rangle \leq 0$, donc on peut conclure en passant à la limite. ■

Proposition I.52 (Caractérisation du cône normal par la projection). Soit $x \in C$. Alors $x^* \in N_C(x)$ si et seulement si $\text{proj}_C(x + x^*) = x$.

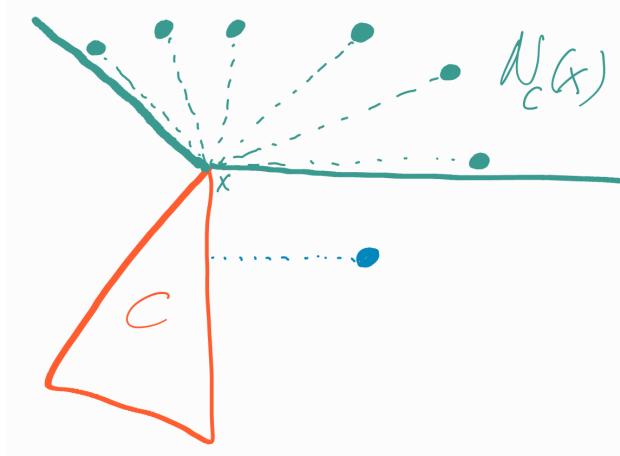


FIGURE I.21 – Le cône normal s'obtient en cherchant la réciproque de la projection.

Démonstration. Il suffit de combiner le résultat de la proposition I.51 avec la caractérisation de la projection par les angles (voir proposition I.22):

$$\begin{aligned}
 & x^* \in N_C(x) \\
 \Leftrightarrow & (\forall c \in C) \quad \langle x^*, c - x \rangle \leq 0 \\
 \Leftrightarrow & (\forall c \in C) \quad \langle (x^* + x) - x, c - x \rangle \leq 0 \\
 \Leftrightarrow & x = \text{proj}_C(x + x^*).
 \end{aligned}$$

■

Proposition I.53 (Cônes tangent et normal sont triviaux à l'intérieur). Soit $x \in C$ convexe fermé. Alors

$$x \in \text{int } C \Leftrightarrow T_C(x) = \mathbb{R}^N \Leftrightarrow N_C(x) = \{0\}.$$

Démonstration. On procède par triple implication.

\Rightarrow : Supposons que $x \in \text{int } C$, montrons que $x \in T_C(x)$. On va en fait simplement montrer que $\mathbb{R}_+(C - x) = \mathbb{R}^N$. Pour cela, prenons $d \in \mathbb{R}^N$ quelconque, et vérifions que $d \in \mathbb{R}_+(C - x)$. On va donc essayer d'écrire $d = \lambda(c - x)$ avec $\lambda \geq 0$ et $c \in C$ bien choisis. Puisque $x \in \text{int } C$ on dispose d'un $\varepsilon > 0$ tel que $\mathbb{B}(x, \varepsilon) \subset C$. On peut alors définir $c := x + \frac{\varepsilon}{\|d\|}$ qui vérifie $c \in \mathbb{B}(x, \varepsilon) \subset C$ par définition, puis $\lambda := \frac{\varepsilon}{\|d\|}$ qui est tel que $d = \lambda(c - x)$. Noter que le cas $d = 0$ n'est pas un problème car la conclusion aurait été triviale en prenant $c = x$ et $\lambda = 0$.

\Rightarrow : Immédiat par polarité, $(\mathbb{R}^N)^* = \{0\}$.

\Rightarrow : Supposons que $N_C(x) = \{0\}$, et supposons par l'absurde que $x \notin \text{int } C$. Alors il existe une suite $x_n \notin C$ telle que $x_n \rightarrow x$. On peut alors définir $p_n := \text{proj}_C(x_n)$ qui, par continuité de la projection (proposition I.25), converge vers $p = \text{proj}_C(x) = x$.

D'après la caractérisation du cône normal par la projection (proposition I.52) on sait que $x_n - p_n \in N_C(p_n)$. Or $N_C(p_n)$ est un cône, et $x_n \neq p_n$ (puisque $x_n \notin C$), donc on a $d_n := \frac{x_n - p_n}{\|x_n - p_n\|} \in N_C(p_n)$. Par compacité, on sait que d_n converge vers un certain d , quitte à considérer une sous-suite. La caractérisation du cône normal par les angles (proposition I.51) avec la caractérisation de la projection par les angles (proposition I.22) nous permet de passer à la limite et de montrer que $d \in N_C(x)$:

$$(\forall c \in C) \quad \langle d, c - x \rangle = \lim_n \langle d_n, c - x_n \rangle = \lim_n \frac{1}{\|x_n - p_n\|} \langle x_n - p_n, c - x_n \rangle \leq 0.$$

On a donc obtenu un $d \neq 0$ tel que $d \in N_C(x)$, ce qui est une contradiction. ■

Exemple I.54 (Cônes tangent et normal). On peut calculer explicitement les cônes tangent et normal à de nombreux convexes simples. On ne s'intéressera qu'à des points x du bord de C , puisque ces cônes sont triviaux à l'intérieur. Voir la feuille de TD correspondante pour plus de détails sur les calculs.

- Si $C = [a, b] \subset \mathbb{R}$ alors pour tout $x \in C$:

$$T_C(x) = \begin{cases} [0, +\infty[& \text{si } x = a, \\ \mathbb{R} & \text{si } a < x < b, \\]-\infty, 0] & \text{si } x = b, \end{cases} \quad \text{et} \quad N_C(x) = \begin{cases}]-\infty, 0] & \text{si } x = a, \\ \{0\} & \text{si } a < x < b, \\ [0, +\infty[& \text{si } x = b. \end{cases}$$

- Si $C = F$ est un s.e.v. et $x \in F$ alors $T_F(x) = F$ et $N_F(x) = F^\perp$.
- Si $C = \mathcal{F}$ est un s.e.a. avec $\mathcal{F} = F + v$ alors $T_{\mathcal{F}}(x) = F$ et $N_{\mathcal{F}}(x) = F^\perp$ pour tout $x \in \mathcal{F}$.
- Si $C = [Ax = b]$ alors $T_{[Ax=b]}(x) = \text{Ker } A$ et $N_{[Ax=b]}(x) = \text{Im } A$ pour toute solution x .
- Si $C = \mathbb{R}_+ a$ est une demi-droite alors pour tout $x \in C$:

$$T_C(x) = \begin{cases} \mathbb{R}_+ a & \text{si } x = 0, \\ \mathbb{R} a & \text{si } x \neq 0, \end{cases} \quad \text{et} \quad N_C(x) = \begin{cases} [\langle a, \cdot \rangle \leq 0] & \text{si } x = 0, \\ [\langle a, \cdot \rangle = 0] & \text{si } x \neq 0. \end{cases}$$

- Si $C = [\langle a, \cdot \rangle \leq b]$ est un demi-espace alors pour tout $x \in C$:

$$T_C(x) = \begin{cases} [\langle a, \cdot \rangle \leq 0] & \text{si } \langle a, x \rangle = b, \\ \mathbb{R}^N & \text{si } \langle a, x \rangle < b, \end{cases} \quad \text{et} \quad N_C(x) = \begin{cases} \mathbb{R}_+ a & \text{si } \langle a, x \rangle = b, \\ \{0\} & \text{si } \langle a, x \rangle < b. \end{cases}$$

- Si $C = \mathbb{B}(0, 1)$ est la boule unité, alors pour tout $x \in C$:

$$T_C(x) = \begin{cases} [\langle x, \cdot \rangle \leq 0] & \text{si } \|x\| = 1, \\ \mathbb{R}^N & \text{si } \|x\| < 1, \end{cases} \quad \text{et} \quad N_C(x) = \begin{cases} \mathbb{R}_+ x & \text{si } \|x\| = 1, \\ \{0\} & \text{si } \|x\| < 1. \end{cases}$$

Proposition I.55 (Cônes tangent et normal d'un produit cartésien). Soient $C \subset \mathbb{R}^N$ et $D \subset \mathbb{R}^M$ convexes fermés non vides, et soit $z = (x, y) \in C \times D$. Alors

$$T_{C \times D}(z) = T_C(x) \times T_D(y) \quad \text{et} \quad N_{C \times D}(z) = N_C(x) \times N_D(y).$$

Démonstration. Nous allons prouver le résultat pour le cône normal. Le résultat pour le cône tangent sera une conséquence directe par polarité (voir proposition I.50) de la formule du polaire pour le produit de deux cônes (voir proposition I.42). Montrons donc que $N_{C \times D}(z) = N_C(x) \times N_D(y)$. Nous allons procéder par double inclusion, et on utilisera la caractérisation du cône normal par les angles (voir proposition I.51).

\subset : Soit $(x^*, y^*) \in N_{C \times D}(x, y)$, et montrons que $(x^*, y^*) \in N_C(x) \times N_D(y)$. Par symétrie il suffit de montrer que $x^* \in N_C(x)$. On va utiliser la caractérisation du cône normal par les angles: prenons $c \in C$ quelconque, observons que $(c, y) \in C \times D$ et concluons que $\langle x^*, c - x \rangle = \langle (x^*, y^*), (c, y) - (x, y) \rangle \leq 0$.

\supset : Soit $(x^*, y^*) \in N_C(x) \times N_D(y)$ et montrons que $(x^*, y^*) \in N_{C \times D}(x, y)$. On utilise encore la caractérisation par les angles: prenons $(c, d) \in C \times D$ quelconque, et observons que $\langle (x^*, y^*), (c, d) - (x, y) \rangle = \langle x^*, c - x \rangle + \langle y^*, d - y \rangle \leq 0$. ■

Exemple I.56 (Cônes tangent et normal d'un produit cartésien). Avec la proposition I.55, on peut calculer facilement les cônes associés à l'orthant, puisque $\mathbb{R}_+^N = \prod_{i=1}^N \mathbb{R}_+$:

$$T_C(x) = \prod_{i=1}^N \begin{cases} \mathbb{R}_+ & \text{si } x_i = 0 \\ \mathbb{R} & \text{si } x_i > 0, \end{cases} \quad \text{et} \quad N_C(x) = \prod_{i=1}^N \begin{cases} \mathbb{R}_- & \text{si } x_i = 0 \\ \{0\} & \text{si } x_i > 0. \end{cases}$$

De la même manière, on peut calculer les cônes associées à des boîtes de la forme

$$\{x \in \mathbb{R}^N \mid a_i \leq x_i \leq b_i\},$$

et en particulier de la boule unité de la norme ℓ^∞ , qui est $B_\infty = \{x \in \mathbb{R}^N \mid |x_i| \leq 1\}$.

Théorème I.57 (Cônes tangent et normal d'un polyèdre). Soient $A \in \mathcal{M}_{M,N}(\mathbb{R})$, $b \in \mathbb{R}^M$, et $C = [Ax \leq b]$. On notera a_i la i -ème ligne de A . Soit $x \in C$, et on définit

$$I := \{i \in [M] \mid \langle a_i, x \rangle = b_i\},$$

que l'on appelle l'ensemble des contraintes actives. Alors

$$T_C(x) = [A_I x \leq 0] \quad \text{et} \quad N_C(x) = \text{cone}(a_i)_{i \in I},$$

où A_I est la sous-matrice dont les lignes sont $(a_i)_{i \in I}$.

Démonstration. Laissé en exercice, voir la feuille de TD correspondante. Une preuve est également disponible en annexe, voir le Lemme C.7. ■

Chapitre II

Analyse convexe non lisse

II.I Fonctions convexes s.c.i. propres

II.I.1 Fonctions à valeurs réelles étendues

Définition II.1 (Réels étendus). On considère l'ensemble des **RÉELS ÉTENDUS** $\mathbb{R} \cup \{+\infty\}$, muni des propriétés suivantes:

- pour tout $x \in \mathbb{R}$, $x < +\infty$,
- pour tout $x \in \mathbb{R} \cup \{+\infty\}$, $x + (+\infty) = +\infty$,
- pour tout $x \in]0, +\infty[$, $x(+\infty) = +\infty$.

Définition II.2 (Domaine et propriété). Soit $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$. On définit son **DOMAINE** par

$$\text{dom } f = \{x \in \mathbb{R}^N \mid f(x) < +\infty\}.$$

On dira que f est **PROPRE** si $\text{dom } f \neq \emptyset$.

Définition II.3 (Fonction indicatrice). Soit $A \subset \mathbb{R}^N$. On définit sa **FONCTION INDICATRICE** par

$$\delta_A(x) = \begin{cases} 0 & \text{si } x \in A, \\ +\infty & \text{si } x \notin A. \end{cases}$$

Il est clair que $\text{dom } \delta_A = A$.

Remarque II.4 (Optimisation sous contraintes). Comme mentionné en introduction, autoriser des fonctions à prendre des valeurs $+\infty$ va nous permettre de transformer des contraintes en fonctions. Par exemple, minimiser $f(x)$ sur la contrainte $x \in C$ est équivalent à minimiser $\hat{f}(x) = f(x) + \delta_C(x)$. Ceci nous permet de concentrer notre analyse sur les fonctions et d'éviter la dichotomie fonctions/contraintes.



FIGURE II.1 – L'épigraphe d'une fonction.

Remarque II.5 (Injection ensemble \hookrightarrow fonction). C'est un simple exercice que de vérifier que $C = D$ si et seulement si $\delta_C = \delta_D$.

Exemple II.6 (Fonctions à valeurs étendues). Voici quelques exemples typiques de fonctions à valeurs réelles étendues.

- $f(x) = \frac{1}{x}$ si $x \in]0, +\infty[$, $f(x) = +\infty$ sinon. On a $\text{dom } f = \mathbb{R}_{++}$.
- $f(x) = -\ln(b - \langle a, x \rangle)$ si $\langle a, x \rangle < b$, $f(x) = +\infty$ sinon. On a $\text{dom } f = [\langle a, x \rangle < b]$.

Proposition II.7 (Propreté d'une indicatrice). Si $A \subset \mathbb{R}^N$ alors δ_A est propre si et seulement si $A \neq \emptyset$.

Démonstration. Immédiat. ■

Proposition II.8 (Domaine de la somme). Soient $f, g : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$. Alors $\text{dom}(f + g) = \text{dom } f \cap \text{dom } g$.

Démonstration. Immédiat. ■

Exemple II.9 (Somme d'indicatrices). Si $A, B \subset \mathbb{R}^N$ alors $\delta_A + \delta_B = \delta_{A \cap B}$.

Définition II.10 (Épigraphe). Soit $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$. On définit son **ÉPIGRAPHE** par

$$\text{epi } f = \{(x, r) \in \mathbb{R}^N \times \mathbb{R} \mid f(x) \leq r\}.$$

Remarque II.11 (Épigraphe et domaine). L'épigraphe est un sous-ensemble de l'espace produit $\mathbb{R}^N \times \mathbb{R}$. Plus particulièrement, $\text{epi } f \subset \text{dom } f \times \mathbb{R}$. En fait l'épigraphe ne dépend que du comportement de f sur son domaine:

$$\begin{aligned} \text{epi } f &= \{\{x\} \times [f(x), +\infty[\mid x \in \text{dom } f\} \\ &= \{(x, r) \in \mathbb{R}^N \times \mathbb{R} \mid x \in \text{dom}(f), f(x) \leq r\}. \end{aligned}$$

Remarque II.12 (Injection fonction \rightarrow ensemble). C'est un simple exercice que de vérifier que $f = g$ si et seulement si $\text{epi } f = \text{epi } g$.

Proposition II.13 (Propreté et épigraphe). Soit $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$. Alors f est propre si et seulement si $\text{epi } f \neq \emptyset$.

Démonstration. On procède par double implication.

\Rightarrow : Si f est propre alors il existe $x \in \text{dom } f$. Donc $f(x) < +\infty$, d'où $(x, f(x)) \in \text{epi } f$.

\Leftarrow : Si $(x, r) \in \text{epi } f$ alors $f(x) \leq r \in \mathbb{R}$ d'où $x \in \text{dom } f$. ■

Proposition II.14 (Épigraphe du sup). Soient $f_i : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$. Alors $\text{epi} \left(\sup_{i \in I} f_i \right) = \bigcap_{i \in I} \text{epi } f_i$.

Démonstration. On procède par double inclusion.

\subset : Si $(x, r) \in \text{epi} \sup f_i$ alors $\sup f_i(x) \leq r$. Par définition du sup cela veut dire que $f_i(x) \leq r$ pour tout i , et donc que $(x, r) \in \text{epi } f_i$ pour tout i .

\supset : Si $(x, r) \in \bigcap_{i \in I} \text{epi } f_i$ alors pour tout i nous avons $(x, r) \in \text{epi } f_i$, c'est-à-dire $f_i(x) \leq r$. En passant cette inégalité au sup, nous voyons bien que $\sup f_i(x) \leq r$ et donc que $(x, r) \in \text{epi} \sup f_i$. ■

Proposition II.15 (Intérieur de l'épigraphe). Soit $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$.

1) Si $(x, r) \in \text{int epi } f$, alors $f(x) < r$.

2) Si f est continue en x , alors $(x, r) \in \text{int epi } f$ si et seulement si $f(x) < r$.

Démonstration. C'est un bon petit exercice de topologie que vous pourrez trouver dans la feuille de TD. Sinon, une preuve est disponible en Annexe, voir le Lemme C.8. ■

II.I.2 Fonctions semi-continues inférieurement

Définition II.16 (Fonction sci). On dit que $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$ est **SEMI-CONTINUE INFÉRIEUREMENT** (sci) en $x \in \mathbb{R}^N$ si

$$(\forall x_n \rightarrow x) \quad f(x) \leq \liminf_{n \rightarrow +\infty} f(x_n).$$

On dira que f est sci si f est sci en tout point $x \in \mathbb{R}^N$.

Remarque II.17 (Continuité et sci). Il est immédiat de voir que f est continue en x si et seulement si f et $-f$ sont sci en x . Le caractère sci permet de détecter quelle est la nature éventuelle de la discontinuité en x . Dans la suite on notera $\text{cont } f$ l'ensemble des points où f est continue.

Proposition II.18 (Sci sur le domaine). Soit $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$. Alors f est sci si et seulement si pour tout $x \in \text{adh dom } f$, f est sci en x .



FIGURE II.2 – Gauche: une fonction sci. Droite: une fonction pas sci.

Démonstration. D'après la définition de sci, il nous suffit de montrer que f est toujours sci en les points $x \notin \text{adh dom } f$. Considérons donc x qui appartient à $\text{adh dom } f^c$; cet ensemble étant ouvert, il existe un voisinage $\mathbb{B}(x, \varepsilon) \subset \text{adh dom } f^c$. Donc pour tout $y \in \mathbb{B}(x, \varepsilon)$, nous avons $f(y) = f(x) = +\infty$. Donc trivialement

$$+\infty = f(x) = \liminf_{x_n \rightarrow x} f(x_n) = +\infty.$$

■

Proposition II.19 (Sci d'une indicatrice). Soit $A \subset \mathbb{R}^N$. Alors δ_A est sci si et seulement si A est fermé.

Démonstration. On procède par double implication.

\Rightarrow : Montrons que A est fermé : pour cela considérons $a_n \in A$ qui converge vers un certain a , et montrons que $a \in A$. Pour cela, il suffit de montrer que $\delta_A(a) = 0$. Or par le caractère sci de δ_A , et puisque $\delta_A(a_n) = 0$ par hypothèse, nous avons

$$\delta_A(a) \leq \liminf_{n \rightarrow +\infty} \delta_A(a_n) = 0.$$

Puisque δ_A est une fonction positive, on conclut.

\Leftarrow : Montrons que δ_A est sci. D'après la proposition II.18 il suffit de montrer qu'elle est sci sur $\text{adh dom } \delta_A = \text{adh } A = A$. Soit donc $a \in A$ et montrons que δ_A est sci en a . Soit donc une suite $x_n \rightarrow a$, et montrons que $0 = \delta_A(a) \leq \liminf_{n \rightarrow +\infty} \delta_A(x_n)$. Or δ_A est positive, donc cette inégalité est toujours vraie. ■

Exemple II.20 (Fonction sci).

- Soit $f(x) = \frac{1}{x} + \delta_{]0, +\infty[}$. Clairement f est continue sur $]0, +\infty[$, donc elle y est sci. De plus en 0 nous voyons que $f(0) = +\infty$ tandis que $\lim_{x \rightarrow 0} f(x) = +\infty$, donc f est bien sci en 0. On en déduit que f est sci.
- Soit $f(x) = x^2 + \delta_{]-1, 1[}$. Alors f n'est pas sci en 1. En effet, $f(1) = +\infty$, mais $\liminf_{x \rightarrow 1} f(x) = 1$.

Proposition II.21 (Sci via l'épigraphe). Soit $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$. Alors f est sci si et seulement si $\text{epi } f$ est fermé.

Démonstration. On procède par double implication.

\Rightarrow : Montrons que $\text{epi } f$ est fermé. Soit $(x_n, r_n) \in \text{epi } f$ une suite qui converge vers un (x, r) , et montrons que $(x, r) \in \text{epi } f$. Par hypothèse nous savons que $f(x_n) \leq r_n$, et puisque f est sci on peut conclure que

$$f(x) \leq \liminf_{n \rightarrow +\infty} f(x_n) \leq \liminf_{n \rightarrow +\infty} r_n = r.$$

\Leftarrow : Montrons que f est sci. D'après la proposition II.18, il suffit de montrer que elle est sci en $x \in \text{adh dom } f$. Considérons une suite x_n quelconque qui converge vers x , et montrons que $f(x) \leq \liminf_{n \rightarrow \infty} f(x_n)$. On pose $r = \liminf_{n \rightarrow \infty} f(x_n)$, et on considère une sous-suite x_{n_k} telle que $\lim_{k \rightarrow \infty} f(x_{n_k}) = r$. Si on pose $r_k = f(x_{n_k})$, on a trivialement que $(x_{n_k}, r_k) \in \text{epi } f$ et par construction cette suite converge vers (x, r) . Or $\text{epi } f$ est fermé donc $(x, r) \in \text{epi } f$, ce qui veut dire que $f(x) \leq r$, qui est bien ce que l'on voulait montrer. ■

Proposition II.22 (Somme de fonctions sci). Soient $f, g : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$. Si f et g sont sci alors $f + g$ est sci.

Démonstration. C'est essentiellement une conséquence de l'inégalité de la liminf pour la somme:

$$(f + g)(x) = f(x) + g(x) \leq \liminf_{x_n \rightarrow x} f(x_n) + \liminf_{x_n \rightarrow x} g(x_n) \leq \liminf(f + g)(x_n).$$

■

Proposition II.23 (Sup de fonctions sci). Soient $f_i : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$ des fonctions sci. Alors $\sup_{i \in I} f_i$ est sci.

Démonstration. Si les f_i sont sci alors les $\text{epi } f_i$ sont fermés (proposition II.21) donc leur intersection est fermée, or $\text{epi sup } f_i = \cap \text{epi } f_i$ (proposition II.14), donc on conclut que $\sup f_i$ est sci. ■

Proposition II.24 (Sous-niveaux d'une fonction sci). Soit $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$. Alors il y a équivalence entre:

- 1) f est sci;
- 2) pour tout $r \in \mathbb{R}$ le sous-niveau $[f \leq r]$ est fermé;
- 3) pour tout $r \in \mathbb{R}$ et tout $\delta > 0$ le sous-niveau local $\delta \mathbb{B} \cap [f \leq r]$ est fermé.

Démonstration.

1) \Rightarrow 2) : Considérons donc une suite $x_n \in [f \leq r]$ qui converge vers un certain x , et montrons que $x \in [f \leq r]$. Or ceci est immédiat par le caractère sci de f : $f(x) \leq \liminf_{n \rightarrow +\infty} f(x_n) \leq r$.

2) \Rightarrow 3) : Trivial.

$3) \Rightarrow 1)$: Montrons que f est sci. Pour cela considérons $x \in \mathbb{R}^N$ et $x_n \rightarrow x$, et montrons que $f(x) \leq \liminf_{n \rightarrow \infty} f(x_n)$. Quitte à prendre une sous-suite, on suppose que cette liminf est une limite. Si cette limite vaut $+\infty$ la conclusion est triviale. Si elle est finie on la note $r \in \mathbb{R}$, et on observe que pour tout $\varepsilon > 0$ on a $x_n \in [f \leq r + \varepsilon]$ à partir d'un certain rang. Par ailleurs, x_n étant bornée (car convergente) on sait qu'il existe $\delta > 0$ tel que $x_n \in \delta \mathbb{B}$. Par fermeture de $\delta \mathbb{B} \cap [f \leq r + \varepsilon]$, on en déduit que $x \in \delta \mathbb{B} \cap [f \leq r + \varepsilon]$. Ceci étant vrai pour tout $\varepsilon > 0$, on conclut que $f(x) \leq r$. ■

Théorème II.25 (de semi-Weierstrass). Soient $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$ sci propre et $C \subset \mathbb{R}^N$ compact tel que $C \cap \text{dom } f \neq \emptyset$. Alors il existe un minimiseur de f sur C , c'est-à-dire

$$(\exists \bar{x} \in C)(\forall x \in C) \quad f(\bar{x}) \leq f(x).$$

Démonstration. Considérons $\inf_C f$ l'infimum de f sur C . Puisque f est propre et $C \cap \text{dom } f \neq \emptyset$, nous savons que $\inf_C f < +\infty$. Par définition de l'inf, nous pouvons trouver une suite $x_n \in C$ telle que $\inf_C f \leq f(x_n) \leq \inf_C f + \frac{1}{n}$. En particulier, nous voyons que $f(x_n) \rightarrow \inf_C f$. Or C est compact, donc il existe une sous-suite convergente: $x_{n_k} \rightarrow \bar{x} \in C$. On conclut alors avec le caractère sci de f que

$$f(\bar{x}) \leq \liminf_{k \rightarrow +\infty} f(x_{n_k}) = \lim_{k \rightarrow +\infty} f(x_{n_k}) = \inf_C f.$$

■

II.I.3 Fonctions coercives et existence de minimiseurs

Définition II.26 (Fonction coercive). On dit que $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$ est **COERCIVE** si

$$\lim_{\|x\| \rightarrow +\infty} f(x) = +\infty.$$

Exemple II.27 (Fonction coercive).

- x^2 est coercive.
- e^x n'est pas coercive.
- δ_A est coercive si et seulement si A est borné.

Proposition II.28 (Sous-niveaux d'une fonction coercive). Soit $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$. Alors f est coercive si et seulement si pour tout $r \in \mathbb{R}^N$ le sous-niveau $[f(x) \leq r]$ est borné.

Démonstration. On procède par double implication.

\Rightarrow : Supposons que f est coercive et raisonnons par l'absurde: supposons qu'il existe $r \in \mathbb{R}$ que le sous-niveau $[f(x) \leq r]$ ne soit pas borné. Alors il existe $\|x_n\| \rightarrow +\infty$ tel que $f(x_n) \leq r$. Mais par définition de la coercivité on a $\lim f(x_n) = +\infty$, on a donc une contradiction.

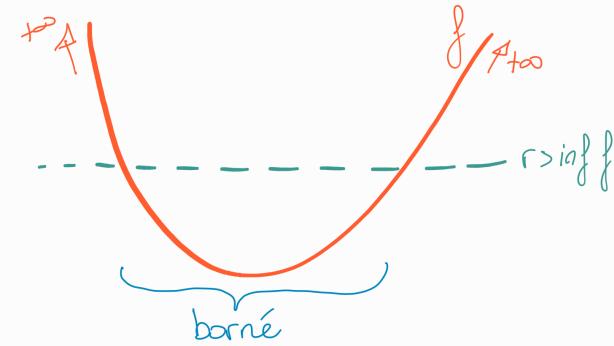


FIGURE II.3 – Le domaine d'une fonction coercive peut être coupé en deux: un compact au voisinage des solutions où un minimiseur va exister, par compacité; et son complémentaire où la fonction tend vers l'infini.

\Leftarrow : Soit x_n une suite telle que $\|x_n\| \rightarrow +\infty$, montrons que $f(x_n) \rightarrow +\infty$. Pour cela, prenons $M \in \mathbb{R}$ quelconque et montrons que $f(x_n) > M$ à partir d'un certain rang. Or on sait que $[f(x) \leq M]$ est borné, et que x_n diverge, donc à partir d'un certain rang on ne peut plus avoir $x_n \in [f(x) \leq M]$. Autrement dit $f(x_n) > M$. ■

Théorème II.29 (des fonctions coercives). Soit $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$ coercive sci propre. Alors il existe un minimiseur de f .

Démonstration. Puisque f est propre il existe $\hat{x} \in \text{dom } f$, et $\hat{r} := f(\hat{x}) \in \mathbb{R}$. On peut donc définir $C := [f(x) \leq \hat{r}]$ qui contient \hat{x} . Puisque f est sci on sait que C est fermé (voir proposition II.24). Puisque f est coercive on sait que C est borné (voir proposition II.28). Donc C est compact, d'intersection non vide avec $\text{dom } f$, on peut donc utiliser le théorème II.25 qui nous donne un minimiseur \bar{x} de f sur C :

$$f(x) \leq \hat{r} \Rightarrow f(\bar{x}) \leq f(x).$$

D'autre part, si x est tel que $f(x) > \hat{r}$, et sachant que $\hat{r} = f(\hat{x}) \geq f(\bar{x})$, on en déduit que $f(x) > f(\bar{x})$. Dans tous les cas on voit que $f(\bar{x}) \leq f(x)$. ■

II.I.4 Fonctions convexes

Définition II.30 (Fonction convexe). On dit que $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$ est **CONVEXE** si

$$(\forall x, y \in \mathbb{R}^N)(\forall t \in]0, 1[) \quad f((1-t)x + ty) \leq (1-t)f(x) + tf(y).$$

On notera $\Gamma_0(\mathbb{R}^N)$ l'ensemble des fonctions convexes sci propres.

Exemple II.31 (Fonctions convexes).

- δ_C est convexe si et seulement si C est convexe.
- $\|x\|^p$ est convexe pour tout $p \geq 0$.

Proposition II.32 (Inégalité de Jensen). Soit $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$ convexe. Soient $\lambda \in \Delta^M$, et $x_1, \dots, x_M \in \mathbb{R}^N$. Alors $f(\sum_{i=1}^M \lambda_i x_i) \leq \sum_{i=1}^M \lambda_i f(x_i)$.

Démonstration. Voir TD, c'est exactement le même argument que pour la preuve de la proposition I.4. ■

Proposition II.33 (Somme, produit, composée de fonctions convexes).

- 1) La somme finie de fonctions convexes est convexe.
- 2) Si f est convexe et $\lambda > 0$ alors λf est convexe.
- 3) Si f est convexe et $b \in \mathbb{R}^N$ alors $f(\cdot - b)$ est convexe.
- 4) Si f est convexe et A est linéaire, alors $f \circ A$ est convexe.

Démonstration.

- 1) Immédiat par définition.
- 2) Idem.
- 3) Idem.
- 4) On a

$$(f \circ A)((1-t)x + ty) = f((1-t)Ax + tAy) \leq (1-t)f(Ax) + tf(Ay).$$

■

Remarque II.34 (Multiplier une fonction par 0). Il est à noter que l'on a évité de parler du cas $\lambda = 0$ pour λf . Si f ne prend que des valeurs finies, il n'y a pas de problème et on peut écrire $0f = 0$. Mais si f prend des valeurs $+\infty$, alors on est confronté au problème que le produit $0 \cdot +\infty$ n'est pas défini. On pourrait utiliser un argument de continuité et dire que $0 \cdot f = \delta_{\text{dom } f}$. En effet pour $x \in \text{dom } f$ on a trivialement $0f(x) = 0$, tandis que pour $x \notin \text{dom } f$ on peut écrire $\lim_{\varepsilon \rightarrow 0} \varepsilon f(x) = \lim_{\varepsilon \rightarrow 0} +\infty = +\infty$. Néanmoins on ne s'aventurera pas dans ces eaux, et on fera toujours attention à ne pas faire ce genre de produit, et à régler d'éventuels problèmes au cas par cas.

Remarque II.35 (Cône des fonctions convexes). La propriété précédente implique que l'ensemble des fonctions convexes finies forme un cône parmi l'ensemble des fonctions de \mathbb{R}^N dans $\overline{\mathbb{R}}$.

Exemple II.36 (Fonctions convexes).

- 1) Si f et g sont convexes et A linéaire, alors $f(x) + g(Ax)$ est convexe.

- 2) Si f est convexe, A est linéaire et $b \in \mathbb{R}^M$, alors $f(x) + \delta_{[Ax=b]}(x)$ est convexe. En effet il suffit de prendre $g = \delta_0(\cdot - b) = \delta_b$.
- 3) Si f est convexe, A est linéaire et $b \in \mathbb{R}^M$, alors $f(x) + \delta_{[Ax \leq b]}(x)$ est convexe. En effet il suffit de prendre $g = \delta_{\mathbb{R}_-^M}(\cdot - b) = \delta_{b+\mathbb{R}_-^M}$.
- 4) $f(x) = \frac{1}{2}\|Ax - b\|^2$ est convexe.
- 5) $f(x) = \langle Ax, x \rangle$ est convexe si $A \succeq 0$.

Proposition II.37 ($\Gamma_0(\mathbb{R}^N)$ est stable par somme). Soient $f, g \in \Gamma_0(\mathbb{R}^N)$ telles que $\text{dom } f \cap \text{dom } g \neq \emptyset$. Alors $f + g \in \Gamma_0(\mathbb{R}^N)$.

Démonstration. La convexité et le caractère sci sont stables par somme (propositions II.33 et II.22). La somme de deux fonctions propres est propre pourvu que leur domaines soient d'intersection non vide (proposition II.8). ■

Proposition II.38 (Convexité et épigraphe). Soit $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$. Alors f est convexe si et seulement si $\text{epi } f$ est convexe.

Démonstration. On procède par double implication.

\Rightarrow : Montrons que $\text{epi } f$ est convexe. Soit $(x, r), (x', r') \in \text{epi } f$, $t \in]0, 1[$ et montrons que $((1-t)x + tx', (1-t)r + tr') \in \text{epi } f$. En effet on a

$$f((1-t)x + tx') \leq (1-t)f(x) + tf(x') \leq (1-t)r + tr'.$$

\Leftarrow : Montrons que f est convexe. Soit $x, x' \in \mathbb{R}^N$ et $t \in]0, 1[$. Si x ou $x' \notin \text{dom } f$ alors l'inégalité de convexité est triviale. Si $x, x' \in \text{dom } f$ alors on peut dire que $(x, r), (x', r') \in \text{epi } f$ avec $r = f(x)$ et $r' = f(x')$. On en déduit que $((1-t)x + tx', (1-t)r + tr') \in \text{epi } f$, qui nous donne exactement l'inégalité de convexité pour f . ■

Proposition II.39 (Sup de fonctions convexes). Soient $f_i : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$ convexes. Alors $\sup_{i \in I} f_i$ est convexe.

Démonstration. L'épigraphe de $\sup_i f_i$ est l'intersection des $\text{epi } f_i$ qui sont convexes (on utilise les propositions II.14 et II.38), donc lui-même convexe. D'où le résultat. ■

Exemple II.40 (Fonction support). Soit $A \subset \mathbb{R}^N$ non vide. On définit sa **fonction support**¹ comme étant

$$\sigma_A(x) = \sup_{a \in A} \langle x, a \rangle.$$

C'est un sup de fonctions linéaires (donc convexes sci), donc σ_A est convexe sci. Elle est par ailleurs propre puisque $\sigma_A(0) = 0$. On en déduit que $\sigma_A \in \Gamma_0(\mathbb{R}^N)$.

¹On parle parfois de fonction d'appui.

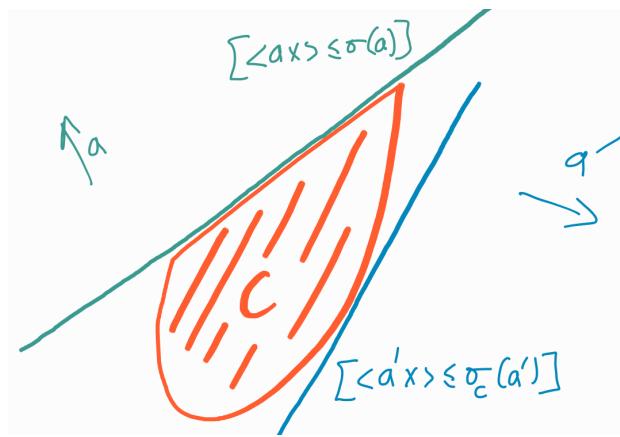


FIGURE II.4 – La fonction support définit les hyperplans asymptotiquement tangents.

Remarque II.41 (Interprétation géométrique de la fonction support). Soit \$C\$ un convexe fermé non vide, et soit \$a \in \mathbb{R}^N\$ une direction. On dit que \$a\$ *supporte asymptotiquement* \$C\$ si on peut trouver un demi-espace de la forme \$[\langle a, x \rangle \leq b]\$ qui soit asymptotiquement tangent à \$C\$: c'est-à-dire que ce demi-espace contient \$C\$ tout entier, et est le plus petit possible. Il est facile de voir que l'inclusion \$C \subset [\langle a, x \rangle \leq b]\$ est vraie si et seulement si \$b \geq \sigma_C(a)\$, ceci n'étant possible que lorsque \$\sigma_C(a) < +\infty\$. D'autre part le demi-espace est d'autant plus petit que l'on prend \$b\$ petit, donc le plus petit \$b\$ que l'on peut prendre est clairement \$b = \sigma_C(a)\$. En conclusion, les directions qui supportent asymptotiquement \$C\$ sont exactement les \$a \in \text{dom } \sigma_C\$, et le demi-espace asymptotiquement tangent dans la direction \$a\$ est exactement \$[\langle a, x \rangle \leq \sigma_C(a)]\$. Noter que l'intersection entre l'hyperplan tangent \$[\langle a, x \rangle = \sigma_C(a)]\$ et \$C\$ peut être vide: si \$C\$ est l'épigraphe de la fonction \$\frac{1}{x}\$ et \$d = (0, -1)\$, alors l'hyperplan tangent dans cette direction est \$\mathbb{R} \times \{0\}\$ qui n'intersecte jamais \$C\$. On verra dans la remarque II.107 comment trouver des directions de support exactes, pour lesquelles cette intersection est non vide.

Proposition II.42 (Sous-niveaux d'une fonction convexe). Soient \$f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}\$ convexe, et \$r \in \mathbb{R}\$. Alors son sous-niveau \$[f(x) \leq r]\$ est convexe.

Démonstration. Immédiat par définition de la convexité. ■

Lemme II.43 (Topologie des fonctions convexes). Soient \$f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}\$ convexe et propre, et \$\bar{x} \in \text{dom } f\$. Alors ces propriétés sont équivalentes:

- \$f\$ est majorée au voisinage de \$\bar{x}\$,
- \$f\$ est continue au voisinage de \$\bar{x}\$,
- \$f\$ est Lipschitzienne au voisinage de \$\bar{x}\$.

Démonstration. C'est un exercice un peu technique mais néanmoins abordable que l'on

peut trouver dans la feuille de TD correspondante. Sinon, voir le Lemme C.9 en Annexe. ■

Théorème II.44 (Continuité des fonctions convexes). Soit $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$ convexe. Alors f est continue sur l'intérieur de son domaine: $\text{cont } f = \text{int dom } f$.

Démonstration. On va procéder par double inclusion.

\subset : Soit \bar{x} où f est continue. Soit $\varepsilon > 0$ quelconque, alors il existe $\mathbb{B}(\bar{x}, \delta)$ sur laquelle f prend des valeurs dans $[f(\bar{x}) - \varepsilon, f(\bar{x}) + \varepsilon]$. Donc f est bornée sur ce voisinage.

\supset : Soit $\bar{x} \in \text{int dom } f$, et montrons que f y est continue. Au vu du lemme II.43, il suffit de montrer que f est localement majorée. Par hypothèse il existe $\delta > 0$ tel que $\mathbb{B}(\bar{x}, \delta) \subset \text{dom } f$. Puisque nous sommes en dimension finie et que les normes y sont équivalentes, on peut supposer que c'est la boule pour la norme 1, c'est-à-dire $\mathbb{B}_1(\bar{x}, \delta)$. Or $\mathbb{B}_1(\bar{x}, \delta) = \bar{x} + \delta \mathbb{B}_1$ où cette boule unité est exactement $\text{co}(e_1, \dots, e_N, -e_1, \dots, -e_N)$ (voir exemple I.16). Donc $\mathbb{B}_1(\bar{x}, \delta) = \text{co}(x_1, \dots, x_{2N})$ où $x_i = \bar{x} \pm \delta e_i$. Dans ce cas pour tout $x = \sum_i \lambda_i x_i \in \mathbb{B}_1(\bar{x}, \delta)$ nous pouvons écrire à l'aide de l'inégalité de Jensen (voir proposition II.32)

$$f(x) = f\left(\sum_i \lambda_i x_i\right) \leq \sum_i \lambda_i f(x_i) \leq \max_i f(x_i) \sum_i \lambda_i = \max_i f(x_i).$$

Or chaque $x_i \in \mathbb{B}_1(\bar{x}, \delta) \subset \text{dom } f$, donc nous avons montré que f est majorée sur cette boule. ■

Corollaire II.45 (Continuité des fonctions convexes finies). Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$ une fonction convexe finie, au sens où $\text{dom } f = \mathbb{R}^N$. Alors f est continue sur \mathbb{R}^N .

Démonstration. On applique le théorème précédent avec $\text{cont}(f) = \text{int } \mathbb{R}^N = \mathbb{R}^N$. ■

Théorème II.46 (Les fonctions convexes sont le sup de fonctions affines). Soit $f \in \Gamma_0(\mathbb{R}^N)$. Alors f est le supremum de fonctions affines: il existe une famille non vide $(h_i)_{i \in I}$ de fonctions affines $h_i : \mathbb{R}^N \rightarrow \mathbb{R}$ telles que $f = \sup_i h_i$.

Démonstration. Soit $\bar{x} \in \mathbb{R}^N$ fixé. Soit $\bar{r} < f(\bar{x})$ quelconque dans \mathbb{R} . Nous allons montrer que

$$\text{il existe } h_{\bar{x}, \bar{r}} : \mathbb{R}^N \rightarrow \mathbb{R} \text{ affine telle que } \bar{r} < h_{\bar{x}, \bar{r}}(\bar{x}) \quad \text{et} \quad h_{\bar{x}, \bar{r}} \leq f. \quad (\text{II.1})$$

Ainsi nous pourrons conclure en considérant la famille $\mathcal{H} = \{h_{\bar{x}, \bar{r}} \mid \bar{x} \in \mathbb{R}^N, \bar{r} < f(\bar{x})\}$. En effet puisque $h_{\bar{x}, \bar{r}} \leq f$ nous aurons $\sup_{h \in \mathcal{H}} h \leq f$. D'autre part nous aurons pour tout $\bar{x} \in \mathbb{R}^N$ que

$$f(\bar{x}) = \lim_{\bar{r} \uparrow f(\bar{x})} \bar{r} \leq \lim_{\bar{r} \uparrow f(\bar{x})} h_{\bar{x}, \bar{r}}(\bar{x}) \leq \sup_{h \in \mathcal{H}} h(\bar{x}).$$

Le reste de la preuve consiste donc à prouver (II.1) pour \bar{x} et \bar{r} donnés.

Considérons $C := \text{epi } f$. Puisque f est convexe sci propre, nous savons que C est convexe fermé non vide. Par ailleurs $(\bar{x}, \bar{r}) \notin \text{epi } f$ par définition de \bar{r} . Donc on peut faire

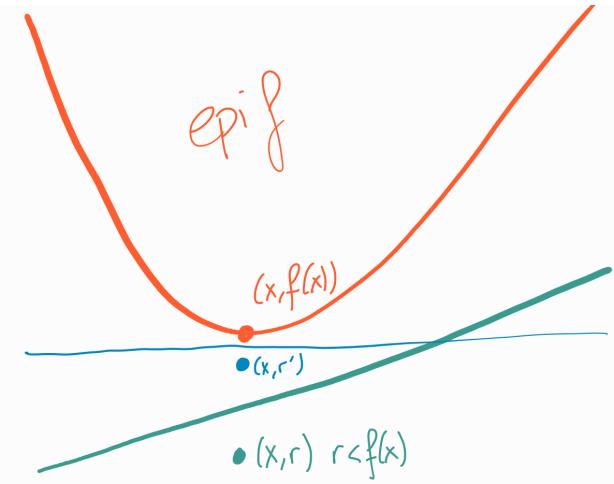


FIGURE II.5 – Avec le théorème de séparation, on peut trouver des minorantes affines arbitrairement proches de l'épigraphe en séparant ce dernier de points situés sous le graphe.

appel au théorème I.26 de séparation forte de Hahn-Banach, pour séparer C et le singleton (\bar{x}, \bar{r}) dans $\mathbb{R}^N \times \mathbb{R}$. Il existe donc un $\hat{\alpha} \in \mathbb{R}^N \times \mathbb{R}$ non nul, que l'on notera $\hat{\alpha} = (\alpha, \beta)$, et un $\varepsilon > 0$ tels que

$$(\forall (x, r) \in \text{epi } f) \quad \langle (\alpha, \beta), (x, r) \rangle \geq \varepsilon + \langle (\alpha, \beta), (\bar{x}, \bar{r}) \rangle.$$

Autrement dit

$$(\forall (x, r) \in \text{epi } f) \quad \langle \alpha, x \rangle + \beta r \geq \varepsilon + \langle \alpha, \bar{x} \rangle + \beta \bar{r}. \quad (\text{II.2})$$

On distingue maintenant trois cas, en fonction du signe de β .

- Cas $\beta < 0$: Ce cas est impossible. En effet si on avait $\beta < 0$ alors en prenant n'importe quel $x \in \text{dom } f$ et une suite $r_n \rightarrow +\infty$, on aura à partir d'un certain rang $r_n \geq f(x)$ et donc $(x, r_n) \in \text{epi } f$. On pourrait donc passer à la limite dans (II.2) pour obtenir $-\infty \geq \varepsilon + \langle \alpha, \bar{x} \rangle + \beta \bar{r}$, ce qui est impossible.
- Cas $\beta > 0$: Quitte à diviser (II.2) par β et renommer α et ε , on peut sans perte de généralité supposer que $\beta = 1$. Nous avons donc (après réorganisation des termes)

$$(\forall (x, r) \in \text{epi } f) \quad r \geq \langle \alpha, \bar{x} - x \rangle + \bar{r} + \varepsilon. \quad (\text{II.3})$$

On décide de poser $h_{\bar{x}, \bar{r}}(x) := \langle \alpha, \bar{x} - x \rangle + \bar{r} + \varepsilon$, et nous allons vérifier que (II.1) est vrai. Pour commencer nous voyons que $h_{\bar{x}, \bar{r}}(\bar{x}) = \bar{r} + \varepsilon > \bar{r}$, qui est la première propriété désirée. Pour la seconde, prenons $x \in \mathbb{R}^N$ quelconque. Si $x \notin \text{dom } f$ alors trivialement $h_{\bar{x}, \bar{r}}(x) < +\infty = f(x)$. Si $x \in \text{dom } f$, alors $(x, f(x)) \in \text{epi } f$ donc (II.3) nous dit exactement que $h_{\bar{x}, \bar{r}}(x) \leq f(x)$.

- Cas $\beta = 0$: Dans ce cas (II.2) devient

$$(\forall (x, r) \in \text{epi } f) \quad \langle \alpha, x \rangle \geq \varepsilon + \langle \alpha, \bar{x} \rangle.$$

Puisque cette inégalité ne dépend plus de r , et puisque $\text{epi } f$ ne dépend que du domaine (revoir remarque II.11), ceci est équivalent à

$$(\forall x \in \text{dom } f) \quad 0 \geq \varepsilon + \langle \alpha, \bar{x} - x \rangle. \quad (\text{II.4})$$

Observons que \bar{x} ne peut pas être dans $\text{dom } f$: si $\bar{x} \in \text{dom } f$ on obtiendrait de (II.4) que $0 \geq \varepsilon$, ce qui serait une contradiction. On décide de poser $h(x) = \varepsilon + \langle \alpha, \bar{x} - x \rangle$, qui est telle que $h(x) \leq 0$ pour $x \in \text{dom } f$, et $h(\bar{x}) = \varepsilon > 0$. Puisque f est propre, il existe $\hat{x} \in \text{dom } f$, et au vu des cas précédents il existe également une fonction affine $\hat{h} := h_{\hat{x}, f(\hat{x})-1}$ telle que $\hat{h} \leq f$. On pose maintenant $h_n := \hat{h} + nh$, telle que :

- si $x \in \text{dom } f$: $h_n(x) = \hat{h}(x) + nh(x) \leq f(x) + 0 = f(x)$,
- si $x \notin \text{dom } f$: $h_n(x) < +\infty = f(x)$,
- si $x = \bar{x}$: $h_n(\bar{x}) = \hat{h}(\bar{x}) + nh(\bar{x}) = \hat{h}(\bar{x}) + n\varepsilon \rightarrow +\infty$.

Si on prend n assez grand pour que $\hat{h}(\bar{x}) + n\varepsilon > \bar{r}$, on voit que l'on a bien trouvé une fonction affine h_n telle que $h_n \leq f$ et $h_n(\bar{x}) > \bar{r}$. Donc (II.1) est vérifiée avec $h_{\bar{x}, \bar{r}} := h_n$, ce qui conclut la preuve.



Corollaire II.47 (Existence de minorante affine). Soit $f \in \Gamma_0(\mathbb{R}^N)$. Alors il existe $h : \mathbb{R}^N \rightarrow \mathbb{R}$ affine telle que $f \geq h$.

Démonstration. Immédiat d'après le théorème II.46.



Remarque II.48 (Fonctions polyédrales). Toute fonction convexe peut s'écrire comme supremum possiblement infini de fonctions affines. Si on se limite à un nombre fini de fonctions affines, on obtient les fonctions convexes *affines par morceaux*, qui s'écrivent

$$f(x) = \max_{i=1, \dots, p} \langle \alpha_i, x \rangle - \beta_i, \quad \alpha_i \in \mathbb{R}^N, \beta_i \in \mathbb{R}.$$

La norme L^1 en est un cas particulier. Plus généralement il est également intéressant de considérer la classe des fonctions *polyédrales*: ce sont les fonctions affines par morceaux auxquelles on ajoute une contrainte polyédrale. Ces fonctions bénéficient de propriétés très sympathiques, dont la lectrice curieuse pourra prendre connaissance dans l'annexe B.I.4.

II.I.5 Fonctions fortement convexes

Définition II.49 (Fonction fortement convexe). Soient $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$ et $\mu > 0$. On dit que f est μ -FORTEMENT CONVEXE si

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

On notera $\Gamma_\mu(\mathbb{R}^N)$ l'ensemble des fonctions μ -fortement convexes sci propres. On dira simplement que f est fortement convexe si il existe $\mu > 0$ tel que f soit μ -fortement convexe.

Proposition II.50 (Caractérisation des fonctions fortement convexes par la norme). Soient $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$ et $\mu > 0$. Alors f est μ -fortement convexe si et seulement si il existe $g : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$ convexe telle que $f = g + \frac{\mu}{2} \|\cdot\|^2$.

Démonstration. Voir TD, ou le Lemme C.10 en Annexe. ■

Exemple II.51 (Fonction fortement convexe).

- $f(x) = \frac{\mu}{2} \|\cdot\|^2$ est μ -fortement convexe;
- $f(x) = \delta_a(x)$ pour $a \in \mathbb{R}^N$ est μ -fortement convexe pour tout $\mu > 0$.
- Si A est une matrice symétrique, alors $f(x) = \frac{1}{2} \langle Ax, x \rangle$ est fortement convexe si et seulement si $A \succ 0$. Dans ce cas $\mu = \lambda_{\min}(A)$.

Proposition II.52 (Somme, produit, composée de fonctions fortement convexes).

- 1) Si f est μ -fortement convexe et g est convexe alors $f + g$ est μ -fortement convexe.
- 2) Si f est μ -fortement convexe et $\lambda > 0$ alors λf est $\mu\lambda$ -fortement convexe.
- 3) Si f est μ -fortement convexe et $b \in \mathbb{R}^N$ alors $f(\cdot - b)$ est μ -fortement convexe.
- 4) Si f est μ -fortement convexe et A est inversible, alors $f \circ A$ est $\mu\lambda_{\min}(A^\top A)$ -fortement convexe.

Démonstration. 1) Immédiat via la proposition II.50.

2) Idem.

3) Idem.

4) On utilise encore la proposition II.50: si $f = g + \frac{1}{2} \|\cdot\|^2$ alors

$$f(x - b) = g(x - b) + \frac{\mu}{2} \|x - b\|^2 = g(x - b) + \frac{\mu}{2} \|b\|^2 + \langle b, x \rangle + \frac{\mu}{2} \|x\|^2,$$

et on conclut en voyant que $g(x - b) + \frac{\mu}{2} \|b\|^2 + \langle b, x \rangle$ est convexe. ■

Théorème II.53 (Unique minimiseur des fonctions fortement convexes). Soit $f \in \Gamma_\mu(\mathbb{R}^N)$. Alors f admet un unique minimiseur.

Démonstration. Commençons par prouver l'existence, à l'aide du théorème II.29. Il suffit donc de montrer que f est coercive. Pour cela on utilise la proposition II.50 pour écrire $f(x) = g(x) + \frac{\mu}{2}\|x\|^2$. Puis on utilise le fait que $g \in \Gamma_0(\mathbb{R}^N)$ pour obtenir une minorante affine $h(x) = \langle \alpha, x \rangle + \beta$ telle que $g \geq h$. Alors

$$f(x) \geq \frac{\mu}{2}\|x\|^2 + \langle \alpha, x \rangle + \beta \geq \frac{\mu}{2}\|x\|^2 - \|\alpha\|\|x\| + \beta \xrightarrow{\|x\| \rightarrow +\infty} +\infty.$$

Prouvons maintenant l'unicité. Soient $x, y \in \operatorname{argmin} f$, et montrons que $x = y$. D'après la définition de forte convexité, nous avons avec par exemple $t = \frac{1}{2}$ que

$$f\left(\frac{x+y}{2}\right) \leq \frac{1}{2}f(x) + \frac{1}{2}f(y) - \mu\frac{1}{8}\|y-x\|^2 = \inf f - \mu\frac{1}{8}\|y-x\|^2.$$

Si x et y sont différents alors nous avons trouvé un point $\frac{x+y}{2}$ en lequel f est strictement inférieure à $\inf f$: contradiction! ■

II.II Sous-différentiel d'une fonction convexe

II.II.1 Sous-différentiel

Définition II.54 (Sous-gradient). Soient $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$ et $x \in \mathbb{R}^N$. On dit que x^* est un **Sous-Gradient** de f en x si

$$(\forall y \in \mathbb{R}^N) \quad f(y) - f(x) - \langle x^*, y - x \rangle \geq 0.$$

Exemple II.55 (Sous-gradient). On voit que x^* est un sous-gradient de f en x si et seulement si la fonction affine

$$h = \langle x^*, \cdot - x \rangle + f(x)$$

est une *minorante exacte* de f en x , c'est-à-dire que $h \leq f$ et $h(x) = f(x)$. On retrouve l'idée de « tangente à la courbe » qui est à la base de la notion de dérivée. La différence essentielle étant qu'il peut exister plus d'une telle minorante exacte lorsque f n'est pas différentiable. Noter que comme pour le gradient, on dispose d'une version « petit o », plus locale, de la définition de sous-gradient!

Lemme II.56 (Sous-gradient via minorantes locales). Soient $f \in \Gamma_0(\mathbb{R}^N)$ et $x \in \mathbb{R}^N$. Si x^* vérifie

$$f(y) - f(x) - \langle x^*, y - x \rangle \geq o(\|y - x\|),$$

alors x^* est un sous-gradient de f en x .

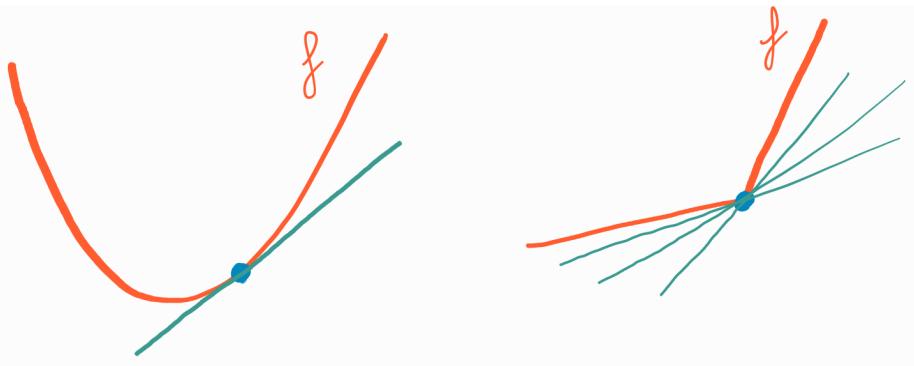


FIGURE II.6 – Sous-gradiants d'une fonction différentiable (gauche) ou pas (droite).

Démonstration. Posons $h(y) := f(x) + \langle x^*, y - x \rangle$, nous devons donc montrer que $f \geq h$. Raisonnons par l'absurde et supposons qu'il existe $y \in \mathbb{R}^N$ tel que $f(y) < h(y)$. Autrement dit il existe $\varepsilon > 0$ tel que $f(y) + \varepsilon \leq h(y)$. Considérons $z_t := (1-t)x + ty$ pour $t \in]0, 1[$ alors par convexité de f :

$$f(z_t) \leq (1-t)f(x) + tf(y) \leq (1-t)f(x) + th(y) - t\varepsilon.$$

On peut donc écrire

$$f(z_t) - h(z_t) \leq (1-t)f(x) + th(y) - t\varepsilon - h(z_t) = -t\varepsilon$$

où dans la dernière égalité nous avons utilisé le caractère affine de h pour écrire $h(z_t) = (1-t)h(x) + th(y)$, ainsi que la définition de h pour laquelle $h(x) = f(x)$. D'autre part notre hypothèse sur x^* veut dire que $f(z_t) - h(z_t) \geq o(\|z_t - x\|) = o(t)$. On obtient donc après division par t que $o(1) \leq -\varepsilon$, ce qui est une contradiction ! ■

Définition II.57 (Sous-différentiel). Soient $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$ et $x \in \mathbb{R}^N$. On définit le **SOUS-DIFFÉRENTIEL** de f en x comme l'ensemble des sous-gradiants de f en x , et on le notera $\partial f(x)$. On notera $\text{dom } \partial f = \{x \in \mathbb{R}^N \mid \partial f(x) \neq \emptyset\}$ son domaine.

Exemple II.58 (Sous-différentiel de la valeur absolue). Soit $f(x) = |x|$. Alors $\partial f(x) = \text{sgn}(x)$, où

$$\text{sgn}(x) = \begin{cases} -1 & \text{si } x < 0, \\ [-1, 1] & \text{si } x = 0, \\ +1 & \text{si } x > 0. \end{cases}$$

Proposition II.59 (Sous-différentiel d'une indicatrice). Soit $C \subset \mathbb{R}^N$ convexe fermé non vide. Alors pour tout $x \in C$ nous avons $\partial \delta_C(x) = N_C(x)$.

Démonstration. On va utiliser la définition de sous-différentiel ainsi que la caractérisation du cône normal par les angles:

$$\begin{aligned} x^* \in \partial\delta_C(x) &\Leftrightarrow (\forall y \in \mathbb{R}^N) \quad \delta_C(y) - \delta_C(x) - \langle x^*, y - x \rangle \geq 0 \\ &\Leftrightarrow (\forall y \in \mathbb{R}^N) \quad \delta_C(y) \geq \langle x^*, y - x \rangle \\ &\Leftrightarrow (\forall y \in C) \quad 0 \geq \langle x^*, y - x \rangle \\ &\Leftrightarrow x^* \in N_C(x). \end{aligned}$$

■

Proposition II.60 (Sous-différentiel et différentiabilité). Soit $f \in \Gamma_0(\mathbb{R}^N)$ et $x \in \text{dom } f$. Alors

- 1) Si f est différentiable en x alors $\partial f(x) = \{\nabla f(x)\}$.
- 2) Si $\partial f(x) = \{x^*\}$, alors f est différentiable en x avec $\nabla f(x) = x^*$.

Démonstration.

- 1) On procède par double inclusion.

\subset : Soit $x^* \in \partial f(x)$, montrons que $x^* - \nabla f(x) = 0$. On se donne donc $d \in \mathbb{R}^N$, et avec $y := x + td$ on écrit

$$\begin{aligned} &\langle \nabla f(x) - x^*, y - x \rangle \\ &= f(y) - f(x) - \langle x^*, y - x \rangle - [f(y) - f(x) - \langle \nabla f(x), y - x \rangle] \\ &\geq 0 - o(\|y - x\|) \\ &= o(\|y - x\|). \end{aligned}$$

Ceci est donc équivalent à ce que

$$\langle \nabla f(x) - x^*, d \rangle \geq o(1).$$

Autrement dit $\langle \nabla f(x) - x^*, d \rangle \geq 0$ pour tout $d \in \mathbb{R}^N$, ce qui implique $x^* - \nabla f(x) = 0$.

\supset : Montrons que $\nabla f(x) \in \partial f(x)$. Par définition de la différentiabilité on a

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle = o(\|y - x\|),$$

on a donc une minorante affine locale et on conclut avec le lemme II.56.

- 2) Ce résultat est admis, et peut être par exemple consulté dans [3, Proposition 17.45]. On en propose une preuve sous l'hypothèse supplémentaire que $x \in \text{cont } f$ en annexe, voir le théorème B.45.

■

Proposition II.61 (Caractérisation géométrique du sous-différentiel). Soit $f \in \Gamma_0(\mathbb{R}^N)$ et $x \in \text{dom } f$. Alors le sous-différentiel s'obtient via les normales à l'épigraphie:

$$\partial f(x) = \{x^* \in \mathbb{R}^N \mid (x^*, -1) \in N_{\text{epi } f}((x, f(x)))\}.$$

Démonstration. On écrit

$$\begin{aligned} x^* \in \partial f(x) &\Leftrightarrow (\forall y \in \mathbb{R}^N) \quad f(y) - f(x) - \langle x^*, y - x \rangle \geq 0 \\ &\Leftrightarrow (\forall y \in \text{dom } f) \quad f(y) \geq f(x) + \langle x^*, y - x \rangle \\ &\Leftrightarrow (\forall y \in \text{dom } f) (\forall r \geq f(y)) \quad r \geq f(x) + \langle x^*, y - x \rangle \\ &\Leftrightarrow (\forall (y, r) \in \text{epi}(f)) \quad r \geq f(x) + \langle x^*, y - x \rangle \\ &\Leftrightarrow (\forall (y, r) \in \text{epi}(f)) \quad 0 \geq (-1)(r - f(x)) + \langle x^*, y - x \rangle \\ &\Leftrightarrow (\forall (y, r) \in \text{epi}(f)) \quad 0 \geq \left\langle \begin{pmatrix} x^* \\ -1 \end{pmatrix}, \begin{pmatrix} y \\ r \end{pmatrix} - \begin{pmatrix} x \\ f(x) \end{pmatrix} \right\rangle \\ &\Leftrightarrow (x^*, -1) \in N_{\text{epi}(f)}((x, f(x))), \end{aligned}$$

où la dernière équivalence vient de la caractérisation du cône normal par les angles, voir la proposition I.51. ■

Remarque II.62 (Caractérisation géométrique du sous-différentiel). Il peut être très instructif d'essayer de comprendre ce résultat géométrique en faisant un dessin. En considérant par exemple une fonction dérivable $f : \mathbb{R} \rightarrow \mathbb{R}$, on sait que le nombre dérivé $f'(x)$ est égal à la pente de la tangente au graphe en $(x, f(x))$. Ceci veut dire que la droite est dirigée par le vecteur $(1, f'(x))$. Donc la droite perpendiculaire à cette tangente, qui n'est rien d'autre que la normale au graphe, est dirigée par le vecteur perpendiculaire $(f'(x), -1)$. On pourra également se demander ce qu'il se passe lorsque la fonction n'est pas dérivable en x , avec par exemple une fonction affine par morceaux.

Proposition II.63 (Domaine du sous-differentiel). Soit $f \in \Gamma_0(\mathbb{R}^N)$. Alors

$$\text{int dom } f \subset \text{dom } \partial f \subset \text{dom } f$$

Démonstration. On montre les deux inclusions.

$\text{dom } \partial f \subset \text{dom } f$: Si $x \in \text{dom } \partial f$ alors on a un $x^* \in \partial f(x)$ tel que $f(x) \leq f(y) - \langle x^*, y - x \rangle$ pour tout $y \in \mathbb{R}^N$. Il suffit de prendre $y \in \text{dom } f \neq \emptyset$ pour conclure.

$\text{int dom } f \subset \text{dom } \partial f$: Soit $x \in \text{int dom } f$. D'après la proposition II.15 on sait que $(x, f(x)) \notin \text{int epi } f$, ce qui veut dire que le cône normal est non trivial : $N_{\text{epi } f}(x, f(x)) \neq \{(0, 0)\}$ (voir la proposition I.53). Donc il existe $(u, v) \in N_{\text{epi } f}(x, f(x))$ tel que $(u, v) \neq (0, 0)$. La caractérisation du cône normal par les angles (proposition I.51) nous dit que

$$(\forall (y, r) \in \text{epi } f) \quad \langle u, y - x \rangle + v(r - f(x)) \leq 0.$$

On fait maintenant une distinction de cas sur le signe de $v \in \mathbb{R}$.

Cas $v > 0$: On prend $y = x \in \text{dom } f$ et $r = f(x) + 1$ dans l'inégalité ci-dessus pour obtenir que $v \leq 0$, ce qui est une contradiction.

Cas $v = 0$: L'inégalité ci-dessus devient

$$(\forall (y, r) \in \text{epi } f) \quad \langle u, y - x \rangle \leq 0.$$

Vu que r est absent de cette inégalité, cela implique que

$$(\forall y \in \text{dom } f) \quad \langle u, y - x \rangle \leq 0.$$

Or par hypothèse il existe un voisinage $\mathbb{B}(x, \delta) \subset \text{dom } f$. On peut donc considérer $y = x + \frac{\delta}{\|u\|}u \in \mathbb{B}(x, \delta)$ qui est donc dans le domaine, et qui une fois injecté dans l'inégalité nous donne $\delta \|u\| \leq 0$. Donc $u = 0$ et $v = 0$ ce qui contredit $(u, v) \neq (0, 0)$.

Cas $v < 0$: Puisque $(u, v) \in N_{\text{epi } f}(x, f(x))$ qui est un cône, on peut diviser par $-v > 0$ pour obtenir $(\frac{u}{-v}, -1) \in N_{\text{epi } f}(x, f(x))$. D'après la caractérisation géométrique du sous-différentiel, on en conclut que $\frac{u}{-v} \in \partial f(x)$. ■

Exemple II.64 (Le sous-différentiel n'est pas toujours bien défini). On n'a pas toujours $\text{dom } f = \text{dom } \partial f$. Par exemple considérons $f(x) = -\sqrt{x} + \delta_{[0, +\infty[}(x)$. Alors il est clair que $0 \in \text{dom } f$, mais il est possible de vérifier que $\partial f(0) = \emptyset$.

II.II.2 Calcul sous-différentiel

Proposition II.65 (Sous-différentiel du produit). Soient $f \in \Gamma_0(\mathbb{R}^N)$ et $\lambda > 0$. Alors $\partial(\lambda f)(x) = \lambda \partial f(x)$.

Démonstration. Immédiat par définition du sous-différentiel. ■

Théorème II.66 (Sous-différentiel de la composée). Soit $g \in \Gamma_0(\mathbb{R}^M)$ et $A \in \mathcal{M}_{M,N}(\mathbb{R})$. Alors

- 1) pour tout $x \in \mathbb{R}^N$, on a $\partial(g \circ A)(x) \supset A^\top \partial g(Ax)$.
- 2) si on suppose que $\text{cont } g \cap \text{Im } A \neq \emptyset$, alors

$$(\forall x \in \mathbb{R}^N) \quad \partial(g \circ A)(x) = A^\top \partial g(Ax).$$

Démonstration. On procède par double inclusion.

\supset : Soit $x^* \in A^\top \partial g(Ax)$, c'est-à-dire que $x^* = A^\top y^*$ où $y^* \in \partial g(Ax)$. Alors

$$(\forall y \in \mathbb{R}^M) \quad g(y) - g(Ax) - \langle y^*, y - Ax \rangle \geq 0$$

Si on prend $y = Ax'$ pour $x' \in \mathbb{R}^N$ quelconque on a

$$(\forall x' \in \mathbb{R}^N) \quad g(Ax') - g(Ax) - \langle y^*, Ax' - Ax \rangle \geq 0,$$

et on conclut en voyant que $\langle y^*, Ax' - Ax \rangle = \langle x^*, x' - x \rangle$.

\subset : Admis. On peut en voir une preuve faisant appel à la dualité dans une section facultative du cours, voir le corollaire B.37. On peut également prouver ce résultat directement avec un théorème de séparation, voir la section C.I.6 en annexe. ■

Exemple II.67 (Sous-différentiel de la composée).

- Si $f(x) = \delta_C(Ax)$ et $\text{Im } A \cap \text{int } C \neq \emptyset$, alors on peut écrire que $\partial f(x) = A^\top N_C(Ax)$. Or f est l'indicatrice de $A^{-1}C$, donc on a obtenu la formule

$$N_{A^{-1}C}(x) = A^\top N_C(Ax).$$

- Sans l'hypothèse de qualification, on ne peut pas garantir en général que la formule est vraie. Par exemple si $f = \delta_C \circ A \in \Gamma_0(\mathbb{R}^2)$, on peut considérer $A \in \mathcal{M}_2(\mathbb{R})$ comme étant la matrice de projection sur les abscisses $A(x_1, x_2) = (x_1, 0)$; et d'autre part $C = \mathbb{B}(e_2, 1)$ où $e_2 = (0, 1)$. On peut alors voir que $\text{Im } A = \mathbb{R}e_1$ et que $\text{cont } \delta_C = \text{int } C$ avec $\text{int } C \cap \text{Im } A = \emptyset$, donc que l'hypothèse de qualification n'est pas vérifiée. Et on peut faire les calculs pour se rendre compte que la formule n'est pas valable. D'une part on a $g(Ax) = \delta_D(x)$ où $D = \mathbb{R}e_2$, donc

$$\partial(g \circ A)(x) = N_D(x) = \begin{cases} \mathbb{R}e_1 & \text{si } x \in \mathbb{R}e_2 \\ \emptyset & \text{sinon.} \end{cases}$$

D'autre part on a

$$\partial g(Ax) = N_C(Ax) = \begin{cases} \mathbb{R}-e_2 & \text{si } x \in \mathbb{R}e_2, \\ \emptyset & \text{sinon} \end{cases} \Rightarrow A^* \partial g(Ax) = \begin{cases} 0 & \text{si } x \in \mathbb{R}e_2, \\ \emptyset & \text{sinon.} \end{cases}$$

On voit bien qu'il y a un problème pour les $x \in \mathbb{R}e_2$.

- L'hypothèse de qualification est une condition nécessaire, mais elle n'est pas suffisante! Dans certains cas elle est superflue. Par exemple si $C = \{b\}$ alors d'une part $f = \delta_{[Ax=b]}$ dont on sait que le sous-différentiel est le cône normal à $[Ax = b]$, qui vaut $\text{Im } A^\top$. D'autre part, la formule nous donne

$$A^\top \partial \delta_b(Ax) = A^\top N_b(Ax) = A^\top \mathbb{R}^N = \text{Im } A,$$

ce qui est la même chose. En fait, tout se passe bien lorsque la fonction f est polyédrale (on pourra consulter le théorème B.18 en annexe).

Proposition II.68 (Sous-différentiel d'une somme séparable). Soit $f(x) = \sum_{i=1}^M f_i(x_i)$, où $f_i \in \Gamma_0(\mathbb{R}^{N_i})$. Alors

$$\partial f(x) = \prod_{i=1}^N \partial f_i(x_i).$$

Démonstration. On observe que $x^* \in \partial f(x)$ si et seulement si

$$(\forall y, x \in \mathbb{R}^{\sum N_i}) \quad \sum_{i=1}^M f_i(y_i) - f_i(x_i) - \langle x_i^*, y_i, x_i \rangle \geq 0.$$

On raisonne maintenant par double inclusion.

\subset : Pour i fixé on doit montrer que $x_i^* \in \partial f_i(x_i)$. Pour cela on pose $y_j = x_j$ lorsque $j \neq i$, de manière à ce que l'inégalité devienne $f_i(y_i) - f_i(x_i) - \langle x_i^*, y_i, x_i \rangle \geq 0$.

\supset : Immédiat par somme. ■

Exemple II.69 (Sous-différentiel de la norme ℓ^1).

- Soit $f(x) = \|x\|_1$. Si on écrit $f(x) = \sum_i f_i(x_i)$ où $f_i(t) = |t|$, alors on voit que

$$\partial\|\cdot\|_1(x) = \prod_{i=1}^N \partial| \cdot |(x_i) = \prod_{i=1}^N \text{sgn}(x_i) =: \text{SGN}(x).$$

- Soit $f(x) = \|Ax\|_1$. Puisque $\|\cdot\|_1$ est continue partout, on en déduit que $\partial f(x) = A^\top \text{SGN}(Ax)$.

Exemple II.70 (Sous-différentiel de la somme). De manière générale, le sous-différentiel de la somme n'est pas égal à la somme des sous-différentiels. Considérons par exemple $f(x) = -\sqrt{x} + \delta_{[0,+\infty[}(x)$ et $g(x) = \delta_{]-\infty,0]}$. On voit que $(f+g)(x) = \delta_0(x)$, et donc que

$$\partial(f+g)(x) = N_0(x) = \begin{cases} \mathbb{R} & \text{si } x = 0 \\ \emptyset & \text{si } x \neq 0. \end{cases}$$

D'autre part nous pouvons calculer

$$\partial f(x) = \begin{cases} \frac{-1}{2\sqrt{x}} & \text{si } x > 0 \\ \emptyset & \text{si } x \leq 0, \end{cases} \quad \text{et} \quad \partial g(x) = \begin{cases} 0 & \text{si } x < 0 \\ \mathbb{R}_+ & \text{si } x = 0 \\ \emptyset & \text{si } x > 0. \end{cases}$$

Ainsi nous voyons que $\partial f(x) + \partial g(x)$ est toujours vide. Ceci étant dit, cet exemple est un peu particulier, en le sens que les deux fonctions qu'on ajoute ne s'emboitent pas bien. Plus précisément leur domaine ne s'intersecte pas très bien. En faisant une hypothèse supplémentaire on peut obtenir la règle de somme espérée.

Proposition II.71 (Sous-différentiel de la somme). Soient $f, g \in \Gamma_0(\mathbb{R}^N)$. Soit $x \in \mathbb{R}^N$ et supposons que f soit différentiable en x . Alors

$$\partial(f+g)(x) = \nabla f(x) + \partial g(x).$$

Démonstration. On procède par double inclusion.

\subset : Soit $x^* \in \partial(f+g)(x)$, montrons que $x^* - \nabla f(x) \in \partial g(x)$. On écrit donc

$$\begin{aligned} & g(y) - g(x) - \langle x^* - \nabla f(x), y - x \rangle \\ &= (f+g)(y) - (f+g)(x) - \langle x^*, y - x \rangle - [f(y) - f(x) - \langle \nabla f(x), y - x \rangle] \\ &\geq 0 - o(\|y - x\|) \\ &= o(\|y - x\|). \end{aligned}$$

On a donc une minorante affine locale et on conclut à l'aide du lemme II.56.

\supset : Soit $x^* \in \partial g(x)$, montrons que $x^* + \nabla f(x) \in \partial(f+g)(x)$. On écrit alors

$$\begin{aligned} & (f+g)(y) - (f+g)(x) - \langle x^* + \nabla f(x), y - x \rangle \\ &= g(y) - g(x) - \langle x^*, y - x \rangle + f(y) - f(x) - \langle \nabla f(x), y - x \rangle \\ &\geq 0 + f(y) - f(x) - \langle \nabla f(x), y - x \rangle \\ &\geq 0, \end{aligned}$$

où l'on utilise dans les dernières inégalités la définition de x^* ainsi que le fait que $\nabla f(x) \in \partial f(x)$ (voir proposition II.60). ■

Le résultat précédent reste assez limité si aucun des fonctions n'est différentiable en x . On est également bloqués si l'une des fonctions est différentiable en certains points seulement (penser à la norme 1) et que x n'en fasse pas partie. Le prochain Théorème est une version beaucoup plus forte: on demande seulement qu'il existe un point quelque part où l'une des fonctions soit continue.

Théorème II.72 (Sous-différentiel de la somme (Moreau-Rockafellar)). Soient $f, g \in \Gamma_0(\mathbb{R}^N)$. Alors

$$(\forall x \in \mathbb{R}^N) \quad \partial(f+g)(x) \supset \partial f(x) + \partial g(x).$$

Si de plus $\text{cont } f \cap \text{dom } g \neq \emptyset$ alors cette inclusion est une égalité.

Démonstration. On procède par double inclusion.

\supset : C'est une conséquence immédiate de la définition de sous-différentiel.

\subset : On admet le résultat. On peut en voir une preuve faisant appel à la dualité dans une section facultative du cours, voir le corollaire B.29. On peut également prouver ce résultat directement avec un théorème de séparation, voir la section C.I.7 dans l'annexe. On peut néanmoins rapidement prouver que le résultat est vrai sous une condition de qualification un peu plus forte, à savoir que $\text{cont } f \cap \text{cont } g \neq \emptyset$.

Dans ce cas, posons $G(x, y) = f(x) + g(y)$ et $A : x \in \mathbb{R}^N \mapsto (x, x) \in \mathbb{R}^{2N}$. Alors on peut voir que $f+g = G \circ A$, et on veut appliquer la règle de dérivation en chaîne du théorème II.66. Pour cela il faut vérifier que $\text{cont } G \cap \text{Im } A \neq \emptyset$. Or on a supposé qu'il existe $\hat{x} \in \text{cont } f \cap \text{cont } g$. Donc G est continue en (\hat{x}, \hat{x}) , qui appartient bien à $\text{Im } A$. On peut donc écrire que $\partial(f+g)(x) = A^\top \partial G(Ax)$. Or G est une somme séparable, donc

d'après la proposition II.68 nous avons que $\partial G(x, y) = \partial f(x) \times \partial g(y)$. Puisque $Ax = (x, x)$ et que $A^\top(x, y) = x + y$, on en déduit que $\partial(f + g)(x) = \partial f(x) + \partial g(x)$. ■

Corollaire II.73 (Sous-différentiel de la somme composée). Soit $f = g \circ A + h$, où $g \in \Gamma_0(\mathbb{R}^M)$, $h \in \Gamma_0(\mathbb{R}^N)$ et $A \in \mathcal{M}_{M,N}(\mathbb{R})$. On suppose que $A(\text{dom } h) \cap \text{cont } g \neq \emptyset$. Alors

$$(\forall x \in \mathbb{R}^N) \quad \partial f(x) = A^\top \partial g(Ax) + \partial h(x).$$

Démonstration. Il suffit de combiner les théorèmes II.72 et II.66. L'hypothèse de qualification dit qu'il existe $x \in \text{dom } h$ tel que $Ax \in \text{cont } g$. Le fait que g soit continue en Ax , et que A soit continue impliquent que $g \circ A$ est continue en x . Autrement dit on a $x \in \text{dom } h \cap \text{cont}(g \circ A) \neq \emptyset$. On peut donc écrire $\partial f = \partial(g \circ A) + \partial h$. Ensuite, on sait que $Ax \in \text{cont } g$ ce qui veut bien dire que $Ax \in \text{cont } g \cap \text{Im } A \neq \emptyset$. On peut donc également écrire que $\partial(g \circ A) = A^* \circ \partial g \circ A$. ■

Remarque II.74 (Problème régulier). Lorsqu'on travaille avec des fonctions non lisses (et c'est notre cas dans ce cours), on a très souvent besoin et envie de dire que les règles de calcul sous-différentiel s'appliquent comme on le veut. Il est alors souvent confortable de simplement faire l'hypothèse que cela marche: on va par exemple dire que la fonction $f = g \circ A + h$ est **régulière** en x si la formule

$$\partial f(x) = A^\top \partial g(Ax) + \partial h(x)$$

est vraie. Pour que f soit régulière il suffit qu'une des conditions suivantes soit vérifiées:

- que $A(\text{dom } h) \cap \text{cont } g \neq \emptyset$, c'est ce que dit le corollaire II.73 ;
- que g soit différentiable en Ax , c'est une conséquence de la proposition II.71 ;
- que g et h soient polyédrales (indicatrice de polyèdres, fonction affine par morceaux comme la norme 1), c'est un résultat non trivial que l'on peut trouver dans l'annexe, voir le théorème B.18.

Théorème II.75 (Sous-différentiel du max). Soit $f = \max_{i=1,\dots,M} f_i$ où $f_i \in \Gamma_0(\mathbb{R}^N)$. Soit $x \in \cap_{i=1}^M \text{cont } f_i$. Alors

$$\partial f(x) = \text{adh co} \bigcup_{i \in I(x)} \partial f_i(x),$$

où $I(x) = \{i \mid f_i(x) = f(x)\}$.

Démonstration. Admis. On pourra en trouver une preuve dans l'annexe, voir le corollaire B.44. ■

Exemple II.76 (Max de fonctions).

- Si f est un max de fonctions f_i qui sont différentiables, alors on a $\partial f(x) = \text{co}\{\nabla f_i(x)\}_{i \in I(x)}$.
- Si f est un max de fonctions affines $f_i(x) = \langle \alpha_i, x \rangle + \beta_i$, alors $\partial f(x) = \text{co}\{\alpha_i\}_{i \in I(x)}$.
- Si $f(x) = |x| = \max\{x, -x\}$, alors on retrouve le fait que $\partial f(x) = \text{sgn}(x)$.

II.II.3 Conditions d'optimalité

Théorème II.77 (de Fermat généralisé). Soit $f \in \Gamma_0(\mathbb{R}^N)$. Alors $x \in \operatorname{argmin} f \Leftrightarrow 0 \in \partial f(x)$.

Démonstration.

$$\begin{aligned} 0 &\in \partial f(x) \\ \Leftrightarrow (\forall y \in \mathbb{R}^N) \quad &f(y) - f(x) - \langle 0, y - x \rangle \\ \Leftrightarrow (\forall y \in \mathbb{R}^N) \quad &f(y) \geq f(x) \\ \Leftrightarrow x &\in \operatorname{argmin}_C f. \end{aligned}$$

■

Corollaire II.78 (CNS de l'optimisation sous contraintes). Soit $f \in \Gamma_0(\mathbb{R}^N)$ et $C \subset \mathbb{R}^N$ convexe fermé. On suppose que $C \cap \operatorname{cont} f \neq \emptyset$. Alors

$$x \in \operatorname{argmin}_C f \Leftrightarrow 0 \in N_C(x) + \partial f(x).$$

Démonstration. On applique le théorème II.77 de Fermat à $f + \delta_C$, et on calcule $\partial(f + \delta_C) = \partial f + N_C$ grâce au fait que $\operatorname{dom} \delta_C \cap \operatorname{cont} f = C \cap \operatorname{cont} f \neq \emptyset$ et au théorème II.72 de Moreau-Rockafellar. ■

Corollaire II.79 (des multiplicateurs de Lagrange). Soit $f \in \Gamma_0(\mathbb{R}^N)$ continue et soit $C = [Ax = b]$ où $A \in \mathcal{M}_{M,N}(\mathbb{R})$ et $b \in \mathbb{R}^M$. Alors x minimise f sur C si et seulement si il existe des multiplicateurs $\lambda_1, \dots, \lambda_M \in \mathbb{R}$ tels que

$$(\forall i = 1, \dots, M) \quad \begin{cases} 0 \in \partial f(x) + \sum_{i=1}^M \lambda_i a_i, \\ \langle a_i, x \rangle = b_i, \end{cases}$$

où l'on a noté a_i la i -ème ligne de A .

Démonstration. D'après le corollaire II.78 (que l'on peut appliquer car f est continue), $x \in C$ est une solution si et seulement si $0 \in N_C(x) + \partial f(x)$. On sait que le cône normal de cet espace affine est $\operatorname{Im} A^\top$ (voir l'exemple I.54) si $Ax = b$, et \emptyset sinon. Cela veut dire qu'il existe $x^* \in \partial f(x)$ tel que $-x^* \in \operatorname{Im} A^\top$. Donc il existe $\lambda \in \mathbb{R}^M$ tel que $-x^* = A^\top \lambda$. Or si on note a_i la i -ème ligne de A alors c'est aussi la i -ème colonne de A , et donc $A^\top \lambda = \sum_i \lambda_i a_i$. ■

Corollaire II.80 (KKT linéaire). Soit $f \in \Gamma_0(\mathbb{R}^N)$ continue et soit $C = [Ax \leq b]$ où $A \in \mathcal{M}_{M,N}(\mathbb{R})$ et $b \in \mathbb{R}^M$. Alors x minimise f sur C si et seulement si il existe des multiplicateurs $\mu_1, \dots, \mu_M \in \mathbb{R}$ tels que

$$(\forall i = 1, \dots, M) \quad \begin{cases} 0 \in \partial f(x) + \sum_{i=1}^M \mu_i a_i, \\ \mu_i(\langle a_i, x \rangle - b_i) = 0, \\ \mu_i \geq 0, \\ \langle a_i, x \rangle \leq b_i. \end{cases}$$

où l'on a noté a_i la i -ème ligne de A .

Démonstration. D'après le corollaire II.78 (que l'on peut appliquer car f est continue), $x \in C$ est une solution si et seulement si $0 \in N_C(x) + \partial f(x)$. On sait calculer le cône normal de ce polyèdre en x (voir le théorème I.57):

$$N_C(x) = \text{cone}(a_i)_{i \in I(x)} \text{ où } I(x) := \{i \in [M] \mid \langle a_i, x \rangle = b_i\}.$$

Donc il existe $\mu \in \mathbb{R}_+^M$ tel que

$$0 \in \sum_{i \in I(x)} \mu_i a_i + \partial f(x).$$

Une autre façon d'écrire cela est de dire que

$$0 \in \sum_{i=1}^M \mu_i a_i + \partial f(x),$$

mais en imposant également le fait que $\mu_j = 0$ pour tout $j \notin I(x)$. Une manière astucieuse d'encoder cette propriété est de demander que

$$(\forall i \in [M]) \quad \mu_i (\langle a_i, x \rangle - b_i) = 0.$$

En effet si $i \in I(x)$ on sait que $(\langle a_i, x \rangle - b_i) = 0$ ce qui veut dire qu'aucune condition ne porte sur μ_i , tandis que si $i \notin I(x)$ alors nécessairement $\mu_i = 0$. En combinant tout cela on obtient le résultat désiré. ■

Remarque II.81 (Combiner égalités et inégalités affines). Si on dispose d'une contrainte combinant égalités et inégalités affines, on ne peut utiliser ni le théorème de Lagrange ou de KKT. Mais comme on l'a vu dans le chapitre précédent, les contraintes d'égalité peuvent se réécrire avec un double de contraintes d'inégalités, puisque chaque égalité est équivalente à deux inégalités. En faisant ainsi on obtient le résultat suivant qui combine bien les corollaires II.79 et II.80, avec des multiplicateurs de Lagrange $\lambda_j \in \mathbb{R}$ pour les contraintes d'égalité et des multiplicateurs de KKT $\mu_i \in \mathbb{R}_+$ pour les contraintes d'inégalité.

Corollaire II.82 (Lagrange-KKT linéaire). Soit $C = [Ax \leq b] \cap [A'x = b']$ et $f \in \Gamma_0(\mathbb{R}^N)$ continue où $A \in \mathcal{M}_{M,N}(\mathbb{R})$, $b \in \mathbb{R}^M$, $A' \in \mathcal{M}_{M',N}(\mathbb{R})$ et $b' \in \mathbb{R}^{M'}$. Alors x minimise f sur C si et seulement si il existe des multiplicateurs $\mu_1, \dots, \mu_M, \lambda_1, \dots, \lambda_{M'} \in \mathbb{R}$ tels que

$$(\forall i = 1, \dots, M)(\forall j = 1, \dots, M') \quad \begin{cases} 0 \in \partial f(x) + \sum_{i=1}^M \mu_i a_i + \sum_{j=1}^{M'} \lambda_j a'_j, \\ \mu_i (\langle a_i, x \rangle - b_i) = 0, \\ \mu_i \geq 0, \\ \langle a_i, x \rangle \leq b_i \text{ et } \langle a'_j, x \rangle = b'_j, \end{cases} \quad (\text{LKKT})$$

où l'on a noté a_i et a'_j la i -ème ligne de A et j -ème ligne de A' .

Démonstration. On commence par réécrire C comme une seule contrainte d'inégalités affines: $C = [\mathcal{A}x \leq \beta]$ où $\mathcal{A} = [A; A'; -A]$ et $\beta = (b; b'; -b')$. On peut alors directement appliquer le théorème II.80 de KKT linéaire pour obtenir le système d'inéquations suivant: $(\forall i = 1, \dots, M)(\forall j = 1, \dots, M')$,

$$\begin{cases} 0 \in \partial f(x) + \sum_{i=1}^M \mu_i a_i + \sum_{j=1}^{M'} \mu'_j a'_j + \sum_{j=1}^{M'} \mu_j''(-a'_j), \\ \mu_i(\langle a_i, x \rangle - b_i) = 0 \text{ et } \mu'_j(\langle a'_j, x \rangle - b'_j) = 0 \text{ et } \mu_j''(\langle a'_j, x \rangle - b'_j) = 0, \\ \mu_i \geq 0 \text{ et } \mu'_j \geq 0 \text{ et } \mu_j'' \geq 0, \\ \langle a_i, x \rangle \leq b_i \text{ et } \langle a'_j, x \rangle \leq b'_j \text{ et } \langle a'_j, x \rangle \geq b'_j. \end{cases}$$

On voit que ce système un peu compliqué contient des informations redondantes. Par exemple la condition d'admissibilité $\langle a'_j, x \rangle \leq b'_j$ et $\langle a'_j, x \rangle \geq b'_j$ est équivalente à ce que $\langle a'_j, x \rangle = b'_j$ pour tout $j = 1, \dots, M'$. Une fois ceci observé, on voit que les conditions de complémentarité $\mu'_j(\langle a'_j, x \rangle - b'_j) = 0$ et $\mu_j''(\langle a'_j, x \rangle - b'_j) = 0$ sont automatiquement vérifiées. Enfin, la somme $\sum_{j=1}^{M'} \mu'_j a'_j + \sum_{j=1}^{M'} \mu_j''(-a'_j)$ peut être écrite de manière équivalente en $\sum_{j=1}^{M'} \lambda'_j a'_j$, en introduisant le changement de variables $\lambda_j = \mu'_j - \mu_j'' \in \mathbb{R}$, qui est équivalent à prendre $\mu'_j = (\lambda_j)_+$ et $\mu_j'' = (\lambda_j)_-$. Notre système devient alors

$$\begin{cases} 0 \in \partial f(x) + \sum_{i=1}^M \mu_i a_i + \sum_{j=1}^{M'} \lambda_j a'_j \\ \mu_i(\langle a_i, x \rangle - b_i) = 0, \\ \mu_i \geq 0, \\ \langle a_i, x \rangle \leq b_i \text{ et } \langle a'_j, x \rangle = b'_j, \end{cases}$$

qui est bien ce que l'on voulait montrer. ■

Remarque II.83 (Le système d'inéquations de Lagrange-KKT). Pour trouver une solution du problème de minimiser une fonction f sous une contrainte polyédrale, il « suffit » de trouver une solution au système d'inéquations de Lagrange-KKT (LKKT). Ce système se décompose en 4 conditions:

- La condition de *stationnarité* $0 \in \partial f(x) + \sum_{i=1}^M \mu_i a_i + \sum_{j=1}^{M'} \lambda_j a'_j$. C'est la condition qui relie la dérivée de la fonction f aux contraintes, c'est la généralisation de la condition de Fermat classique $f'(x) = 0$.
- La condition d'*admissibilité* $\langle a_i, x \rangle \leq b_i$ et $\langle a'_j, x \rangle = b'_j$. C'est la condition qui garantit que $x \in C$, c'est-à-dire que x est admissible.
- La condition de *positivité* $\mu_i \geq 0$, qui ne concerne que les multiplicateurs associés aux inégalités. Elle reflète le caractère unilatéral des contraintes d'inégalité, et vient du fait que le cône normal à une contrainte d'inégalité est une demi-droite (là où le cône normal à une contrainte d'égalité est une droite).

- La condition de complémentarité $\mu_i(\langle a_i, x \rangle - b_i) = 0$, qui ne concerne que les multiplicateurs associés aux inégalités. C'est cette condition qui traduit le fait que seules les contraintes actives en x contribuent la stationnarité, les contraintes inactives auront forcément un multiplicateur nul. On se convaincra que les contraintes d'égalité sont toujours actives, donc n'ont pas besoin de cette condition.

II.III Conjuguée de Fenchel

II.III.1 Définitions et calcul de la conjuguée

Définition II.84 (Conjuguée de Fenchel). Soit $f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$ propre. On définit sa **CONJUGUÉE** (de Fenchel) par

$$f^*(x^*) = \sup_{x \in \mathbb{R}^N} \langle x^*, x \rangle - f(x).$$

Remarque II.85. Calculer f^* n'est pas évident en général car il faut résoudre un problème de minimisation. Par exemple $f^*(0) = -\inf f$ qui n'est pas trivial à obtenir ! Cependant on va pouvoir calculer la conjuguée pour des fonctions élémentaires, et les combiner grâce à des règles de calcul.

Exemple II.86 (Conjuguée de fonctions élémentaires).

- Si $f(x) = \delta_C(x)$ alors $f^*(x^*) = \sigma_C(x^*)$.
- Si K est un cône convexe fermé alors $\delta_K^* = \delta_{K^*}$, car on a vu en TD que $\sigma_K = \delta_{K^*}$.
- Si $f = \delta_0$ alors $f^* \equiv 0$. De même si $f \equiv 0$ alors $f^* = \delta_0$.
- Si $f(x) = \frac{1}{2}\|x\|^2$ alors $f^* = f$. On notera souvent q cette fonction.
- Si $f(x) = \frac{1}{p}\|x\|_p^p$ alors $f^*(x^*) = \frac{1}{q}\|x^*\|_q^q$ où $\frac{1}{p} + \frac{1}{q} = 1$.
- Si $f(x) = \frac{1}{2}\langle Ax, x \rangle$ où A est symétrique définie positive, alors $f^*(x^*) = \frac{1}{2}\langle A^{-1}x^*, x^* \rangle$.

Remarque II.87 (Biconjuguée). Sur beaucoup de ces exemples, on voit que $f^{**} = f$. C'est un résultat qu'on verra plus tard, sous condition que $f \in \Gamma_0(\mathbb{R}^N)$.

Proposition II.88 (Conjuguée est convexe sci propre). Si $f \in \Gamma_0(\mathbb{R}^N)$ alors $f^* \in \Gamma_0(\mathbb{R}^N)$.

Démonstration. Montrons que f^* est convexe sci propre.

- f^* convexe : par définition, f^* est un sup de fonctions convexes $x^* \mapsto \langle x^*, x \rangle - f(x)$.
- f^* sci : idem, c'est un sup de fonctions sci.
- f^* est à valeurs dans $\overline{\mathbb{R}}$: il nous faut vérifier que $f^* > -\infty$. Puisque f est propre, il existe $\hat{x} \in \mathbb{R}^N$ tel que $f(\hat{x}) = r \in \mathbb{R}$. Alors pour tout $x^* \in \mathbb{R}^N$ on voit que

$$f^*(x^*) \geq \langle x^*, \hat{x} \rangle - f(\hat{x}) \geq \langle x^*, \hat{x} \rangle - r > -\infty.$$

- f^* est propre : puisque $f \in \Gamma_0(\mathbb{R}^N)$ elle admet une minorante affine (voir corollaire II.47). Cette minorante s'écrit $h(x) = \langle \alpha, x \rangle + \beta$, et on voit alors que

$$f^*(\alpha) = \sup_x \langle \alpha, x \rangle - f(x) \leq \sup_x \langle \alpha, x \rangle - h(x) = \beta < +\infty.$$

■

Proposition II.89 (Calcul de la conjuguée). Soit $f \in \Gamma_0(\mathbb{R}^N)$.

- 1) Si $g(x) = f(x) + c$ où $c \in \mathbb{R}$, alors $g^*(x^*) = f^*(x^*) - c$.
- 2) Si $g(x) = f(x - a)$ où $a \in \mathbb{R}^N$, alors $g^*(x^*) = f^*(x^*) + \langle a, x^* \rangle$.
- 3) Si $g(x) = f(x) + \langle a, x \rangle$ où $a \in \mathbb{R}^N$, alors $g^*(x^*) = f^*(x^* - a)$.
- 4) Si $g(x) = \lambda f(x)$ où $\lambda > 0$, alors $g^*(x^*) = \lambda f^*(\lambda^{-1}x^*)$.
- 5) Si $g(x) = f(\lambda x)$ où $\lambda \in \mathbb{R}^*$, alors $g^*(x^*) = f^*(\lambda^{-1}x^*)$.
- 6) Si $g(x) = f(Ax)$ où A matrice inversible, alors $g^*(x^*) = f^*((A^\top)^{-1}x^*)$.

Démonstration. Immédiat d'après la définition de la conjuguée et un changement de variables approprié. C'est un bon exercice de TD pour commencer à travailler sur la notion. ■

Exemple II.90 (Conjuguée de fonctions affines et diracs).

- Si $f(x) = \langle a, x \rangle - b$, alors $f^* = \delta_a + b$ car $0^* = \delta_0$. Donc la conjuguée d'une fonction affine est un dirac.
- Si $f(x) = \delta_a(x) + b$, alors $f^*(x^*) = \langle a, x^* \rangle - b$. Donc la conjuguée d'un dirac est une fonction affine.

Une conséquence de cet exemple est que la biconjuguée d'une fonction affine est elle-même.

Exemple II.91 (Conjuguée d'un moindre carré). Soit $f(x) = \frac{1}{2}\|Ax - b\|^2$, où A est inversible. Alors $f(x) = q(Ax - b) = g(Ax)$ où $g(y) = q(y - b)$. Avec la règle de composée par une application linéaire,

$$f^*(x^*) = g^*(A^{\top -1}x^*).$$

Avec la règle de translation, on voit que $g^*(y^*) = q^*(y^*) + \langle y^*, b \rangle$. De plus $q^* = q$. Alors

$$f^*(x^*) = q(A^{\top -1}x^*) + \langle A^{\top -1}x^*, b \rangle = \frac{1}{2}\|A^{\top -1}x^*\|^2 + \langle x^*, A^{-1}b \rangle.$$

Remarque II.92 (Conjuguée de la somme). Malheureusement on n'a pas $(f + g)^* = f^* + g^*$. Pour le voir il suffit de prendre $f = g$ et de se rendre compte que $(2f)^* \neq 2f^*$ puisque $(2f)^* = 2f^*(\frac{1}{2})$. Par contre pour des sommes séparables cela se passe bien:

Proposition II.93 (Conjuguée de la somme séparable). Soit $f(x) = \sum_{i=1}^M f_i(x_i)$, où $f_i \in \Gamma_0(\mathbb{R}^{N_i})$. Alors

$$f^*(x^*) = \sum_{i=1}^M f_i^*(x_i^*).$$

Démonstration. On utilise le fait que f est une fonction de \mathbb{R}^N avec $N = \sum_i N_i$, qui est muni du produit scalaire $\langle x, y \rangle = \sum_{i=1}^M \langle x_i, y_i \rangle$, pour écrire

$$\begin{aligned} f^*(x^*) &= \sup_x \langle x^*, x \rangle - f(x) \\ &= \sup_x \sum_i \langle x_i^*, x_i \rangle - f_i(x_i) \\ &= \sum_i \sup_{x_i} \langle x_i^*, x_i \rangle - f_i(x_i) \\ &= \sum_i f_i^*(x_i). \end{aligned}$$

■

Proposition II.94 (Calcul pratique de la conjuguée). Soit $f \in \Gamma_0(\mathbb{R}^N)$. Pour tout $x^* \in \mathbb{R}^N$, si on trouve x tel que $x^* \in \partial f(x)$, alors

$$f^*(x^*) = \langle x^*, x \rangle - f(x).$$

Démonstration. On a $f^*(x^*) = -\inf_x f(x) - \langle x^*, x \rangle$. Si $x^* \in \partial f(x)$ alors $0 \in \partial f(x) - \nabla[\langle x^*, \cdot \rangle](x)$. Par la règle de somme simple (proposition II.71) on obtient $0 \in \partial[f - \langle x^*, \cdot \rangle](x)$, et avec le théorème II.77 de Fermat on conclut que x minimise $f - \langle x^*, \cdot \rangle$. ■

Exemple II.95 (Conjuguée de la norme ℓ^1). On peut calculer (voir TD) que la conjuguée de la valeur absolue est $\delta_{[-1,1]}$. On peut alors étendre ce résultat à la norme ℓ^1 , qui est une somme séparable de valeurs absolues:

$$\|\cdot\|_1^*(x^*) = \sum_{i=1}^N \delta_{[-1,1]}(x_i^*) = \delta_{\mathbb{B}_\infty}(x^*),$$

où $\mathbb{B}_\infty = [-1, 1]^N$ est la boule unité de la norme ℓ^∞ . Ce lien entre normes ℓ^1 et ℓ^∞ n'est pas une coïncidence, on en rediscutera dans l'exemple II.101.

II.III.2 Propriétés duales de la conjuguée

Proposition II.96 (La conjugaison est décroissante). Soient $f, g \in \Gamma_0(\mathbb{R}^N)$. Si $f \leq g$ alors $g^* \leq f^*$.

Démonstration. Il suffit d'appliquer la définition

$$g^*(x^*) = \sup_x \langle x^*, x \rangle - g(x) \leq \sup_x \langle x^*, x \rangle - f(x) = f^*(x^*).$$

■

Théorème II.97 (de Fenchel-Young). Soit $f \in \Gamma_0(\mathbb{R}^N)$. Alors l'inégalité de Fenchel-Young a lieu:

$$(\forall x, x^* \in \mathbb{R}^N) \quad f(x) + f^*(x^*) \geq \langle x^*, x \rangle.$$

De plus, cette inégalité est une inégalité si et seulement si $x^* \in \partial f(x)$.

Démonstration. L'inégalité découle directement de la définition:

$$f^*(x^*) = \sup \langle x^*, \cdot \rangle - f \geq \langle x^*, x \rangle - f(x).$$

On procède ensuite par double inégalité

\Rightarrow : Si $f^*(x^*) = \langle x^*, x \rangle - f(x)$, alors cela veut dire d'après la définition de f^* que x maximise $\langle x^*, \cdot \rangle - f$, ou autrement dit que x minimise $f - \langle x^*, \cdot \rangle$. Puisque les formes linéaires sont différentiables, on peut utiliser le théorème II.77 de Fermat avec la règle de somme simple (proposition II.71) pour en déduire que $0 \in \partial f(x) - x^*$. D'où $x^* \in \partial f(x)$.

\Leftarrow : Si $x^* \in \partial f(x)$, alors on conclut avec la proposition II.94.

■

Exemple II.98 (Inégalité de Young). Si $\frac{1}{p} + \frac{1}{q} = 1$ avec $p > 1$, alors avec $f(x) = \frac{1}{p} \|x\|_p^p$ l'inégalité de Fenchel-Young devient

$$\langle x, y \rangle \leq \frac{1}{p} \|x\|_p^p + \frac{1}{q} \|y\|_q^q.$$

Théorème II.99 (de la biconjuguée). Soit $f \in \Gamma_0(\mathbb{R}^N)$. Alors $f^{**} = f$.

Démonstration. On procède par double inégalité.

\leq : Soit $x \in \mathbb{R}^N$, montrons que $f^{**}(x) \leq f(x)$. Pour tout $x^* \in \mathbb{R}^N$, l'inégalité de Fenchel-Young nous dit que

$$f(x) \geq \langle x^*, x \rangle - f^*(x^*).$$

Si on prend le sup sur x^* dans cette inégalité, on obtient

$$f(x) \geq \sup_{x^*} \langle x^*, x \rangle - f^*(x^*) = f^{**}(x).$$

\geq : Montrons que $f^{**} \geq f$. Puisque $f \in \Gamma_0(\mathbb{R}^N)$ alors on sait qu'elle peut s'écrire comme un sup de fonctions affines, d'après le théorème II.46. Autrement dit, il existe une famille de fonctions affines $(h_i)_{i \in I}$ telle que $f = \sup_i h_i$. En particulier nous avons $f \geq h_i$, donc si on applique deux fois la décroissance de la conjuguée (voir proposition II.96), on obtient successivement $f^* \leq h_i^*$ puis $f^{**} \geq h_i^{**}$. En passant au sup, on en déduit que $f^{**} \geq \sup_i h_i^{**}$. Or on a vu dans l'exemple II.90 que la biconjuguée d'une fonction affine est elle-même, c'est-à-dire que $h_i^{**} = h_i$. On conclut alors que $f^{**} \geq \sup_i h_i = f$.

■

Exemple II.100 (Dualité indicatrice / support). Si $C \subset \mathbb{R}^N$ est convexe fermé non vide, alors nous avons vu que $\delta_C^* = \sigma_C$. Le théorème de la biconjuguée nous permet de dire que l'on a également $\sigma_C^* = \delta_C$.

Exemple II.101 (Conjuguée de la norme). Soit $\|\cdot\|$ une norme quelconque sur \mathbb{R}^N , et notons \mathbb{B} sa boule unité. Considérons sa norme duale, dont on rappelle la définition:

$$\|x^*\|_* = \sup\{|\langle x^*, x \rangle| \mid x \in \mathbb{B}\}.$$

C'est un simple exercice (voir TD) que de montrer que $\|\cdot\|_* = \sigma_{\mathbb{B}}$. Si on utilise le fait que $\|\cdot\|_{**} = \|\cdot\|$ on en déduit que $\|\cdot\| = \sigma_{\mathbb{B}_*}$, où \mathbb{B}_* est la boule unité de la norme duale. Si on passe cette égalité à la conjuguée, nous en déduisons que $\|\cdot\|^* = \delta_{\mathbb{B}^*}$. Autrement dit, la conjuguée d'une norme est l'indicatrice de la boule unité duale. Ceci explique pourquoi la conjuguée de la norme ℓ^1 est l'indicatrice de la boule ℓ^∞ , comme on l'a vu dans l'exemple II.95.

Proposition II.102 (Sous-différentiel de la conjuguée (Formule de Legendre-Fenchel)). Si $f \in \Gamma_0(\mathbb{R}^N)$ alors $\partial f^* = (\partial f)^{-1}$. Autrement dit

$$x \in \partial f^*(x^*) \Leftrightarrow x^* \in \partial f(x).$$

Démonstration. On utilise le théorème II.97 de Fenchel-Young et le théorème II.99 de la biconjuguée:

$$\begin{aligned} & x \in \partial f^*(x^*) \\ \Leftrightarrow & f^*(x^*) + f^{**}(x) = \langle x^*, x \rangle \\ \Leftrightarrow & f^*(x^*) + f(x) = \langle x^*, x \rangle \\ \Leftrightarrow & x^* \in \partial f(x). \end{aligned}$$

■

Corollaire II.103 (Domaine et image du sous-differentiel de la conjuguée). Soit $f \in \Gamma_0(\mathbb{R}^N)$. Alors

$$\text{dom } \partial f^* = \text{Im } \partial f \quad \text{et} \quad \text{dom } \partial f = \text{Im } \partial f^*,$$

où l'on a noté $\text{Im } \partial f := \bigcup_{x \in \mathbb{R}^N} \partial f(x)$.

Démonstration. Immédiat via la formule de Legendre-Fenchel. ■

Dans le théorème qui suit, on note $C_L^{1,1}(\mathbb{R}^N)$ l'ensemble des fonctions L -lisses, c'est-à-dire les fonctions différentiables sur \mathbb{R}^N dont le gradient est L -Lipschitzien.

Théorème II.104 (Dualité lisse / fortement convexe). Soit $f \in \Gamma_0(\mathbb{R}^N)$, et $\mu, L > 0$. Alors

$$f \in \Gamma_\mu(\mathbb{R}^N) \Leftrightarrow f^* \in C_{\mu^{-1}}^{1,1} \quad \text{et} \quad f \in C_L^{1,1}(\mathbb{R}^N) \Leftrightarrow f^* \in \Gamma_{L^{-1}}(\mathbb{R}^N).$$

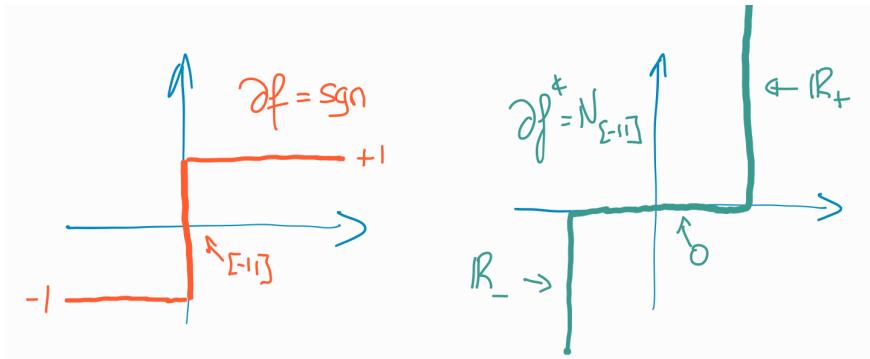


FIGURE II.7 – Le graphe de ∂f^* est exactement l'inverse du graphe de ∂f .

Démonstration. Admis. On peut en trouver une preuve dans la section B.II.4 en annexe, voir le corollaire B.55. ■

Exemple II.105 (Dualité lisse / fortement convexe). Soit $f(x) = \frac{1}{2}\langle Ax, x\rangle$ où A est une matrice symétrique définie positive, et notons μ, L ses plus petite et grande valeur propres, respectivement. En passant par la hessienne, on voit que $f \in \Gamma_\mu(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$. Si maintenant on prend sa conjuguée, on sait d'après l'exemple II.86 que f^* est la fonction quadratique associée à A^{-1} , dont la plus petite valeur propre est L^{-1} et la plus grande est μ^{-1} . Ceci illustre pourquoi $f^* \in \Gamma_{L^{-1}}(\mathbb{R}^N) \cap C_{\mu^{-1}}^{1,1}(\mathbb{R}^N)$.

Exemple II.106 (Régularisation de Moreau-Yosida). Une technique classique en optimisation consiste à remplacer une fonction f non lisse qui nous embête par une fonction $\hat{f} \simeq f$ qui est lisse. La régularisation de Moreau-Yosida consiste à exploiter la dualité lisse/fortement convexe, et le fait qu'il est très facile de rendre une fonction fortement convexe: il suffit de lui rajouter $\frac{\mu}{2}\|x\|^2$. On procède alors ainsi:

$$f \longrightarrow f^* \longrightarrow f^* + \frac{\mu}{2}\|\cdot\|^2 \longrightarrow \left(f^* + \frac{\mu}{2}\|\cdot\|^2\right)^* =: f_\mu.$$

Durant ce processus on a rendu f^* fortement convexe, puis on a pris la conjuguée afin d'obtenir f_μ qui est une fonction μ^{-1} -lisse. Intuitivement on se doute que lorsque $\mu \sim 0$ alors f_μ est proche de f , et il est possible de le prouver (c'est un exercice du TD).

II.III.3 Interlude: preuve de résultats jusque-là admis

On prend ici un instant pour prouver quelques résultats importants que nous avons laissé en suspens dans le chapitre précédent.

Remarque II.107 (Interprétation géométrique de la fonction support II). Nous avons vu dans la remarque II.41 que $\text{dom } \sigma_C$ correspond aux directions de support asymptotique.

On dira qu'une direction $a \in \mathbb{R}^N$ supporte (exactement) C si elle supporte asymptotiquement C , et si $[\langle a, x \rangle = \sigma_C(a)] \cap C \neq \emptyset$. On veut donc que $a \in \text{dom } \sigma_C$ et qu'il existe $c \in C$ tel que $\langle a, c \rangle = \sigma_C(a)$. Puisque $c \in C$ on a $\delta_C(c) = 0$, donc le théorème II.97 de Fenchel-Young nous dit que ceci est équivalent à l'existence de $c \in C$ tel que $a \in N_C(c)$. La formule de Legendre-Fenchel nous dit que c'est équivalent à $c \in \partial\sigma_C(a)$. En d'autres termes, les directions qui supportent C sont exactement $\text{dom } \partial\sigma_C$. Pour toute direction de support $a \in \text{dom } \partial\sigma_C$, l'hyperplan tangent associé est $H(a) := [\langle a, x \rangle = \sigma_C(a)]$, et le sous-différentiel de σ_C peut s'écrire de plusieurs façons:

$$\begin{aligned}\partial\sigma_C(a) &= \{c \in C \mid a \in N_C(c)\} \\ &= \{c \in C \mid \sigma_C(a) = \langle a, c \rangle\} \\ &= H^+(a) \cap C \\ &= \{c \in C \mid \text{proj}_C(c + a) = c\}.\end{aligned}$$

On dit en général que $\partial\sigma_C(a)$ est la *face exposée* de C par a , notée $F_C(a)$, une terminologie que l'on comprend facilement lorsqu'on fait un dessin. D'après ce qui précède, $F_C = N_C^{-1}$.

Corollaire II.108. *Le théorème I.19 sur la caractérisation des convexes comme intersection de demi-espaces est vrai.*

Démonstration. Soit $\mathcal{D} = \text{dom } F_C$, et pour $d \in \mathcal{D}$ on note $H^+(d) = [\langle d, x \rangle \leq \sigma_C(d)]$. Montrons que $C = \cap_{d \in \mathcal{D}} H^+(d)$ par double inclusion.

\subset : Si $x \in C$, alors pour tout $d \in \mathcal{D}$ il est trivial par définition de σ_C de voir que $\langle d, x \rangle \leq \sigma_C(d)$, et donc que $x \in H^+(d)$.

\supset : Soit $x \in \cap_{d \in \mathcal{D}} H^+(d)$ et posons $p = \text{proj}_C(x)$. D'après la caractérisation du cône normal par la projection (voir proposition I.52) nous avons $x - p \in N_C(p)$, donc $p \in F_C(x - p)$ et donc $\sigma_C(x - p) = \langle x - p, p \rangle$ (voir remarque II.107). Or par hypothèse $x \in H^+(x - p)$, donc $\langle x - p, x \rangle \leq \sigma_C(x - p)$. Ceci équivaut à $\|x - p\|^2 \leq 0$, c'est-à-dire que $x = p \in C$. ■

Corollaire II.109. *Le théorème I.39 sur la caractérisation des cônes comme intersection de demi-espaces linéaires est vrai.*

Démonstration. Soit K un cône fermé non vide. Nous avons vu dans la preuve du corollaire II.108 que l'on peut écrire $K = \cap_{d \in \mathcal{D}} H^+(d)$. Il nous suffit donc de montrer que ces demi-espaces sont linéaires, c'est à dire que $\sigma_K(d) = 0$ pour tout $d \in \mathcal{D} = \text{dom } \partial\sigma_K$. Puisque K est un cône, on peut montrer (voir TD) que $\sigma_K = \delta_{K^*}$. Or si $d \in \text{dom } \partial\sigma_K \subset \text{dom } \sigma_K = \text{dom } \delta_{K^*}$ cela veut dire que $\sigma_K(d) = \delta_{K^*}(d) = 0$. ■

Corollaire II.110. *Le théorème I.45 du cône bipolaire est vrai.*

Démonstration. On a vu dans l'exemple II.86 que si K est un cône convexe fermé, alors $\delta_K^* = \delta_{K^*}$. Si on applique deux fois ce résultat on voit que $\delta_{K^{**}} = \delta_{K^*}^* = \delta_K^{**}$. Puisque $\delta_K \in \Gamma_0(\mathbb{R}^N)$ on peut appliquer le théorème de la biconjuguée, et déduire que $\delta_{K^{**}} = \delta_K$. Ceci est équivalent à dire que $K^{**} = K$. ■

II.III.4 Dualité de Fenchel-Rockafellar

Tout au long de cette section, on considérera le problème d'optimisation

$$\min_{x \in \mathbb{R}^N} f(x) + g(Ax), \quad (\text{P})$$

où $f \in \Gamma_0(\mathbb{R}^N)$, $g \in \Gamma_0(\mathbb{R}^M)$ et $A \in \mathcal{M}_{M,N}(\mathbb{R})$. On dira que ce problème est **régulier** si il est régulier en toute solution: pour toute solution $x \in \operatorname{argmin} g + g \circ A$, la règle de calcul sous-différentiel s'applique: $\partial(f + g \circ A)(x) = \partial f(x) + A^* \partial g(Ax)$.

Remarque II.111 (Problème régulier). On a vu dans le corollaire II.73 qu'une condition suffisante pour qu'un problème soit régulier est que $\operatorname{cont} g \cap A(\operatorname{dom} f) \neq \emptyset$. On a aussi évoqué dans la remarque II.74 qu'une autre condition suffisante est le fait que les fonctions f et g soient polyédrales, ou que g soit différentiable.

Lemme II.112 (du Lagrangien). Soit L le Lagrangien associé à (P):

$$L : \mathbb{R}^N \times \mathbb{R}^M \longrightarrow \overline{\mathbb{R}}, \quad L(x, u) := f(x) + \langle u, Ax \rangle - g^*(u).$$

Alors

$$\sup_{u \in \mathbb{R}^M} L(x, u) = f(x) + g(Ax) \quad \text{et} \quad \inf_{x \in \mathbb{R}^N} L(x, u) = -f^*(-A^*u) - g^*(u).$$

Démonstration. On écrit

$$\sup_u L(x, u) = f(x) + \sup_u \langle u, Ax \rangle - g^*(u) = f(x) + g^{**}(Ax) = f(x) + g(Ax).$$

et

$$-\inf_x L(x, u) = g^*(u) + \sup_x -\langle A^*u, x \rangle - f(x) = g^*(u) + f^*(-A^*u).$$



Définition II.113 (Problème dual). On dit que le **PROBLÈME DUAL** de (P) est

$$\min_{u \in \mathbb{R}^M} f^*(-A^*u) + g^*(u). \quad (\text{D})$$

On dira que (x, u) est une **SOLUTION PRIMALE-DUALE** si x est solution de (P) et u est une solution de (D).

Remarque II.114 (Problème bidual). Si on prend le problème dual du dual, on retombe sur le problème primal! En effet, le problème «bidual» est

$$\min_{x \in \mathbb{R}^N} g^{**}(-A^{**}x) + f^{**}(x) = g(Ax) + f(x).$$

Remarque II.115 (Dualité des problèmes primal/dual). Les problèmes (P) et (D) échangent des propriétés importantes.

- 1) La dimension du problème: le problème primal est dans \mathbb{R}^N , le dual est dans \mathbb{R}^M . Pour de nombreux problèmes, N correspond à la taille du modèle, le nombre de paramètres, tandis que M correspond au nombre de données, le nombre de contraintes. Donc dans un régime où $N \ll M$ (big data) ou $M \ll N$ (deep learning) on trouvera plus facile de minimiser l'un ou l'autre de ces problèmes.
- 2) La nature du problème: lissité vs. forte convexité. Par exemple, si f et g sont fortement convexes alors le problème dual sera lisse. De même, si g est lisse alors le problème dual sera fortement convexe.

Exemple II.116 (Problème dual).

- Le problème de minimiser f sur une contrainte d'égalités affines $[Ax = b]$ est un cas particulier de (P), en prenant $g = \delta_b$. Dans ce cas le problème dual s'écrit

$$\min_{u \in \mathbb{R}^M} f^*(-A^*u) + \langle b, u \rangle.$$

On voit que si f est fortement convexe alors ce problème dual est lisse, et sans contraintes ! on pourra donc envisager de le résoudre avec des méthodes d'optimisation classique.

- Le problème de minimiser une somme séparée $f(x) + g(y)$ sous une contrainte linéaire $[Ax + By = c]$ est équivalent à minimiser $F(z) + G(\Phi z)$, en prenant $F = f \oplus g$, $G = \delta_c$ et $\Phi = [A|B]$. Dans ce cas le problème dual associé est

$$\min_u F^*(-\Phi^*u) + G^*(u) = f^*(-A^*u) + g^*(-B^*u) + \langle c, u \rangle.$$

- Le problème de minimiser une somme $f(x) + g(x)$ est un cas particulier de (P) en prenant $A = I$. Dans ce cas le problème dual s'écrit

$$\min_{u \in \mathbb{R}^N} f^*(-u) + g^*(u).$$

Proposition II.117 (Conditions d'optimalité primale-duale). Supposons que le problème soit régulier. Soit $(x, u) \in \mathbb{R}^N \times \mathbb{R}^M$. Alors les propriétés suivantes sont équivalentes:

- 1) (x, u) est une solution primale-duale;

- 2) (x, u) vérifie la condition nécessaire d'optimalité primale-duale: $\begin{cases} x \in \partial f^*(-A^*u), \\ u \in \partial g(Ax). \end{cases}$

Démonstration. Nous utiliserons le Lagrangien L introduit dans le lemme II.112, et nous noterons $p(x) := f(x) + g(Ax)$ et $d(u) = g^*(u) + f^*(-A^*u)$ les fonctions primale et duale. On va procéder en plusieurs étapes.

- Montrons que ii) $\Leftrightarrow (x, u)$ est un minimax de L . Si on utilise le théorème II.77 de Fermat avec la règle de somme simple (proposition II.71) alors on voit que

$$\begin{aligned} & \left\{ \begin{array}{l} x \in \partial f^*(-A^*u) \\ u \in \partial g(Ax) \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} -A^*u \in \partial f(x) \\ Ax \in \partial g^*(u) \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} 0 \in \partial f(x) + A^*u \\ 0 \in \partial g^*(u) - Ax \end{array} \right. \\ & \Leftrightarrow \left\{ \begin{array}{l} x \in \operatorname{argmin} f + \langle A^*u, \cdot \rangle \\ u \in \operatorname{argmin} g^* - \langle Ax, \cdot \rangle \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} x \in \operatorname{argmin} L(\cdot, u) \\ u \in \operatorname{argmax} L(x, \cdot). \end{array} \right. \end{aligned}$$

- Montrons que ii) \Rightarrow i) et $\inf p = -\inf d$. D'après le point précédent, on sait que x minimise $L(\cdot, u)$ et u maximise $L(x, \cdot)$. Or on a vu dans le lemme II.112 que $p(x) = \sup L(x, \cdot)$ et $-d(u) = \inf L(\cdot, u)$. On en déduit donc que

$$\left\{ \begin{array}{l} x \in \operatorname{argmin} L(\cdot, u) \\ u \in \operatorname{argmax} L(x, \cdot) \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} L(x, u) = \inf L(\cdot, u) = -d(u) \\ L(x, u) = \sup L(x, \cdot) = p(x), \end{array} \right.$$

et en particulier que $p(x) = -d(u)$. On peut alors conclure que (x, u) est solution primale-duale:

$$(\forall x' \in \mathbb{R}^N) \quad p(x) = -d(u) \leq L(x', u) \leq \sup L(x', \cdot) = p(x'),$$

$$(\forall u' \in \mathbb{R}^M) \quad d(u) = -p(x) \leq -L(x, u') \leq -\inf L(\cdot, u') = d(u'),$$

et au passage on voit que $\inf p = p(x) = -d(u) = -\inf d$.

- Montrons que i) \Rightarrow ii). Puisque x est solution primale et que l'on a supposé le problème régulier, on peut utiliser le corollaire II.73 pour obtenir $0 \in \partial f(x) + A^*\partial g(Ax)$. Cette condition d'optimalité nous dit qu'il existe $u' \in \partial g(Ax)$ tel que $0 \in \partial f(x) + A^*u'$. En réorganisant cette inclusion, on voit que $x \in \partial f^*(-A^*u')$, ce qui veut dire que (x, u') vérifie la CNO ii). D'après un point précédent, on en déduit que $\inf p = -\inf d$. On peut maintenant s'intéresser à notre solution primale-duale (x, u) . On écrit (on utilise le lemme II.112)

$$-\inf d = -d(u) = \inf L(\cdot, u) \leq L(x, u) \leq \sup L(x, \cdot) = p(x) = \inf p.$$

Or on vient de voir que $\inf p = -\inf d$, donc toutes les inégalités sont des égalités ! En particulier $\inf L(\cdot, u) = L(x, u)$, ce qui veut dire que $x \in \operatorname{argmin} L(\cdot, u)$; et $L(x, u) = \sup L(x, \cdot)$, ce qui veut dire que $u \in \operatorname{argmax} L(\cdot, u)$. D'après le premier point de la preuve, on conclut que (x, u) vérifie la CNO.



Théorème II.118 (de représentation primale-duale). Supposons que le problème (P) admette une solution, et qu'il soit régulier. Alors le problème dual (D) admet une solution, et on dispose de la formule de représentation primale-duale: pour n'importe quelle solution u de (D),

$$\operatorname{argmin} p = \partial f^*(-A^*u) \cap A^{-1}\partial g^*(u).$$

Démonstration. Commençons par vérifier l'existence d'une solution duale. L'existence d'une solution primale x et l'hypothèse de régularité nous permet d'écrire la condition d'optimalité $0 \in \partial f(x) + A^* \partial g^*(Ax)$, ce qui nous donne l'existence d'un $u \in \partial g(Ax)$ tel que $0 \in \partial f(x) + Au$. On voit donc que (x, u) vérifie la condition nécessaire d'optimalité primale-duale, et donc que u est une solution duale d'après la proposition II.117. Maintenant considérons x' quelconque, et utilisons encore la proposition II.117 pour écrire que x' est solution primale si et seulement si (x', u) est solution primale-duale, si et seulement si

$$\begin{cases} x' \in \partial f^*(-A^*u) \\ u \in \partial g(Ax') \end{cases} \Leftrightarrow \begin{cases} x' \in \partial f^*(-A^*u) \\ Ax' \in \partial g^*(u) \end{cases} \Leftrightarrow \begin{cases} x' \in \partial f^*(-A^*u) \\ x' \in A^{-1}\partial g^*(u). \end{cases}$$

D'où le résultat. ■

Exemple II.119 (Le cas des problèmes fortement convexes). Si f est fortement convexe et g continue, alors (P) admet une unique solution \bar{x} que l'on peut calculer avec la formule de représentation primale-duale. En effet f^* est lisse ce qui veut dire que $\partial f^* = \nabla f^*$, et donc

$$\bar{x} = \nabla f^*(-A^*\bar{u}) \cap A^{-1}\partial g^*(\bar{u}) = \nabla f^*(-A^*\bar{u}).$$

Autrement dit, si on sait résoudre le problème dual (D) et calculer une solution \bar{u} , on obtient automatiquement une solution primale avec la formule $\bar{x} = \nabla f^*(-A^*\bar{u})$.

On termine cette section avec quelques exemples d'application du théorème de représentation primale-duale. C'est un peu hors programme, mais cela peut être intéressant à lire pour comprendre l'intérêt de ce résultat.

Corollaire II.120 (Théorème de représentation linéaire). *Considérons le problème de minimiser $\frac{\mu}{2}\|x\|^2 + g(Ax)$, où $g \in \Gamma_0(\mathbb{R}^M)$ et $A \in \mathcal{M}_{M,N}(\mathbb{R})$. Supposons que le problème soit régulier, et soit $\bar{x} \in \mathbb{R}^N$ son unique solution. Alors on peut écrire*

$$\bar{x} = \sum_{j=1}^M \alpha_j a_j,$$

où $\alpha \in \mathbb{R}^M$ et a_j^* est la j -ième ligne de A .

Démonstration. Le problème admet une unique solution puisqu'il est fortement convexe. On peut donc appliquer le théorème II.118 de représentation primale duale, qui nous dit en particulier que $\bar{x} \in \partial f^*(-A^*u)$, où $f(x) = \frac{\mu}{2}\|x\|^2$ et $u \in \mathbb{R}^M$. Puisque $f^*(v) = \frac{1}{2\mu}\|v\|^2$, on a $\nabla f^*(-A^*u) = \frac{-1}{\mu}A^*u$, où A^*u est égal à $\sum_{j=1}^M u_j a_j$. D'où le résultat avec $\alpha_j = \frac{-1}{\mu}u_j$. ■

Remarque II.121 (Théorème de représentation linéaire et astuce du noyau (Kernel Trick)). Ce résultat² est très important en machine learning, où l'on cherche souvent à minimiser

²Apparu pour la première fois dans *Learning with kernels*, B. Scholkopf and A. J. Smola, MIT Press, 2002.

un risque empirique régularisé de la forme

$$\frac{\mu}{2} \|x\|^2 + D(Ax; y),$$

où D joue le rôle d'une distance. En pratique, notamment dans le régime du deep learning, N représente le nombre de paramètres et est bien supérieur au nombre de données M . Dans certains cas, on a même moralement $N = +\infty$ car on travaille sur un espace de modèles de dimension infinie. C'est typiquement le cas pour les méthodes à noyaux Gaussiens, qui sont efficaces pour de nombreux problèmes simples. Dans ce contexte il est impossible d'implémenter directement des vecteurs x d'un espace de Hilbert de dimension infinie.

Mais grâce à ce théorème de représentation, on sait que notre solution vit dans un espace vectoriel de dimension M engendré par les M lignes de A . Il suffit de résoudre le problème d'optimisation associé dans \mathbb{R}^M , où l'inconnue est ici $\alpha \in \mathbb{R}^M$. Dans le cadre de ce cours, on comprend que cela revient à résoudre un problème dual. Dans le cadre du machine learning, cette technique s'appelle l'astuce du noyau (Kernel Trick).

Corollaire II.122 (Théorème de représentation parcimonieuse). *Considérons le problème de minimiser $\mu\|x\|_1 + g(Ax)$, où $g \in \Gamma_0(\mathbb{R}^M)$ et $A \in \mathcal{M}_{M,N}(\mathbb{R})$. Supposons que le problème soit régulier, et qu'il admette une unique solution $\bar{x} \in \mathbb{R}^N$. Alors on peut écrire*

$$\bar{x} = \sum_{i \in I} \alpha_i e_i,$$

où $I \subset \{1, \dots, N\}$ est de taille $|I| \leq M$, $\alpha_i \in \mathbb{R}$ et e_i est le i -ième vecteur de la base canonique.

Remarque II.123 (Théorème de représentation parcimonieuse). Ce résultat énonce que les solutions du problème sont parcimonieuses: quand bien même \bar{x} vit dans \mathbb{R}^N , il ne dispose que de M coordonnées non nulles. Ce résultat est fortement exploité en sciences des données, lorsque on cherche une solution d'un système linéaire $Ax = b$ qui soit la plus parcimonieuse possible.

Démonstration du corollaire II.122. Notons $I = \text{supp}(\bar{x}) := \{i \in \{1, \dots, N\} \mid \bar{x}_i \neq 0\}$ le support de \bar{x} . L'objectif de la preuve est de montrer que $|I| \leq M$. Nous allons faire cela de manière indirecte, en montrant que la famille $(a_i)_{i \in I} \subset \mathbb{R}^M$ est libre, où a_i est la i -ème colonne de A . En effet une famille libre dans \mathbb{R}^M ne peut pas avoir plus d'éléments que M .

Le théorème II.118 de représentation primale duale nous dit que il existe $u \in \mathbb{R}^M$ tel que $\bar{x} \in \partial f^*(-A^*u) \cap A^{-1}\partial g^*(u)$, avec ici $f(x) = \mu\|x\|_1$. Quitte à diviser notre problème par μ , on peut supposer sans perte de généralité que $\mu = 1$. Notre hypothèse d'unicité dit que \bar{x} doit être l'unique vecteur dans cette intersection. Supposons donc par l'absurde que la famille $(a_i)_{i \in I} \subset \mathbb{R}^M$ est liée, et montrons que l'on peut construire une autre solution distincte de \bar{x} .

Que $(a_i)_{i \in I} \subset \mathbb{R}^M$ soit liée veut dire qu'il existe $d \in \mathbb{R}^N$ non nul tel que $Ad = 0$ et $\text{supp}(d) \subset I$. Considérons $x' = \bar{x} + td$ pour $t \in \mathbb{R}$. Il est clair que $Ax' = A\bar{x} + tAd = A\bar{x} \in \partial g^*(u)$. Donc il nous reste à montrer que $x' \in \partial f^*(-A^*u)$ pour un certain t bien choisi. Dans ce problème on a $f(x) = \|x\|_1$ donc $f^* = \delta_{\mathbb{B}_\infty}$ et $\partial f^*(-A^*u) = N_{\mathbb{B}_\infty}(-A^*u)$. Notons que on sait déjà que $\bar{x} \in N_{\mathbb{B}_\infty}(-A^*u)$ donc ce cône normal est non vide. De plus c'est le cône normal à un ensemble produit, donc il nous faut donc montrer que $x'_i \in N_{[-1,1]}(-\langle a_i, u \rangle)$ pour tout $i \leq N$.

- Si $i \notin I$: cela veut dire que $\bar{x}_i = 0$ mais aussi $d_i = 0$, donc $x'_i = 0$. Donc c'est gagné puisque on a toujours 0 dans un cône non vide.
- Si $i \in I$: on sait déjà que $0 \neq \bar{x}_i \in N_{[-1,1]}(-\langle a_i, u \rangle)$. Cela veut donc dire que ce cône normal n'est pas réduit à $\{0\}$, et même plus précisément que $-\langle a_i, u \rangle = \text{sgn}(\bar{x}_i)$. Existe-t-il un t tel que $-\langle a_i, u \rangle = \text{sgn}(\bar{x}_i + td_i)$? La réponse est oui: il suffit de prendre t suffisamment petit, par exemple $t = \min_{d_i \neq 0} \frac{|\bar{x}_i|}{2|d_i|}$ qui est tel que $\text{sgn}(\bar{x}_i + td_i) = \text{sgn}(\bar{x}_i)$.



Chapitre III

Algorithmes d'éclatement pour l'optimisation convexe

III.I Algorithmes élémentaires

III.I.1 Algorithme du Gradient

On considère ici le problème de minimiser une fonction $f \in \Gamma_0(\mathbb{R}^N)$ lisse.

Définition III.1 (Algorithme du gradient (1847)). Soit $f \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$. L'**ALGORITHME DU GRADIENT** génère une suite $(x_n)_{n \in \mathbb{N}} \subset \mathbb{R}^N$ telle que

$$x_{n+1} = x_n - \lambda \nabla f(x_n), \quad (\text{G})$$

où $\lambda > 0$ est appelé le pas de l'algorithme.

Remarque III.2 (Gradient et sous-niveaux). L'algorithme du gradient peut s'interpréter géométriquement, en regardant les sous-niveaux de la fonction f . En effet, étant donné un point x_n et le sous-niveau associé $S_{x_n} = [f(x) \leq f(x_n)]$, on peut montrer que le cône normal au sous-niveau est la demi-droite engendrée par le gradient (admis, cf Proposition B.49 en Annexe) :

$$N_{S_{x_n}}(x_n) = \mathbb{R}_+ \nabla f(x_n).$$

Ainsi, on peut voir que $-\nabla f(x_n)$ pointe vers l'intérieur du sous-niveau. Une stratégie raisonnable est donc de suivre cette direction, ou du moins pendant un certain temps, afin de rentrer dans l'intérieur du sous-niveau, et ainsi faire décroître f . Le pas λ sert justement à contrôler combien de temps on suit cette direction ; on veut qu'il soit le plus grand possible (pour se déplacer le plus loin possible de x_n) mais pas trop grand non plus (pour éviter de ressortir du sous-niveau).

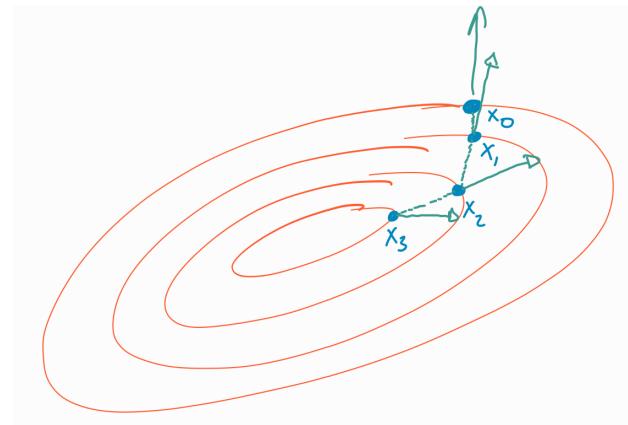


FIGURE III.1 – Quelques itérés de l’algorithme du gradient. On suit à chaque pas la direction donnée par le gradient précédent.

Remarque III.3 (Flot gradient). Le flot gradient associé à f est l’équation différentielle ordinaire générant¹ des trajectoires $x(t) \in \mathbb{R}^N$ satisfaisant

$$\dot{x}(t) + \nabla f(x(t)) = 0.$$

On peut se convaincre que ce système dynamique est important du point de vue de la minimisation de f . Par exemple, si $x(t)$ converge vers \bar{x} avec une vitesse évanescante ($\dot{x}(t) \rightarrow 0$), alors on voit en passant à la limite que $\nabla f(\bar{x}) = 0$, c’est-à-dire que \bar{x} est une solution. On peut aussi voir que f décroît le long de la trajectoire, puisque

$$\frac{d}{dt}(f \circ x)(t) = \langle \nabla f(x(t)), \dot{x}(t) \rangle = -\|\dot{x}(t)\|^2 \leq 0.$$

Ceci n’est pas surprenant puisque, à chaque temps t , nous avons que le vecteur vitesse $\dot{x}(t) = -\nabla f(x(t))$ qui pointe vers l’intérieur du sous-niveau de f .

Remarque III.4 (Discrétisation du flot gradient : schéma explicite). Si on réalise une *discrétisation explicite* du flot gradient, on obtient

$$\frac{x_{n+1} - x_n}{\lambda} + \nabla f(x_n) = 0.$$

En réarrangeant les termes, on obtient exactement $x_{n+1} = x_n - \lambda \nabla f(x_n)$, c’est-à-dire l’algorithme du gradient. noter que l’on parle de discrétisation *explicite* parce que x_{n+1} se calcule explicitement comme une fonction de x_n .

¹Noter que f étant lisse alors ∇f est Lipschitz, donc cette EDO admet des solutions d’après le Théorème de Cauchy-Lipschitz !

Théorème III.5 (Convergence du Gradient). Soit $f \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$, telle que $\operatorname{argmin} f \neq \emptyset$. Soit $(x_n)_{n \in \mathbb{N}}$ générée par l'algorithme du gradient, avec un pas $0 < \lambda < 2/L$. Alors x_n converge vers $\bar{x} \in \operatorname{argmin} f$ lorsque $n \rightarrow +\infty$.

Démonstration. C'est un cas particulier du Théorème III.27 que l'on verra plus loin. ■

Exemple III.6 (Algorithme de Landweber). Si $f(x) = \frac{1}{2}\|Ax - b\|^2$, alors $\nabla f(x) = A^\top(Ax - b)$ et $\nabla^2 f(x) = A^\top A$. En particulier, f est L -lisse, avec $L = \operatorname{Lip}(\nabla f) = \|A^\top A\| = \|A\|^2$. Donc avec un pas $0 < \lambda < 2/\|A\|^2$, la suite générée par

$$x_{n+1} = x_n - \lambda A^\top(Ax_n - b)$$

converge vers un minimiseur de f . Cet algorithme est connu sous le nom d'*algorithme de Landweber*.

III.I.2 Algorithme Proximal

Définition III.7 (Prox). Soit $f \in \Gamma_0(\mathbb{R}^N)$. On définit l'**OPÉRATEUR PROXIMAL** de f comme étant la fonction $\operatorname{prox}_f : \mathbb{R}^N \rightarrow \mathbb{R}^N$ définie par

$$\operatorname{prox}_f(x) = \operatorname{argmin}_{y \in \mathbb{R}^N} f(y) + \frac{1}{2}\|y - x\|^2.$$

Définition III.8 (Algorithme Proximal (1976)). Soit $f \in \Gamma_0(\mathbb{R}^N)$. L'**ALGORITHME PROXIMAL** génère une suite $(x_n)_{n \in \mathbb{N}} \subset \mathbb{R}^N$ telle que

$$x_{n+1} = \operatorname{prox}_{\lambda f}(x_n), \quad (\text{G})$$

où $\lambda > 0$ est appelé le pas de l'algorithme.

Proposition III.9 (Caractérisation du prox par le sous-différentiel). Soit $f \in \Gamma_0(\mathbb{R}^N)$, et $\lambda > 0$. Soient $x, p \in \mathbb{R}^N$. Alors

$$p = \operatorname{prox}_{\lambda f}(x) \Leftrightarrow x - p \in \lambda \partial f(p).$$

Démonstration. D'après le théorème II.77 de Fermat

$$p = \operatorname{prox}_{\lambda f}(x) \Leftrightarrow 0 \in \partial \left(\lambda f + \frac{1}{2}\|\cdot\|^2 \right)(p),$$

où $\partial \left(\lambda f + \frac{1}{2}\|\cdot\|^2 \right)(p) = \lambda \partial f(p) + p - x$ par la règle de somme simple (proposition II.71). Ceci est donc équivalent à $x - p \in \lambda \partial f(p)$. ■

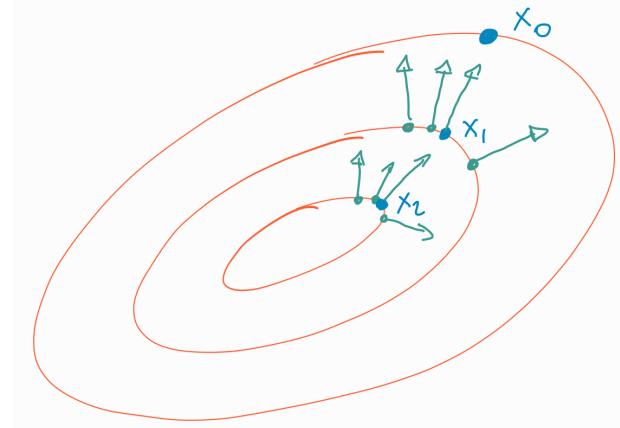


FIGURE III.2 – Quelques itérés de l'algorithme proximal. On suit à chaque pas la direction donnée par le gradient en le futur point.

Remarque III.10 (Discrétisation du flot gradient : schéma implicite). Si on réalise une *discrétisation implicite* du flot gradient, on obtient

$$\frac{x_{n+1} - x_n}{\lambda} + \nabla f(x_{n+1}) = 0.$$

En réarrangeant les termes, on obtient exactement $x_{n+1} = x_n - \lambda \nabla f(x_{n+1})$, c'est-à-dire l'algorithme du gradient *implicite*. Si on réarrange les termes on voit que $x_n - x_{n+1} = \lambda \nabla f(x_{n+1})$, ce qui équivaut à dire d'après la proposition précédente que $x_{n+1} = \text{prox}_{\lambda f}(x_n)$. On voit ainsi que l'algorithme proximal n'est rien d'autre que l'algorithme du gradient implicite. Mais il est important de noter que l'algorithme proximal n'a aucunement besoin que f soit différentiable !

Théorème III.11 (Convergence du proximal). Soit $f \in \Gamma_0(\mathbb{R}^N)$, telle que $\text{argmin } f \neq \emptyset$. Soit $(x_n)_{n \in \mathbb{N}}$ générée par l'algorithme proximal, avec un pas $0 < \lambda$. Alors x_n converge vers $\bar{x} \in \text{argmin } (f)$ lorsque $n \rightarrow +\infty$.

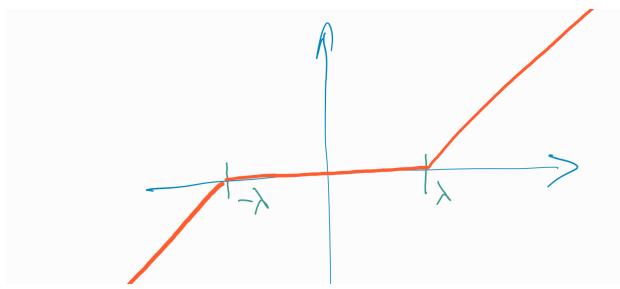
Démonstration. C'est un cas particulier du Théorème III.27 que l'on verra plus loin. ■

III.I.3 Calcul proximal

Ici on s'intéresse à comment calculer le prox d'une fonction en pratique.

Exemple III.12 (Exemples simples de prox).

- Si $f(x) = \delta_C(x)$ où C est convexe fermé non vide, alors $\text{prox}_{\lambda f}(x) = \text{proj}_C(x)$. On peut donc voir l'opérateur proximal comme une généralisation de la notion de projection pour les fonctions.

FIGURE III.3 – Le graphe de soft_λ .

- En particulier, si $f(x) = \delta_0(x)$, alors $\text{prox}_{\lambda f}(x) \equiv 0$; et si $f(x) = 0$ alors $\text{prox}_{\lambda f}(x) = x$.
- Si $f(x) = |x|$, alors on peut calculer (voir TD) que $\text{prox}_{\lambda f}(x) = \text{soft}_\lambda(x)$, où

$$\text{soft}_\lambda(x) = \begin{cases} x + \lambda & \text{if } x \leq -\lambda, \\ 0 & \text{if } x \in [-\lambda, \lambda], \\ x - \lambda & \text{if } x \geq \lambda. \end{cases}$$

L'opérateur soft_λ est appelé le *seuillage doux* (soft thresholding en VO). Il peut se voir comme un filtre passe-haut : il met à zéro les faibles valeurs, et laisse passer (tout en les atténuant) les valeurs plus élevées.

- Si $f(x) = \frac{1}{2}\|x\|^2$, alors $\text{prox}_{\lambda f}(x) = \frac{1}{1+\lambda}x$.
- Si $f(x) = \frac{1}{2}\langle Ax, x \rangle$, où $A \in \mathcal{M}_N(\mathbb{R})$ est une matrice symétrique semi-définie positive, alors $\text{prox}_{\lambda f}(x) = (I + \lambda A)^{-1}(x)$.

Remarque III.13 (Fonction prox-amicale). Par définition, calculer le prox nécessite de résoudre un problème d'optimisation qui fait intervenir f . Donc c'est à priori aussi dur à calculer que de minimiser f elle-même. En général, calculer $\text{prox}_f(x)$ de façon explicite est impossible. Mais lorsque une formule explicite existe (comme pour les exemples ci-dessus), alors on dira que la fonction f est *prox-amicale*.

Proposition III.14 (Règles de calcul du prox). Soient $g \in \Gamma_0(\mathbb{R}^N)$ et $\lambda > 0$.

- 1) Si $f(x) = g(x) + c$, où $c \in \mathbb{R}$, alors $\text{prox}_{\lambda f}(x) = \text{prox}_{\lambda g}(x)$.
- 2) Si $f(x) = g(x + a)$, où $a \in \mathbb{R}^N$, alors $\text{prox}_{\lambda f}(x) = \text{prox}_{\lambda g}(x + a) - a$.
- 3) Si $f(x) = g(x) + \langle a, u \rangle$, où $a \in \mathbb{R}^N$, alors $\text{prox}_{\lambda f}(x) = \text{prox}_{\lambda g}(x - \lambda a)$.
- 4) Si $f(x) = g(ax)$, où $a \in \mathbb{R}_*$, alors $\text{prox}_{\lambda f}(x) = \frac{1}{a} \text{prox}_{\lambda a^2 g}(ax)$.
- 5) Si $f(x) = g(Ax)$ où $A \in \mathcal{M}_N(\mathbb{R})$ est orthogonale, alors $\text{prox}_{\lambda f}(x) = A^\top \text{prox}_{\lambda g}(Ax)$.

Démonstration. Voir TD.

Proposition III.15 (Prox de la somme séparable). Soit $f(x) = \sum_{i=1}^M f_i(x_i)$, où $f_i \in \Gamma_0(\mathbb{R}^{N_i})$. Alors

$$\text{prox}_{\lambda f}(x) = \left(\text{prox}_{\lambda f_i}(x_i) \right)_{i=1}^M.$$

Démonstration. On utilise la définition du prox et on utilise le fait que l'argmin porte sur une somme séparable:

$$\begin{aligned} \text{prox}_{\lambda f}(x) &= \underset{y \in \mathbb{R}^N}{\operatorname{argmin}} f(y) + \frac{1}{2\lambda} \|y - x\|^2 \\ &= \underset{y_1, \dots, y_M \in \mathbb{R}^{N_1} \times \mathbb{R}^{N_M}}{\operatorname{argmin}} \sum_{i=1}^M f_i(y_i) + \frac{1}{2\lambda} \sum_{i=1}^M \|y_i - x_i\|^2 \\ &= \prod_{i=1}^M \underset{y_i \in \mathbb{R}^{N_i}}{\operatorname{argmin}} f_i(y_i) + \frac{1}{2\lambda} \sum_{i=1}^M \|y_i - x_i\|^2 \\ &= \prod_{i=1}^M \text{prox}_{\lambda f_i}(x_i). \end{aligned}$$

Remarque III.16 (Prox de la somme). Malheureusement il n'existe aucune formule permettant de calculer le prox d'une somme $f + g$ en fonction de prox_f et prox_g . En particulier les formules « naïves » suivantes ne marchent pas :

$$\text{prox}_{f+g} \neq \text{prox}_f + \text{prox}_g \quad \text{et} \quad \text{prox}_{f+g} \neq \text{prox}_f \circ \text{prox}_g.$$

On peut le voir en prenant $f = \delta_C$ et $g = \delta_D$, tels que $\text{prox}_{f+g} = \text{proj}_{C \cap D}$ qui est différent de $\text{proj}_C + \text{proj}_D$ et $\text{proj}_C \circ \text{proj}_D$. C'est par exemple évident à voir si on prend pour C et D deux droites linéaires de \mathbb{R}^2 .

Exemple III.17 (Prox de la norme L^1). Si $f(x) = \|x\|_1$ alors son opérateur proximal applique le seuillage doux soft_λ coordonnées par coordonnées. On peut noter $\text{Soft}_\lambda := \text{prox}_{\lambda \|\cdot\|_1}$, qui vérifie donc

$$\text{Soft}_\lambda(x) = (\text{soft}_\lambda(x_i))_{i=1}^N,$$

autrement dit $\text{Soft}_\lambda(x)$ est le vecteur défini par

$$(\text{Soft}_\lambda(x))_i = \begin{cases} x_i + \lambda & \text{if } x_i \leq -\lambda, \\ 0 & \text{if } x_i \in [-\lambda, \lambda], \\ x_i - \lambda & \text{if } x_i \geq \lambda. \end{cases}$$

Théorème III.18 (de décomposition de Moreau). Soit $f \in \Gamma_0(\mathbb{R}^N)$. Alors pour tout $x \in \mathbb{R}^N$,

$$x = \text{prox}_f(x) + \text{prox}_{f^*}(x) \quad \text{et} \quad \langle \text{prox}_f(x), \text{prox}_{f^*}(x) \rangle = f(\text{prox}_f(x)) + f^*(\text{prox}_{f^*}(x)).$$

Démonstration. Soient $x \in \mathbb{R}^N$ et $p = \text{prox}_f(x)$. Alors d'après la caractérisation du prox par le sous-différentiel et la formule de Legendre, on a

$$x - p \in \partial f(p) \Leftrightarrow p \in \partial f^*(x - p) \Leftrightarrow x - (x - p) \in \partial f^*(x - p) \Leftrightarrow x - p = \text{prox}_{f^*}(x).$$

On en déduit la première égalité. Ensuite on observe que $\text{prox}_{f^*}(x) = x - \text{prox}_f(x) \in \partial f(\text{prox}_f(x))$. Donc on peut utiliser Fenchel-Young (version égalité) et en déduire la deuxième égalité. ■

Exemple III.19 (Théorèmes de décomposition). Si $f(x) = \delta_F$, où F est un sous-espace vectoriel, alors on retrouve un Théorème de décomposition bien connu sur F et son orthogonal. En effet, $f^* = \delta_{F^\perp}$ et $\text{prox}_f = \text{proj}_F$ et $\text{prox}_{f^*} = \text{proj}_{F^\perp}$. Donc on obtient

$$x = \text{proj}_F(x) + \text{proj}_{F^\perp}(x) \quad \text{et} \quad \langle \text{proj}_F(x), \text{proj}_{F^\perp}(x) \rangle = \delta_F(\text{proj}_F(x)) + \delta_{F^\perp}(\text{proj}_{F^\perp}(x)) = 0.$$

Il est à noter que ce résultat se généralise à tout cône convexe fermé $K \subset \mathbb{R}^N$! En effet avec $f = \delta_K$ on a $f^* = \delta_{K^*}$ et on obtient

$$x = \text{proj}_K(x) + \text{proj}_{K^*}(x) \quad \text{et} \quad \langle \text{proj}_K(x), \text{proj}_{K^*}(x) \rangle = \delta_K(\text{proj}_K(x)) + \delta_{K^*}(\text{proj}_{K^*}(x)) = 0.$$

Proposition III.20 (Prox de la conjuguée : Formule de Moreau). Soit $f \in \Gamma_0(\mathbb{R}^N)$ et $\lambda > 0$. Alors

$$\text{prox}_{\lambda f^*}(x) = x - \lambda \text{prox}_{\frac{1}{\lambda} f}(\frac{x}{\lambda}).$$

Démonstration. C'est une application du théorème de Moreau et des règles de calcul du prox. ■

Remarque III.21 (Prox de la composée). Si $f(x) = g(Ax)$ alors il n'y a pas en général de formule exprimant explicitement prox_f en fonction de prox_g et A . Néanmoins c'est possible lorsque A est orthogonale (voir Proposition III.14.v)), et un peu plus généralement lorsque A est semi-orthogonale :

Proposition III.22 (Prox avec matrice semi-orthogonale). Soient $g \in \Gamma_0(\mathbb{R}^N)$, $A \in \mathcal{M}_{M,N}(\mathbb{R})$ et $f(x) = g(Ax)$. Supposons que A soit semi-orthogonale : $\exists \mu > 0$ tel que $AA^\top = \mu I$. Alors

$$\begin{aligned} \text{prox}_{\lambda f}(x) &= \left(I - \frac{1}{\mu} A^\top A \right) x + \frac{1}{\mu} A^\top \text{prox}_{\lambda \mu g}(Ax), \\ \text{prox}_{\frac{1}{\lambda} f^*}(x) &= A^\top \text{prox}_{\frac{1}{\lambda \mu} g^*} \left(\mu^{-1} Ax \right). \end{aligned}$$

Démonstration. Commençons par prouver la deuxième formule. Posons $v = \text{prox}_{\frac{1}{\lambda\mu}g^*}(\mu^{-1}Ax)$, $p = A^\top v$, et montrons que $p = \text{prox}_{\frac{1}{\lambda}f^*}(x)$. Par définition de v et la caractérisation du prox par le sous-différentiel, nous avons

$$(\mu^{-1}Ax) - v \in \frac{1}{\lambda\mu}\partial g^*(v) \Leftrightarrow \lambda(Ax - \mu v) \in \partial g^*(v) \Leftrightarrow v \in \partial g(\lambda Ax - \lambda\mu v).$$

Or $p = A^\top v$ et $Ap = AA^\top v = \mu v$ donc

$$p \in A^\top \partial g(\lambda Ax - \lambda Ap) \subset \partial f(\lambda x - \lambda p) \Leftrightarrow \lambda x - \lambda p \in \partial f^*(p) \Leftrightarrow x - p \in \frac{1}{\lambda}\partial f^*(p).$$

D'après la caractérisation du prox par le sous-différentiel, ceci est effectivement équivalent à $p = \text{prox}_{\frac{1}{\lambda}f^*}(x)$. Nous allons maintenant déduire la première formule à l'aide de la formule de Moreau :

$$\begin{aligned} \text{prox}_{\lambda f} &= I - \lambda \text{prox}_{\frac{1}{\lambda}f^*} \circ \frac{1}{\lambda}I \quad (\text{Proposition III.20}) \\ &= I - \lambda \left(A^\top \circ \text{prox}_{\frac{1}{\mu\lambda}g^*} \circ \frac{1}{\mu}I \right) \circ \frac{1}{\lambda}I \quad (\text{Formule précédente}) \\ &= I - \lambda A^\top \text{prox}_{\frac{1}{\mu\lambda}g^*} \circ \frac{1}{\mu\lambda}A \\ &= I - \lambda A^\top \left[I - \frac{1}{\mu\lambda} \text{prox}_{\lambda\mu g} \circ \lambda\mu I \right] \circ \frac{1}{\mu\lambda}A \quad (\text{Proposition III.20}) \\ &= I - \lambda A^\top \frac{1}{\mu\lambda}A + \lambda A^\top \frac{1}{\mu\lambda} \text{prox}_{\lambda\mu g} \circ \lambda\mu \frac{1}{\mu\lambda}A \\ &= (I - \frac{1}{\mu}A^\top A) + \frac{1}{\mu} \text{prox}_{\lambda\mu g} \circ A. \end{aligned}$$



III.II Algorithmes d'éclatement

Considérons un problème d'optimisation convexe quelconque, faisant intervenir des fonctions lisses f_i , des fonctions h_j qui sont non lisses mais prox-amicales, et des opérateurs linéaires :

$$(P) \quad \min_{x \in \mathbb{R}^N} f_1(A_1x) + \cdots + f_p(A_px) + h_1(B_1x) + \cdots + h_p(B_px).$$

Notre objectif est d'avoir un algorithme capable de résoudre ce problème, tel qu'à chaque itération nous n'avons besoin que de :

- évaluer le gradient des fonctions lisses: ∇h_j ;

- évaluer le prox des fonctions non lisses: $\text{prox}_{\lambda f_i}$;
- évaluer les opérateurs linéaires A_i, B_j , et leur transposées.

Un tel algorithme sera qualifié d'*algorithme d'éclatement*, car il sera capable de décomposer le problème en ses composantes les plus élémentaires. On parlera d'*éclatement total* pour un algorithme capable de traiter n'importe quelle somme de fonctions, et d'*éclatement composite* total si en plus l'algorithme arrive à traiter les opérateurs linéaires sans avoir à inverser les matrices.

Il faut noter que l'éclatement des fonctions lisses est trivial, car la règle de dérivation en chaîne nous dit que

$$\nabla(h_1 \circ B_1 + \cdots + h_q \circ B_q)(x) = B_1^\top \nabla h_1(B_1 x) + \cdots + B_q^\top \nabla h_q(B_q x).$$

Donc sans perte de généralité on se focalisera sur les problèmes de la forme

$$(P) \quad \min_{x \in \mathbb{R}^N} f_1(A_1 x) + \cdots + f_p(A_p x) + h(x).$$

III.II.1 Éclatement simple : Algorithme du Gradient Proximal

On considère ici le problème

$$(P) \quad \min_{x \in \mathbb{R}^N} f(x) + h(x),$$

où $f, h \in \Gamma_0(\mathbb{R}^N)$ et $h \in C_L^{1,1}(\mathbb{R}^N)$. On définit l'*algorithme du gradient proximal*, qui consiste à réaliser une étape de type gradient par rapport à la partie lisse h , suivie d'une étape de type proximal par rapport à la partie non lisse f .

Définition III.23 (Algorithme Gradient-Proximal (1979)). Soient $f, h \in \Gamma_0(\mathbb{R}^N)$ avec $h \in C_L^{1,1}(\mathbb{R}^N)$, et $\lambda > 0$. L'*algorithme gradient proximal (GP)* génère une suite $(x_n)_{n \in \mathbb{N}} \subset \mathbb{R}^N$ telle que

$$x_{n+1} = \text{prox}_{\lambda f}(x_n - \lambda \nabla h(x_n)).$$

Remarque III.24 (GP généralise gradient et prox). Noter que si $f = 0$ alors (GP) est exactement l'*algorithme du gradient* appliqué à h ; et que si $h = 0$ alors (GP) est exactement l'*algorithme proximal* appliqué à f .

Remarque III.25 (Discréétisation du flot gradient : explicite-implicite). L'*algorithme gradient proximal* peut s'interpréter comme une discréétisation explicite-implicite du flot gradient. En effet si on applique la caractérisation du prox, on voit que

$$x_n - \lambda \nabla h(x_n) - x_{n+1} \in \lambda \partial g(x_{n+1}),$$

ce qui équivaut à

$$\frac{x_{n+1} - x_n}{\lambda} + \partial g(x_{n+1}) + \nabla h(x_n) \ni 0.$$

Autrement dit, on a exploité le fait que $\partial(f + h)(x) = \partial f(x) + \nabla h(x)$, et on a discrétisé de manière explicite par rapport à h , et implicite par rapport à g . C'est une combinaison des approches explicite (algorithme gradient) et implicite (algorithme proximal) vues précédemment.

Exemple III.26. Nous présentons ici quelques cas particuliers importants de l'algorithme du gradient-proximal.

- Si $f(x) = \delta_C(x)$ où $C \subset \mathbb{R}^N$ est convexe fermé non vide, alors on s'intéresse à minimiser une fonction lisse h sur la contrainte C . Dans ce cas $\text{prox}_{\lambda f} = \text{proj}_C$ et l'algorithme gradient-proximal devient

$$x_{n+1} = \text{proj}_C(x_n - \lambda \nabla h(x_n)),$$

connu sous le nom d'*algorithme gradient projeté*.

- Si $f(x) = \delta_C(x)$ et $h(x) = \frac{1}{2} \text{dist}_D(x)^2$, alors on s'intéresse à minimiser $\text{dist}_D(x)$ sur la contrainte D . Autrement dit, on cherche à trouver des éléments dans $C \cap D$, connu sous le nom de *problème de faisabilité*. Dans ce cas $\nabla h(x) = x - \text{proj}_C(x)$ (admis), et l'algorithme gradient-proximal avec $\lambda = 1$ devient

$$x_{n+1} = \text{proj}_C(\text{proj}_D(x_n)),$$

connu sous le nom d'*algorithme de projection alternée*.

- Si $f(x) = \alpha \|x\|_1$, alors on s'intéresse à minimiser $\alpha \|x\|_1 + h(x)$. Dans ce cas $\text{prox}_{\lambda f} = \text{soft}_{\alpha\lambda}$, et l'algorithme gradient-proximal devient

$$x_{n+1} = \text{soft}_{\alpha\lambda}(x_n - \lambda \nabla h(x_n)),$$

connu sous le nom d'*algorithme du gradient seuillé* (en anglais: *ISTA* pour *Iterative Soft Thresholding Algorithm*).

Théorème III.27 (Convergence du Gradient-Proximal). Soient $f, h \in \Gamma_0(\mathbb{R}^N)$ avec $h \in C_L^{1,1}(\mathbb{R}^N)$, et on suppose que $\text{argmin } f + h \neq \emptyset$. Soit $(x_n)_{n \in \mathbb{N}}$ générée par l'algorithme (GP), avec un pas $0 < \lambda < \frac{2}{L}$. Alors x_n converge vers $\bar{x} \in \text{argmin } (f + h)$ lorsque $n \rightarrow +\infty$.

Démonstration. C'est une conséquence du Théorème III.35, voir Section B.III.2. ■

III.II.2 Éclatement total : Algorithme de Davis-Yin

On considère ici le problème

$$(P) \quad \min_{x \in \mathbb{R}^N} f(x) + g(x) + h(x),$$

où $f, g, h \in \Gamma_0(\mathbb{R}^N)$ et $h \in C_L^{1,1}(\mathbb{R}^N)$. On a maintenant *deux* fonctions non lisses à traiter. Si l'on veut un algorithme d'éclatement pour ce problème, alors il doit s'exprimer en fonction de prox_f , prox_g et ∇h .

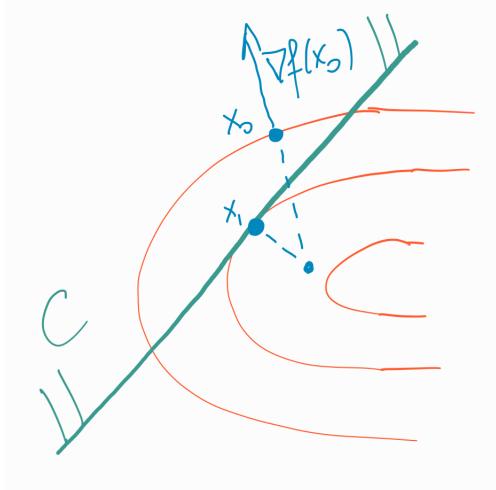


FIGURE III.4 – Quelques itérés de l'algorithme du gradient projeté.

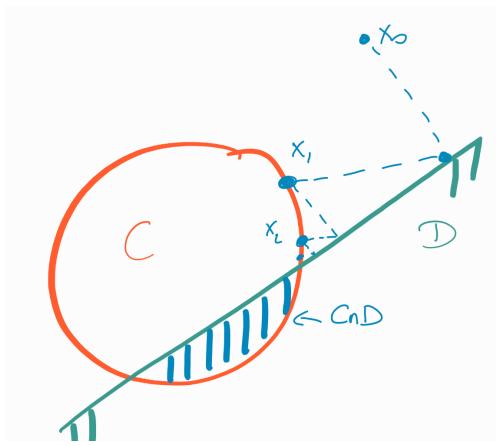


FIGURE III.5 – Quelques itérés de l'algorithme de projection alternée.

III.II.2.i) À la recherche d'un algorithme

Vu comment fonctionne l'algorithme du gradient-proximal, qui enchaîne une étape du gradient puis une étape proximale, il serait naturel de rajouter ici une troisième étape proximale, c'est-à-dire un algorithme qui s'écrirait par exemple:

$$x_{n+1} = \text{prox}_{\lambda f}(\text{prox}_{\lambda g}(x_n - \lambda \nabla h(x_n))).$$

Malheureusement, cet algorithme *ne fonctionne pas* en général : concaténer deux opérateurs proximaux ne conduit pas à résoudre notre problème, comme on l'a déjà vu dans la remarque III.16.

Pour trouver un algorithme, une autre approche consiste à regarder la discrétisation du flot gradient:

Remarque III.28 (Discrétisation du flot gradient : explicite-implicite-implicite). Considérons le flot gradient associé à notre problème :

$$\dot{x}(t) + \partial f(x(t)) + \partial g(x(t)) + \nabla h(x(t)) \ni 0.$$

D'une part, il est clair que l'on va discréteriser $\dot{x}(t)$ par une différence finie $\frac{x_{n+1}-x_n}{\lambda}$. D'autre part, il est aussi clair que l'on va faire une discrétisation explicite par rapport à h , en remplaçant $\nabla h(x(t))$ par $\nabla h(x_n)$. En ce qui concerne f et g , il va falloir passer par des pas implicites, où l'on va remplacer $\partial f(x(t))$ par $\partial f(x_{n+1})$, et idem pour g . Le problème est que l'on ne peut pas être implicite en même temps par rapport à f et g , car sinon ce sera équivalent à calculer le prox de $f + g$, ce que l'on ne veut pas ! Une idée simple pour résoudre ce problème est de faire les choses en deux temps : d'abord on est implicite par rapport à g , puis par rapport à f (l'ordre n'importe pas). Concrètement on considère la double discrétisation suivante :

$$\begin{aligned} \frac{\hat{x}_n - x_n}{\lambda} + \partial f(x_n) + \partial g(\hat{x}_n) + \nabla h(x_n) &\ni 0, \\ \frac{x_{n+1} - x_n}{\lambda} + \partial f(x_{n+1}) + \partial g(\hat{x}_n) + \nabla h(x_n) &\ni 0. \end{aligned} \tag{III.1}$$

Tel quel, il n'est pas clair que cet algorithme ne dépend que des prox de f et g , mais avec un peu de travail on peut le prouver.

Lemme III.29. *Si la suite $(x_n)_{n \in \mathbb{N}}$ est définie à travers les relations (III.1), alors elle vérifie les relations:*

$$\begin{cases} x_{n+1} = \text{prox}_{\lambda f}(x_n - \lambda \nabla h(x_n) - \lambda u_n), \\ u_{n+1} = \text{prox}_{\frac{1}{\lambda} g^*} \left(u_n + \frac{1}{\lambda} [2x_{n+1} - x_n - \lambda \nabla h(x_{n+1}) + \lambda \nabla h(x_n)] \right). \end{cases}$$

Démonstration. D'après les relations (III.1), il existe des sous-gradients $v_n \in \partial f(x_n)$ et $u_n \in \partial g(\hat{x}_n)$ tels que

$$\hat{x}_n - x_n + \lambda v_n + \lambda u_n + \lambda \nabla h(x_n) = 0 \tag{III.2}$$

$$x_{n+1} - x_n + \lambda v_{n+1} + \lambda u_n + \lambda \nabla h(x_n) = 0. \tag{III.3}$$

Si on réorganise (III.3), on voit que

$$x_{n+1} + \lambda v_{n+1} = x_n - \lambda \nabla h(x_n) - \lambda u_n \in x_{n+1} + \lambda \partial f(x_{n+1}).$$

D'après la caractérisation du prox par le sous-différentiel, cela équivaut à la première relation

$$x_{n+1} = \text{prox}_{\lambda f}(x_n - \lambda \nabla h(x_n) - \lambda u_n).$$

Passons à la seconde relation. Puisque $u_n \in \partial g(\hat{x}_n)$ alors $\hat{x}_n \in \partial g^*(u_n)$ et

$$u_n + \frac{1}{\lambda} \hat{x}_n \in u_n + \frac{1}{\lambda} \partial g^*(u_n) \Leftrightarrow u_n = \text{prox}_{\frac{1}{\lambda} g^*}(u_n + \frac{1}{\lambda} \hat{x}_n).$$

Si on fait l'égalité entre les équations de (III.3) et (III.2) pour laquelle on remplace n par $n+1$, on obtient

$$\hat{x}_{n+1} - 2x_{n+1} + x_n + \lambda u_{n+1} - \lambda u_n + \lambda \nabla h(x_{n+1}) - \lambda \nabla h(x_n) = 0. \quad (\text{III.4})$$

Donc en utilisant (III.4) on conclut que

$$u_{n+1} = \text{prox}_{\frac{1}{\lambda} g^*}(u_{n+1} + \frac{1}{\lambda} \hat{x}_{n+1}) = \text{prox}_{\frac{1}{\lambda} g^*}(u_n + \frac{1}{\lambda} [2x_{n+1} - x_n - \lambda \nabla h(x_{n+1}) + \lambda \nabla h(x_n)]).$$



III.II.2.ii) L'algorithme de Davis-Yin

Définition III.30 (Algorithme de Davis-Yin (2017)). Soient $f, g, h \in \Gamma_0(\mathbb{R}^N)$ avec $h \in C_L^{1,1}(\mathbb{R}^N)$, et $\lambda > 0$. L'algorithme de Davis-Yin (DY) génère une suite $(x_n)_{n \in \mathbb{N}} \subset \mathbb{R}^N$ telle que

$$\begin{cases} x_{n+1} = \text{prox}_{\lambda f}(x_n - \lambda \nabla h(x_n) - \lambda u_n), \\ u_{n+1} = \text{prox}_{\frac{1}{\lambda} g^*}\left(u_n + \frac{1}{\lambda} [2x_{n+1} - x_n - \lambda \nabla h(x_{n+1}) + \lambda \nabla h(x_n)]\right). \end{cases} \quad (\text{DY})$$

Remarque III.31 (Formulation primale de Davis-Yin). En jouant avec la formule de décomposition de Moreau, on peut trouver une écriture équivalente de (DY):

$$\begin{cases} x_{n+1} = \text{prox}_{\lambda f}(y_n) \\ \hat{x}_{n+1} = \text{prox}_{\lambda g}(2x_{n+1} - y_n - \nabla h(x_{n+1})) \\ y_{n+1} = y_n + \hat{x}_{n+1} - x_{n+1} \end{cases} \quad (\text{DY}')$$

On dit que la formulation (DY') est *primale*, tandis que la formulation (DY) est *primales-duales*, car elles font intervenir des variables primales x_n ou \hat{x}_n , associées au prox de f et/ou g , et/ou des variables duales u_n , associées au prox de g^* . Elles peuvent simplifier l'implémentation, notamment dans des cas où il est plus simple de calculer le prox de g^* que celui de g , et que l'utilisateur ne veut pas écrire les règles de calcul (représentation de Moreau, etc). Prouver cette équivalence est un très bon exercice de calcul proximal. On pourra en trouver une preuve en annexe (voir le lemme B.66).

Exemple III.32 (Proximal-Gradient est un cas particulier de Davis-Yin).

- Lorsque $g = 0$, alors $g^* = \delta_0$ et $\text{prox}_{\lambda g^*} = \text{proj}_0$, donc on déduit de (DY) que $u_n = 0$ et $x_{n+1} = \text{prox}_{\lambda f}(x_n - \lambda \nabla h(x_n))$.
- Lorsque $f = 0$, alors $\text{prox}_{\lambda f} = I$ et on déduit de (DY') que $x_{n+1} = y_n = \hat{x}_n$, ce qui nous donne $\hat{x}_{n+1} = \text{prox}_{\lambda g}(\hat{x}_n - \lambda \nabla h(\hat{x}_n))$.

Exemple III.33 (Douglas-Rachford est un cas particulier de Davis-Yin). Lorsque il n'y a pas de terme lisse ($h = 0$), Davis-Yin permet de minimiser la somme de deux fonctions non lisses f et g , avec

$$\begin{cases} x_{n+1} = \text{prox}_{\lambda f}(x_n - \lambda u_n), \\ u_{n+1} = \text{prox}_{\frac{1}{\lambda} g^*} \left(u_n + \frac{1}{\lambda} [2x_{n+1} - x_n] \right). \end{cases}$$

Si on pose $y_n = x_n - \lambda u_n$, cet algorithme est équivalent (voir TD) à

$$\begin{cases} x_{n+1} = \text{prox}_{\lambda f}(y_n) \\ \hat{x}_{n+1} = \text{prox}_{\lambda g}(2x_{n+1} - y_n) \\ y_{n+1} = y_n + \hat{x}_{n+1} - x_{n+1}, \end{cases} \quad (\text{DR})$$

méthode connue sous le nom d'*algorithme de Douglas-Rachford* (1979). Si on élimine la variable \hat{x}_n , on voit que (DR) peut se réécrire (encore un simple exercice):

$$\begin{cases} x_{n+1} = \text{prox}_{\lambda f}(y_n) \\ y_{n+1} = \frac{1}{2}y_n + \frac{1}{2} \text{refl}_{\lambda f}(\text{refl}_{\lambda g}(y_n)), \end{cases} \quad (\text{DR})$$

où $\text{refl}_f(y) := 2\text{prox}_f(y) - y$ est l'opérateur de *réflexion* de f en x . Ce terme de réflexion prend tout son sens lorsque on considère la réflexion d'une indicatrice, voir exemple suivant.

Exemple III.34 (L'algorithme des réflexions alternées). Soient $C, D \subset \mathbb{R}^N$ convexes fermés non vides, et on s'intéresse au problème de faisabilité associé (trouver un point dans l'intersection $C \cap D$). On considère $\text{refl}_C(x) := 2\text{proj}_C(x) - x$ la réflexion par rapport à C . Alors l'algorithme de Douglas-Rachford appliqué à $f = \delta_C$ et $g = \delta_D$ est

$$\begin{cases} x_{n+1} = \text{proj}_C(y_n) \\ y_{n+1} = \frac{1}{2}y_n + \frac{1}{2} \text{refl}_C(\text{refl}_D(y_n)). \end{cases}$$

C'est un algorithme totalement différent de l'algorithme de projection alternées ! Comprendre les différences entre ces deux algorithmes est un sujet de recherche actif.

On peut maintenant énoncer le résultat de convergence pour l'algorithme de Davis-Yin.

Théorème III.35 (Convergence de Davis-Yin). Soient $f, g, h \in \Gamma_0(\mathbb{R}^N)$ avec $h \in C_L^{1,1}(\mathbb{R}^N)$. On suppose que $f + g + h$ admet un minimiseur non dégénéré : $\exists x$ tel que $0 \in \partial f(x) + \partial g(x) + \nabla h(x)$. Soit $(x_n)_{n \in \mathbb{N}}$ une suite générée par l'algorithme de Davis-Yin (DY), avec un pas $0 < \lambda < 2/L$. Alors x_n converge vers $\bar{x} \in \operatorname{argmin}(f + g + h)$, lorsque $n \rightarrow +\infty$.

Démonstration. Voir la section B.III.2 en Annexe. ■

Preuve du Théorème III.27 sur l'algorithme Gradient-Proximal. Comme on l'a vu à l'exemple III.32, (GP) est un cas particulier de (DY), donc on peut faire appel au Théorème III.35. Pour cela il faut en vérifier les hypothèses. Celle sur le pas est immédiatement vérifiée. Celle sur l'existence d'un minimiseur non dégénéré aussi, car on suppose l'existence d'un minimiseur et h étant lisse on peut appliquer la règle de somme simple (proposition II.71) qui nous dit que $\partial(f + h)(x) = \partial f(x) + \nabla h(x)$. ■

Remarque III.36 (Éclatement total). L'algorithme de Davis-Yin permet de minimiser une somme $f + g + h$. Mais on peut également l'utiliser pour minimiser n'importe quelle somme finie $g_1 + \dots + g_p + h$! Des détails à ce sujet sont donnés dans l'annexe B.III.1.i), ainsi qu'en TD.

III.II.3 Éclatement composite total : Algorithme de Yan

On considère ici le problème

$$(P) \quad \min_{x \in \mathbb{R}^N} f(x) + g(Ax) + h(x),$$

où $f, h \in \Gamma_0(\mathbb{R}^N)$, $g \in \Gamma_0(\mathbb{R}^M)$, $h \in C_L^{1,1}(\mathbb{R}^N)$ et $A \in \mathcal{M}_{M,N}(\mathbb{R})$. On a maintenant un opérateur linéaire en plus à traiter. Si l'on veut un algorithme d'éclatement pour ce problème, alors il doit s'exprimer en fonction de prox_f , prox_g , ∇h , mais aussi de A et A^\top et ce de manière explicite.

III.II.3.i) A la recherche d'un algorithme

La difficulté principale ici est que $\operatorname{prox}_{g \circ A}$ ne se calcule pas explicitement en général. Si c'était le cas, on pourrait juste appliquer (DY). Il existe un cas particulier où l'on peut s'en sortir, c'est celui où A est semi-orthogonale : si $AA^\top = \mu I$, alors on disposerait de la formule explicite (voir Proposition III.22) :

$$\operatorname{prox}_{\frac{1}{\lambda}(g \circ A)^*} = A^\top \circ \operatorname{prox}_{\frac{1}{\lambda\mu}g^*} \circ \frac{1}{\mu}A. \quad (\text{III.5})$$

Évidemment, toute matrice n'est pas semi-orthogonale. Mais il s'avère qu'on peut toujours augmenter une matrice quelconque pour la rendre semi-orthogonale :

Lemme III.37 (Augmentation semi-orthogonale). Soit $A \in \mathcal{M}_{M,N}(\mathbb{R})$ et $\mu \geq \|A\|^2$. Alors il existe $C \in \mathcal{M}_M(\mathbb{R})$ telle que la matrice $B = [A \ C] \in \mathcal{M}_{M,N+M}(\mathbb{R})$ vérifie $BB^\top = \mu I$.

Démonstration. Puisque $\mu \geq \|A\|^2 = \|AA^\top\|$, alors $\mu I - AA^\top$ est une matrice semi-définie positive. De plus elle est clairement symétrique, donc on peut définir sa racine carrée (on diagonalise puis on prend la racine carrée des valeurs propres, qui sont positives). On pose donc $C = (\mu I - AA^\top)^{1/2}$, puis $B = [A \ C]$. On calcule alors

$$BB^\top = AA^\top + CC^\top = AA^\top + C^2 = AA^\top + (\mu I - AA^\top) = \mu I.$$

■

On peut alors se servir de cette nouvelle matrice semi-orthogonale B pour construire un nouveau problème équivalent à (P), avec l'aide d'une variable auxiliaire que l'on va contraindre à être égale à zéro.

Lemme III.38. Le problème (P) ci-dessus est équivalent à

$$\min_{X=(x,x') \in \mathbb{R}^{N+M}} F(X) + G(BX) + H(X), \quad (\hat{P})$$

où

- $F(X) = f(x) + \delta_0(x')$, telle que $\text{prox}_{\lambda F}(X) = (\text{prox}_{\lambda f}(x), 0)$ et $\partial F(X) = \partial f(x) \times \mathbb{R}^M$;
- $B = [A \ C]$, où $C = (\mu I - AA^\top)^{1/2}$ avec $\mu \geq \|A\|^2$, telle que $BX = Ax + Cx'$;
- $G(y) = g(y)$;
- $H(X) = h(x)$, telle que $\nabla H(X) = (\nabla h(x), 0)$.

L'algorithme de Davis-Yin (DY) appliqué à (\hat{P}) génère une suite $(x_n)_{n \in \mathbb{N}}$ vérifiant

$$\begin{cases} x_{n+1} = \text{prox}_{\lambda f}(x_n - \lambda \nabla h(x_n) - \lambda A^\top w_n) \\ w_{n+1} = \text{prox}_{\sigma g^*}(w_n + \sigma A [2x_{n+1} - x_n + \lambda \nabla h(x_n) - \lambda \nabla h(x_{n+1})]), \end{cases} \quad (\text{Yan})$$

avec $\sigma = \frac{1}{\lambda \mu}$.

Démonstration. Le fait que ces problèmes soient équivalents est immédiat, puisque on constraint la nouvelle variable x' à être égale à zéro. Le calcul du prox de F est immédiat car c'est une somme directe, et idem pour le gradient de H . On s'attaque maintenant à l'expression de l'algorithme de Davis-Yin dans ce cas. Le but est de montrer que l'algorithme ne dépend pas de C explicitement, en exploitant le fait que on impose la seconde variable x' d'être nulle.

On note $X_n = (x_n, x'_n)$ et $U_n = (u_n, u'_n)$ la suite générée par (DY) au problème (\hat{P}) , ce qui donne :

$$\begin{cases} X_{n+1} = \text{prox}_{\lambda F}(X_n - \lambda \nabla H(X_n) - \lambda U_n), \\ U_{n+1} = \text{prox}_{\frac{1}{\lambda}(G \circ B)^*} \left(U_n + \frac{1}{\lambda} [2X_{n+1} - X_n - \lambda \nabla H(X_{n+1}) + \lambda \nabla H(X_n)] \right). \end{cases}$$

La première relation $X_{n+1} = \text{prox}_{\lambda F}(X_n - \lambda \nabla H(X_n) - \lambda U_n)$ se traduit par

$$\begin{cases} x_{n+1} = \text{prox}_{\lambda f}(x_n - \lambda \nabla h(x_n) - \lambda u_n) \\ x'_{n+1} = \text{proj}_0(x'_n - \lambda 0 - \lambda u'_n) = 0 \end{cases} \quad (\text{III.6})$$

Quitte à imposer $x'_0 = 0$, on en déduit que $x'_n \equiv 0$. On passe ensuite à la seconde relation, où on va exploiter le fait que B est semi-orthogonale (en utilisant (III.5)), et où on utilisera la notation $\sigma = 1/(\lambda\mu)$:

$$\begin{aligned} U_{n+1} &= \text{prox}_{\frac{1}{\lambda}(G \circ B)^*} \left(U_n + \frac{1}{\lambda} [2X_{n+1} - X_n - \lambda \nabla H(X_{n+1}) + \lambda \nabla H(X_n)] \right) \\ &= B^\top \left[\text{prox}_{\sigma g^*} \left(\frac{1}{\mu} BU_n + \sigma B [2X_{n+1} - X_n - \lambda \nabla H(X_{n+1}) + \lambda \nabla H(X_n)] \right) \right]. \end{aligned}$$

On pose maintenant w_{n+1} tel que $U_{n+1} = B^\top w_{n+1}$ dans l'équation ci-dessus. Autrement dit,

$$w_{n+1} = \text{prox}_{\sigma g^*} \left(\frac{1}{\mu} BU_n + \sigma B [2X_{n+1} - X_n - \lambda \nabla H(X_{n+1}) + \lambda \nabla H(X_n)] \right). \quad (\text{III.7})$$

Si on retourne dans la première étape de l'algorithme (III.6), on voit qu'il y a u_n , qui est la première composante de $U_n = (u_n, u'_n)$. Or on a maintenant que $U_n = B^\top w_n = (A^\top w_n, Cw_n)$. On en déduit que $u_n = A^\top w_n$, et la première étape de l'algorithme devient

$$x_{n+1} = \text{prox}_{\lambda f}(x_n - \lambda \nabla h(x_n) - \lambda A^\top w_n).$$

On s'intéresse maintenant à ce qu'il y a dans le prox dans (III.7). Pour commencer, on a $BX_n = Ax_n + Cx'_n = Ax_n$ puisque $x'_n = 0$ (idem pour BX_{n+1}). Ensuite, on a $B\nabla H(X_n) = A\nabla h(x_n) + C0 = A\nabla h(x_n)$ (idem pour $B\nabla H(X_{n+1})$). Enfin, on a un terme $(1/\mu)BU_n$ un peu embêtant. On pose $\hat{w}_n = (1/\mu)BU_n$, et il est immédiat de voir que l'on a

$$B^\top \hat{w}_n = \frac{1}{\mu} B^\top BU_n = \frac{1}{\mu} B^\top BB^\top w_n = B^\top w_n.$$

Or notre hypothèse que $BB^\top = \mu I$ implique aussi que BB^\top est injective, et donc que B^\top est injective. On en déduit donc que $w_n = \hat{w}_n = (1/\mu)BU_n$! On conclut alors que

$$w_{n+1} = \text{prox}_{\sigma g^*} (w_n + \sigma A [2x_{n+1} - x_n - \lambda \nabla h(x_{n+1}) + \lambda \nabla h(x_n)]).$$



III.II.3.ii) L'algorithme de Yan

Définition III.39 (Algorithme de Yan (2018)). Soient $f, h \in \Gamma_0(\mathbb{R}^N)$, $g \in \Gamma_0(\mathbb{R}^M)$, $h \in C_L^{1,1}(\mathbb{R}^N)$ et $A \in \mathcal{M}_{M,N}(\mathbb{R})$. L'algorithme de Yan génère une suite $(x_n)_{n \in \mathbb{N}} \subset \mathbb{R}^N$ qui vérifie (Yan), où $\lambda, \sigma > 0$ sont appelés respectivement les pas primal et dual de l'algorithme.

Exemple III.40 (Cas particuliers de Yan).

- Lorsque $A = I$ et que l'on prend $\sigma = \frac{1}{\lambda}$, on retrouve l'algorithme de Davis-Yin pour minimiser $f + g + h$. On rappelle que Davis-Yin généralise lui-même les algorithmes Gradient-Proximal (lorsque f ou $g = 0$), de Douglas-Rachford (lorsque $h = 0$), qui eux-mêmes généralisent les algorithmes proximal et du gradient.
- Lorsque $g = 0$, on se retrouve à minimiser $f + h$. Puisque $g^* = \delta_0$ on a $\text{prox}_{\sigma g^*} = 0$, donc $w_n \equiv 0$ et on voit que l'algorithme de Yan devient l'algorithme Gradient Proximal.
- Lorsque $f = 0$, on se retrouve à minimiser $g(Ax) + h(x)$. Dans ce cas $\text{prox}_{\lambda f} = I$ et l'algorithme de Yan se simplifie en

$$\begin{cases} x_{n+1} = x_n - \lambda \nabla h(x_n) - \lambda A^\top w_n \\ w_{n+1} = \text{prox}_{\sigma g^*}(w_n + \sigma A [x_{n+1} - \lambda \nabla h(x_{n+1}) - \lambda A^\top w_n]). \end{cases} \quad (\text{LV})$$

Cet algorithme est connu sous le nom d'*algorithme de Loris-Verhoeven* (2011).

- Lorsque $h = 0$, on se retrouve à minimiser $f(x) + g(Ax)$. Dans ce cas $\nabla h = 0$, et l'algorithme de Yan se simplifie en

$$\begin{cases} x_{n+1} = \text{prox}_{\lambda f}(x_n - \lambda A^\top w_n) \\ w_{n+1} = \text{prox}_{\sigma g^*}(w_n + \sigma A [2x_{n+1} - x_n]). \end{cases} \quad (\text{CP})$$

Cet algorithme est connu sous le nom d'*algorithme de Chambolle-Pock* (2011).

Exemple III.41 (Cas particulier : Chen-Teboulle). Supposons que l'on veuille minimiser $f(x) + g(Ax)$. Ce problème est équivalent à minimiser $F(X) + G(\Phi X)$, où $X = (x, y)$, $\Phi X = Ax - y$, $G = \delta_0$ et $F(X) = f(x) + g(y)$. Si on applique l'algorithme de Chambolle-Pock (**CP**) à ce problème avec $\sigma = \lambda$, on obtient (après avoir introduit la variable supplémentaire \hat{w}_n) :

$$\begin{cases} w_n = \hat{w}_n + \lambda(Ax_n - y_n) \\ x_{n+1} = \text{prox}_{\lambda f}(x_n - \lambda A^\top w_n) \\ y_{n+1} = \text{prox}_{\lambda g}(y_n + \lambda w_n) \\ \hat{w}_{n+1} = \hat{w}_n + \lambda(Ax_{n+1} - y_{n+1}). \end{cases} \quad (\text{III.8})$$

Cet algorithme est connu sous le nom d'*algorithme de Chen-Teboulle* (1994). Elle fait partie de ce que l'on appelle des méthodes *prédiction-correction* : ici la variable w_n fait une prédiction basée sur l'état (x_n, y_n) , puis après avoir mis à jour cet état on fait une correction en calculant \hat{w}_{n+1} avec (x_{n+1}, y_{n+1}) .

Théorème III.42 (Convergence de Yan). Soient $f, h \in \Gamma_0(\mathbb{R}^N)$, $g \in \Gamma_0(\mathbb{R}^M)$, $h \in C_L^{1,1}(\mathbb{R}^N)$ et $A \in \mathcal{M}_{M,N}(\mathbb{R})$. On suppose que le problème associé admet une solution non dégénérée : $\exists x$ tel que $0 \in \partial f(x) + A^\top \partial g(Ax) + \nabla h(x)$. Soit $(x_n)_{n \in \mathbb{N}}$ générée par l'algorithme de Yan, avec $0 < \lambda < \frac{2}{L}$ et $\sigma \leq \frac{1}{\lambda \|A\|^2}$. Alors x_n converge vers $\bar{x} \in \arg\min (f + g \circ A + h)$ lorsque $n \rightarrow +\infty$.

Démonstration. Soit $\mu := 1/(\lambda\sigma)$ qui d'après notre hypothèse sur σ vérifie $\mu \geq \|A\|^2$. On peut donc introduire $C = (\mu I - A^\top A)^{1/2}$ et considérer le problème augmenté (\hat{P}) du Lemme III.38, avec les fonctions F, G, H correspondantes. D'après le Lemme III.38, notre suite x_n provient d'une suite $X_n = (x_n, 0)$ générée par l'algorithme de Davis-Yin appliqué à (\hat{P}) . On doit donc simplement vérifier les hypothèses du Théorème III.35 sur la convergence de Davis-Yin appliquée à $F + G + H$. Tout d'abord l'hypothèse sur le pas λ : il faut $\lambda < 2/\text{Lip}(\nabla H)$. Or $\nabla H(X) = (\nabla h(x), 0)$, donc il est facile de voir que $\text{Lip}(\nabla H) = \text{Lip}(\nabla h) = L$. Donc notre hypothèse $\lambda < 2/L$ est suffisante. Ensuite, il y a l'hypothèse de solution non-dégénérée. On suppose ici qu'il existe x tel que $0 \in \partial f(x) + A^\top u + \nabla h(x)$, avec $u \in \partial g(Ax)$. Posons $X = (x, 0)$, et montrons que $0 \in \partial F(X) + \partial(G \circ B)(X) + \nabla H(X)$, ce qui voudra dire que X est un minimiseur non-dégénéré de $F + G + H$. Nous avons $BX = Ax + C0 = Ax$, ce qui nous permet d'écrire que

$$(A^\top u, Cu) = B^\top u \in B^\top \partial g(Ax) = B^\top \partial g(BX) \subset \partial(G \circ B)(X).$$

Nous avons donc

$$\begin{cases} 0 \in \partial f(x) + A^\top u + \nabla h(x), \\ 0 \in \mathbb{R}^M + Cu + 0. \end{cases}$$

Puisque $\partial F(X) = \partial f(x) \times \mathbb{R}^M$, $(A^\top u, Cu) \in \partial(G \circ B)(X)$ et $\nabla H(X) = (\nabla h(x), 0)$, cela implique bien que

$$0 \in \partial F(X) + \partial(G \circ B)(X) + \nabla H(X).$$

■

Remarque III.43 (Éclatement composite total). L'algorithme de Yan permet de minimiser une somme $f + g \circ A + h$. Mais on peut également l'utiliser pour minimiser n'importe quelle somme composite finie $g_1 \circ A_1 + \dots + g_p \circ A_p + h$! Des détails à ce sujet sont donnés dans l'annexe B.III.1.ii), ainsi qu'en TD.

Annexe A

Annexe: Quelques éléments de modélisation mathématique

Dans ce chapitre, nous proposons de modéliser certains des problèmes présentés en introduction. Plus précisément, nous montrons (avec plus ou moins de détails) comment il est possible de passer d'un problème formulé naïvement en un problème d'optimisation.

A.I Le transport optimal

Todo

A.II La classification

A.II.1 Présentation du problème.

On suppose que l'on dispose d'un certain type de données, et on veut être capable de les **classer** en deux groupes. Ce type de problème peut être très facile à réaliser pour un humain, mais toute la question est de savoir comment automatiser cette prise de décision pour l'implémenter sur une machine.

Une façon de modéliser ce problème est la suivante: on se retrouve face à une donnée $x \in \mathbb{R}^N$ et on veut lui attribuer une classe. Pour simplifier, lorsqu'on a un problème à deux classes comme dans les exemples ci-dessus, on dit souvent que les deux classes sont $\{-1, +1\}$. Par exemple -1 pourrait désigner « chien » et $+1$ désignerait « chat ».

Ce que l'on souhaite donc implémenter est ce que l'on appelle un **classifieur** binaire, c'est-à-dire une fonction $c : \mathbb{R}^N \rightarrow \{-1, +1\}$ qui prend en entrée une donnée $x \in \mathbb{R}^N$, et qui donne en sortie une étiquette ± 1 . Évidemment on ne va pas prendre n'importe quelle fonction $\mathbb{R}^N \rightarrow \{-1, +1\}$, on veut que notre classifieur fasse un « bon » travail en classant les données qu'on lui fournit. Comment faire cela ? En pratique on suit le processus suivant:



FIGURE A.1 – Un potichien et un potichat se trouvent sur ces images. Saurez-vous les distinguer?

1) Constitution d'une base de données. On se procure

- Une famille de données $\{x_1, \dots, x_m\} \subset \mathbb{R}^N$ (par exemple tout un tas de photos de chats et de chiens)
- La famille des étiquettes correspondantes $\{y_1, \dots, y_m\} \subset \{-1, +1\}$, où chaque y_i a été bien choisi, en général par un·e humain·e.

Par exemple, en reprenant la convention ci-dessus, si x_{47} est une photo de chien, alors $y_{47} = -1$ (rappelons que toute photo peut être vue comme un vecteur de \mathbb{R}^N où N correspond au nombre de pixels de la photo).

- 2) *La phase d'entraînement.* A partir de cette base de données, on va construire un classifieur $c : \mathbb{R}^N \rightarrow \{-1, +1\}$, en demandant à ce que ce classifieur vérifie, pour tout point de notre base de données, $c(x_i) = y_i$. Ainsi, on espère que si le classifieur fonctionne bien sur notre base de données, alors il fonctionnera également lorsqu'on lui présentera de nouvelles données.
- 3) *La phase de test.* Une fois que l'on aura trouvé notre classifieur, il faudra bien tester si il marche bien! Pour cela, on se constituera une nouvelle base de données $\{(\hat{x}_i, \hat{y}_i)\}$ similaire à celle mentionnée plus haut, et on regardera si l'étiquette prédite $c(\hat{x}_i)$ est bien égale à \hat{y}_i . On pourra par exemple compter le nombre de fois où le classifieur trouve la bonne réponse, et donc en déduire un pourcentage de succès estimé.

Remarque A.1 (Train/test split). Noter qu'en pratique on ne constitue pas deux bases de données. On construit une grosse base de données une fois pour toute, puis on la divise en deux parties: les données dites « d'entraînement » qui vont servir à construire le classifieur c , et d'autre part les données de « test » qui vont permettre d'évaluer si notre classifieur marche bien sur de nouvelles données.

Maintenant la question reste de déterminer comment construire notre classifieur. Dans le reste de cette section nous allons nous focaliser sur des classificateurs **linéaires** (en fait c'est va être affine mais la coutume est de parler de classification linéaire).

A.II.2 Hyperplans séparateurs

Étant donnés des points avec des étiquettes ± 1 , on s'intéresse à la famille des hyperplans qui séparent l'espace en deux, de manière à ce que les points avec étiquette $+1$ soient d'un côté, et ceux avec étiquette -1 de l'autre. Tout hyperplan \mathcal{H} de \mathbb{R}^N s'écrit sous la forme

$$\mathcal{H} = \{x \in \mathbb{R}^2 \mid \langle a, x \rangle = b\}, \quad (\text{A.1})$$

où $a \in \mathbb{R}^N$ et $b \in \mathbb{R}$ sont à choisir. On notera par la suite $w = (a, b) \in \mathbb{R}^N \times \mathbb{R}$ le vecteur de paramètres décrivant l'hyperplan \mathcal{H} .

Définition A.2 (Hyperplan séparateur). On dira que un hyperplan \mathcal{H}_w sépare les données si $\langle a, x_i \rangle \leq b$ pour les i tels que $y_i = -1$ et $\langle a, x_i \rangle \geq b$ pour les i tels que $y_i = +1$.

On dira que \mathcal{H}_w sépare strictement les données si de plus on a $\langle a, x_i \rangle \neq b$ pour tout $i = 1, \dots, m$.

Notez que dans la définition on pourrait inverser les inégalités, mais cela ne changerait pas grand-chose car il suffirait de multiplier $w = (a, b)$ par -1 . On va donc garder cette convention.

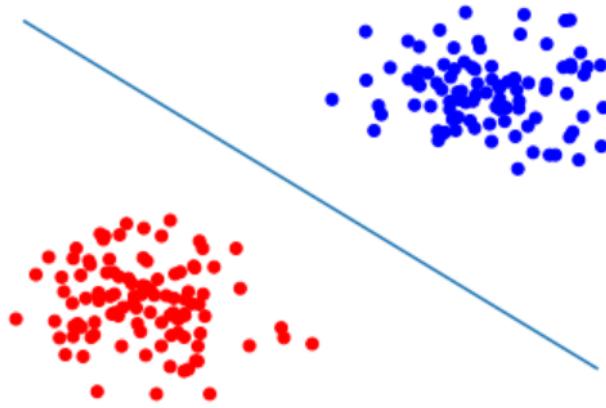


FIGURE A.2 – Deux nuages de points, strictement séparés par un hyperplan de \mathbb{R}^2 (c'est-à-dire: une droite).

Il est évident que tout hyperplan n'est pas séparateur. Il n'est d'ailleurs pas clair non plus qu'il en existe toujours ! Afin de poursuivre notre modélisation, il va falloir commencer par répondre à la question: quelles conditions sur les paramètres $w = (a, b)$ permettent de garantir que l'hyperplan associé (notons-le \mathcal{H}_w) sépare bien nos deux nuages ? La proposition suivante répond à cette question.

Proposition A.3 (Caractérisation des hyperplans séparateurs). Soit $\{x_1, \dots, x_m\} \subset \mathbb{R}^N$ des données, et soit $\{y_1, \dots, y_m\} \subset \{\pm 1\}$ des étiquettes. On définit la matrice $\Phi \in \mathcal{M}_{M,N+1}(\mathbb{R})$ par

$$\Phi := \begin{pmatrix} -y_1 x_1^\top & y_1 \\ \vdots & \vdots \\ -y_i x_i^\top & y_i \\ \vdots & \vdots \\ -y_m x_m^\top & y_m \end{pmatrix}.$$

Alors

- 1) \mathcal{H}_w sépare les données si et seulement si $\Phi w \leq 0$;
- 2) \mathcal{H}_w sépare strictement les données si et seulement si $\Phi w < 0$.

Démonstration. Par définition, \mathcal{H}_w sépare les données si et seulement si les quantités y_i et $\langle a, x_i \rangle - b$ ont le même signe. Ceci est équivalent à dire que

$$(\forall i \in \{1, \dots, m\}) \quad -y_i (\langle a, x_i \rangle - b) \leq 0.$$

C'est alors un simple exercice de vérifier que les coefficients du vecteur $\Phi w \in \mathbb{R}^M$ sont exactement ces quantités:

$$\Phi w = \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} -y_1 x_1^\top & y_1 \\ \vdots & \vdots \\ -y_i x_i^\top & y_i \\ \vdots & \vdots \\ -y_m x_m^\top & y_m \end{pmatrix} \begin{pmatrix} -y_1 (\langle a, x_1 \rangle - b) \\ \vdots \\ -y_i (\langle a, x_i \rangle - b) \\ \vdots \\ -y_m (\langle a, x_m \rangle - b) \end{pmatrix}.$$

On applique le même raisonnement pour la séparation stricte. ■

Nous voyons donc que l'ensemble des hyperplans séparateurs est un cône polyédral¹ $[\Phi w \leq 0]$ défini par les données du problème via la matrice Φ ! On peut donc déjà établir qu'il existe un hyperplan séparateur si et seulement si $[\Phi w \leq 0]$ est non vide. Si c'est le cas, on pourrait prendre n'importe quel paramètre $w = (a, b)$ dans ce cône, et définir un classifieur via²

$$c(x) = \text{sgn}(\langle a, x \rangle - b).$$

¹Pour être précis cet ensemble d'hyperplans est paramétré par un cône polyédral.

²Ici l'auteur vous enfume un peu, en ignorant volontairement le cas où $\langle a, x \rangle - b = 0$, dont le signe n'est pas bien défini, et qui correspond au cas où l'hyperplan ne sépare pas strictement. On cherche en général à éviter que ce soit le cas, on en reparlera plus loin.

Il est évident que, par définition, ce classifieur va bien marcher sur nos données de test, et renvoyer les bonnes étiquettes. Mais rien ne garantit qu'il va bien performer sur de nouvelles données de *test*, puisque Φ n'a été défini qu'à partir des données d'entraînement. Pour ce faire, on va essayer de trouver parmi tous les hyperplans séparateurs celui qui généralise le mieux sur de nouvelles données.

A.II.3 Trouver un bon hyperplan séparateur

Ici soyons clairs: il n'y a pas *une* bonne façon de procéder. Nous présentons simplement *une* approche, qui possède de bonnes bases théoriques.

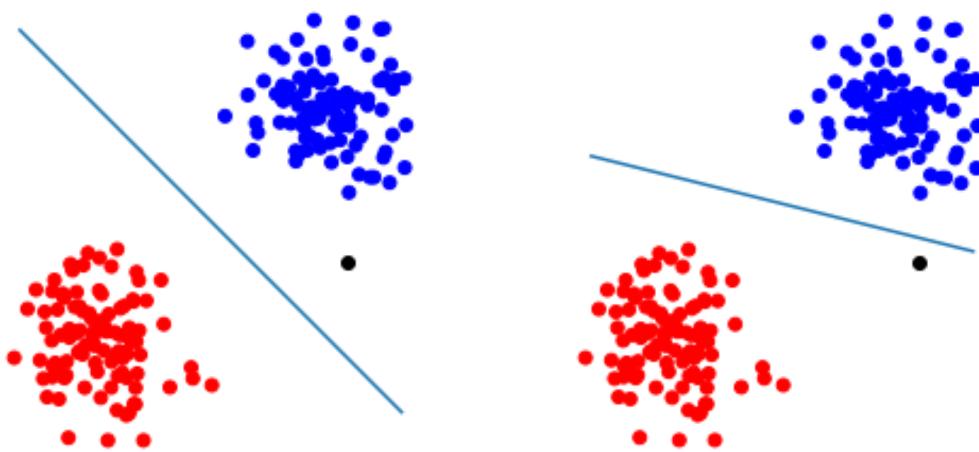


FIGURE A.3 – Deux nuages de points strictement séparés, et un nouveau point (noir) apparaît. A quelle classe appartient-il? Quel hyperplan séparateur fournit la meilleure réponse?

Le point de départ de notre raisonnement va être basé sur l'observation du cas illustré dans la figure A.3. On se convainc que sur cet exemple le nouveau point de test appartient probablement à la catégorie bleue. Or les deux hyperplans renvoient des réponses de classification différentes! On peut donc en déduire à priori que l'hyperplan de gauche est meilleur car il fournit une meilleure réponse.

Si on y regarde de plus près, la raison pour laquelle il est meilleur est qu'il se trouve **loin** des données. En effet, si un nouveau point vient à apparaître, on imagine qu'il va rester près d'un des deux nuages. Donc si l'hyperplan est loin des nuages, il sera également loin de ce nouveau point, et on évitera le problème d'un point qui apparaît trop près de l'hyperplan, en qui on n'aurait pas trop confiance.

Nous sommes donc maintenant prêts à passer à une nouvelle étape de notre modélisation: nous allons *choisir* parmi les hyperplans séparateurs celui qui maximise la distance aux données. Notons $\text{dist}(x; \mathcal{H})$ la distance entre un point $x \in \mathbb{R}^N$ et un hyperplan $\mathcal{H} \subset \mathbb{R}^N$. On peut donc dire que la distance entre un hyperplan \mathcal{H} et des données $\{x_1, \dots, x_m\}$ est

égale à la distance entre \mathcal{H} et le point qui en est le plus proche. Ceci est équivalent à mesurer le minimum de ces distances:

$$\min_{i=1,\dots,m} \text{dist}(x_i; \mathcal{H}).$$

En d'autres termes, la quantité ci-dessus représente la distance entre l'hyperplan \mathcal{H} et le nuage de points formé par les données.

Puisque notre objectif est de choisir la droite la plus éloignée des données, on veut donc *maximiser* cette distance aux données. Autrement dit, on veut trouver le paramètre $w = (a, b) \in \mathbb{R}^{N+1}$ qui maximise cette distance, tout en respectant la contrainte de bien séparer les données (cf. section précédente):

$$\max_{w \in \mathbb{R}^{N+1}} \min_{i=1,\dots,m} \text{dist}(x_i; \mathcal{H}_w), \quad \text{sous la contrainte que } \Phi w \leq 0. \quad (\text{SVM}')$$

Ce problème est assez affreux, car il consiste à maximiser sur un polyèdre un *minimum* de fonctions convexes. Or, autant un max de fonctions convexes est convexe, autant un minimum de fonctions convexes n'a que peu de chance d'être convexe ou concave. Mais heureusement, avec un peu de travail on peut montrer que ce problème est *équivalent* à un autre problème d'optimisation, qui lui est convexe.

Théorème A.4 (Problème SVM). *Supposons que le problème de classification soit*

- *non trivial, au sens que toutes les étiquettes n'ont pas la même valeur;*
- *régulier, au sens où il existe un hyperplan qui sépare strictement les données.*

Alors le problème (SVM') est équivalent au problème suivant:

$$\min_{w=(a,b) \in \mathbb{R}^N \times \mathbb{R}} \frac{1}{2} \|a\|^2, \quad \text{sous la contrainte que } \Phi w \leq -e, \quad (\text{SVM})$$

où $e = (1, \dots, 1) \in \mathbb{R}^m$. En particulier, si w^* est une solution de (SVM), alors c'est une solution de (SVM').

Démonstration. Voir la section A.II.5. ■

On se retrouve donc un problème qui a une très jolie structure: c'est la minimisation d'une fonction quadratique convexe sous une contrainte polyédrale. La résolution du problème SVM en utilisant les résultats du cours fait l'objet d'un TP dédié.

A.II.4 Pour aller plus loin

Nous sommes loin d'avoir fait un tour d'horizon exhaustif sur les problèmes de classification, qui méritent un cours à eux seuls. Un point sur lequel on peut s'attarder est que l'on a supposé que les données étaient strictement séparables. Mais que faire si les données ne sont même pas séparables ?

- 1) Dans certains cas, seules quelques données posent problème: on a quelques données mal classifiées, ou les deux nuages de points se recouvrent un peu mais pas trop. Dans ce cas il est possible de reformuler le problème en autorisant quelques petites violations: on permet à des points d'être dans le mauvais demi-espace, mais on veut minimiser leur distance à l'hyperplan délimitant. On parle alors de SVM à marge **souple**. Pour une brève explication avec des exemples, voir la section 10.2 SVM à marge souple de ce [cours](#).
- 2) Si les données ne sont pas du tout proches d'être linéairement séparables, on pourrait imaginer les séparer par des surfaces non linéaires. C'est le principe caché derrière les *méthodes à noyau*, où l'on remplace les données x_i par des représentations $\phi(x_i)$, avec une fonction ϕ bien choisie, qui envoie typiquement les données dans un espace de dimension plus grande. L'idée est qu'en grande dimension, « tout est linéairement séparable ». Pour une brève explication avec des exemples, voir la section 10. SVM à noyau de ce [cours](#).
- 3) Si on ne sait pas quelle fonction ϕ prendre, on pourrait la paramétriser et chercher à optimiser en même temps la valeur de ces paramètres. C'est le principe des méthodes à réseaux de neurones.

A.II.5 Preuve du Théorème A.4 sur le problème SVM

La preuve repose essentiellement sur le lemme technique suivant:

Lemme A.5. Soient $f, g : \mathbb{R}^N \rightarrow \mathbb{R}$, et supposons que elles vérifient les hypothèses suivantes:

- *Positive homogénéité* : $(\forall x \in \mathbb{R}^N)(\forall \lambda \geq 0) \quad f(\lambda x) = \lambda f(x), g(\lambda x) = \lambda g(x);$
- *Condition de Slater* : $[f > 0] \neq \emptyset;$
- *L'implication* $f(x) > 0 \Rightarrow g(x) > 0$ a lieu.

Considérons les deux problèmes suivants:

$$\min_{x \in \mathbb{R}^N} g(x) \text{ sous la contrainte que } f(x) \geq 1. \quad (\text{P})$$

$$\max_{x \in \mathbb{R}^N} \frac{f(x)}{g(x)} \text{ sous la contrainte que } f(x) \geq 0. \quad (\text{P}')$$

Alors:

- 1) si x^* est une solution de (P), alors λx^* est solution de (P') pour tout $\lambda > 0$;
- 2) si \hat{x} est une solution de (P') et si (P) admet une solution, alors il existe $\lambda > 0$ tel que $\lambda \hat{x}$ soit solution de (P), avec $\lambda = 1/f(\hat{x})$.

En particulier, si (P) admet une solution alors ces deux problèmes sont équivalents.

Démonstration. On procède en deux temps.

1) : Soit x^* une solution de (P) , et montrons que λx^* est solution de (P') . Puisque f et g sont positivement homogènes, on peut supposer sans perte de généralité que $\lambda = 1$.

Tout d'abord, prouvons que x^* sature la contrainte, c'est-à-dire que $f(x^*) = 1$. Par définition, on sait que $f(x^*) \geq 1$. Supposons donc par l'absurde que $f(x^*) > 1$. Dans ce cas, il est possible de prendre un $\lambda \in]0, 1[$ tel que l'on ait toujours $\lambda f(x^*) > 1$. Puisque f et g sont positivement homogènes, cela implique d'une part que $f(\lambda x^*) > 1$, autrement dit que $\lambda x^* \in [f \geq 1]$; et cela implique d'autre part que $g(\lambda x^*) = \lambda g(x^*)$. Or $f(x^*) > 0$, donc d'après nos hypothèses, on a forcément $g(x^*) > 0$. Ceci nous permet d'obtenir que $g(\lambda x^*) < g(x^*)$. On voit donc que λx^* est meilleur que x^* , ce qui contredit l'optimalité de x^* en temps que solution de (P) .

Maintenant, prouvons que x^* est une solution du problème de minimiser g sous la contrainte d'égalité $[f = 1]$. Pour ce faire, notons $\ell \in \mathbb{R}$ le minimum de g sur $[f \geq 1]$, et ℓ' le minimum de g sur $[f = 1]$. Puisque $[f = 1] \subset [f \geq 1]$, on a forcément $\ell \leq \ell'$. Mais puisque $f(w^*) = 1$ d'après la question précédente, on également $\ell' \leq g(w^*) = \ell$. D'où $\ell' = g(w^*)$, ce qui avec $w^* \in [f = 1]$ permet d'obtenir la propriété désirée.

Nous sommes maintenant prêts à conclure que x^* est solution de (P') . Le fait que x^* minimise $g(x)$ sur $[f = 1]$ est équivalent à dire que x^* maximise $1/g(x)$ sur $[f = 1]$. Par ailleurs sur cette contrainte on a f qui est constante et égale à 1. Donc on peut aussi dire que x^* maximise $\frac{f(x)}{g(x)}$ sur $[f = 1]$. Notons ℓ le maximum de cette fraction sur $[f = 1]$, et ℓ' le maximum de cette fraction sur $[f \geq 0]$. Puisque $[f = 1] \subset [f \geq 0]$, on voit immédiatement que $\ell \leq \ell'$. De plus, l'optimalité de x^* veut dire que $\frac{f(x^*)}{g(x^*)} = \ell$. Comme $f(x^*) = 1 \geq 0$, il nous suffit donc pour conclure de montrer que $\ell = \ell'$. Supposons donc par l'absurde que $\ell < \ell'$. Dans ce cas il existe un $w \in [f \geq 0]$ tel que $\ell < \frac{f(w)}{g(w)}$. On a vu dans une question précédente que $f(w^*) > 0$, donc on a également $g(x^*) > 0$ d'après nos hypothèses. Cela veut dire que $\ell > 0$, et donc que l'on a également $f(w) > 0$, et $g(w) > 0$. On peut donc définir $\lambda = 1/f(w)$, tel que $f(\lambda w) = 1$ (donc $\lambda w \in [f = 1]$) et $g(\lambda w) = \lambda g(w) > \ell$. Ceci contredit donc l'optimalité de x^* .

2) : Soit \hat{x} une solution de (P') , et montrons que $\lambda \hat{x}$ est solution de (P) pour un certain $\lambda > 0$.

Tout d'abord, observons le fait que \hat{x} ne sature pas la contrainte, c'est-à-dire que $f(\hat{x}) > 0$. En effet, d'après la condition de Slater il existe un x tel que $f(x) > 0$, et d'après les hypothèses cela implique $g(x) > 0$. Donc la valeur maximale du problème de (P') est forcément > 0 .

Montrons maintenant qu'il existe λ tel que $\lambda \hat{x}$ minimise g sur la contrainte $[f = 1]$. Soit $\lambda = 1/f(\hat{x})$ qui est > 0 d'après la question précédente. La fraction f/g est invariable lorsque on multiplie \hat{x} par λ (positive homogénéité de f, g). On en déduit que $\lambda \hat{x}$ est aussi solution de (P') . Mais on a également par construction que $f(\lambda \hat{x}) = 1$. On en déduit donc que $\lambda \hat{x}$ maximise f/g sur $[f = 1]$. Mais sur $[f = 1]$ on a ... $f = 1$! Donc $\lambda \hat{x}$ maximise $1/g$ sur $[f = 1]$. Et maximiser $1/g$, c'est équivalent à minimiser g .

Nous sommes maintenant prêts à conclure. Soit $\lambda \hat{x}$ le vecteur précédent qui minimise

g sur $[f = 1]$. Considérons également x^* , une solution de (P) , qui existe par hypothèse. On a vu dans le point 1) que x^* minimise g sur $[f = 1]$. Ceci veut dire que $g(x^*) = g(\tilde{x})$. Or $f(\tilde{x}) \geq 1$, donc \tilde{x} est également une solution de (P) . ■

Nous sommes maintenant prêts à prouver le Théorème A.4.

Démonstration du Théorème A.4. On commence par expliciter la valeur de $\text{dist}(x_i; \mathcal{H}_w)$. Puisque on sait projeter sur un hyperplan (voir l'exemple ??), on peut écrire

$$\text{dist}(x_i; \mathcal{H}_w) = \|x_i - \text{proj}_{\mathcal{H}_w}(x_i)\| = \left\| \frac{\langle a, x_i \rangle - b}{\|a\|^2} a \right\| = \frac{|\langle a, x_i \rangle - b|}{\|a\|}.$$

Or dans le problème (SVM') on a la contrainte $\Phi w \leq 0$ qui veut dire que le signe de $\langle a, x_i \rangle - b$ est le même que celui de y_i , donc on peut écrire que

$$|\langle a, x_i \rangle - b| = y_i(\langle a, x_i \rangle - b).$$

Ainsi (SVM') est équivalent à

$$\max_{w \in \mathbb{R}^{N+1}} \min_{i=1, \dots, m} \frac{y_i(\langle a, x_i \rangle - b_i)}{\|a\|}, \quad \text{sous la contrainte que } \Phi w \leq 0.$$

Un petit détail technique à relever est qu'on a utilisé le fait que $a \neq 0$ plus haut. Mais c'est forcément vrai, car si on avait $a = 0$ alors la condition de séparation des données deviendrait $y_i b \leq 0$ ce qui est incompatible avec notre hypothèse que les étiquettes ne sont pas toutes les mêmes !

Définissons maintenant pour tout $w = (a, b) \in \mathbb{R}^N \times \mathbb{R}$ les deux fonctions $f(w) = \min_{i=1, \dots, m} y_i(\langle a, x_i \rangle - b)$ et $g(w) = \|a\|$. Il est clair que (SVM') revient à maximiser $\frac{f}{g}$ sous la contrainte $[\Phi w \leq 0]$. On peut d'ailleurs aussi voir que la relation $\Phi w \leq 0$ est exactement la même chose que $y_i(\langle a, x_i \rangle - b) \geq 0$ pour tout i , et donc équivalente à $f(w) \geq 0$. Donc (SVM') revient à maximiser $\frac{f}{g}$ sous la contrainte $[f \geq 0]$.

On va maintenant faire appel au Lemme A.5, qui nous suggère de regarder le problème de minimiser g sous la contrainte $[f \geq 1]$. Avec nos notations, la contrainte se réécrit $[\Phi w \leq -e]$, et il est facile de voir que minimiser $g(w) = \|a\|$ est équivalent à minimiser $\frac{1}{2} \|a\|^2$. On obtient donc notre deuxième problème (SVM) . Le but du reste de la preuve consiste donc à vérifier que le Lemme A.5 s'applique, pour pouvoir dire que les deux problèmes sont équivalents.

Premièrement il est clair dans notre cas que f et g sont positivement homogènes, à savoir qu'elles vérifient $f(\lambda w) = \lambda f(w)$ pour tout $w \in \mathbb{R}^{N+1}$ et $\lambda \geq 0$. On note également que notre hypothèse de problème régulier veut dire qu'il existe w tel que $f(w) > 0$, c'est-à-dire que la condition de Slater est vérifiée.

Montrons maintenant que si $f(w) > 0$ alors $g(w) > 0$. Supposons que $f(w) > 0$. Donc $y_i(\langle a, x_i \rangle - b) > 0$ pour tout i . Supposons par l'absurde que $g(w) \leq 0$, c'est-à-dire que $w = (0, b)$. Alors $f(w) = \min_i -y_i b$. Or $f(w) > 0$ et $|y_i| = 1$, ce qui veut dire que $b \neq 0$,

et que $-y_i = \text{sgn}(b)$. Donc les y_i ont toutes les mêmes valeurs, ce qui contredit notre hypothèse.

Pour conclure, il nous reste à vérifier que le problème (SVM) admet une solution. Pour ce faire, nous allons prouver que g est continue et coercive sur $K := \{w \in \mathbb{R}^{N+1} \mid f(w) \geq 1\}$. La continuité est triviale. Pour ce qui concerne la coercivité, considérons une suite $w^n = (a^n, b^n) \in K$ telle que $\|w^n\| \rightarrow +\infty$, et montrons que $g(w^n) \rightarrow +\infty$. On considère deux cas. Si $\|a^n\| \rightarrow +\infty$, la conclusion est immédiate. Sinon, cela veut dire qu'il existe une sous-suite (qu'on notera a^k par simplicité) telle que a^k reste bornée. On va voir que c'est en fait impossible. En effet si on pose $M := \sup_k \|a^k\| \in \mathbb{R}$ et $D := \max_i \|x_i\| \in \mathbb{R}$, on obtient alors

$$\begin{aligned} 1 \leq f(w^k) &= \min_i y_i(\langle a^k, x_i \rangle - b^k) \quad \text{car } w^n \in K \\ &\leq MD + \min_i -y_i b^k \quad \text{par Cauchy-Schwarz} \\ &= MD - |b^k| \quad (\text{hypothèse que le problème est non trivial}). \end{aligned}$$

Or $\|w^k\| \rightarrow +\infty$ tandis que a^k est bornée. Cela veut dire que $|b^k| \rightarrow +\infty$. En passant à la limite, on obtient alors une contradiction : $1 \leq -\infty$. ■

A.III Le traitement du signal et de l'image

Todo

Annexe B

Annexe: Pour aller plus loin

B.I Résultats avancés sur les polyèdres

Dans cette section on prouve certains résultats sur les polyèdres que l'on a énoncé au fil du cours. Mis à part le Lemme de Farkas sur le polaire d'un cône polyédral, ce ne sont pas des résultats à connaître, bien qu'ils soient selon moi intéressants et également peuvent aider à comprendre les liens entre polyèdres, polytopes, cônes polyédraux, et cônes à base finie.

B.I.1 Cônes et lemme de Farkas

Ici on prouve formellement certains résultats que l'on a laissé en exercice dans les feuilles de TD.

Lemme B.1. Soit $K = \text{cone}(a_1, \dots, a_p)$ un cône à base finie. Pour tout $x \in K$, il existe une sous-famille libre $\mathcal{A} \subset \{a_1, \dots, a_p\}$ telle que $x \in \text{cone}(\mathcal{A})$.

Démonstration. La preuve est triviale si les vecteurs forment déjà une famille libre. Considérons maintenant le cas où les vecteurs a_i sont liés. Donnons nous un $x \in K$, que l'on peut donc écrire sous la forme $x = \sum_{i=1}^p \lambda_i a_i \in C$ où $\lambda_i \geq 0$. Puisque les a_i sont supposés liés, il existe μ_i non tous nuls tels que

$$\sum_{i=1}^p \mu_i a_i = 0.$$

On peut alors écrire que, pour tout $t \in \mathbb{R}$,

$$x = x + t \left(\sum_{i=1}^p \mu_i a_i \right) = \sum_{i=1}^p (\lambda_i + t\mu_i) a_i.$$

Puisque les μ_i sont non tous nuls, il doit exister j tel que $\mu_j \neq 0$. On peut alors choisir de

prendre $t = \frac{-\lambda_j}{\mu_j}$, tel que $\lambda_j + t\mu_j = 0$ et donc

$$x = \sum_{i=1, i \neq j}^p b_i a_i, \quad b_i := \lambda_i + t\mu_i.$$

Pour conclure, on aimerait pouvoir dire que les b_i restants sont positifs, afin de pouvoir dire que x est dans l'enveloppe conique des a_i restants. Or pour l'instant on ne connaît pas le signe des b_i . Il faut donc légèrement modifier nos arguments.

On va commencer par poser $b_i(t) := \lambda_i + t\mu_i$, et $m(t) = \min_i b_i(t)$. Notons que m est le minimum de fonctions continues, donc est elle-même continue. Notre objectif est de trouver un $t \in \mathbb{R}$ tel que $m(t) = 0$: ainsi on aura d'une part que tous les $b_i(t)$ sont positifs, et d'autre part l'un d'entre eux sera nul. Pour ce faire, nous allons utiliser le théorème des valeurs intermédiaires. D'une part, on sait qu'il existe un b_j non nul. Quitte à multiplier tous les b_i par -1 , on peut supposer sans perte de généralité que $b_j < 0$. Puisque $b_j(t)$ tend vers $-\infty$ lorsque $t \rightarrow +\infty$, on sait qu'il existe un $t_j > 0$ tel que $b_j(t_j) < 0$, et donc en particulier $m(t_j) < 0$. D'autre part, on voit que $m(0) = \sum_i \lambda_i$. Si cette somme est nulle, alors puisque $\lambda \geq 0$ on a forcément les λ_i tous nuls. Donc, par définition, $x = 0$. Dans ce cas l'énoncé est trivial, donc on peut supposer sans perte de généralité que $x \neq 0$, et donc que $m(0) > 0$. On peut donc appliquer le théorème des accroissements finis, et obtenir un $\hat{t} \in]0, t_j[$ tel que $m(\hat{t}) = 0$.

En particulier nous avons que $b_i := b_i(\hat{t}) \geq 0$ et il existe k tel que $b_k = 0$. et donc

$$x = \sum_{i=1, i \neq k}^p b_i a_i \in \text{cone}(\{a_i\}_{i \neq k}).$$

On a donc une combinaison positive avec un élément en moins. On peut alors répéter tout cet argument pour faire décroître le nombre d'éléments, et ce jusqu'à tomber sur une famille libre. ■

Proposition B.2. Soit $K = \text{cone}(a_1, \dots, a_p)$ un cône à base finie. Alors K est fermé.

Démonstration. Pour commencer, prouvons le résultat lorsque les vecteurs a_i sont indépendants. Définissons $u : (\lambda_1, \dots, \lambda_p) \mapsto \sum_i \lambda_i a_i$. On voit que c'est une application linéaire. De plus, comme les a_i sont indépendants, l'application u est injective. On note \hat{u} la restriction de u à l'arrivée sur $u(\mathbb{R}^p) = \text{Vect}(a_1, \dots, a_p)$. Alors \hat{u} est clairement une application linéaire bijective. En particulier, \hat{u} et \hat{u}^{-1} sont continues. Or on peut écrire $\text{cone}(a_1, \dots, a_p) = \hat{u}(\mathbb{R}_+^p)$, où \mathbb{R}_+^p est l'orthant positif (qui est fermé). D'où le fait que $\text{cone}(a_1, \dots, a_p)$ soit fermé.

Soit $x_k \in \text{cone}(a_1, \dots, a_p)$ une suite qui tend vers un certain $x \in \mathbb{R}^d$. On doit montrer que $x \in \text{cone}(a_1, \dots, a_p)$. D'après le Lemme B.1, pour tout $k \in \mathbb{N}$ il existe une sous-famille libre \mathcal{A}_k telle que $x_k \in \text{cone}(\mathcal{A}_k)$. Cette famille change pour chaque itération ; mais il n'y a qu'un nombre fini de sous-familles possibles, donc forcément une de ces sous-familles apparaît un nombre infini de fois. Soit donc x_{k_n} une sous-suite telle que les familles \mathcal{A}_{k_n}

soient constantes et égale à \mathcal{A} . On a donc $x_{k_n} \in \text{cone}(\mathcal{A})$. Or \mathcal{A} est libre donc $\text{cone}(\mathcal{A})$ est fermé d'après ce que l'on a prouvé précédemment; et x_{k_n} converge vers x . On peut donc conclure que $x \in \text{cone}(\mathcal{A}) \subset \text{cone}(a_1, \dots, a_p)$. ■

Proposition B.3 (Polaire d'un cône polyédral - Lemme de Farkas). *La Proposition I.47 sur le Lemme de Farkas est vraie.*

Démonstration. 1) \subset : Soit $x^* \in \text{cone}(a_i)^*$, et montrons que $x^* \in [Ax \leq 0]$. Donc il faut montrer que $\langle a_i, x^* \rangle \leq 0$ pour tout $i \in [p]$. Or $x^* \in \text{cone}(a_i)^*$ et $a_i \in \text{cone}(a_i)$ donc c'est bien vrai.

\supset : Soit $x^* \in [Ax \leq 0]$, et montrons que $x^* \in \text{cone}(a_i)^*$. Soit $x \in \text{cone}(a_i)$, on doit montrer que $\langle x^*, x \rangle \leq 0$. Par définition de l'enveloppe conique, ceci est égal à $\sum_i \lambda_i \langle x^*, a_i \rangle$, où $\lambda_i \geq 0$. Or notre hypothèse sur x^* dit que $\langle x^*, a_i \rangle \leq 0$, donc on a bien ce que l'on voulait.

- 2) En appliquant le Théorème I.45 du bipolaire, on obtient que $[Ax \leq 0]^* = \text{cone}(A^*)^{**}$. Or on sait d'après la Proposition B.2 que $\text{cone}(A^*)$ est fermé, donc il est égal à son bipolaire. ■

B.I.2 Théorème de Weyl: cônes finis = cônes polyédraux

On va prouver le Théorème de Weyl en se reposant sur le Lemme de Farkas. Je me suis inspiré de l'approche utilisée dans [11].

Lemme B.4. *Soit $\pi : \mathbb{R}^{N+1} \rightarrow \mathbb{R}^N$ la projection canonique définie par $\pi(x_0, x_1, \dots, x_N) = (x_1, \dots, x_N)$. Soit $K' \subset \mathbb{R}^{N+1}$ un cône polyédral. Alors $\pi(K')$ est aussi un cône polyédral.*

Démonstration. Supposons que $K' = [A'x \leq 0]$, pour une matrice $A' \in \mathcal{M}_{M, N+1}(\mathbb{R})$. On note $a'_1, \dots, a'_M \in \mathbb{R}^{N+1}$ les lignes de A , avec $a'_i = (a_{i0}, a_i) \in \mathbb{R} \times \mathbb{R}^N$. On voit que $x' = (x_0, x) \in \mathbb{R} \times \mathbb{R}^N$ appartient à K' si et seulement si

$$\langle a'_i, x' \rangle \leq 0 \Leftrightarrow a_{i0}x_0 + \langle a_i, x \rangle \leq 0.$$

On décompose $\{1, \dots, M\}$ en trois ensembles disjoints:

$$I_0 = \{i \mid a_{i0} = 0\}, \quad I_+ = \{i \mid a_{i0} > 0\}, \quad I_- = \{i \mid a_{i0} < 0\}.$$

Pour $i \in I_0$ notre condition s'écrit $\langle a_i, x \rangle \leq 0$. Pour $j \in I_+$, quitte à diviser par a_{j0} et renommer $\hat{a}_j = a_j/a_{j0}$, notre condition devient

$$x_0 \leq -\langle \hat{a}_j, x \rangle.$$

Pour $k \in I_-$, quitte à diviser par a_{k0} et renommer $\hat{a}_k = a_k/a_{k0}$, notre condition devient

$$x_0 \geq -\langle \hat{a}_k, x \rangle.$$

Maintenant on s'intéresse à $\pi(K')$, et on voit que $x \in \pi(K')$ si et seulement si il existe $x_0 \in \mathbb{R}$ tel que $(x_0, x) \in K'$, autrement dit

$$\Leftrightarrow \exists x_0 \in \mathbb{R}, \quad \begin{cases} \langle a_i, x \rangle \leq 0 \text{ pour } i \in I_0 \\ x_0 \leq -\langle \hat{a}_j, x \rangle \text{ pour } j \in I_+ \\ x_0 \geq -\langle \hat{a}_k, x \rangle \text{ pour } k \in I_- \end{cases}$$

Qu'il existe un x_0 qui puisse s'intercaler entre ces inégalités est équivalent à ce que les produits scalaires de I_+ soient tous plus grands que ceux de I_- . Autrement dit:

$$x \in \pi(K') \Leftrightarrow \begin{cases} \langle a_i, x \rangle \leq 0 \text{ pour } i \in I_0 \\ \langle \hat{a}_k - \hat{a}_j, x \rangle \leq 0 \text{ pour } j \in I_+, k \in I_- \end{cases}$$

On voit donc bien que $\pi(K')$ est défini par des inégalités linéaires, c'est donc bien un cône polyédral. ■

Lemme B.5. Soit $\pi : \mathbb{R}^M \times \mathbb{R}^N \rightarrow \mathbb{R}^N$ la projection canonique définie par $\pi(y, x) = x$. Soit $K' \subset \mathbb{R}^M \times \mathbb{R}^N$ un cône polyédral. Alors $\pi(K')$ est aussi un cône polyédral.

Démonstration. Il suffit de voir que cette projection est une composition de M petites projections de codimension 1 telles que l'on a étudiées dans le Lemme B.4. ■

Proposition B.6 (Weyl). Si $K \subset \mathbb{R}^N$ est un cône à base finie, alors K est un cône polyédral.

Démonstration. Soit $K \subset \mathbb{R}^N$ un cône à base finie: il s'écrit $K = \text{cone}(a_1, \dots, a_M)$ et on note A la matrice dont la i -ème colonne est a_i . Autrement dit,

$$K = \text{cone}(A) = \left\{ Ay \in \mathbb{R}^N \mid y \geq 0 \right\}.$$

On introduit la matrice $B = (I; A)$ et le cône à base finie $K' = \text{cone}(B)$. Autrement dit par définition on a

$$K' = \left\{ \begin{pmatrix} y \\ Ay \end{pmatrix} \in \mathbb{R}^M \times \mathbb{R}^N \mid y \geq 0 \right\}.$$

D'autre part, nous voyons que $z = (y, x)$ appartient à K' si et seulement si

$$Ay - x = 0 \quad \text{et} \quad y \geq 0,$$

ce qui est équivalent à

$$Ay - x \leq 0 \quad \text{et} \quad Ay - x \geq 0 \quad \text{et} \quad y \leq 0.$$

On voit donc que K' est défini par des inégalités linéaires, c'est donc un cône polyédral. Or il est facile de voir que $K = \pi(K')$, où $\pi : \mathbb{R}^M \times \mathbb{R}^N \rightarrow \mathbb{R}^N$ est la projection canonique définie par $\pi(y, x) = x$. On peut donc conclure avec le Lemme B.5. ■

Théorème B.7 (Weyl). *Un cône est polyédral si et seulement si il est à base finie.*

Démonstration. Nous avons déjà vu une implication dans la Proposition B.6, prouvons maintenant l'autre implication. Soit K un cône polyédral, et montrons qu'il est à base finie. Par définition nous avons donc $K = [Ax \leq 0]$ pour une certaine matrice $A \in \mathcal{M}_{M,N}(\mathbb{R})$. D'après le Lemme de Farkas (Proposition I.47), nous savons que $K^* = \text{cone}(A^*)$. En particulier K^* est un cône à base finie, on peut donc utiliser la Proposition B.6 pour en déduire que K^* est un cône polyédral: donc il existe une matrice B telle que $K^* = [By \leq 0]$. Mais du coup on peut encore appliquer le Lemme de Farkas pour obtenir $K = K^{**} = [By \leq 0]^* = \text{cone}(B^*)$, c'est-à-dire que K est un cône à base finie. ■

B.I.3 Théorème de Motzkin et de Minkowski sur les polyèdres

Lemme B.8. *Soit $C' \subset \mathbb{R}^{N+1}$ un polyèdre. Soit $\pi(C') := \{x \in \mathbb{R}^N \mid (1, x) \in C'\}$. Alors $\pi(C')$ est un polyèdre.*

Démonstration. Si C' est un polyèdre, alors il existe $A' \in \mathcal{M}_{M,N+1}(\mathbb{R})$ et $b \in \mathbb{R}^M$ tels que $C' = [A'x' \leq b]$. On peut décomposer $A' = [a_0 | A]$ et $x' = (x_0, x)$ afin de voir que

$$C' = [x_0a_0 + Ax \leq b].$$

Ainsi, $x \in \pi(C')$ si et seulement si $(1, x) \in C'$, ce qui équivaut à $a_0 + Ax \leq b$ ou encore $Ax \leq (b - a_0)$. Donc $\pi(C') = [Ax \leq (b - a_0)]$, c'est donc bien un polyèdre. ■

Théorème B.9 (Motzkin). *Soit $C \subset \mathbb{R}^N$. Alors C est un polyèdre si et seulement si il peut s'écrire comme la somme $C = P + K$ d'un polytope P et d'un cône polyédral K .*

Démonstration. Soit $C \subset \mathbb{R}^N$ un polyèdre, que l'on écrit $C = [Ax \leq b]$ avec $A \in \mathcal{M}_{M,N}(\mathbb{R})$ et $b \in \mathbb{R}^M$. On introduit $K' \subset \mathbb{R}^{N+1}$ via

$$K' := [A'x' \leq 0], \quad A' = \begin{pmatrix} -1 & 0 \\ -b & A \end{pmatrix}.$$

Autrement dit, $K' = \{x' = (x_0, x) \mid x_0 \geq 0, Ax \leq x_0b\}$.

Par définition, K' est un cône polyédral. En vertu du Théorème B.7 de Weyl nous savons que K' est un cône à base finie. Donc on peut exhiber une telle base et écrire $K' = \text{cone}(b'_1, \dots, b'_q)$. Par la suite on va noter $b'_i = (b_{i0}, b_i) \in \mathbb{R} \times \mathbb{R}^N$. Il est important de noter que, au vu de la définition de K' (le fait que $x_0 \geq 0$), nous avons forcément $b_{i0} \geq 0$. Maintenant nous allons réordonner les b'_i et les séparer en deux groupes: ceux pour lesquels $b_{i0} > 0$ (que l'on renommera b'_1, \dots, b'_p) et ceux pour lesquels $b_{i0} = 0$ (que l'on renommera b'_{p+1}, \dots, b'_q). Pour le premier groupe, quitte à diviser les vecteurs b'_i par $b_{i0} > 0$ (cela ne changera pas le fait que $K' = \text{cone}(b'_1, \dots, b'_q)$), on peut supposer que

$b_{i0} = 1$. Il est alors facile de voir que

$$\begin{aligned}
 & x \in C \\
 \Leftrightarrow & (1, x) \in K' \\
 \Leftrightarrow & (1, x) \in \text{cone}(b'_1, \dots, b'_p, b'_{p+1}, \dots, b'_q) \\
 \Leftrightarrow & \exists \lambda_1, \dots, \lambda_p \geq 0, \exists \mu_{p+1}, \dots, \mu_q \geq 0, \quad (1, x) = \sum_{i=1}^p \lambda_i b'_i + \sum_{j=p+1}^q \mu_j b'_j \\
 \Leftrightarrow & \exists \lambda_1, \dots, \lambda_p \geq 0, \exists \mu_{p+1}, \dots, \mu_q \geq 0, \quad 1 = \sum_{i=1}^p \lambda_i \quad \text{et} \quad x = \sum_{i=1}^p \lambda_i b_i + \sum_{j=p+1}^q \mu_j b_j \\
 \Leftrightarrow & x \in \text{co}(b_1, \dots, b_p) + \text{cone}(b_{p+1}, \dots, b_q).
 \end{aligned}$$

Nous avons donc montré que $C = \text{co}(b_1, \dots, b_p) + \text{cone}(b_{p+1}, \dots, b_q)$; c'est donc bien la somme d'un polytope et d'un cône polyédral.

Passons maintenant à la réciproque, plus facile. Considérons un polytope $P = \text{co}(a_1, \dots, a_p)$, un cône polyédral K , et soit $C = P + K$. D'après le Théorème B.7 de Weyl nous pouvons dire que K est un cône à base finie, donc $K = \text{cone}(b_1, \dots, b_q)$. On introduit $C' \subset \mathbb{R} \times \mathbb{R}^N$ via

$$C' = \text{cone}(a'_1, \dots, a'_p, b'_1, \dots, b'_q), \quad a'_i = \begin{pmatrix} 1 \\ a_i \end{pmatrix}, \quad b'_j = \begin{pmatrix} 0 \\ b_j \end{pmatrix}.$$

On peut alors voir que $C = \pi(C')$, où π est la fonction introduite dans le Lemme B.8. Or C' est un cône polyédral, donc en particulier un polyèdre; donc ce Lemme nous permet de conclure que C est un polyèdre. ■

Théorème B.10 (Minkowski). Soit $C \subset \mathbb{R}^N$. Alors C est un polytope si et seulement si C est un polyèdre borné.

Démonstration. Supposons que C soit un polytope. Qu'il soit borné est facile à vérifier. De plus $C = C + \{0\}$ où $\{0\}$ est un cône polyédral, donc le Théorème B.9 de Motzkin nous dit que C est un polyèdre.

Supposons que C soit un polyèdre borné. En utilisant le Théorème B.9 de Motzkin, nous pouvons écrire que $C = P + K$ où P est un polytope et K est un cône polyédral. Or ce dernier est nécessairement réduit à $\{0\}$. Si ce n'était pas le cas, on aurait $d \in K$ non nul et $P + td \subset C$ pour tout $t \in \mathbb{R}$, ce qui contredirait le caractère borné de C . ■

B.I.4 Fonctions polyédrales

Ici on fait un lien entre ensembles et fonctions que l'on a peu exploré dans le cours: quel est l'analogue fonctionnel des polyèdres? On pourrait penser à une fonction dont le graphe est polyédral, c'est-à-dire affine par morceaux.

Définition B.11. On dit que $f \in \Gamma_0(\mathbb{R}^N)$ est **AFFINE PAR MORCEAUX** s'il existe une famille finie h_1, \dots, h_p de fonctions affines telles que $f = \max_{i=1, \dots, p} h_i$.

Un autre exemple naturel serait de dire que l'indicatrice d'un ensemble polyédral est polyédral. Mais clairement une telle fonction n'est pas affine par morceaux... On va donc réunir les deux dans une même définition:

Définition B.12 (Fonction polyédrale). On dit que $f \in \Gamma_0(\mathbb{R}^N)$ est **POLYÉDRALE** si on peut écrire $f = g + h$ où $g \in \Gamma_0(\mathbb{R}^N)$ est l'indicatrice d'un polyèdre, et $h \in \Gamma_0(\mathbb{R}^N)$ est affine par morceaux. Autrement dit, si il existe des familles $a_1, \dots, a_q \in \mathbb{R}^N$, $\alpha_1, \dots, \alpha_p \in \mathbb{R}^N$, $b \in \mathbb{R}^q$, $\beta \in \mathbb{R}^p$ tels que

$$f(x) = \delta_{[Ax \leq b]}(x) + \max_{i=1, \dots, p} \langle \alpha_i, x \rangle - \beta_i, \quad (\text{B.1})$$

où les a_j^\top forment les lignes de A . On dira que (B.1) est la représentation standard de f .

Moralement, les fonctions polyédrales sont les fonctions dont le domaine est polyédral et qui sur leur domaine s'expriment comme une fonction affine par morceaux.

Exemple B.13. Voici quelques exemples immédiats de fonctions polyédrales:

- Les indicatrices de polyèdres (prendre $\alpha = 0$, $\beta = 0$ dans la définition, ce qui donne $h = 0$).
- Les fonctions convexes affines par morceaux (prendre $A = 0$, $b = 0$ dans la définition, ce qui donne $g = 0$).
- La valeur absolue, qui est affine par morceaux puisque elle vaut $\max\{t, -t\}$.

On dispose de règles de calcul qui préservent le caractère polyédral.

Proposition B.14.

- 1) La somme de deux fonctions polyédrales est polyédrale (si propre).
- 2) Le max de deux fonctions polyédrales est polyédral (si propre).
- 3) La composée d'une fonction polyédrale avec une matrice est polyédrale (si propre).

Démonstration. Dans cette preuve les P sont des ensembles polyédraux et les h sont des fonctions affines.

- 1) Soient $f = \delta_P + \max h_i$ et $f' = \delta_{P'} + \max h'_j$. Alors

$$f(x) + f'(x) = \delta_{P \cap P'} + \max_{i,j} h_i + h'_j$$

est bien polyédrale.

2) Soient $f = \delta_P + \max h_i$ et $f' = \delta_{P'} + \max h'_j$. Alors

$$f(x) + f'(x) = \delta_{P \cap P'} + \max_{i,j} h_i, h'_j$$

est bien polyédrale.

3) Soit $f = \delta_P + \max h_i$ où $P = [Ax \leq b]$ et $h_i(x) = \langle \alpha_i, x \rangle - \beta_i$. Soit $\Phi \in \mathcal{M}_{N,M}(\mathbb{R})$. Alors

$$f(\Phi y) = \delta_{[Ax \leq b]}(\Phi y) + \max_i \langle \alpha_i, \Phi y \rangle - \beta_i = \delta_{[A\Phi y \leq b]}(y) + \max_i \langle \Phi^* \alpha_i, y \rangle - \beta_i$$

ce qui montre que $f \circ \Phi$ est polyédrale. ■

Exemple B.15 (Norme L^1). La norme $\|x\|_1$ est polyédrale. En effet elle est la somme de fonctions $x \mapsto |x_i|$, qui peuvent se voir comme la composée de la valeur absolue avec la projection sur la i -ème coordonnée.

Sans surprise, on dispose d'une caractérisation via l'épigraphe:

Proposition B.16 (Épigraphe d'une fonction polyédrale). Soit $f \in \Gamma_0(\mathbb{R}^N)$. Elle est polyédrale si et seulement si $\text{epi } f$ est un polyèdre. En particulier, lorsque f s'écrit sous la forme standard (B.1), alors $\text{epi } f = [\mathcal{A}(x, r) \leq \mathcal{B}]$ où

$$\mathcal{A} = \begin{pmatrix} \alpha_1^\top & -1 \\ \vdots & \vdots \\ \alpha_p^\top & -1 \\ a_1^\top & 0 \\ \vdots & \vdots \\ a_q^\top & 0 \end{pmatrix}, \quad \mathcal{B} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \\ b_1 \\ \vdots \\ b_q \end{pmatrix}.$$

Démonstration. On procède par double implication.

\Rightarrow : Supposons que f soit polyédrale et admette la représentation standard (B.1). Clairement $\text{dom } h = \mathbb{R}^N$ donc $\text{dom } f = [Ax \leq b]$. On peut donc écrire son épigraphe comme (on rappelle la remarque II.11 sur l'épigraphe et le domaine):

$$\text{epi } f = \{(x, r) \in \mathbb{R}^{N+1} \mid Ax \leq b, \text{ et } \langle \alpha_i, x \rangle - \beta_i \leq r \forall i \leq p\}.$$

Nous pouvons réécrire cela en propriétés dans l'espace produit:

$$\begin{aligned} & (x, r) \in \text{epi } f \\ \Leftrightarrow & \langle a_i, x \rangle - r \leq \beta_i, \quad \langle a_j, x \rangle \leq b_j \quad \forall i \leq p, j \leq q \\ \Leftrightarrow & \langle (a_i, -1), (x, r) \rangle \leq \beta_i, \quad \langle (a_j, 0), (x, r) \rangle \leq b_j \quad \forall i \leq p, j \leq q. \end{aligned}$$

Si on note \hat{A} la matrice dont les p lignes sont les $(a_i, -1)$, et $\hat{\alpha}$ la matrice dont les q lignes sont les $(a_j, 0)$, alors on voit bien que

$$(x, r) \in \text{epi } f \Leftrightarrow \hat{A}(x, r) \leq b \text{ et } \hat{\alpha}(x, r) \leq \beta.$$

Donc $\text{epi } f$ est l'intersection de deux polyèdres, c'est donc un polyèdre.

\Leftarrow : Supposons que $\text{epi } f$ soit un polyèdre. On a donc $\text{epi } f = [\mathcal{A}(x, r) \leq b]$, et on note $(a_i, t_i) \in \mathbb{R}^N \times \mathbb{R}$ la i -ème ligne de \mathcal{A} , ce qui veut dire que

$$\text{epi } f = \{(x, r) \mid \langle a_i, x \rangle + t_i r \leq b_i \forall i\}.$$

Observons immédiatement que l'on a forcément $t_i \leq 0$. En effet si on avait un $t_i > 0$, alors on pourrait prendre $x \in \text{dom } f$ et pour tout $r \geq f(x)$ on aurait $(x, r) \in \text{epi } f$ et donc en particulier que $\langle a_i, x \rangle + t_i r \leq b_i$. Or ceci est impossible, car en prenant $r \rightarrow +\infty$ et en utilisant $t_i > 0$ on obtiendrait que $+\infty \leq b_i$. On a donc bien $t_i \leq 0$, et quitte à tout diviser par $|t_i|$ on peut supposer que les t_i valent 0 ou -1 .

On peut donc maintenant diviser les lignes de \mathcal{A} en deux groupes: les lignes pour lesquelles t_i vaut 0 et celles où $t_i = -1$. Histoire d'avoir des notations claires, on va changer de notation et pour les lignes telles que $t_i = -1$ on va noter α_i au lieu de a_i et β_i au lieu de b_i . En résumé, nous avons donc

$$\text{epi } f = \{(x, r) \mid \langle \alpha_i, x \rangle - \beta_i \leq r, \text{ et } \langle a_j, x \rangle \leq b_j \forall i, j\}.$$

Définissons g comme l'indicatrice du polyèdre P défini par les inégalités $\langle a_j, x \rangle \leq b_j$, et h comme le maximum des fonctions affines $h_i(x) = \langle \alpha_i, x \rangle - \beta_i$. Alors il est clair que $\text{dom}(g + h) = P$, et que

$$\begin{aligned} \text{epi}(g + h) &= \{(x, r) \mid h(x) \leq r, \text{ et } x \in P\} \\ &= \{(x, r) \mid \max_i \langle \alpha_i, x \rangle - \beta_i \leq r, \text{ et } \langle a_j, x \rangle \leq b_j \forall j\} \\ &= \text{epi } f. \end{aligned}$$

Si ces deux épigraphes coïncident, alors $f = g + h$ et on a bien prouvé que f est polyédrale. ■

Proposition B.17 (Sous-différentiel d'une fonction polyédrale). Soit $f \in \Gamma_0(\mathbb{R}^N)$ une fonction polyédrale, et considérons sa représentation standard (B.1). Alors $\text{dom } \partial f = \text{dom } f$, et

$$\partial f(x) = \begin{cases} \text{co}(\alpha_i)_{i \in I(x)} + \text{cone}(a_j)_{j \in J(x)} & \text{si } Ax \leq b, \\ \emptyset & \text{sinon,} \end{cases}$$

où $I(x)$ est l'ensemble des indices atteignant le max dans h et $J(x)$ est l'ensemble des contraintes actives de $[Ax \leq b]$:

$$I(x) = \{i \leq p \mid \langle \alpha_i, x \rangle - \beta_i = h(x)\} \quad \text{et} \quad J(x) = \{j \leq q \mid \langle a_j, x \rangle = b_j\}.$$

Démonstration. Soit $x \in \mathbb{R}^N$. Si $x \notin \text{dom } f$ alors clairement $\partial f(x) = \emptyset$, cf. proposition II.63. Considérons donc $x \in \text{dom } f$. Nous allons utiliser la caractérisation du sous-différentiel via l'épigraphe, vue dans la proposition II.61:

$$\partial f(x) = \{x^* \in \mathbb{R}^N \mid (x^*, -1) \in N_{\text{epi } f}(x, f(x))\}.$$

Puisque f est polyédrale, son épigraphe est un polyèdre $\text{epi } f = [\mathcal{A}(x, r) \leq \mathcal{B}]$, où \mathcal{A}, \mathcal{B} sont connus grâce à la proposition B.16. De plus on sait calculer le cône normal à ce polyèdre grâce au théorème I.57, qui nous dit que c'est le cône engendré par les lignes de \mathcal{A} qui saturent la contrainte:

$$N_{\text{epi } f}(x, f(x)) = \text{cone}((\alpha_i, -1))_{i \in \hat{I}(x, f(x))} + \text{cone}((a_j, 0))_{j \in \hat{J}(x, f(x))},$$

où $\hat{I}(x, f(x)) = \{i \leq p \mid \langle \alpha_i, x \rangle - f(x) = \beta_i\}$ et $\hat{J}(x, f(x)) = \{j \leq q \mid \langle a_j, x \rangle = b_j\}$. Il est assez immédiat de voir que $\hat{I}(x, f(x)) = I(x)$ et $\hat{J}(x, f(x)) = J(x)$, où $I(x)$ et $J(x)$ ont été définis dans l'énoncé, puisque $f(x) = h(x)$ lorsque $x \in \text{dom } f$.

On utilise maintenant la caractérisation du sous-différentiel via l'épigraphe pour dire que $x^* \in \partial f(x)$ si et seulement si $(x^*, -1) \in N_{\text{epi } f}(x, f(x))$, ce qui est équivalent dans notre contexte à:

$$\exists \mu_i \geq 0, \exists \lambda_j \geq 0 \text{ tels que } (x^*, -1) = \sum_{i \in I(x)} \mu_i (\alpha_i, -1) + \sum_{j \in J(x)} \lambda_j (a_j, 0).$$

Si on regarde attentivement ce qui se passe sur la dernière coordonnée, on voit qu'il faut $\sum_i \mu_i = 1$ et donc tout ceci est équivalent à

$$\exists \mu_i \geq 0, \exists \lambda_j \geq 0 \text{ tels que } x^* = \sum_{i \in I(x)} \mu_i \alpha_i + \sum_{j \in J(x)} \lambda_j a_j \text{ et } \sum_{i \in I(x)} \mu_i = 1.$$

Autrement dit, c'est équivalent à

$$x^* \in \text{co}(\alpha_i)_{i \in I(x)} + \text{cone}(a_j)_{j \in J(x)}.$$

■

Les fonctions polyédrales se comportent très bien avec les règles de calcul sous-différentiel, sans qu'il soit nécessaire de faire une hypothèse de qualification comme pour le théorème II.72. En d'autres termes les problèmes polyédraux sont toujours réguliers, au sens de la remarque II.74.

Théorème B.18 (Calcul pour fonctions polyédrales).

- 1) Si $f \in \Gamma_0(\mathbb{R}^M)$ est polyédrale et $\Phi \in \mathcal{M}_{M,N}(\mathbb{R})$, alors $\partial(f \circ \Phi) = \Phi^* \circ \partial f \circ \Phi$.
- 2) Si $f, g \in \Gamma_0(\mathbb{R}^N)$ sont polyédrales, alors $\partial(f + g) = \partial f + \partial g$.

Démonstration.

- 1) Supposons que f admette la décomposition standard (B.1). Alors

$$f(\Phi x) = \delta_{[A\phi x \leq b]}(x) + \max_i \langle \Phi^* \alpha_i, x \rangle - \beta_i.$$

On voit en particulier que $f \circ \Phi$ est polyédrale, et on peut donc utiliser la proposition B.17 précédente qui nous indique que

$$\partial(f \circ \Phi)(x) = \text{co}(\Phi^* \alpha_i)_{i \in \hat{I}(x)} + \text{cone}(\Phi^* a_j)_{j \in \hat{J}(x)},$$

où $\hat{I}(x) = \{i \leq p \mid \langle \Phi^* \alpha_i, x \rangle - \beta_i = h(\Phi x)\}$ et $\hat{J}(x) = \{j \leq q \mid \langle \Phi^* a_j, x \rangle = b_j\}$. Il est clair que $\hat{I}(x) = I(\Phi x)$ et $\hat{J}(x) = J(\Phi x)$, où $I(y) = \{i \leq p \mid \langle \alpha_i, y \rangle - \beta_i = h(y)\}$ et $J(y) = \{j \leq q \mid \langle a_j, y \rangle = b_j\}$. Il également clair que $\text{co}(\Phi^* \alpha_i)_{i \in \hat{I}(x)} = \Phi^* \text{co}(\alpha_i)_{i \in I(\Phi x)}$ et $\text{cone}(\Phi^* a_j)_{j \in \hat{J}(x)} = \Phi^* \text{cone}(a_j)_{j \in J(\Phi x)}$. On peut donc conclure que

$$\partial(f \circ \Phi)(x) = \Phi^* \text{co}(\alpha_i)_{i \in I(\Phi x)} + \Phi^* \text{cone}(a_j)_{j \in J(\Phi x)} = \Phi^* \partial f(\Phi x),$$

où dans la dernière égalité nous avons de nouveau utilisé la proposition B.17.

- 2) On va écrire la somme comme une composition. En effet $f + g = F \circ \Phi$ où $F(x, y) = f(x) + g(y)$ et $\Phi x = (x, x)$. Il est clair que F est polyédrale puisque somme de fonctions polyédrales composées avec des projections linéaires. Donc on peut utiliser le point précédent pour écrire que

$$\partial(f + g) = \partial(F \circ \Phi) = \Phi^* \circ \partial F \circ \Phi.$$

D'autre part, F est une somme séparée donc $\partial F = \partial f \times \partial g$ d'après la proposition II.68. Enfin on a par définition de Φ que $\Phi^*(u, v) = u + v$, on peut donc conclure que

$$\partial(f + g)(x) = \Phi^* \circ \partial F \circ \Phi(x) = \Phi^* \circ \partial F(x, x) = \Phi^*(\partial f(x), \partial g(x)) = \partial f(x) + \partial g(x).$$



On termine cette section en se posant la question de ce qu'est la conjuguée d'une fonction polyédrale. On dispose déjà, dans certains cas particuliers, d'outils pour y répondre:

Exemple B.19.

- 1) Si $f(x) = \max_i \langle \alpha_i, x \rangle$ est une fonction linéaire par morceaux, alors clairement f est la fonction support de $C = \text{co}(\alpha_i)$, et donc on sait via le théorème de la biconjuguée que $f^* = \sigma_C^* = \delta_C$, et donc f^* est bien polyédrale puisque indicatrice d'un polytope.
- 2) Si $f = \delta_C$ où C est un polytope, alors avec l'exemple ci-dessus et le théorème de la biconjuguée on comprend bien que f^* est une fonction linéaire par morceaux, et donc polyédrale.

- 3) Si $f = \delta_K$ où K est un cône polyédral, alors $f^* = \delta_{K^*}$. On sait via la proposition B.3 que le cône polaire K^* est à base finie, et le théorème B.7 de Weyl nous permet de conclure que K^* est aussi un polyèdre, et donc que f^* est polyédrale.
- 4) Si $f = \delta_P$ où P est un polyèdre, alors on peut utiliser le théorème B.9 de structure de Motzkin pour décomposer $P = C + K$ où K est un cône polyédral et C est un polytope. On a alors $f^* = \sigma_P = \sigma_C + \sigma_K$, où σ_C est une fonction linéaire par morceaux et σ_K est l'indicatrice du cône polyédral K^* . Donc f^* est bien polyédrale.

Pour une fonction polyédrale générale, la conjuguée est encore une fonction polyédrale, mais on ne dispose pas en général d'une formule pour exprimer la conjuguée en fonction de la fonction initiale. Au fond c'est le même problème que le théorème B.9 de décomposition de Motzkin : on sait qu'on peut décomposer un polyèdre en un cone plus un polytope, mais on ne dispose en général pas de forme close pour construire cette décomposition.

Théorème B.20 (Conjuguée d'une fonction polyédrale). *Si $f \in \Gamma_0(\mathbb{R}^N)$ est polyédrale, alors f^* aussi.*

Démonstration. On commence cette preuve avec une formulation alternative pour f . On peut se servir de l'épigraphé pour écrire

$$f(x) = \inf\{r \in \mathbb{R} \mid (x, r) \in \text{epi } f\}.$$

Or f est polyédrale donc son épigraphé est polyédral (proposition B.16). D'après le théorème B.9 de décomposition de Motzkin, on peut donc écrire que $\text{epi } f = \text{co}(\alpha_i, \beta_i) + \text{cone}(a_j, b_j)$, où $\alpha_1, \dots, \alpha_p, a_1, \dots, a_q \in \mathbb{R}^N$ et $\beta_i, b_j \in \mathbb{R}$. Ainsi,

$$f(x) = \inf\{r \in \mathbb{R} \mid \exists \lambda \in \mathbb{R}_+^q, \exists \mu \in \Delta^p, (x, r) = \sum_i \mu_i (\alpha_i, \beta_i) + \sum_j \lambda_j (a_j, b_j)\}.$$

Autrement dit

$$f(x) = \inf\{\sum_i \mu_i \beta_i + \sum_j \lambda_j b_j \mid \lambda \in \mathbb{R}_+^q, \mu \in \Delta^p, x = \sum_i \mu_i \alpha_i + \sum_j \lambda_j a_j\}.$$

Il est alors possible de voir cette formule comme $f = \inf_{\nu \in \mathcal{N}} f_\nu$, où l'on définit $\nu = (\mu, \lambda)$, $\mathcal{N} = \Delta^p \times \mathbb{R}_+^q$, et f_ν de manière appropriée. Plus exactement, définissons A la matrice dont les colonnes sont les a_j ; α la matrice dont les colonnes sont les α_i ; qui sont telles que $\sum_i \mu_i \alpha_i + \sum_j \lambda_j a_j = \alpha \mu + A \lambda$. Alors en définissant f_ν comme un dirac

$$f_{(\mu, \lambda)}(x) = \delta_{\{\alpha \mu + A \lambda\}}(x) + \langle \mu, \beta \rangle + \langle \lambda, b \rangle,$$

on voit bien que $f(x) = \inf_{\nu \in \mathcal{N}} f_\nu(x)$. On peut maintenant invoquer la formule de la conjuguée d'une inf (voir le TD ou le lemme B.42), ainsi que le fait que la conjuguée d'un

dirac est une fonction affine (cf. exemple II.90) pour écrire

$$\begin{aligned}
 f^*(u) &= \sup_{\nu \in \mathcal{N}} f_\nu^*(u) \\
 &= \sup_{\nu \in \mathcal{N}} \langle \alpha\mu + A\lambda, u \rangle - \langle \mu, \beta \rangle - \langle \lambda, b \rangle \\
 &= \sup_{\nu \in \mathcal{N}} \langle \mu, \alpha^*u - \beta \rangle + \langle \lambda, A^*u - b \rangle \\
 &= \sup_{\mu \in \Delta^p} \langle \mu, \alpha^*u - \beta \rangle + \sup_{\lambda \in \mathbb{R}_+^q} \langle \lambda, A^*u - b \rangle \\
 &= \sigma_{\Delta^p}(\alpha^*u - \beta) + \sigma_{\mathbb{R}_+^q}(A^*u - b).
 \end{aligned}$$

D'une part, nous savons que la fonction support du cône \mathbb{R}_+^q est l'indicatrice de son cône polaire \mathbb{R}_-^q . Autrement dit

$$\sigma_{\mathbb{R}_+^q}(A^*u - b) = \delta_{\mathbb{R}_-^q}(A^*u - b) = \delta_{[A^*u \leq b]}(u),$$

c'est-à-dire que cette fonction est l'indicatrice d'un polyèdre. D'autre part, Δ^p est l'enveloppe convexe de la base canonique $(e_i)_{i=1}^p$, et on sait que la fonction support d'un ensemble ou de son enveloppe convexe sont la même chose. Autrement dit

$$\sigma_{\Delta^p}(\alpha^*u - \beta) = \sigma_{(e_i)_{i=1}^p}(\alpha^*u - \beta) = \max_{i=1,\dots,p} \langle e_i, \alpha^*u - \beta \rangle = \max_{i=1,\dots,p} \langle \alpha_i, u \rangle - \beta_i,$$

c'est-à-dire que cette fonction est affine par morceaux. On peut donc conclure que f^* est polyédrale. ■

B.II Résultats avancés sur la conjuguée

Dans cette section facultative on propose quelques développements concernant la notion de conjuguée, qui est centrale à l'analyse convexe.

- Dans la section B.II.1, on répond à la question de savoir ce que vaut la conjuguée d'une somme $f + g$. Cela nous permettra au passage de prouver le théorème II.72 sur la règle de somme pour le sous-différentiel.
- Dans la section B.II.2, on répond à la question de savoir ce que vaut la composée $f \circ A$ lorsque A n'est pas inversible. Cela nous permettra au passage de prouver le théorème II.66 sur la règle de dérivation en chaîne pour le sous-différentiel.
- Dans la section B.II.3 on introduit la notion de dérivée directionnelle (que l'on peut voir comme une généralisation des dérivées partielles pour des fonctions non différentiables). On montrera qu'elle est en dualité avec la notion de sous-différentiel (voir le théorème B.43). Cela nous permettra de prouver le théorème II.75 sur le sous-différentiel d'un max de fonctions, et la propriété II.60 qui montre que f est différentiable dès lors que son sous-différentiel est un singleton.

- La section B.II.4 est dédiée à la preuve du théorème II.104 sur la dualité entre les fonctions lisses et les fonctions fortement convexes.

B.II.1 Inf-convolution

Définition B.21. Soient $f, g \in \Gamma_0(\mathbb{R}^N)$. On définit leur **INF-CONVOLUTION** par

$$f \# g(x) = \inf_{y \in \mathbb{R}^N} f(y) + g(x - y).$$

Si pour tout $x \in \text{dom}(f \# g)$ cet infimum est atteint, on dira que $f \# g$ est exacte.

Proposition B.22. Soient $f, g \in \Gamma_0(\mathbb{R}^N)$. Alors $f \# g$ est propre, et $\text{dom } f \# g = \text{dom } f + \text{dom } g$.

Démonstration. Il suffit de montrer la formule concernant le domaine.

\subset : Soit $x \in \text{dom } f \# g$. Alors il existe $y \in \mathbb{R}^N$ tel que $f(y) + g(x - y) < +\infty$. Donc $y \in \text{dom } f$ et $x \in y + \text{dom } g$.

\supset : Soient $x \in \text{dom } f, z \in \text{dom } g$. Alors $f \# g(x + y) \leq f(x) + g(x + z - x) = f(x) + g(z) < +\infty$. ■

Proposition B.23. Soient $f, g \in \Gamma_0(\mathbb{R}^N)$. Alors $f \# g$ est convexe.

Démonstration. C'est une conséquence directe du lemme B.24 suivant avec $H(x, y) = f(y) + g(x - y)$. ■

Lemme B.24 (Convexité d'une marginale). Soit $H : \mathbb{R}^N \times \mathbb{R}^M \rightarrow \overline{\mathbb{R}}$ convexe, et $h(x) := \inf_{y \in \mathbb{R}^M} H(x, y)$. Alors h est convexe.

Démonstration. Montrons que h est convexe, pour cela on prend $x, x' \in \text{dom } h$ et $t \in]0, 1[$. Par définition de h , pour tout $\varepsilon > 0$ fixé il doit exister y, y' tels que

$$h(x) \leq H(x, y) \leq h(x) + \varepsilon \quad \text{et} \quad h(x') \leq H(x', y') \leq h(x') + \varepsilon.$$

On peut alors écrire

$$\begin{aligned} h((1-t)x + tx') &\leq H((1-t)x + tx', (1-t)y + ty') \\ &\leq (1-t)H(x, y) + tH(x', y') \\ &\leq (1-t)h(x) + th(x') + \varepsilon, \end{aligned}$$

et on conclut en prenant $\varepsilon \rightarrow 0$. ■

Proposition B.25. Soient $f, g \in \Gamma_0(\mathbb{R}^N)$, tels que $\text{cont } f^* \cap \text{dom } g^* \neq \emptyset$. Alors $f \# g$ est sci et exacte.

Démonstration. Montrons que $h := f \# g$ est sci. On va utiliser la caractérisation par les sous-niveaux de la proposition II.24. Fixons donc $r \in \mathbb{R}$, et montrons que $[h \leq r]$ est fermé. Ceci est équivalent à montrer que $[h \leq r] \cap \delta\mathbb{B}$ est fermé pour tout $\delta > 0$. On se fixe donc $\delta > 0$, et on va montrer que $C := [h \leq r] \cap \delta\mathbb{B}$ est fermé. Considérons donc une suite $x_n \in C$ telle que $x_n \rightarrow x$ et montrons que $x \in C$.

Par définition de l'inf-convolution, il existe pour tout $n \in \mathbb{N}$ un $y_n \in \mathbb{R}^N$ tel que

$$f(y_n) + g(x_n - y_n) \leq h(x_n) + \frac{1}{n} \leq r + \frac{1}{n}. \quad (\text{B.2})$$

On voudrait passer à la limite, donc on va montrer que y_n est bornée pour utiliser un argument de compacité. Cette suite est bornée si et seulement si ses coefficients sont bornés, ce qui est équivalent à montrer que $\langle y_n, u \rangle$ est majoré pour tout $u \in \{\pm e_i\}_{i=1}^N \subset \delta\mathbb{B}$. Par ailleurs notre hypothèse de qualification nous donne l'existence d'un $\hat{x}^* \in \text{cont } f^*$, ce qui veut dire que f^* est localement bornée (voir lemme II.43): il existe $\varepsilon > 0$ tel que sur $\mathbb{B}(\hat{x}^*, \varepsilon)$ la fonction f^* soit majorée par M . On écrit alors

$$\begin{aligned} \varepsilon \langle y_n, u \rangle &= \langle y_n, \hat{x}^* \rangle - \langle y_n, \hat{x}^* - \varepsilon u \rangle \\ &= \langle y_n, \hat{x}^* \rangle + \langle x_n - y_n, \hat{x}^* - \varepsilon u \rangle - \langle x_n, \hat{x}^* - \varepsilon u \rangle \\ &\leq g(y_n) + g^*(\hat{x}^*) + f(x_n - y_n) + f^*(\hat{x}^* - \varepsilon u) + \|x_n\| \|\hat{x}^* - \varepsilon u\|, \end{aligned}$$

où dans la dernière inégalité nous avons utilisé Fenchel-Young et Cauchy-Schwarz. Si on utilise la définition de y_n on obtient

$$\varepsilon \langle y_n, u \rangle \leq r + \frac{1}{n} + g^*(\hat{x}^*) + f^*(\hat{x}^* - \varepsilon u) + \|x_n\| \|\hat{x}^* - \varepsilon u\|,$$

et on voit bien que ce terme est borné. En effet $g^*(\hat{x}^*) < +\infty$ par la condition de qualification, $\|x_n\| \leq \delta$ par définition de C , et tous les autres termes sont indépendants de n .

On peut donc conclure: y_n est bornée donc quitte à prendre une sous-suite on peut supposer qu'elle converge vers un y , donc en passant à la limite dans (B.2) on obtient

$$(f \# g)(x) \leq f(y) + g(x - y) \leq \liminf f(y_n) + g(x_n - y_n) \leq r.$$

Pour terminer il nous reste à vérifier que l'inf-convolution est exacte. Pour cela il suffit de reprendre cette preuve avec $x_n \equiv x$ et $r = (f \# g)(x)$ et de voir que le y obtenu rend l'inf-convolution exacte. ■

Proposition B.26. Soient $f, g \in \Gamma_0(\mathbb{R}^N)$. Si $\text{cont } f^* \cap \text{dom } g^* \neq \emptyset$, alors $f \# g \in \Gamma_0(\mathbb{R}^N)$ et est exacte.

Démonstration. On combine les propositions B.22, B.25 et B.23. ■

Proposition B.27. Soient $f, g \in \Gamma_0(\mathbb{R}^N)$. Alors $(f \# g)^* = f^* + g^*$.

Démonstration. Il suffit d'appliquer les définitions:

$$\begin{aligned}
 (f\#g)^* &= \sup_x \langle x^*, x \rangle - (f\#g)(x) \\
 &= \sup_x \langle x^*, x \rangle - \inf_y f(y) + g(x-y) \\
 &= \sup_x \sup_y \langle x^*, x \rangle - f(y) - g(x-y) \\
 &= \sup_x \sup_y \langle x^*, x-y \rangle - g(x-y) + \langle x^*, y \rangle - f(y) \\
 &= \sup_z \sup_y \langle x^*, z \rangle - g(z) + \langle x^*, y \rangle - f(y) \\
 &= g^*(x^*) + f^*(x^*).
 \end{aligned}$$

■

Théorème B.28 (Attouch-Brezis). Soient $f, g \in \Gamma_0(\mathbb{R}^N)$, tels que $\text{cont } f \cap \text{dom } g \neq \emptyset$. Alors $(f+g)^* = f^*\#g^*$ et cette inf-convolution est exacte.

Démonstration. On utilise le théorème de la biconjuguée pour écrire que $f = f^{**}$ et $g = g^{**}$. Ainsi la proposition B.27 nous permet d'écrire

$$f+g = f^{**} + g^{**} = (f^*\#g^*)^*,$$

ce qui en prenant la conjuguée implique

$$(f+g)^* = (f^*\#g^*)^{**}.$$

Or on suppose que $\text{cont } f \cap \text{dom } g = \text{cont } f^{**} \cap \text{dom } g^{**} \neq \emptyset$. Donc d'après la proposition B.26 nous avons que $f^*\#g^* \in \Gamma_0(\mathbb{R}^N)$, et que cette inf-convolution est exacte. ■

Corollaire B.29. Le théorème II.72 de Moreau-Rockafellar est vrai.

Démonstration. Soit $x^* \in \partial(f+g)(x)$, montrons que $x^* \in \partial f(x) + \partial g(x)$. D'après le théorème de Fenchel-Young

$$(f+g)(x) + (f+g)^*(x^*) = \langle x^*, x \rangle.$$

L'hypothèse de qualification nous permet d'invoquer le théorème B.28 d'Attouch-Brézis, qui nous dit que $(f+g)^* = f^*\#g^*$. Puisque cette inf-convolution est exacte, on sait qu'il existe y^* tel que $(f+g)^* = f(y^*) + g(x^* - y^*)$. Ainsi

$$f(x) + f^*(y^*) + g(x) + g^*(x^* - y^*) = \langle x^*, x \rangle.$$

Or l'inégalité de Fenchel-Young nous dit que

$$\begin{aligned}
 f(x) + f^*(y^*) &\geq \langle y^*, x \rangle \\
 g(x) + g^*(x^* - y^*) &\geq \langle x^* - y^*, x \rangle.
 \end{aligned}$$

Puisque la somme de ces inégalités donne une égalité, on en déduit que ces deux inégalités sont des égalités. Ceci implique via le théorème de Fenchel-Young que $y^* \in \partial f(x)$ et $x^* - y^* \in \partial g(x)$. On conclut en observant que $x^* = y^* + (x^* - y^*)$. ■

B.II.2 Inf-composition

Définition B.30. Soient $f \in \Gamma_0(\mathbb{R}^N)$ et $A \in \mathcal{M}_{M,N}(\mathbb{R})$. On définit leur **INF-COMPOSITION** comme étant la fonction $A \triangleright f : \mathbb{R}^M \rightarrow \overline{\mathbb{R}}$ définie par

$$(A \triangleright f)(y) = \inf \{f(x) \mid Ax = y\}.$$

Si pour tout $y \in \text{dom}(A \triangleright f)$ cet infimum est atteint, alors on dira que cette inf-composition est exacte.

Proposition B.31. Soient $f \in \Gamma_0(\mathbb{R}^N)$ et $A \in \mathcal{M}_{M,N}(\mathbb{R})$. Alors $A \triangleright f$ est propre, avec $\text{dom}(A \triangleright f) = A(\text{dom } f)$.

Démonstration. Il suffit de montrer la formule pour le domaine. Par définition, $y \in \text{dom}(A \triangleright f)$ si et seulement si il existe $x \in \text{dom } f$ tel que $Ax = y$. Autrement dit, si et seulement si $y \in A(\text{dom } f)$. ■

Proposition B.32. Soient $f \in \Gamma_0(\mathbb{R}^N)$ et $A \in \mathcal{M}_{M,N}(\mathbb{R})$. Alors $A \triangleright f$ est convexe.

Démonstration. C'est une conséquence directe du lemme B.24, en posant $H(x, y) = f(x) + \delta_G(x, y)$, où G est le graphe de A , c'est-à-dire

$$G = \{(x, y) \mid Ax = y\}$$

qui est un sous-espace affine, donc convexe. ■

Proposition B.33. Soient $f \in \Gamma_0(\mathbb{R}^N)$ et $A \in \mathcal{M}_{M,N}(\mathbb{R})$, tels que $\text{cont } f^* \cap \text{Im } A^* \neq \emptyset$. Alors $A \triangleright f$ est sci et exacte.

Démonstration. On procède comme pour la preuve de la proposition B.25. Soit $r \in \mathbb{R}$ et $\delta > 0$, il nous suffit de montrer que le sous-niveau local $C := [A \triangleright f \leq r] \cap \delta\mathbb{B}$ est fermé. Considérons donc une suite $y_n \in C$ qui converge vers un y , et montrons que $y \in C$. Puisque $\delta\mathbb{B}$ est fermée, il suffit de montrer que $(A \triangleright f)(y) \leq r$. Par définition de l'inf-composition, il existe pour tout $n \in \mathbb{N}$ un $x_n \in \mathbb{R}^N$ tel que $Ax_n = y_n$ et

$$f(x_n) \leq (A \triangleright f)(y_n) + \frac{1}{n} \leq r + \frac{1}{n}.$$

Afin de passer à la limite dans cette inégalité, il nous faut justifier que x_n admet une valeur d'adhérence, donc que cette suite est bornée. Soit $u \in \mathbb{B}$ quelconque, et montrons donc que $\langle x_n, u \rangle$ est majorée. Pour cela on utilise notre hypothèse pour obtenir un $\hat{y} \in \text{cont } f^* \cap$

$\text{Im } A^*$; en particulier on peut écrire $\hat{y} = A^* \hat{x}$, et obtenir un $\varepsilon > 0$ tel que f^* est majorée sur $\mathbb{B}(\hat{y}, \varepsilon)$ (on rappelle le lemme II.43). On écrit alors

$$\begin{aligned}\varepsilon \langle x_n, u \rangle &= \langle x_n, \hat{y} + \varepsilon u \rangle - \langle x_n, \hat{y} \rangle \\ &= \langle x_n, \hat{y} + \varepsilon u \rangle - \langle y_n, \hat{x} \rangle \\ &\leq f(x_n) + f^*(\hat{y} + \varepsilon u) + \|y_n\| \|\hat{x}\|\end{aligned}$$

On voit alors que tous ces termes sont majorés: $f(x_n)$ est majoré par $r + \frac{1}{n}$, f^* est majorée sur $\mathbb{B}(\hat{y}, \varepsilon)$, $\|y_n\| \leq \delta$ par définition, et $\|\hat{x}\|$ est juste constant. On peut donc dire que x_n est bornée, et quitte à prendre une sous-suite qu'elle converge vers un certain x . Cette limite vérifie $Ax = \lim_n Ax_n = \lim_n y_n = y$, ce qui nous permet de conclure que

$$(A \triangleright f)(y) = \inf_{Ax' = y} f(x') \leq f(x) \leq \liminf_n f(x_n) \leq \liminf_n r + \frac{1}{n} = r.$$

Pour voir que cette inf-composition est exacte, il suffit de considérer le cas particulier où $r = (A \triangleright f)(y)$, ainsi l'inégalité ci-dessus devient une égalité. ■

Proposition B.34. Soient $f \in \Gamma_0(\mathbb{R}^N)$ et $A \in \mathcal{M}_{M,N}(\mathbb{R})$. Si $\text{cont } f^* \cap \text{Im } A^* \neq \emptyset$, alors $(A \triangleright f) \in \Gamma_0(\mathbb{R}^M)$ et est exacte.

Démonstration. Il suffit de combiner les propositions B.31, B.32, B.33. ■

Proposition B.35. Soient $f \in \Gamma_0(\mathbb{R}^N)$ et $A \in \mathcal{M}_{M,N}(\mathbb{R})$. Alors $(A \triangleright f)^* = f^* \circ A^*$.

Démonstration. Il suffit d'appliquer les définitions:

$$\begin{aligned}(A \triangleright f)^*(y^*) &= \sup_y \langle y^*, y \rangle - (A \triangleright f)(y) \\ &= \sup_y \langle y^*, y \rangle - \inf_{x, Ax=y} f(x) \\ &= \sup_y \sup_{x, Ax=y} \langle y^*, y \rangle - f(x) \\ &= \sup_x \langle y^*, Ax \rangle - f(x) \\ &= \sup_x \langle A^* y^*, x \rangle - f(x) \\ &= f^*(A^* y^*).\end{aligned}$$

■

Théorème B.36. Soient $g \in \Gamma_0(\mathbb{R}^M)$ et $A \in \mathcal{M}_{M,N}(\mathbb{R})$, tels que $\text{cont } g \cap \text{Im } A \neq \emptyset$. Alors $(g \circ A)^* = A^* \triangleright g^*$, et cette inf-composition est exacte.

Démonstration. On applique les propositions B.34 et B.35, avec le théorème II.99 sur la biconjuguée :

$$(g \circ A)^* = (g^{**} \circ A^{**})^* = ((A^* \triangleright g^*)^*)^* = (A^* \triangleright g^*)^{**} = (A^* \triangleright g^*).$$

Corollaire B.37. *Le théorème II.66 sur la règle de la chaîne est vrai.*

Démonstration. Soit $x^* \in \partial(g \circ A)(x)$, et montrons que $x^* = A^*y^*$ où $y^* \in \partial g(Ax)$. D'après le théorème II.97 de Fenchel-Young,

$$(g \circ A)(x) + (g \circ A)^*(x^*) = \langle x^*, x \rangle.$$

Or d'après le théorème B.36, nous savons (en utilisant l'exactitude de l'inf-composition) qu'il existe y^* tel que $A^*y^* = x^*$ et

$$(g \circ A)^*(x^*) = (A^* \triangleright g^*)(x^*) = g^*(y^*).$$

On en déduit donc que

$$g(Ax) + g^*(y^*) = \langle x^*, x \rangle = \langle A^*y^*, x \rangle = \langle y^*, Ax \rangle.$$

On conclut alors avec le théorème II.97 de Fenchel-Young. ■

B.II.3 Dualité entre sous-différentiel et dérivée directionnelle

B.II.3.i) Dérivée directionnelle

Définition B.38 (Dérivée directionnelle). Soient $f \in \Gamma_0(\mathbb{R}^N)$, $x \in \text{dom } f$ et $d \in \mathbb{R}^N$. On définit la **DÉRIVÉE DIRECTIONNELLE** de f en x dans la direction d par

$$df(x; d) := \lim_{t \rightarrow 0^+} \frac{f(x + td) - f(x)}{t}.$$

Proposition B.39 (Formule globale de la dérivée directionnelle). Soient $f \in \Gamma_0(\mathbb{R}^N)$, $x \in \text{dom } f$ et $d \in \mathbb{R}^N$. Alors

$$df(x; d) = \inf_{t > 0} \frac{f(x + td) - f(x)}{t}.$$

En particulier, la dérivée directionnelle $df(x; d)$ est bien définie dans $[-\infty, +\infty]$.

Démonstration. Définissons $\phi(t) := \frac{f(x+td)-f(x)}{t}$ sur $]0, +\infty[$, il suffit alors de montrer que ϕ est croissante. Or montrer que ϕ est croissante revient à montrer que $\phi(t+s) \geq \phi(t)$ pour tout $t, s > 0$. En réécrivant cette inégalité, on obtient

$$f(x + td) \leq \frac{t}{t+s} f(x + (t+s)d) + \frac{s}{t+s} f(x)$$

qui est bien vraie d'après la définition de la convexité, puisque

$$x + td = \frac{t}{t+s}(x + (t+s)d) + \frac{s}{t+s}(x).$$

On peut donc en déduire que $df(x; d)$ est bien définie comme l'inf d'une partie de \mathbb{R} . ■

Proposition B.40 (Régularité par rapport aux directions). Soient $f \in \Gamma_0(\mathbb{R}^N)$ et $x \in \text{cont } f$. Alors $df(x; \cdot) \in \Gamma_0(\mathbb{R}^N)$ et est continue.

Démonstration. On pose $\phi(d) = df(x; d)$.

- A valeurs dans $\overline{\mathbb{R}}$: Puisque $x \in \text{cont } f \subset \text{dom } \partial f$ (voir théorème II.44 et proposition II.63), il existe un $x^* \in \partial f(x)$ qui nous permet d'écrire pour tout $d \in \mathbb{R}^N$ que

$$df(x; d) = \lim_t \frac{f(x + td) - f(x)}{t} \geq \lim_t \frac{\langle x^*, (x + td) - x \rangle}{t} = \langle x^*, d \rangle > -\infty.$$

- propre : ϕ est propre car $\phi(0) = 0$.
- convexe : En utilisant la convexité de f , on peut écrire que

$$f(x + t[(1-\alpha)d + \alpha d']) = f((1-\alpha)(x + td) + \alpha(x + td')) \leq (1-\alpha)f(x + td) + \alpha f(x + td').$$

Donc on en déduit que

$$\begin{aligned} & \phi((1-\alpha)d + \alpha d') \\ &= \lim_t \frac{f(x + t[(1-\alpha)d + \alpha d']) - f(x)}{t} \\ &= \lim_t \frac{(1-\alpha)f(x + td) + \alpha f(x + td') - f(x)}{t} \\ &= (1-\alpha) \lim_t \frac{f(x + td) - f(x)}{t} + \alpha \lim_t \frac{f(x + td') - f(x)}{t} \\ &= (1-\alpha)\phi(d) + \alpha\phi(d'). \end{aligned}$$

- Continue : Soit $d \in \mathbb{R}^N$, montrons que $\phi(d) < +\infty$. Puisque f est continue en x , elle est M -Lipschitzienne au voisinage de x (voir le lemme II.43) donc

$$\phi(d) = \lim_{t \rightarrow 0} \frac{f(x + td) - f(x)}{t} \leq \lim_{t \rightarrow 0} \frac{M\|x + td - x\|}{t} = M\|d\| < +\infty.$$

On a donc montré que $\text{dom } \phi = \mathbb{R}^N$. Or elle est convexe et propre, donc $\text{cont } \phi = \text{int dom } \phi = \mathbb{R}^N$ d'après le théorème II.44. On en déduit que ϕ est continue, et en particulier qu'elle est sci.



Proposition B.41 (Caractérisation analytique du sous-différentiel). Soient $f \in \Gamma_0(\mathbb{R}^N)$ et $x \in \text{dom } f$. Alors

$$\partial f(x) = \{x^* \in \mathbb{R}^N \mid (\forall d \in \mathbb{R}^N) \quad \langle x^*, d \rangle \leq df(x; d)\}.$$

Démonstration. On procède par double inclusion.

\subset : Soit $x^* \in \partial f(x)$. Pour $t > 0$ quelconque, on peut écrire

$$\langle x^*, d \rangle = \frac{1}{t} \langle x^*, (x + td) - x \rangle \leq \frac{1}{t} (f(x + td) - f(x))$$

où l'inégalité vient de la définition de $\partial f(x)$. En prenant l'inf sur $t > 0$, et avec la formule globale de la dérivée directionnelle (proposition B.39) on conclut que $\langle x^*, d \rangle \leq df(x; d)$.

\supset : Soit x^* dont la forme linéaire minore $df(x; \cdot)$. Pour montrer que $x^* \in \partial f(x)$, on prend $y \in \mathbb{R}^N$ quelconque, et on écrit

$$\langle x^*, y - x \rangle \leq df(x; y - x) \leq f(y) - f(x),$$

où la dernière inégalité vient de la formule globale de la dérivée directionnelle avec $t = 1$. ■

Lemme B.42 (Conjuguée de l'inf). Soit $(f_i)_{i \in I}$ une famille de fonctions $\mathbb{R}^N \rightarrow \overline{\mathbb{R}}$. Alors $(\inf_I f_i)^* = \sup_i f_i^*$.

Démonstration.

$$\begin{aligned} (\inf_I f_i)^*(u) &= \sup_x \langle u, x \rangle - \inf_i f_i(x) = \sup_x \sup_i \langle u, x \rangle - f_i(x) \\ &= \sup_i \sup_x \langle u, x \rangle - f_i(x) = \sup_i f_i^*(u). \end{aligned}$$

■

Théorème B.43 (Formule du max pour la dérivée directionnelle). Soient $f \in \Gamma_0(\mathbb{R}^N)$ et $x \in \text{cont } f$. Alors

$$(\forall d \in \mathbb{R}^N) \quad df(x; d) = \sigma_{\partial f(x)}(d).$$

Démonstration. Pour montrer le résultat, on pose $\phi(d) = df(x; d)$ et on calcule sa conjuguée, à l'aide de la formule globale de la dérivée directionnelle (proposition B.39) et du lemme

B.42 sur la conjuguée d'un inf:

$$\begin{aligned}
 \phi^*(x^*) &= \sup_{d \in \mathbb{R}^N} \langle x^*, d \rangle - \phi(d) \\
 &= \sup_{d \in \mathbb{R}^N} \langle x^*, d \rangle - \inf_{t > 0} \frac{f(x + td) - f(x)}{t} \\
 &= \sup_{t > 0} \sup_{d \in \mathbb{R}^N} \langle x^*, d \rangle - \frac{f(x + td) - f(x)}{t} \\
 &= \sup_{t > 0} \sup_{y \in \mathbb{R}^N} \langle x^*, \frac{y - x}{t} \rangle - \frac{f(y) - f(x)}{t} \\
 &= \sup_{t > 0} \sup_{y \in \mathbb{R}^N} \frac{-f(y) + f(x) + \langle x^*, y - x \rangle}{t} \\
 &= \sup_{t > 0} \sup_{y \in \mathbb{R}^N} \frac{f(x) - \langle x^*, x \rangle + \langle x^*, y \rangle - f(y)}{t} \\
 &= \sup_{t > 0} \frac{f(x) - \langle x^*, x \rangle + f^*(x^*)}{t}.
 \end{aligned}$$

D'après le théorème II.97 de Fenchel-Young, on sait que ce numérateur est toujours positif, et nul si et seulement si $x^* \in \partial f(x)$. On en déduit donc que $\phi^*(x^*) = 0$ si $x^* \in \partial f(x)$ et $\phi(x^*) = +\infty$ sinon. Autrement dit, $\phi^* = \delta_{\partial f(x)}$, et $\phi^{**} = \sigma_{\partial f(x)}$. Pour pouvoir conclure avec le théorème II.99 de la biconjuguée, il nous suffit de vérifier que $\phi \in \Gamma_0(\mathbb{R}^N)$, ce qui est vrai d'après la proposition B.40. ■

B.II.3.ii) Application: calcul du sous-différentiel du max

Corollaire B.44. *Le théorème II.75 sur le sous-différentiel du max est vrai.*

Démonstration. Commençons par prouver que la dérivée directionnelle du max est le max des dérivées directionnelles:

$$df(x; d) = \max_{i \in I(x)} df_i(x; d).$$

D'une part, on a par définition du max et de $I(x)$ que

$$\max_{i \in I(x)} df_i(x, d) = \max_{i \in I(x)} \lim_{t \rightarrow 0} \frac{f_i(x + td) - f_i(x)}{t} \leqslant \max_{i \in I(x)} \lim_{t \rightarrow 0} \frac{f(x + td) - f(x)}{t} = df(x, d).$$

D'autre part, on a

$$df(x, d) = \lim_{t \rightarrow 0^+} \frac{f(x + td) - f(x)}{t}.$$

Pour chaque $t > 0$, il existe par définition du max un $i_t \in \{1, \dots, M\}$ tel que $f(x + td) = f_{i_t}(x + td)$. Puisque $\{1, \dots, M\}$ est fini, il doit exister une suite $t_n \rightarrow 0$ et un unique $i \in \{1, \dots, M\}$ tel que $i_{t_n} \equiv i$. Autrement dit, pour tout n , $f(x + t_nd) = f_i(x + t_nd)$. Notons que $i \in I(x)$: en effet on suppose que f_i est continue en x (donc f l'est aussi), donc on peut passer à la limite lorsque $t_n \rightarrow 0$ et voir que $f(x) = f_i(x)$. On a donc montré que

$$df(x, d) = \lim_{t \rightarrow 0^+} \frac{f(x + td) - f(x)}{t} = df_i(x; d).$$

Ceci prouve donc que $df(x; d) = \max_{i \in I(x)} df_i(x; d)$. On peut maintenant appliquer la formule du max (théorème B.43) à f_i et f pour écrire

$$\sigma_{\partial f(x)} = df(x; \cdot) = \max_{i \in I(x)} df_i(x; \cdot) = \max_{i \in I(x)} \sigma_{\partial f_i(x)} = \sigma_{\cup_{i \in I(x)} \partial f_i(x)} = \sigma_C,$$

où $C = \text{adh co } \cup_{i \in I(x)} \partial f_i(x)$, et nous avons utilisé un résultat qui dit que la fonction support d'un ensemble est égal à la fonction support de son enveloppe convexe fermée (voir TD). On a donc montré que $\sigma_C = \sigma_{\partial f(x)}$. En passant à la conjuguée, et puisque C et $\partial f(x)$ sont convexes fermés non vides, on en déduit que $\delta_C = \delta_{\partial f(x)}$, et donc que $C = \partial f(x)$. ■

B.II.3.iii) Application: caractérisation de la différentiabilité via le sous-différentiel

Théorème B.45 (Caractérisation de la différentiabilité via le sous-différentiel).

Soient $f \in \Gamma_0(\mathbb{R}^N)$ et $x \in \text{cont } f$. Alors les propriétés suivantes sont équivalentes:

- 1) f est différentiable en x ;
- 2) f est Gâteaux différentiable en x , au sens où $df(x; \cdot)$ est linéaire;
- 3) $\partial f(x)$ est un singleton.

Démonstration. On procède par triple implication:

1) \Rightarrow 2) : Immédiat par définition de la différentiabilité

$$\begin{aligned} df(x; d) &= \lim_{t \rightarrow 0^+} \frac{f(x + td) - f(x)}{t} = \lim_{t \rightarrow 0^+} \frac{\langle \nabla f(x), td \rangle + o(\|td\|)}{t} \\ &= \langle \nabla f(x), d \rangle + \lim_{t \rightarrow 0^+} \frac{o(\|td\|)}{t} = \langle \nabla f(x), d \rangle. \end{aligned}$$

2) \Rightarrow 3) : Si $df(x; \cdot)$ est linéaire alors par le théorème de représentation de Riesz il existe x^* tel que $df(x; d) = \langle x^*, d \rangle$. D'après la caractérisation analytique du sous-différentiel (proposition B.41) on voit immédiatement que $x^* \in \partial f(x)$. Pour montrer qu'il est unique, on prend $u \in \partial f(x)$ quelconque, et on écrit

$$\langle u, u - x^* \rangle \leq df(x; u - x^*) = \langle x^*, u - x^* \rangle,$$

où dans la première égalité on a encore utilisé la caractérisation analytique du sous-différentiel. L'inégalité que nous avons obtenue est équivalente à $\|u - x^*\|^2 \leq 0$, et donc que $u = x^*$.

3) \Rightarrow 1) : Supposons que $\partial f(x) = \{x^*\}$. Puisque $x \in \text{cont } f$ on peut appliquer la formule du max (théorème B.43) qui nous dit que

$$df(x; d) = \sigma_{x^*}(d) = \langle x^*, d \rangle.$$

En particulier, 2) est vérifié. Prouvons maintenant 1), c'est-à-dire que f est différentiable en x . Pour cela on se donne une suite x_n quelconque qui tend vers x , et on veut montrer que

$$f(x_n) = f(x) + \langle x^*, x_n - x \rangle + o(\|x_n - x\|).$$

Autrement dit, montrons que

$$\lim_{n \rightarrow +\infty} \frac{f(x_n) - f(x) - \langle x^*, x_n - x \rangle}{\|x_n - x\|} = 0.$$

Posons $t_n := \|x_n - x\|$ que l'on peut supposer non nul sans perte de généralité, et $d_n = \frac{1}{t_n}(x_n - x)$, alors la limite précédente est égale à

$$\lim_{n \rightarrow +\infty} \frac{f(x + t_n d_n) - f(x) - \langle x^*, t_n d_n \rangle}{t_n} = \lim_{n \rightarrow +\infty} \frac{f(x + t_n d_n) - f(x)}{t_n} - \langle x^*, d_n \rangle.$$

Par construction $\|d_n\| = 1$, donc avec un argument de compacité on peut supposer sans perte de généralité que d_n converge vers un certain $d \in \mathbb{R}^N$. On a donc $\langle x^*, d_n \rangle$ qui converge vers $\langle x^*, d \rangle$. D'autre part on peut utiliser notre hypothèse que $x \in \text{cont } f$ pour dire que f est M -Lipschitzienne au voisinage de x (on rappelle le lemme II.43), ce qui nous permet d'écrire pour n grand que

$$\begin{aligned} \frac{f(x + t_n d_n) - f(x)}{t_n} &= \frac{f(x + t_n d) - f(x)}{t_n} + \frac{f(x + t_n d_n) - f(x + t_n d)}{t_n} \\ &\leq \frac{f(x + t_n d) - f(x)}{t_n} + M\|d_n - d\|. \end{aligned}$$

Si on passe à la limite lorsque $n \rightarrow +\infty$, on fait apparaître la définition de la dérivée directionnelle $df(x; d)$ qui vaut exactement $\langle x^*, d \rangle$. On conclut alors que

$$\begin{aligned} &\lim_{n \rightarrow +\infty} \left| \frac{f(x + t_n d_n) - f(x)}{t_n} - \langle x^*, d_n \rangle \right| \\ &\leq \left| \lim_{n \rightarrow +\infty} \frac{f(x + t_n d) - f(x)}{t_n} - \langle x^*, d_n \rangle \right| + \lim_{n \rightarrow +\infty} M\|d_n - d\| \\ &= |df(x; d) - \langle x^*, d \rangle| \\ &= 0. \end{aligned}$$



B.II.3.iv) Cônes tangent et normal à un sous-niveau

Proposition B.46 (Topologie du sous-différentiel). Soit $f \in \Gamma_0(\mathbb{R}^N)$.

- 1) Pour tout $x \in \mathbb{R}^N$, $\partial f(x)$ est convexe fermé.
- 2) Pour tout $x \in \text{cont } f$, $\partial f(x)$ est convexe compact non vide.

Démonstration.

- 1) Immédiat par définition du sous-différentiel et linéarité du produit scalaire.
- 2) Si $x \in \text{cont } f$ alors $x \in \text{int dom } f$ d'après la théorème II.44. Donc $x \in \text{dom } \partial f$ d'après proposition II.63, d'où le fait que $\partial f(x)$ soit non vide. Il reste à montrer qu'il est borné. On utilise le lemme II.43 qui nous indique que f est M -Lipschitzienne sur une boule $\mathbb{B}(x, \delta)$. Soit $x^* \in \partial f(x)$, et montrons que $\|x^*\| \leq M$. Si on applique la définition de sous-gradient avec $y = x + tx^*$, où $t = \frac{\delta}{\|x^*\|}$ alors $y \in \mathbb{B}(x, \delta)$ et

$$\langle x^*, y - x \rangle \leq f(y) - f(x) \leq M\|y - x\|$$

que l'on réécrit

$$t\|x^*\|^2 \leq Mt\|x^*\|$$

qui après division par $t\|x^*\|$ nous donne le résultat désiré. Observons qu'on a divisé par $\|x^*\|$, mais que si $x^* = 0$ alors la borne désirée est trivialement vérifiée.



Proposition B.47 (Cône tangent au sous-niveau). Soit $f \in \Gamma_0(\mathbb{R}^N)$, $\bar{x} \in \text{cont}(f)$, et $S = [f(x) \leq f(\bar{x})]$ le sous-niveau associé. Si $0 \notin \partial f(\bar{x})$, alors

$$T_S(\bar{x}) = \{d \in \mathbb{R}^N \mid df(\bar{x}; d) \leq 0\}.$$

Démonstration. On procède par double inclusion, et on note $D := \{d \in \mathbb{R}^N \mid df(\bar{x}; d) \leq 0\}$.
 \subset : Soit $d \in T_S(\bar{x})$, alors $d = \lim d_n$ avec $d_n = \lambda_n(c_n - \bar{x})$ où $c_n \in S$, c'est-à-dire $f(c_n) \leq f(\bar{x})$. On utilise la convexité de f pour écrire

$$df(\bar{x}; c_n - \bar{x}) = \lim_{t \rightarrow 0} \frac{f(\bar{x} + t(c_n - \bar{x}))}{t} \leq \lim_{t \rightarrow 0} \frac{(1-t)f(\bar{x}) + tf(c_n) - f(\bar{x})}{t} \leq 0.$$

Ainsi $c_n - \bar{x} \in D$. Or $df(\bar{x}; \cdot)$ est sous-linéaire et continue (voir les Théorème B.43 et Proposition B.40), donc d_n puis d appartiennent à D .

\supset : Soit $d \in D$. Prenons $x^* \in \partial f(\bar{x})$ quelconque (on sait qu'il existe grâce à la proposition B.46), et posons $d_n := d - \frac{1}{n}x^*$. Puisque $d_n \rightarrow d$ et que $T_S(\bar{x})$ est fermé (Proposition I.50), il suffit de montrer que $d_n \in T_S(\bar{x})$. Via la formule du max (Théorème B.43) on peut écrire

$$df(\bar{x}; d_n) = \sigma_{\partial f(\bar{x})}(d - \frac{1}{n}x^*) \leq \sup_{y^* \in \partial f(\bar{x})} \langle y^*, d \rangle - \inf_{z^* \in \partial f(\bar{x})} \langle z^*, \frac{1}{n}x^* \rangle \leq df(\bar{x}; d) - \frac{1}{n}\|x^*\|^2.$$

Or $df(\bar{x}; d) \leq 0$ et $x^* \neq 0$ par hypothèse, donc $df(\bar{x}; d_n) < 0$. En utilisant la formule globale de la Proposition B.39, on voit qu'il existe $t > 0$ tel que $f(\bar{x} + td_n) < f(\bar{x})$. Autrement dit, tel que $c := \bar{x} + td_n \in S$. On peut donc écrire $d_n = \lambda(c - \bar{x})$ avec $\lambda = t^{-1}$, et conclure que $d_n \in T_S(\bar{x})$. ■

Lemme B.48. Soit $C \subset \mathbb{R}^N$ compact convexe tel que $0 \notin C$. Alors $\text{cone}(C)$ est fermé.

Démonstration. D'après la Proposition I.35, nous avons $\text{cone}(C) = \mathbb{R}_+ \text{co}(C)$. Or C est convexe donc $\text{cone}(C) = \mathbb{R}_+ C$. Pour montrer que $\text{cone}(C)$ est fermé, considérons une suite $d_n \in \text{cone}(C)$ qui converge vers un $d \in \mathbb{R}^N$, et montrons que $d \in \text{cone}(C)$.

Par définition, $d_n = \lambda_n c_n$ où $\lambda_n \geq 0$ et $c_n \in C$. Puisque C est compact, et quitte à prendre une sous-suite, on peut dire que c_n converge vers un certain $c \in C$, qui est non nul d'après nos hypothèses. On observe que $\|d_n\| = \lambda_n \|c_n\|$, et donc $\lambda_n = \frac{\|d_n\|}{\|c_n\|} \rightarrow \frac{\|d\|}{\|c\|}$. On conclut donc que $d = \lambda c \in \text{cone}(C)$. ■

Proposition B.49 (Cône normal au sous-niveau). Soit $f \in \Gamma_0(\mathbb{R}^N)$, $\bar{x} \in \text{cont}(f)$, et $S = [f(x) \leq f(\bar{x})]$ le sous-niveau associé. Si $0 \notin \partial f(\bar{x})$, alors

$$N_S(\bar{x}) = \text{cone}(\partial f(\bar{x})).$$

Démonstration. Il s'agit de passer au polaire dans la formule du cône tangent de la Proposition B.47. On écrit, en utilisant la formule du max du Théorème II.75:

$$\begin{aligned} T_S(\bar{x}) &= \{d \in \mathbb{R}^N \mid df(\bar{x}; d) \leq 0\} \\ &= \{d \in \mathbb{R}^N \mid \sigma_{\partial f(\bar{x})}(d) \leq 0\} \\ &= \{d \in \mathbb{R}^N \mid (\forall x^* \in \partial f(\bar{x})) \langle x^*, d \rangle \leq 0\} \\ &= \{d \in \mathbb{R}^N \mid (\forall x^* \in \text{cone}(\partial f(\bar{x}))) \langle x^*, d \rangle \leq 0\} \\ &= (\text{cone}(\partial f(\bar{x})))^*. \end{aligned}$$

Or $\partial f(\bar{x})$ est convexe compact non vide d'après la Proposition B.46, donc $\text{cone}(\partial f(\bar{x}))$ est fermé d'après le Lemme B.48. On peut donc appliquer le Théorème du bidual I.45 et conclure que $N_S(\bar{x}) = (\text{cone}(\partial f(\bar{x})))^{**} = \text{cone}(\partial f(\bar{x}))$. ■

B.II.4 Dualité entre fonctions lisses et fortement convexes

L'objectif de cette section est de prouver le Théorème II.104.

Proposition B.50 (Monotonicité du sous-différentiel). Soit $f \in \Gamma_0(\mathbb{R}^N)$. Alors ∂f est monotone:

$$(\forall x, y \in \mathbb{R}^N)(\forall y^* \in \partial f(y))(\forall x^* \in \partial f(x)) \quad \langle y^* - x^*, y - x \rangle \geq 0.$$

Démonstration. Si on applique deux fois la définition de sous-gradient

$$\begin{aligned} f(y) - f(x) &\geq \langle x^*, y - x \rangle \\ f(x) - f(y) &\geq \langle y^*, x - y \rangle \end{aligned}$$

et que l'on fait la somme, on obtient bien

$$0 \geq \langle x^* - y^*, y - x \rangle.$$

■

On dispose d'une réciproque à la proposition B.50 lorsque la fonction est différentiable.

Proposition B.51 (Convexité via monotonie du gradient). Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$ différentiable. Alors f est convexe si et seulement si ∇f est monotone:

$$(\forall x, y \in \mathbb{R}^N) \quad \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0.$$

Démonstration. Si f est convexe alors ∂f est monotone, mais puisque f est différentiable nous avons $\partial f = \{\nabla f\}$. Montrons maintenant la réciproque. On commence par se donner $x, y \in \mathbb{R}^N$ fixés et on pose, pour tout $t \in \mathbb{R}$, le vecteur $z_t = (1-t)x + ty$ et la fonction $g : \mathbb{R} \rightarrow \mathbb{R}$ définie par $g(t) = f(z_t)$. Par composition cette fonction est différentiable, et on calcule que $g'(t) = \langle \nabla f(z_t), y - x \rangle$, avec en particulier $g'(0) = \langle \nabla f(x), y - x \rangle$. On peut alors écrire, pour tout $t > 0$, que

$$\begin{aligned} g'(t) - g'(0) &= \langle \nabla f(z_t) - \nabla f(x), y - x \rangle \\ &= \frac{1}{t} \langle \nabla f(z_t) - \nabla f(x), z_t - x \rangle \geq 0, \end{aligned}$$

où dans la dernière inégalité nous avons utilisé la monotonie de ∇f . Autrement dit, $g'(0) \leq g'(t)$ pour tout $t > 0$. Or g est continue sur $[0, 1]$ et dérivable sur $]0, 1[$ donc on peut utiliser le théorème des accroissements finis pour obtenir un $c \in]0, 1[$ tel que

$$g'(c) = \frac{g(1) - g(0)}{1 - 0} = g(1) - g(0) = f(y) - f(x).$$

On a donc prouvé que

$$(\forall x, y \in \mathbb{R}^N) \quad f(y) - f(x) \geq g'(0) = \langle \nabla f(x), y - x \rangle.$$

Si on considère à nouveau x, y quelconques, et $t \in]0, 1[$, on peut écrire avec cette inégalité

$$\begin{aligned} f(x) &\geq f(z_t) + \langle \nabla f(z_t), x - z_t \rangle \\ f(y) &\geq f(z_t) + \langle \nabla f(z_t), y - z_t \rangle. \end{aligned}$$

En sommant $(1-t)$ fois la première inégalité avec t fois la seconde, on obtient

$$(1-t)f(x) + tf(y) \geq f(z_t) + \langle \nabla f(z_t), (1-t)x + ty - z_t \rangle = f(z_t) = f((1-t)x + ty),$$

et ceci prouve bien que f est convexe. ■

Proposition B.52 (Forte monotonie du sous-différentiel). Soit $f \in \Gamma_\mu(\mathbb{R}^N)$. Alors ∂f est μ -fortement monotone:

$$(\forall x, y \in \mathbb{R}^N)(\forall y^* \in \partial f(y))(\forall x^* \in \partial f(x)) \quad \langle y^* - x^*, y - x \rangle \geq \mu \|y - x\|^2.$$

Démonstration. D'après la proposition II.50 on peut décomposer $f = \mu q + g$ où $g \in \Gamma_0(\mathbb{R}^N)$ et $q = \frac{1}{2}\|\cdot\|^2$. Puisque q est différentiable, on peut utiliser la règle de la somme simple (proposition II.71) pour écrire $\partial f(x) = \partial g(x) + \mu x$. Ainsi, pour $x^* \in \partial f(x)$ et $y^* \in \partial f(y)$ on a $x^* - \mu x \in \partial g(x)$ et $y^* - \mu y \in \partial g(y)$, ce qui nous permet de dire que

$$\langle y^* - x^*, y - x \rangle = \langle (y^* - \mu y) - (x^* - \mu x), y - x \rangle + \mu \|y - x\|^2 \geq \|y - x\|^2,$$

où dans la dernière inégalité nous avons utilisé la monotonie de ∂g , voir la proposition B.50. ■

Théorème B.53 (Conjuguée d'une fonction convexe lisse). Soit $f \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$. Alors $f^* \in \Gamma_\mu(\mathbb{R}^N)$ avec $\mu = L^{-1}$.

Démonstration. Puisque $f \in \Gamma_0(\mathbb{R}^N)$, on sait d'après la proposition II.88 que $f^* \in \Gamma_0(\mathbb{R}^N)$. Donc il suffit de montrer que f^* est fortement convexe. Dans cette preuve on note $q(x) = \frac{1}{2}\|x\|^2$, et $h(x) = Lq(x) - f(x)$. Puisque f est différentiable, et que q aussi, alors h est différentiable, et on a $\nabla h(x) = Lx - \nabla f(x)$. De plus ∇f est L -Lipschitz, donc on peut écrire

$$\begin{aligned} \langle \nabla h(y) - \nabla h(x), y - x \rangle &= L\|y - x\|^2 - \langle \nabla f(y) - \nabla f(x), y - x \rangle \\ &\geq L\|y - x\|^2 - \|y - x\| \|\nabla f(y) - \nabla f(x)\| \\ &\geq L\|y - x\|^2 - L\|y - x\| \|y - x\| = 0. \end{aligned}$$

Donc ∇h est monotone, donc h est convexe en vertu de la proposition B.51. En particulier $h \in \Gamma_0(\mathbb{R}^N)$.

On va maintenant calculer f^* en fonction de h . On part du fait que $f = Lq - h$ et que $h^{**} = h$ pour écrire

$$f = Lq - h^{**} = Lq - \sup_x \langle \cdot, x \rangle - h^*(x) = \inf_x Lq - \langle x, \cdot \rangle - h^*(x).$$

Donc $f = \inf_x \varphi_x$ où $\varphi_x(u) = Lq(u) - \langle x, u \rangle - h^*(x)$. On peut donc utiliser le lemme B.42

sur la conjuguée de l'inf pour écrire

$$\begin{aligned}
 f^* &= (\inf_x \varphi_x)^* = \sup_x \varphi_x^* = \sup_x (Lq - \langle x, \cdot \rangle - h^*(x))^* \\
 &= \sup_x (Lq - \langle x, \cdot \rangle)^* + h^*(x) \quad \text{car } h^*(x) \text{ est une constante} \\
 &= \sup_x (Lq)^*(\cdot + x) + h^*(x) \\
 &= \sup_x \frac{1}{L} q(\cdot + x) + h^*(x) \\
 &= \sup_x \frac{1}{L} q(\cdot) + \frac{1}{L} q(x) + \frac{1}{L} \langle \cdot, x \rangle + h^*(x) \\
 &= \frac{1}{L} q + \sup_x \frac{1}{L} \langle \cdot, x \rangle - (h^* - \frac{1}{L} q)(x) \\
 &= \frac{1}{L} q + (h^* - \frac{1}{L} q)^*.
 \end{aligned}$$

On a donc écrit f^* comme une somme entre $\frac{1}{L}q$ et $g := (h^* - \frac{1}{L}q)^*$. Pour conclure il suffit d'appliquer le critère de la proposition II.50, et pour cela il faut vérifier que g est convexe. Or g est une conjuguée, donc par définition c'est un sup de fonctions affines, donc convexe. ■

Théorème B.54 (Conjuguée d'une fonction fortement convexe). Soit $f \in \Gamma_\mu(\mathbb{R}^N)$. Alors $f^* \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$ avec $L = \mu^{-1}$.

Démonstration. Pour commencer, observons que la forte convexité de f implique que $\text{dom } f^* = \mathbb{R}^N$. En effet, pour tout $x^* \in \mathbb{R}^N$ on a

$$f^*(x^*) = \sup_{x \in \mathbb{R}^N} \langle x^*, x \rangle - f(x) = - \inf_{x \in \mathbb{R}^N} f(x) - \langle x^*, x \rangle.$$

Or f étant fortement convexe et la forme linéaire convexe, on sait que $f - \langle x^*, \cdot \rangle$ est fortement convexe (voir II.52). Donc cette fonction admet bien un minimiseur d'après le théorème II.53, ce qui garantit que $f^*(x^*) < +\infty$.

Nous avons donc montré que f^* est définie partout. Ceci implique que f^* est continue sur \mathbb{R}^N , c'est une conséquence du théorème II.44. Maintenant soit $x^* \in \mathbb{R}^N$ et montrons que f^* est différentiable en x^* . On peut déjà utiliser la proposition II.63 pour voir que $\text{dom } \partial f^* = \mathbb{R}^N$, ce qui implique que $\partial f^*(x^*) \neq \emptyset$. Montrons que ce sous-différentiel est en fait un singleton. Pour cela, on prend $x, y \in \partial f^*(x^*)$ (qui est non vide), et on utilise la formule de Legendre-Fenchel (proposition II.102) pour voir que $x^* \in \partial f(x)$ et $x^* \in \partial f(y)$. En utilisant la proposition B.52 sur la forte monotonie du sous-différentiel, on voit que

$$\langle x^* - x^*, y - x \rangle \geq \mu \|y - x\|^2.$$

Le terme de gauche étant nul, on en déduit que $y = x$. Donc $\partial f^*(x^*)$ est un singleton. On déduit alors du théorème B.45 que f^* est différentiable en x^* , et donc sur tout \mathbb{R}^N .

Pour conclure, il nous reste à montrer que ∇f^* est Lipschitzien. On se donne $x^*, y^* \in \mathbb{R}^N$, et on pose $x = \nabla f^*(x^*)$ et $y = \nabla f^*(y^*)$. La formule de Legendre-Fenchel (proposition II.102) nous dit que $x^* \in \partial f(x)$ et $y^* \in \partial f(y)$. On peut donc appliquer la proposition B.52 sur la forte monotonie du sous-différentiel:

$$\langle y^* - x^*, y - x \rangle \geq \mu \|y - x\|^2.$$

Après avoir utilisé l'inégalité de Cauchy-Schwarz, on voit que

$$\|\nabla f^*(y^*) - \nabla f^*(x^*)\| = \|y - x\| \leq \frac{1}{\mu} \|y^* - x^*\|,$$

et donc que ∇f^* est Lipschitzienne. ■

Corollaire B.55. *Le théorème II.104 sur la dualité entre fonctions lisses et fortement convexes est vrai.*

Démonstration. Il suffit de combiner les théorèmes B.53 et B.54. ■

B.III Résultats avancés sur les algorithmes

On donne ici quelques compléments sur les algorithmes pour résoudre des problèmes d'optimisation, ainsi que quelques preuves manquantes.

- Dans le chapitre III nous avons abouti à un algorithme d'éclatement permettant de résoudre des problèmes faisant intervenir une fonction lisse et *deux* fonctions non lisses, avec possiblement un opérateur linéaire. Dans la section B.III.1 on vérifie que si l'on sait gérer deux fonctions non lisses alors on peut aussi gérer n'importe quel nombre (fini) de fonctions non lisses.
- Dans la section B.III.2 nous prouvons le résultat principal manquant dans le chapitre III, à savoir la preuve de convergence de l'algorithme de Davis-Yin.
- Dans les sections B.III.3 et B.III.4 nous présentons des algorithmes dits lagrangiens pour minimiser des fonctions sous contraintes linéaires. Leur intérêt est que leur définition est très naturelle, et que toutes ces méthodes sont duales de tous les algorithmes vus précédemment: gradient, proximal, gradient proximal, Douglas-Rachford, Davis-Yin. Ceci illustre que la dualité peut aussi s'appliquer aux algorithmes !

B.III.1 Éclatement Total

Dans la section B.III.1.i) nous montrons que l'algorithme de Davis-Yin peut être adapté pour traiter n'importe quelle somme finie de fonctions lisses et non lisses. Dans la section B.III.1.ii) nous montrons que l'algorithme de Yan peut être adapté pour traiter n'importe quelle somme finie de fonctions lisses et non lisses, possiblement composées avec des opérateurs linéaires. Dans chaque cas, l'astuce consiste à déplacer notre problème dans un espace produit pour avoir une fonction séparable (pour laquelle le calcul du prox est trivial), puis à rajouter une contrainte imposant que toutes nos nouvelles variables soient égales (pour laquelle la projection se calcule aisément).

B.III.1.i) Algorithme d'Éclatement Total via Davis-Yin

Si on dispose d'une somme finie quelconque de fonctions non lisses, on peut toujours se ramener à deux fonctions non lisses, quitte à introduire des variables supplémentaires.

Lemme B.56. Soient $g_1, \dots, g_p, h \in \Gamma_0(\mathbb{R}^N)$, avec $h \in C_L^{1,1}(\mathbb{R}^N)$. Alors le problème

$$(P) \quad \min_{x \in \mathbb{R}^N} g_1(x) + \dots + g_p(x) + h(x)$$

est équivalent au problème

$$(\hat{P}) \quad \min_{X=(x_1, \dots, x_p) \in (\mathbb{R}^N)^p} F(X) + G(X) + H(X),$$

où

- $F(X) = \delta_F(X)$ où $F = \{X \in (\mathbb{R}^N)^p \mid x_1 = \dots = x_p\}$, avec $\text{prox}_{\lambda F}(X) = \left(\frac{1}{p} \sum_{j=1}^p x_j\right)_{i=1}^p$;
- $G(X) = g_1(x_1) + \dots + g_p(x_p)$, avec $\text{prox}_{\lambda G}(X) = (\text{prox}_{\lambda g_i}(x_i))_{i=1}^p$;
- $H(X) = h(\bar{x})$ où $\bar{x} = \frac{1}{p} \sum_{j=1}^p x_j$, avec $\nabla H(X) = \frac{1}{p}(\nabla h(\bar{x}), \dots, \nabla h(\bar{x}))$.

Démonstration. Le fait que les problèmes sont équivalents est immédiat puisque la contrainte $X \in F$ impose que toutes les variables x_i soient égales. Le prox de G se calcule car c'est une somme directe, et le calcul de la projection sur F est un simple exercice (voir TD). ■

On dispose alors d'un algorithme d'éclatement total:

Exemple B.57 (Algorithme d'éclatement total). Supposons que l'on veuille minimiser $g_1(x) + \dots + g_p(x) + h(x)$, où $g_1, \dots, g_p, h \in \Gamma_0(\mathbb{R}^N)$, avec $h \in C_L^{1,1}(\mathbb{R}^N)$. Après réécriture du problème à l'aide du Lemme B.56, l'algorithme de Davis-Yin dans sa forme primale (DY') prend la forme :

$$\begin{cases} x_{n+1} = \frac{1}{p} \sum_{j=1}^p Y_{n,j} \\ \hat{X}_{n+1,i} = \text{prox}_{\lambda g_i}(2x_{n+1} - Y_{n,i} - \frac{\lambda}{p} \nabla h(x_{n+1})) \\ Y_{n+1,i} = Y_{n,i} + \hat{X}_{n+1,i} - x_{n+1}, \end{cases} \quad (\text{ET})$$

où (ET) renvoie à « Éclatement Total ».

Théorème B.58 (Convergence de l'éclatement total). Soient $g_1, \dots, g_p, h \in \Gamma_0(\mathbb{R}^N)$, avec $h \in C_L^{1,1}(\mathbb{R}^N)$. On suppose que $g_1 + \dots + g_p + h$ admet un minimiseur non dégénéré :

$$(\exists x \in \mathbb{R}^N) \quad 0 \in \partial g_1(x) + \dots + \partial g_p(x) + \nabla h(x).$$

Soit $(x_n)_{n \in \mathbb{N}}$ généré par (ET), avec un pas $0 < \lambda < \frac{2p}{L}$. Alors x_n converge vers $\bar{x} \in \arg\min(g_1 + \dots + g_p + h)$, lorsque $n \rightarrow +\infty$.

Démonstration. Le Théorème III.35 garantit que x_n va converger vers une solution, pourvu que deux hypothèses soient vérifiées : un pas suffisamment petit, et l'existence d'une solution non dégénérée.

Tout d'abord, il est facile de calculer que $\text{Lip}(\nabla H) = \frac{1}{p} \text{Lip}(\nabla h) = \frac{L}{p}$. En effet,

$$\|\nabla H(Y) - \nabla H(X)\|^2 = \sum_{i=1}^p \frac{1}{p^2} \|\nabla h(\bar{y}) - \nabla h(\bar{x})\|^2 = \frac{1}{p} \|\nabla h(\bar{y}) - \nabla h(\bar{x})\|^2 \leq \frac{L^2}{p} \|\bar{y} - \bar{x}\|^2,$$

où $\bar{x} = (1/p) \sum_i x_i$. Si on note $M = (1/p)[I \dots I]$ telle que $MX = \bar{x}$ et $\|M\|^2 \leq (1/p)$, alors on peut écrire

$$\frac{L^2}{p} \|\bar{y} - \bar{x}\|^2 = \frac{L^2}{p} \|MY - MX\|^2 \leq \frac{L^2}{p} \|M\|^2 \|Y - M\|^2 \leq \frac{L^2}{p^2} \|Y - M\|^2.$$

Donc il nous faut prendre un pas $\lambda < \frac{2p}{L}$.

Ensuite il nous faut l'existence d'une solution non dégénérée pour $F + G + H$. C'est-à-dire que l'on veut un X tel que $0 \in \partial f(X) + \partial g(X) + \nabla h(X)$. On a fait l'hypothèse qu'il existe x et $u_i \in \partial g_i(x)$ tels que $0 = \sum_i u_i + \nabla h(x)$. Posons alors $X = (x, \dots, x)$ et vérifions qu'il est le vecteur que l'on recherche. D'une part, $X \in F$ par définition, donc $\partial F(X) = N_F(X) = F^\perp$, où il est facile de calculer que $F^\perp = \{y \mid \sum_i y_i = 0\}$. Ensuite $\partial G(X) = \prod_i \partial g_i(x)$ et $\nabla H(X) = \prod_i (1/p) \nabla h(x)$. Donc il suffit de montrer qu'il existe $y \in F^\perp$ tel que pour tout i on ait $0 \in y_i + \partial g_i(x) + (1/p) \nabla h(x)$. Vu qu'on dispose de $u_i \in \partial g_i(x)$, on va poser $y_i := -u_i - (1/p) \nabla h(x)$ qui vérifie par définition l'inclusion désirée, et il nous suffit de vérifier que $y \in F^\perp$. Et en effet $\sum_i y_i = -\sum_i u_i - (1/p) \sum_i \nabla h(x) = -\sum_i u_i - \nabla h(x) = 0$. ■

B.III.1.ii) Algorithme d'Éclatement Composite Total via Yan

On montre que l'algorithme de Yan (Yan) est un algorithme d'éclatement composite total : si on dispose d'une somme quelconque de fonctions non lisses composées avec des applications linéaires, alors on peut se ramener au cas de deux fonctions non lisses.

Lemme B.59 (Réduction d'un problème d'éclatement composite total). Soient $f, h \in \Gamma_0(\mathbb{R}^N)$, $h \in C_L^{1,1}(\mathbb{R}^N)$, $g_i \in \Gamma_0(\mathbb{R}^{M_i})$, et $A_i \in \mathcal{M}_{M_i, N}(\mathbb{R})$. Alors le problème

$$(P) \quad \min_{x \in \mathbb{R}^N} f(x) + g_1(A_1 x) + \cdots + g_p(A_p x) + h(x)$$

est équivalent au problème

$$(\hat{P}) \quad \min_{x \in \mathbb{R}^N} f(x) + G(Ax) + h(x),$$

où

- $A = [A_1; \dots; A_p]$, telle que $Ax = (A_1 x, \dots, A_p x)$ et $A^\top y = \sum_{j=1}^p A_j^\top y_j$;
- $G(y) = \sum_{j=1}^p g_j(y_j)$, telle que $G^*(w) = \sum_{j=1}^p g_j^*(w_j)$ et $\text{prox}_{\sigma G^*}(w) = (\text{prox}_{\sigma g_j^*}(w_j))_{j=1}^p$.

Démonstration. La preuve est immédiate, il s'agit juste de réécrire le problème. Le calcul de la conjuguée et du prox de G vient du fait que G est une somme directe. ■

Exemple B.60 (Algorithme d'éclatement composite total). Supposons que l'on veuille minimiser $f(x) + g_1(A_1 x) + \cdots + g_p(A_p x) + h(x)$. Après réécriture du problème à l'aide du Lemme B.59, l'algorithme de Yan prend la forme :

$$\begin{cases} x_{n+1} = \text{prox}_{\lambda f}(x_n - \lambda \nabla h(x_n) - \lambda \sum_{j=1}^p A_j^\top w_{n,j}) \\ w_{n+1,i} = \text{prox}_{\sigma g_i^*}(w_{n,i} + \sigma A_i [2x_{n+1} - x_n + \lambda \nabla h(x_n) - \lambda \nabla h(x_{n+1})]), \end{cases} \quad (\text{ECT})$$

où (ECT) réfère à « Éclatement Composite Total ».

Théorème B.61 (Convergence de l'éclatement composite total). Soient $f, h \in \Gamma_0(\mathbb{R}^N)$, $h \in C_L^{1,1}(\mathbb{R}^N)$, $g_i \in \Gamma_0(\mathbb{R}^{M_i})$, et $A_i \in \mathcal{M}_{M_i, N}(\mathbb{R})$. On suppose que le problème associé admet un minimiseur non dégénéré :

$$(\exists x \in \mathbb{R}^N) \quad 0 \in \partial f(x) + A_1^\top \partial g_1(A_1 x) + \cdots + A_p^\top \partial g_p(A_p x) + \nabla h(x).$$

Soit $(x_n)_{n \in \mathbb{N}}$ générée par (ECT), avec des pas $0 < \lambda < \frac{2}{L}$, et $\sigma \leq \frac{1}{\lambda \sum_{j=1}^p \|A_j\|^2}$. Alors x_n converge vers $\bar{x} \in \operatorname{argmin} f + \sum_i (g_i \circ A_i) + h$ lorsque $n \rightarrow +\infty$.

Démonstration. Ici notre suite est générée par l'algorithme de Yan appliqué au problème réduit du Lemme B.59, qui consiste à minimiser $f + G \circ A + h$. Il nous faut donc vérifier les hypothèses du Théorème III.42 appliquée à ce problème. Commençons par les conditions sur les pas. Concernant λ c'est direct puisque la fonction lisse h est la même. Concernant σ , il nous faut $\sigma \leq 1/(\lambda \|A\|^2)$. Or il est facile de vérifier que $\|A\|^2 \leq \sum_i \|A_i\|^2$:

$$\|Ax\|^2 = \sum_{i=1}^p \|A_i x\|^2 \leq \sum_{i=1}^p \|A_i\|^2 \|x\|^2 = (\sum_{i=1}^p \|A_i\|^2) \|x\|^2.$$

Donc notre choix pour σ est suffisant. Ensuite il faut vérifier l'existence d'un minimiseur non-dégénéré pour $f + G \circ A + h$. Partons de notre hypothèse, qui est qu'il existe x tel que

$$0 \in \partial f(x) + A_1^\top \partial g_1(A_1 x) + \cdots + A_p^\top \partial g_p(A_p x) + \nabla h(x).$$

Donc il existe $u_i \in \partial g_i(A_i x)$ tels que

$$0 \in \partial f(x) + A_1^\top u_1 + \cdots + A_p^\top u_p + \nabla h(x).$$

Par définition de A^\top , nous avons $A_1^\top u_1 + \cdots + A_p^\top u_p = A^\top u$ où $u = (u_1, \dots, u_p)$. De plus, $\partial G(Ax) = \prod_i \partial g_i(A_i x) \ni u$. On en déduit que $0 \in \partial f(x) + A^\top \partial G(Ax) + \nabla h(x)$, ce qui veut dire que x est un minimiseur non-dégénéré pour $f + G \circ A + h$. ■

B.III.2 Preuve de convergence de Davis-Yin

Nous prouvons maintenant la convergence de l'algorithme de Davis-Yin. Il en découlera la convergence des algorithmes Gradient, Proximal, Proximal-Gradient (car c'en sont des cas particuliers), ainsi que de l'algorithme de Yan (qui en est aussi un cas particulier).

B.III.2.i Préliminaires

Cet algorithme combine des opérateurs proximaux et des gradients de fonctions lisses. Nous allons donc commencer par établir deux résultats les concernant :

Proposition B.62 (Gradient est cocoercif). Soit $f \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$. Alors ∇f est $1/L$ -cocoercif :

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

Démonstration. Il suffit d'utiliser le fait que $f^* \in \Gamma_\mu(\mathbb{R}^N)$ avec $\mu = 1/L$, et le fait que ∂f^* est μ -fortement monotone (voir la proposition B.52 dans la section B.II.4):

$$(\forall y^*, x^* \in \mathbb{R}^N) (\forall y \in \partial f^*(y^*), x \in \partial f^*(x^*)) \quad \langle y - x, y^* - x^* \rangle \geq \mu \|y^* - x^*\|^2.$$

On conclut avec la formule de Legendre-Fenchel : $x^* = \nabla f(x) \Leftrightarrow x \in \partial f^*(x^*)$. ■

Proposition B.63 (Prox est 1-cocoercif). Soit $f \in \Gamma_0(\mathbb{R}^N)$. Alors prox_f est 1-cocoercif :

$$(\forall y, x \in \mathbb{R}^N) \quad \langle \text{prox}_f(y) - \text{prox}_f(x), y - x \rangle \geq \|\text{prox}_f(y) - \text{prox}_f(x)\|^2.$$

En particulier prox_f est 1-Lipschitzien.

Démonstration. Soient $y, x \in \mathbb{R}^N$, et on note $p_x = \text{prox}_f(x)$, $p_y = \text{prox}_f(y)$. La caractérisation du prox par le sous-différentiel nous dit que $x - p_x \in \partial f(p_x)$, donc par définition du sous-différentiel :

$$f(p_y) - f(p_x) - \langle x - p_x, p_y - p_x \rangle \geq 0.$$

Idem pour p_y , qui nous donne une autre inégalité

$$f(p_x) - f(p_y) - \langle y - p_y, p_x - p_y \rangle \geq 0$$

telle qu'en faisant la somme de ces deux inégalités on obtienne

$$\langle p_x - x - p_y + y, p_y - p_x \rangle \geq 0$$

qui est équivalente à l'inégalité désirée. La Lipschitzianité s'obtient en appliquant Cauchy-Schwarz. ■

Corollaire B.64. La proposition I.25 sur la continuité de la projection est vraie.

Démonstration. C'est une conséquence de la proposition précédente, en utilisant le fait que la projection est l'opérateur proximal de l'indicatrice (voir exemple III.12). ■

Nous allons devoir prouver la convergence d'une suite. Pour cela nous allons utiliser un lemme relativement simple mais néanmoins extrêmement utile:

Lemme B.65 (Opial). Soit $S \subset \mathbb{R}^N$, et $(x_n)_{n \in \mathbb{N}}$ une suite. On fait l'hypothèse que

- 1) S est non vide;

- 2) pour tout $x \in S$, la suite $(\|x_n - x\|)_{n \in \mathbb{N}}$ est convergente;
- 3) pour toute valeur d'adhérence x_∞ de x_n , on a $x_\infty \in S$.

Alors x_n converge vers $\bar{x} \in S$, lorsque $n \rightarrow +\infty$.

Démonstration. Tout d'abord, la suite x_n est bornée. En effet $S \neq \emptyset$ donc on peut prendre $x \in S$ et écrire $\|x_n\| \leq \|x\| + \|x_n - x\|$. Puisque $\|x_n - x\|$ converge par hypothèse, elle est bornée; donc x_n est bornée. Par Weierstrass, x_n admet donc une sous-suite convergente : $x_{n_k} \rightarrow x_\infty$ lorsque $k \rightarrow +\infty$. Par hypothèse, on sait que $x_\infty \in S$ car c'est une valeur d'adhérence. On peut donc utiliser la première hypothèse pour dire que $\|x_n - x_\infty\|$ converge, et on note $\ell \in \mathbb{R}$ la limite de cette suite. Si on regarde la sous-suite $\|x_{n_k} - x_\infty\|$, elle doit aussi converger vers ℓ . Or on sait que $x_{n_k} - x_\infty \rightarrow 0$, donc forcément $\ell = 0$. On en déduit que $\|x_n - x_\infty\| \rightarrow 0$, ce qui veut dire que x_n tend vers x_∞ . ■

On termine ces préliminaires en exhibant la forme primale de l'algorithme de Davis-Yin, qui sera plus simple à analyser.

Lemme B.66 (Forme primale de Davis-Yin). *L'algorithme de Davis-Yin (DY) est équivalent à sa forme primale (DY').*

Démonstration. On rappelle ici les formes primale-duale et primale de Davis-Yin:

$$\begin{cases} x_{n+1} = \text{prox}_{\lambda f}(x_n - \lambda \nabla h(x_n) - \lambda u_n), \\ u_{n+1} = \text{prox}_{\frac{1}{\lambda} g^*} \left(u_n + \frac{1}{\lambda} [2x_{n+1} - x_n - \lambda \nabla h(x_{n+1}) + \lambda \nabla h(x_n)] \right) \end{cases} \quad (\text{DY})$$

$$\begin{cases} x_{n+1} = \text{prox}_{\lambda f}(y_n) \\ \hat{x}_{n+1} = \text{prox}_{\lambda g}(2x_{n+1} - y_n - \nabla h(x_{n+1})) \\ y_{n+1} = y_n + \hat{x}_{n+1} - x_{n+1}. \end{cases} \quad (\text{DY}')$$

Si on pose $y_n = x_n - \lambda \nabla h(x_n) - \lambda u_n$ alors on voit immédiatement que la première ligne de (DY) devient celle de (DY'). On veut maintenant trouver une relation de récurrence sur y_n . Par définition on a

$$y_{n+1} = x_{n+1} - \lambda \nabla h(x_{n+1}) - \lambda u_{n+1}$$

donc il nous faut développer $-\lambda u_{n+1}$. D'après (DY), u_{n+1} s'obtient en appliquant un opérateur proximal au vecteur

$$u_n + \frac{1}{\lambda} [2x_{n+1} - x_n - \lambda \nabla h(x_{n+1}) + \lambda \nabla h(x_n)] = \frac{1}{\lambda} [2x_{n+1} - y_n - \lambda \nabla h(x_{n+1})].$$

Si on invoque la formule de Moreau

$$\lambda \text{prox}_{\frac{1}{\lambda} g^*} \left(\frac{1}{\lambda} X \right) = X - \text{prox}_{\lambda g}(X)$$

alors on peut écrire

$$\begin{aligned}-\lambda u_{n+1} &= -\lambda \text{prox}_{\frac{1}{\lambda}g^*}\left(\frac{1}{\lambda}[2x_{n+1} - y_n - \lambda \nabla h(x_{n+1})]\right) \\ &= \text{prox}_{\lambda g}(2x_{n+1} - y_n - \lambda \nabla h(x_{n+1})) - [2x_{n+1} - y_n - \lambda \nabla h(x_{n+1})].\end{aligned}$$

Ainsi si on pose $\hat{x}_{n+1} := \text{prox}_{\lambda g}(2x_{n+1} - y_n - \lambda \nabla h(x_{n+1}))$ on conclut que

$$\begin{aligned}y_{n+1} &= x_{n+1} - \lambda \nabla h(x_{n+1}) - \lambda u_{n+1} \\ &= x_{n+1} - \lambda \nabla h(x_{n+1}) + \hat{x}_{n+1} - 2x_{n+1} + y_n + \lambda \nabla h(x_{n+1}) \\ &= y_n + \hat{x}_{n+1} - x_{n+1}.\end{aligned}$$



B.III.2.ii) Théorie des opérateurs de point fixe

Maintenant il nous reste à étudier l'algorithme de Davis-Yin lui-même. Nous allons voir que c'est un algorithme de *point fixe*, et ainsi utiliser des arguments classiques de la théorie des points fixes. Dans la suite, pour une application $T : \mathbb{R}^N \rightarrow \mathbb{R}^N$, on notera $\text{Fix } T$ l'ensemble de ses points fixes, c'est-à-dire

$$\text{Fix } T = \{y \in \mathbb{R}^N \mid Ty = y\}.$$

Nous allons montrer que

- 1) étudier l'algorithme de Davis-Yin revient à étudier un certain opérateur T ;
- 2) les points fixes de T sont liés aux minimiseurs de notre fonction $f + g + h$;
- 3) l'opérateur T a de bonnes propriétés, au sens où il est 1-Lipschitzien (et même un peu plus).

Lemme B.67 (Points fixes de Davis-Yin). *L'algorithme de Davis-Yin (DY') peut s'écrire $x_{n+1} = \text{prox}_{\lambda f}(y_n)$ et $y_{n+1} = Ty_n$, où*

$$T = I - P_g + P_f \circ (2P_g - I - \lambda \nabla h \circ P_g).$$

où $P_g = \text{prox}_{\lambda g}$ et $P_f = \text{prox}_{\lambda f}$. De plus, $\text{prox}_{\lambda g}(\text{Fix } T)$ est exactement l'ensemble des minimiseurs non dégénérés de $f + g + h$, c'est-à-dire

$$\text{prox}_{\lambda g}(\text{Fix } T) = \{x \in \mathbb{R}^N \mid 0 \in \partial f(x) + \partial g(x) + \nabla h(x)\}.$$

Démonstration. La réécriture de l'algorithme est juste une manière d'écrire (DY') de façon condensée. En effet

$$y_{n+1} = y_n + \hat{x}_{n+1} - x_{n+1} = I(y_n) - \text{prox}_{\lambda g}(y_n) + \hat{x}_{n+1},$$

où $\hat{x}_{n+1} = \text{prox}_{\lambda f}(2 \text{prox}_{\lambda g}(y_n) - y_n - \lambda \nabla h(\text{prox}_{\lambda g}(y_n)))$. Montrons maintenant le lien entre points fixes de T et minimiseurs non dégénérés, par une double inclusion.

\supset : Soit x un minimiseur non dégénéré : alors il existe $u_f \in \partial f(x)$ et $u_g \in \partial g(x)$ tels que $0 = u_f + u_g + \nabla h(x)$. Posons $y = x + \lambda u_g$. Alors par définition nous avons $y - x \in \partial g(x)$, ce qui veut dire que $x = P_g(y)$. Il nous reste donc à montrer que $y \in \text{Fix } T$, afin de valider l'inclusion désirée. Autrement dit, montrons que $Ty = y$. On a

$$2P_g y - y - \lambda \nabla h(P_g y) = 2x - y - \lambda \nabla h(x) = x - \lambda \nabla h(x) - \lambda u_g = x + \lambda u_f,$$

où dans la dernière égalité on utilise le fait que par définition, $0 = u_f + u_g + \nabla h(x)$. Notons que $2P_g y - y - \lambda \nabla h(P_g y) - x \in \lambda \partial f(x)$, ce qui veut dire que $x = P_f(x)$. Donc

$$P_f \circ (2P_g - I - \lambda \nabla h \circ P_g)(y) = P_f(x + \lambda u_f) = x,$$

où dans la dernière égalité on utilise le fait que $(x + \lambda u_f) - x \in \lambda \partial f(x)$. Donc on a

$$Ty = y - P_g y + P_f(x + \lambda u_f) = y - x + x = y.$$

\subset : Soit $y \in \text{Fix } T$, et $x = P_g y$. Montrons que x est un minimiseur non dégénéré de $f + g + h$. On utilise le fait que y soit un point fixe pour écrire

$$y = Ty = y - P_g y + P_f [2P_g(y) - y - \lambda \nabla h(P_g y)] = y - x + P_f [2x - y - \lambda \nabla h(x)].$$

Si on pose $p_f = P_f[\dots]$, alors on a $y = y - x + p_f$. Donc clairement $p_f = P_g y = x$. Maintenant, puisque $x = P_g y$ alors on a $u_f := (y - x)/\lambda \in \partial g(x)$. De même, puisque $x = P_f[\dots]$, alors

$$u_g := ((2x - y - \lambda \nabla h(x)) - x)/\lambda = (x - y)/\lambda - \nabla h(x) \in \partial g(x).$$

Alors on a bien

$$u_f + u_g + \nabla h(x) = (y - x)/\lambda + (x - y)/\lambda - \nabla h(x) + \nabla h(x) = 0.$$

■

Lorsqu'on travaille avec un algorithme de point fixe, une propriété essentielle à montrer est que l'opérateur sous-jacent est (fermement) non expansif:

Lemme B.68 (T est non-expansif). *L'opérateur T défini dans le Lemme B.67 vérifie :*

$$(\forall y, x \in \mathbb{R}^N) \quad \|Ty - Tx\|^2 \leq \|y - x\|^2 - \frac{2 - \lambda L}{2} \|(I - T)y - (I - T)x\|^2.$$

En particulier T est 1-Lipschitzien si $\lambda \leq 2/L$.

Démonstration. On garde les notations $P_g = \text{prox}_{\lambda g}$ et $P_f = \text{prox}_{\lambda f}$, et on introduit en plus: $P_g^* = \text{prox}_{(\lambda g)^*}$, $U = \lambda \nabla h \circ P_g$ et $V = 2P_g - I - U$, tels que (on utilise la décomposition de Moreau pour la première égalité) :

$$P_g^* = I - P_g, \quad V + 2P_g^* = I - U \quad \text{et} \quad T = P_f V + P_g^*.$$

On commence par développer le carré en utilisant le fait que $T = P_f V + P_g^*$:

$$\|Ty - Tx\|^2 = \|P_f Vy - P_f Vx\|^2 + \|P_g^* y - P_g^* x\|^2 + 2\langle P_f Vy - P_f Vx, P_g^* y - P_g^* x \rangle.$$

Par définition P_f et P_g^* sont des opérateurs proximaux, donc ils sont 1-cocoercifs (Proposition B.63), ce qui nous permet d'écrire

$$\begin{aligned} & \|Ty - Tx\|^2 \\ & \leq \langle P_f Vy - P_f Vx, Vy - Vx \rangle + \langle P_g^* y - P_g^* x, y - x \rangle + 2\langle P_f Vy - P_f Vx, P_g^* y - P_g^* x \rangle \\ & = \langle P_g^* y - P_g^* x, y - x \rangle + \langle P_f Vy - P_f Vx, (V + 2P_g^*)y - (V + 2P_g^*)x \rangle \\ & = \langle P_g^* y - P_g^* x, y - x \rangle + \langle P_f Vy - P_f Vx, (I - U)y - (I - U)x \rangle \\ & = \langle (P_g^* + P_f V)y - (P_g^* + P_f V)x, y - x \rangle - \langle P_f Vy - P_f Vx, Uy - Ux \rangle \\ & = \langle Ty - Tx, y - x \rangle - \langle P_f Vy - P_f Vx, Uy - Ux \rangle. \end{aligned}$$

Si on multiplie par deux, et qu'on utilise l'identité $2\langle a, b \rangle = \|a\|^2 + \|b\|^2 - \|b - a\|^2$ avec $a = Ty - Tx$ et $b = y - x$, alors on obtient après réorganisation des termes :

$$\|Ty - Tx\|^2 \leq \|y - x\|^2 - \|(I - T)y - (I - T)x\|^2 - 2\langle P_f Vy - P_f Vx, Uy - Ux \rangle.$$

Il nous reste à étudier ce dernier produit scalaire. Puisque $-P_f V = P_g^* - T = (I - T) - P_g$, alors on a

$$-2\langle \dots \rangle = 2\langle (I - T)y - (I - T)x, Uy - Ux \rangle - 2\langle P_g y - P_g x, Uy - Ux \rangle.$$

D'une part on utilise la définition de U et le fait que ∇h est $1/L$ -cocoercif (Proposition B.62) pour écrire

$$\begin{aligned} \langle P_g y - P_g x, Uy - Ux \rangle &= \lambda \langle P_g y - P_g x, \nabla h(P_g y) - \nabla h(P_g x) \rangle \\ &\geq \frac{\lambda}{L} \|\nabla h(P_g y) - \nabla h(P_g x)\|^2 = \frac{1}{\lambda L} \|Uy - Ux\|^2. \end{aligned}$$

D'autre part, on peut utiliser l'inégalité de Young $\langle a, b \rangle \leq \frac{\varepsilon}{2}\|a\|^2 + \frac{1}{2\varepsilon}\|b\|^2$ avec $a = (I - T)y - (I - T)x$, $b = Uy - Ux$ pour écrire

$$\langle (I - T)y - (I - T)x, Uy - Ux \rangle \leq \frac{\varepsilon}{2}\|(I - T)y - (I - T)x\|^2 + \frac{1}{2\varepsilon}\|Uy - Ux\|^2.$$

En combinant toutes ces inégalités, on obtient au final

$$\begin{aligned}
 & \|Ty - Tx\|^2 \\
 & \leq \|y - x\|^2 - \|(I - T)y - (I - T)x\|^2 \\
 & \quad + \varepsilon \|(I - T)y - (I - T)x\|^2 + \frac{1}{\varepsilon} \|Uy - Ux\|^2 - \frac{2}{\lambda L} \|Uy - Ux\|^2 \\
 & = \|y - x\|^2 + (\varepsilon - 1) \|(I - T)y - (I - T)x\|^2 + \left(\frac{1}{\varepsilon} - \frac{2}{\lambda L} \right) \|Uy - Ux\|^2.
 \end{aligned}$$

Ceci est vrai pour tout $\varepsilon > 0$, donc si on prend $\varepsilon = \frac{\lambda L}{2}$, alors le terme en $\|Uy - Ux\|^2$ disparait, et on a bien $\varepsilon - 1 = -\frac{2-\lambda L}{2}$. ■

B.III.2.iii) Preuve du Théorème principal

On peut maintenant prouver notre résultat principal, le Théorème III.35 sur la convergence de Davis-Yin.

Preuve du Théorème III.35. On va commencer par prouver que la suite y_n converge vers un point fixe de T . Pour cela on va utiliser le Lemme d'Opial B.65 avec l'ensemble $S = \text{Fix } T$. il nous faut donc en vérifier les hypothèses.

- 1) Tout d'abord, le théorème fait l'hypothèse qu'il existe une solution non dégénérée x ; d'après le Lemme B.67 cela veut dire qu'il existe un point fixe y tel que $x = P_g y$. En particulier, on voit que S est non vide.
- 2) Soit $y \in \text{Fix } T$, montrons que $\|y_n - y\|$ converge. Puisque $Ty = y$, et avec les Lemmes B.67 et B.68, on peut écrire

$$\begin{aligned}
 \|y_{n+1} - y\|^2 &= \|Ty_n - Ty\|^2 \leq \|y_n - y\|^2 - \frac{2 - \lambda L}{2} \|(I - T)y_n - (I - T)y\|^2 \\
 &\leq \|y_n - y\|^2 - \frac{2 - \lambda L}{2} \|Ty_n - y_n\|^2.
 \end{aligned}$$

Dans le Théorème on fait l'hypothèse que $\lambda < 2/L$, ce qui implique que $\alpha := (2 - \lambda L)/2 > 0$ et donc que $\|y_{n+1} - y\|^2 < \|y_n - y\|^2$. Cette suite est donc décroissante, et elle est minorée, donc elle converge bien.

- 3) Supposons que l'on aie une sous-suite convergente $y_{n_k} \rightarrow y_\infty$, et montrons que $y_\infty \in \text{Fix } T$. D'une part, nous savons que T est Lipschitzienne (Lemme B.68) donc continue, donc $Ty_{n_k} \rightarrow Ty_\infty$. D'autre part, nous avons montré plus haut que

$$\alpha \|Ty_n - y_n\|^2 \leq \|y_n - y\|^2 - \|y_{n+1} - y\|^2.$$

En sommant cette inégalité sur $n \in \mathbb{N}$, et en notant que le membre de droite est une somme télescopique, on obtient

$$\alpha \sum_{n \in \mathbb{N}} \|Ty_n - y_n\|^2 \leq \|y_0 - y\|^2 < +\infty.$$

Donc la suite $\|Ty_n - y_n\|^2$ est sommable, donc elle tend vers 0. En particulier, la sous-suite $\|Ty_{n_k} - y_{n_k}\|^2$ tend vers 0. Or nous avons vu que $Ty_{n_k} - y_{n_k}$ tend vers $Ty_\infty - y_\infty$. On en déduit donc que $Ty_\infty = y_\infty$.

Avec le Lemme d'Opial, nous pouvons donc conclure que y_n converge vers un point fixe de T , que nous noterons \bar{y} . Ensuite rappelons que dans l'algorithme nous avons $x_{n+1} = \text{prox}_{\lambda g}(y_n)$, et le prox est continu car Lipschitzien (Proposition B.63). Donc x_n converge vers $\bar{x} = \text{prox}_{\lambda g}(\bar{y})$. Et d'après le Lemme B.67, cela veut dire que \bar{x} est un minimiseur (non dégénéré) de $f + g + h$. ■

B.III.3 Méthodes Lagrangiennes

Dans cette section on présente des méthodes classiques appelées méthodes Lagrangiennes, qui ont pour objectif de résoudre des problèmes d'optimisation sous contrainte linéaire. Nous allons voir que ces méthodes sont équivalentes (via dualité) aux algorithmes du gradient et proximal.

B.III.3.i Méthode du Lagrangien

Dans cette section, on s'intéresse à un problème d'optimisation sous contrainte linéaire:

$$\min_{x \in \mathbb{R}^N} f(x) \text{ tel que } Ax = b, \quad (\text{P})$$

avec $f \in \Gamma_0(\mathbb{R}^N)$, $A \in \mathcal{M}_{M,N}(\mathbb{R})$ et $b \in \mathbb{R}^M$. Il est à noter que ce problème peut se réécrire comme la minimisation de $p(x) := f(x) + g(Ax)$, avec $g = \delta_b$. On va ici discuter d'une méthode dite du *Lagrangien*, qui fait intervenir le Lagrangien que l'on a vu à la Section II.III.4 (dans notre cas précis, $g^*(u) = \langle b, u \rangle$):

$$L(x; u) = f(x) + \langle u, Ax - b \rangle.$$

En utilisant la proposition II.117, on voit que résoudre (P) est équivalent à

$$\min_{x \in \mathbb{R}^N} \max_{u \in \mathbb{R}^M} L(x; u). \quad (\text{B.3})$$

On veut donc minimiser L par rapport à x , et le maximiser par rapport à u . Une idée simple est d'alterner des étapes de minimisation par rapport à x , et de maximisation par rapport à u . Puisque L est non lisse par rapport à x , on va simplement minimiser $L(\cdot, u)$ pour un u donné. Puisque L est lisse par rapport à u , on va faire une étape de type gradient pour $-L(x, \cdot)$.

Définition B.69 (Méthode du Lagrangien). Soient $f \in \Gamma_0(\mathbb{R}^N)$, $A \in \mathcal{M}_{M,N}(\mathbb{R})$ et $b \in \mathbb{R}^M$. Soit L le Lagrangien associé défini en (B.3). Alors la méthode du Lagrangien génère une suite $(x_n, u_n)_{n \in \mathbb{N}} \subset \mathbb{R}^N \times \mathbb{R}^M$ telle que

$$\begin{cases} x_{n+1} \in \underset{x \in \mathbb{R}^N}{\operatorname{argmin}} L(x; u_n), \\ u_{n+1} = u_n + \lambda \nabla_u L(x_{n+1}; u_n), \end{cases}$$

où λ est le pas de la méthode, et $\nabla_u L(x_{n+1}; u_n)$ est une notation pour $\nabla L(x_{n+1}; \cdot)(u_n)$.

Remarque B.70 (Implémentation de la méthode du Lagrangien). La première étape de l'algorithme est équivalente à minimiser une perturbation linéaire de f :

$$x_{n+1} \in \underset{x \in \mathbb{R}^N}{\operatorname{argmin}} f(x) + \langle A^\top u_n, x \rangle.$$

Il faut noter que cette étape n'est pas toujours bien définie (prendre par exemple $f = 0$)! Donc pour que l'algorithme fonctionne il faudra faire des hypothèses, la plus simple étant que f est fortement convexe, ce qui implique (Théorème II.53) que cet argmin est bien défini. La deuxième étape de l'algorithme peut se calculer explicitement, et est équivalente à

$$u_{n+1} = u_n + \lambda(Ax_{n+1} - b).$$

On va maintenant montrer qu'appliquer la méthode du Lagrangien au problème primal (P), c'est la même chose que d'appliquer l'algorithme du gradient au problème dual, qui s'écrit ici

$$\min_{u \in \mathbb{R}^M} d(u) = f^*(-A^\top u) + \langle b, u \rangle. \quad (\text{D})$$

Théorème B.71 (Convergence du Lagrangien). Soient $f \in \Gamma_\mu(\mathbb{R}^N)$, $A \in \mathcal{M}_{M,N}(\mathbb{R})$ et $b \in \mathbb{R}^M$. On suppose que le problème associé (P) admette une solution, où f est continue. Soit $(x_n, u_n)_{n \in \mathbb{N}}$ générée par la méthode du Lagrangien. Alors

1) La suite u_n réalise un algorithme du gradient appliqué au problème dual (D). C'est-à-dire :

$$\begin{cases} x_{n+1} = \nabla f^*(-A^\top u_n) \\ u_{n+1} = u_n - \lambda \nabla d(u_n). \end{cases}$$

2) Si $0 < \lambda < 2\mu/\|A\|^2$, alors u_n converge vers un minimiseur de d , et x_n converge vers un minimiseur de p .

Afin de prouver ce résultat, on aura besoin d'un lemme:

Lemme B.72 (Lemme du Lagrangien). Soient $f \in \Gamma_\mu(\mathbb{R}^N)$, $x \in \mathbb{R}^N$ et $u \in \mathbb{R}^M$. Si $x_+ = \underset{x \in \mathbb{R}^N}{\operatorname{argmin}} f(x) + \langle u, Ax - b \rangle$, alors $x_+ = \nabla f^*(-A^\top u)$.

Démonstration. Si on applique la condition d'optimalité à la définition de x_+ , on obtient que $0 \in \partial f(x_+) + A^\top u$, ce qui d'après la formule de Legendre nous donne $x_+ \in \partial f^*(-A^\top u)$. On conclut en utilisant la dualité entre fonctions fortement convexes et lisses. ■

Preuve du théorème B.71. D'après le Lemme B.72, nous avons effectivement $x_{n+1} = \nabla f^*(-A^\top u_n)$. Ensuite on peut calculer $\nabla d(u) = -A\nabla f^*(-A^\top u) + b$, et voir que

$$u_n - \lambda \nabla d(u_n) = u_n + \lambda(A\nabla f^*(-A^\top u) - b) = u_n + \lambda(Ax_{n+1} - b).$$

On s'intéresse maintenant à la convergence de u_n , et on va faire appel au Théorème III.5. Nous avons que d est lisse, avec

$$\text{Lip}(\nabla d) = \text{Lip}(A \circ \nabla f^* \circ A^\top) \leq \|A\|^2 \text{Lip}(\nabla f^*) = \|A\|^2 \frac{1}{\mu}.$$

Donc notre choix de pas est valide. Ensuite il faut prouver que $\operatorname{argmin} d \neq \emptyset$. Nous avons fait l'hypothèse qu'il existe x tel que $0 \in \partial(f + \delta_b \circ A)(x)$ et tel que f y soit continue. Donc par Moreau-Rockafellar, $0 \in \partial f(x) + \partial(\delta_b \circ A)(x)$. De plus on a vu (cf. exemple II.67) que $\partial(\delta_b \circ A)(x) = A^\top \partial_b(Ax)$. On en déduit que le problème primal admet une solution non dégénérée, ce qui nous permet d'utiliser le Théorème de représentation primale-duale II.118. Ceci implique premièrement que le problème dual admet une solution ; donc l'algorithme du gradient sur le dual va converger vers une solution du dual, d'où la convergence de u_n vers $\bar{u} \in \operatorname{argmin} d$. Puisque ∇f^* est continue et $x_{n+1} = \nabla f^*(-A^\top u_n)$, on obtient en passant à la limite que x_n converge vers $\bar{x} := \nabla f^*(-A^\top \bar{u})$. Deuxièmement, on a la représentation primale-duale $\operatorname{argmin} p = \partial f^*(-A^\top \bar{u}) \cap A^{-1} \partial g^*(\bar{u})$ qui dans notre cas particulier devient $\operatorname{argmin} p = \bar{x} \cap A^{-1}b$. Ceci nous permet de conclure que $\bar{x} \in \operatorname{argmin} p$. ■

B.III.3.ii) Méthode du Lagrangien Augmenté

Pour que la méthode du Lagrangien fonctionne, on a eu besoin de faire l'hypothèse que f est fortement convexe, afin que le pas de minimisation soit bien défini. Lorsque f n'est pas fortement convexe mais que A est injective, on peut « forteconvexifier » le Lagrangien en introduisant le *Lagrangien augmenté* :

$$L_\lambda(x; u) = f(x) + \langle u, Ax - b \rangle + \frac{\lambda}{2} \|Ax - b\|^2.$$

C'est un exercice simple que de voir, en suivant les arguments de la proposition II.117, que résoudre (P) est équivalent à

$$\min_{x \in \mathbb{R}^N} \max_{u \in \mathbb{R}^M} L_\lambda(x; u). \quad (\text{B.4})$$

On voit d'ailleurs bien que si A est injective alors $L_\lambda(\cdot; u)$ est fortement convexe, et donc admet un minimiseur. On peut alors définir :

Définition B.73 (Méthode du Lagrangien augmenté). Soient $f \in \Gamma_0(\mathbb{R}^N)$, $A \in \mathcal{M}_{M,N}(\mathbb{R})$ et $b \in \mathbb{R}^M$. Soit L_λ le Lagrangien associé défini en (B.4). Alors la méthode du Lagrangien augmenté génère une suite $(x_n, u_n)_{n \in \mathbb{N}} \subset \mathbb{R}^N \times \mathbb{R}^M$ telle que

$$\begin{cases} x_{n+1} \in \operatorname{argmin}_{x \in \mathbb{R}^N} L_\lambda(x; u_n), \\ u_{n+1} = u_n + \lambda(Ax_{n+1} - b), \end{cases}$$

où λ est le pas de la méthode.

De manière analogue, nous allons montrer que qu'appliquer la méthode du Lagrangien augmenté au problème primal (P), c'est la même chose que d'appliquer l'algorithme proximal au problème dual (D).

Lemme B.74 (du Lagrangien Augmenté). Soient $f \in \Gamma_0(\mathbb{R}^N)$, $A \in \mathcal{M}_{M,N}(\mathbb{R})$ et $b \in \mathbb{R}^M$. Soient $x \in \mathbb{R}^N$ et $u \in \mathbb{R}^M$. Si

$$x_+ = \operatorname{argmin}_{x \in \mathbb{R}^N} f(x) + \langle u, Ax - b \rangle + \frac{\lambda}{2} \|Ax - b\|^2 \quad \text{et} \quad u_+ = u + \lambda(Ax_+ - b),$$

alors $x_+ \in \partial f^*(-A^\top u_+)$ et $u_+ = \operatorname{prox}_{\lambda d}(u)$, où $d(u) = f^*(-A^\top u) + \langle b, u \rangle$.

Démonstration. Si on applique le Théorème de Fermat à la définition de x_+ , on obtient

$$0 \in \partial f(x_+) + A^\top u + \lambda A^\top (Ax_+ - b) = \partial f(x_+) + A^\top [u + \lambda(Ax_+ - b)] = \partial f(x_+) + A^\top u_+,$$

donc on en déduit que $x_+ \in \partial f^*(-A^\top u_+)$. Ensuite si on prend la définition de u_+ , on voit que

$$u - u_+ = \lambda(b - Ax_+) \in \lambda(b - A\partial f^*(-A^\top u_+)) = \lambda\partial(\langle b, \cdot \rangle + f^* \circ (-A^\top))(u_+) = \lambda\partial d(u_+),$$

ce qui nous permet de conclure que $u_+ = \operatorname{prox}_{\lambda d}(u)$, via la caractérisation du prox par le sous-différentiel. ■

Théorème B.75 (Convergence du Lagrangien Augmenté). Soient $f \in \Gamma_0(\mathbb{R}^N)$, $A \in \mathcal{M}_{M,N}(\mathbb{R})$ injective et $b \in \mathbb{R}^M$. On suppose que le problème associé (P) admette une solution, où f est continue. Soit $(x_n, u_n)_{n \in \mathbb{N}}$ générée par la méthode du Lagrangien Augmenté. Alors

1) La suite u_n réalise un algorithme proximal appliqué au problème dual (D). C'est-à-dire :

$$\begin{cases} x_{n+1} \in \partial f^*(-A^\top u_n) \\ u_{n+1} = \operatorname{prox}_{\lambda d}(u_n). \end{cases}$$

2) Si $\lambda > 0$, alors u_n converge vers un minimiseur de d , et toute valeur d'adhérence de x_n est un minimiseur de p .

Démonstration. Le premier point est une conséquence directe du Lemme B.74. On s'intéresse maintenant à la suite duale u_n , et on va appliquer le Théorème III.11 sur la convergence de l'algorithme proximal. Tout d'abord on voit qu'on n'a pas besoin d'hypothèses sur le pas λ . Ensuite on suppose qu'il existe une solution de (P) où f est continue. En utilisant les mêmes arguments que dans la preuve du Théorème B.71, on en déduit que le problème primal admet une solution non dégénérée, et donc que l'on peut utiliser le Théorème de représentation primale-duale. D'une part on obtient que $\operatorname{argmin} d \neq \emptyset$ et donc que u_n converge vers $\bar{u} \in \operatorname{argmin} d$. Ensuite en passant à la limite dans $u_{n+1} = u_n + \lambda(Ax_{n+1} - b)$, on obtient que Ax_n tend vers b . On en déduit que toute valeur d'adhérence de x_n vérifie la contrainte $[Ax = b]$. Soit \bar{x} une telle valeur d'adhérence. En passant à la limite dans $x_{n+1} \in \partial f^*(-A^\top u_n)$ (voir TD) on obtient que $\bar{x} \in \partial f^*(-A^\top \bar{u})$. On en déduit via le Théorème de représentation primale-duale que $\bar{x} \in \operatorname{argmin} p$. ■

B.III.4 Méthodes Lagrangiennes alternées

Dans cette section nous considérons des problèmes un peu plus composites, et nous allons appliquer en série des étapes de Lagrangien ou de Lagrangien augmenté. Cela nous permettra de donner une éclairage nouveau à l'algorithme de Davis-Yin, qui on le verra se définit très naturellement dans sa forme Lagrangienne.

On considère donc un problème un peu plus général, où l'on minimise une somme directe sous une contrainte linéaire :

$$\min_{x \in \mathbb{R}^{N_1}, y \in \mathbb{R}^{N_2}, z \in \mathbb{R}^{N_3}} f(x) + g(y) + h(z) \quad \text{tel que} \quad Ax + By + Cz = d, \quad (\text{P})$$

où $f \in \Gamma_0(\mathbb{R}^{N_1})$, $g \in \Gamma_0(\mathbb{R}^{N_2})$, $h \in \Gamma_0(\mathbb{R}^{N_3})$, $A \in \mathcal{M}_{M,N_1}(\mathbb{R})$, $B \in \mathcal{M}_{M,N_2}(\mathbb{R})$, $C \in \mathcal{M}_{M,N_3}(\mathbb{R})$, $d \in \mathbb{R}^M$. On va associer à ce problème un Lagrangien où l'on va faire explicitement mention des variables primales :

$$L(x, y, z; u) = f(x) + g(y) + h(z) + \langle u, Ax + By + Cz \rangle, \quad (\text{B.5})$$

ainsi qu'un Lagrangien augmenté

$$L_\lambda(x, y, z; u) = f(x) + g(y) + h(z) + \langle u, Ax + By + Cz \rangle + \frac{\lambda}{2} \|Ax + By + Cz - d\|^2. \quad (\text{B.6})$$

Remarque B.76 (Un problème pas si général que ça). Il faut noter que ce problème est exactement équivalent à celui considéré dans les sections précédentes :

$$\min_{X \in \mathbb{R}^N} \psi(X) \quad \text{tel que} \quad \Phi X = d,$$

où $X = (x, y, z)$, $\psi(X) = f(x) + g(y) + h(z)$, $\Phi = [A \ B \ C]$ et $N = N_1 + N_2 + N_3$. On voit que les Lagrangiens qu'on vient de définir ne sont rien d'autre que

$$L(X; u) = \psi(X) + \langle u, \Phi X \rangle \quad \text{et} \quad L_\lambda(X; u) = \psi(X) + \langle u, \Phi X \rangle + \frac{\lambda}{2} \|\Phi X - d\|^2.$$

On peut d'ailleurs calculer son problème dual $\psi(-\Phi^\top u) + \langle d, u \rangle$, et obtenir ainsi

$$\min_{u \in \mathbb{R}^M} f^*(-A^\top u) + g^*(-B^\top u) + h^*(-C^\top u) + \langle d, u \rangle. \quad (\text{D})$$

On va résoudre le problème (P) en appliquant les mêmes idées que précédemment : on va minimiser le Lagrangien par rapport à (x, y, z) , puis faire un pas de gradient par rapport à u . La clé ici est qu'on va exploiter le fait que les variables sont séparées : on ne va pas minimiser le Lagrangien par rapport aux trois variables en même temps mais *l'une après l'autre*, afin d'éclater le problème. C'est pour cela que l'on parlera de méthodes Lagragiennes *alternées*. En pratique lorsque une fonction est fortement convexe on fera un pas de Lagrangien, et si la fonction n'est que convexe on fera un pas de Lagrangien Augmenté.

B.III.4.i) Méthode du Lagragien semi-augmenté

Ici on considère le cas particulier de (P) où $h = \delta_0$ et g est fortement convexe. Autrement dit, on dispose de f qui est convexe, de g qui est fortement convexe, et la variable z disparaît. Au vu de la section précédente nous allons effectuer un pas de Lagrangien augmenté par rapport à f , puis un pas de Lagrangien par rapport à g .

Définition B.77 (Méthode du Lagrangien semi-augmenté). Soient $f \in \Gamma_\mu(\mathbb{R}^{N_1})$, $g \in \Gamma_0(\mathbb{R}^{N_2})$, $A \in \mathcal{M}_{M,N_1}(\mathbb{R})$, $B \in \mathcal{M}_{M,N_2}(\mathbb{R})$, et $d \in \mathbb{R}^M$. Soient L et L_λ les Lagrangiens (augmenté) associés, définis en (B.5) et (B.6). Alors la méthode du Lagrangien semi-augmenté génère une suite $(x_n, y_n, u_n)_{n \in \mathbb{N}} \subset \mathbb{R}^{N_1+N_2+M}$ telle que

$$\begin{cases} x_{n+1} \in \underset{x \in \mathbb{R}^{N_1}}{\operatorname{argmin}} L(x, y_n; u_n), \\ y_{n+1} \in \underset{y \in \mathbb{R}^{N_2}}{\operatorname{argmin}} L_\lambda(x_{n+1}, y; u_n), \\ u_{n+1} = u_n + \lambda(Ax_{n+1} + By_{n+1} - d), \end{cases}$$

où λ est le pas de la méthode.

Ici notre algorithme traite une somme convexe-fortement convexe. Si on passe au dual, on aura une somme convexe-lisse ; et on peut montrer que sur le dual notre algo n'est rien d'autre qu'un algorithme du Gradient-Proximal.

Théorème B.78 (Convergence du Lagrangien Semi-Augmenté). Soient $f \in \Gamma_\mu(\mathbb{R}^{N_1})$, $g \in \Gamma_0(\mathbb{R}^{N_2})$, $A \in \mathcal{M}_{M,N_1}(\mathbb{R})$, $B \in \mathcal{M}_{M,N_2}(\mathbb{R})$ injective, et $d \in \mathbb{R}^M$. On suppose que le problème associé (P) admette une solution, où f et g sont continues. Soit $(x_n, y_n, u_n)_{n \in \mathbb{N}}$ générée par la méthode du Lagrangien Augmenté. Alors

- 1) La suite u_n réalise un algorithme du gradient proximal appliqué au problème dual $F + G$, où $F(u) = f^*(-A^\top u)$ et $G(u) = g^*(-B^\top u) + \langle d, u \rangle$. C'est-à-dire :

$$\begin{cases} x_{n+1} = \nabla f^*(-A^\top u_n) \\ u_{n+1} = \text{prox}_{\lambda G}(u_n - \lambda \nabla F(u_n)). \end{cases}$$

- 2) Si $0 < \lambda < 2\mu/\|A\|^2$, alors u_n converge vers un minimiseur de $F + G$, et toute valeur d'adhérence de x_n est une solution de (P).

Démonstration. Voir le Théorème plus général B.83. ■

Exemple B.79 (Algorithme de Tseng). Considérons le problème de minimiser $f(x) + g(x)$, où $f, g \in \Gamma_0(\mathbb{R}^N)$ et f est μ -fortement convexe. On peut réécrire ce problème comme

$$\min_{x, y \in \mathbb{R}^N} f(x) + g(y) \text{ tel que } Ax + By = 0,$$

avec $A = I$ et $B = -I$, afin de forcer $x = y$. L'algorithme du Lagrangien semi-augmenté s'écrit alors dans ce cas

$$\begin{cases} x_{n+1} = \nabla f^*(-u_n) \\ u_{n+1} = \text{prox}_{\lambda g^*}(u_n + \lambda x_{n+1}). \end{cases}$$

Cet algorithme est connu sous le nom d'*algorithme de Tseng*. Ce n'est rien d'autre que l'algorithme du Gradient-Proximal appliqué au problème dual $f^*(-u) + g^*(u)$.

B.III.4.ii) Méthode du Lagragien augmenté alterné (ADMM)

Ici on considère le cas particulier de (P) où $h = \delta_0$. Autrement dit, on dispose de f et g qui sont convexes, et la variable z disparait. Au vu de la section précédente nous allons effectuer un pas de Lagrangien augmenté par rapport à f , puis un autre pas de Lagrangien augmenté par rapport à g .

Définition B.80 (Méthode du Lagrangien augmenté alterné). Soient $f \in \Gamma_0(\mathbb{R}^{N_1})$, $g \in \Gamma_0(\mathbb{R}^{N_2})$, $A \in \mathcal{M}_{M, N_1}(\mathbb{R})$, $B \in \mathcal{M}_{M, N_2}(\mathbb{R})$, et $d \in \mathbb{R}^M$. Soit L_λ le Lagrangien augmenté associé, défini en (B.6). Alors la méthode du Lagrangien augmenté alterné génère une suite $(x_n, y_n, u_n)_{n \in \mathbb{N}} \subset \mathbb{R}^{N_1+N_2+M}$ telle que

$$\begin{cases} x_{n+1} \in \underset{x \in \mathbb{R}^{N_1}}{\text{argmin}} L_\lambda(x, y_n; u_n), \\ y_{n+1} \in \underset{y \in \mathbb{R}^{N_2}}{\text{argmin}} L_\lambda(x_{n+1}, y; u_n), \\ u_{n+1} = u_n + \lambda(Ax_{n+1} + By_{n+1} - d), \end{cases}$$

où λ est le pas de la méthode.

Cette méthode s'appelle *ADMM (Alternating Directions of Multipliers Method)* en anglais, et était assez populaire dans les années 2000-2010. Ici cette méthode traite une somme convexe-convexe. Si on passe au dual, on aura une somme convexe-convexe ; et on peut montrer que sur le dual notre méthode n'est rien d'autre qu'un algorithme de Douglas-Rachford (**DR**).

Théorème B.81 (Convergence du Lagrangien augmenté alterné). Soient $f \in \Gamma_0(\mathbb{R}^{N_1})$, $g \in \Gamma_0(\mathbb{R}^{N_2})$, $A \in \mathcal{M}_{M,N_1}(\mathbb{R})$ injective, $B \in \mathcal{M}_{M,N_2}(\mathbb{R})$ injective, et $d \in \mathbb{R}^M$. On suppose que le problème associé (**P**) admette une solution, où f et g sont continues. Soit $(x_n, y_n, u_n)_{n \in \mathbb{N}}$ générée par la méthode du Lagrangien Augmenté. Alors

- 1) La suite u_n réalise un algorithme de Douglas-Rachford (**DR**) appliqué au problème dual $F + G$, où $F(u) = f^*(-A^\top u) + \langle d, u \rangle$ et $G(u) = g^*(-B^\top u)$. C'est-à-dire :

$$\begin{cases} x_n \in \partial f^*(-A^\top \hat{u}_n), y_n \in \partial g^*(-B^\top u_n) \\ u_n = \text{prox}_{\lambda G}(v_n) \\ \hat{u}_n = \text{prox}_{\lambda F}(2u_n - v_n) \\ v_{n+1} = v_n + \hat{u}_n - u_n \end{cases}$$

- 2) Si $\lambda > 0$, alors u_n converge vers un minimiseur de $F + G$, et toute valeur d'adhérence de (x_n, y_n) est une solution de (**P**).

Démonstration. Voir le Théorème plus général [B.83](#). ■

B.III.4.iii) Méthode du Lagrangien deux-tiers augmenté alterné

Ici on considère le cas particulier de (**P**) où h est fortement convexe. Autrement dit, on dispose de f et g qui sont convexes, et de h fortement convexe. Au vu de la section précédente nous allons effectuer un pas de Lagrangien augmenté par rapport à f , puis un autre pas de Lagrangien augmenté par rapport à g , et enfin un pas de Lagrangien « classique » par rapport à h .

Définition B.82 (Méthode du Lagrangien deux-tiers augmenté alterné). Soient $f \in \Gamma_0(\mathbb{R}^{N_1})$, $g \in \Gamma_0(\mathbb{R}^{N_2})$, $h \in \Gamma_\mu(\mathbb{R}^{N_3})$, $A \in \mathcal{M}_{M,N_1}(\mathbb{R})$, $B \in \mathcal{M}_{M,N_2}(\mathbb{R})$, $C \in \mathcal{M}_{M,N_3}(\mathbb{R})$, et $d \in \mathbb{R}^M$. Soient L et L_λ les Lagrangiens (augmenté) associés, définis en [\(B.5\)](#) et [\(B.6\)](#). Alors la méthode du Lagrangien deux-tiers-augmenté alterné génère une suite $(x_n, y_n, z_n, u_n)_{n \in \mathbb{N}} \subset$

$\mathbb{R}^{N_1+N_2+N_3+M}$ telle que

$$\begin{cases} z_{n+1} = \underset{z \in \mathbb{R}^{N_3}}{\operatorname{argmin}} L(x_n, y_n, z; u_n), \\ x_{n+1} \in \underset{x \in \mathbb{R}^{N_1}}{\operatorname{argmin}} L_\lambda(x, y_n, z_{n+1}; u_n), \\ y_{n+1} \in \underset{y \in \mathbb{R}^{N_2}}{\operatorname{argmin}} L_\lambda(x_{n+1}, y, z_{n+1}; u_n), \\ u_{n+1} = u_n + \lambda(Ax_{n+1} + By_{n+1} + Cz_{n+1} - d), \end{cases}$$

où λ est le pas de la méthode.

Ici cette méthode traite une somme convexe-convexe-fortement convexe. Si on passe au dual, on aura une somme convexe-convexe-lisse ; et on peut montrer que sur le dual notre méthode n'est rien d'autre qu'un algorithme de Davis-Yin.

Théorème B.83 (Convergence du Lagrangien deux-tiers-augmenté alterné). Soient $f \in \Gamma_0(\mathbb{R}^{N_1})$, $g \in \Gamma_0(\mathbb{R}^{N_2})$, $h \in \Gamma_\mu(\mathbb{R}^{N_3})$, $A \in \mathcal{M}_{M,N_1}(\mathbb{R})$ injective, $B \in \mathcal{M}_{M,N_2}(\mathbb{R})$ injective, $C \in \mathcal{M}_{M,N_3}(\mathbb{R})$, et $d \in \mathbb{R}^M$. On suppose que le problème associé (P) admette une solution où f , g et h sont continues. Soit $(x_n, y_n, z_n, u_n)_{n \in \mathbb{N}}$ générée par la méthode du Lagrangien Augmenté. Alors

- 1) La suite u_n réalise un algorithme de Davis-Yin (DY) appliqué au problème dual $F + G + H$, où $F(u) = f^*(-A^\top u) + \langle d, u \rangle$, $G(u) = g^*(-B^\top u)$, et $H(u) = h^*(-C^\top u)$. C'est-à-dire :

$$\begin{cases} x_n \in \partial f^*(-A^\top \hat{u}_n), \quad y_n \in \partial g^*(-B^\top u_n), \quad z_{n+1} = \nabla h^*(-C^\top u_n) \\ u_n = \operatorname{prox}_{\lambda G}(v_n) \\ \hat{u}_n = \operatorname{prox}_{\lambda F}(2u_n - v_n - \lambda \nabla H(u_n)) \\ v_{n+1} = v_n + \hat{u}_n - u_n \end{cases}$$

- 2) Si $0 < \lambda < 2\mu/\|A\|^2$, alors u_n et \hat{u}_n convergent vers une solution de (D), et toute valeur d'adhérence de (x_n, y_n, z_n) est une solution de (P).

Démonstration. La définition de $z_{n+1} = \underset{z \in \mathbb{R}^{N_3}}{\operatorname{argmin}} L(x_n, y_n, z; u_n)$ peut se réécrire

$$z_{n+1} = \underset{z \in \mathbb{R}^{N_3}}{\operatorname{argmin}} h(z) + \langle u_n, Cz - (d - Ax_{n+1} - By_{n+1}) \rangle,$$

donc en appliquant le Lemme B.72, on obtient que $z_{n+1} = \nabla h^*(-C^\top u_n)$. Ensuite, la définition de $x_{n+1} = \underset{x \in \mathbb{R}^{N_1}}{\operatorname{argmin}} L_\lambda(x, y_n, z_{n+1}; u_n)$ peut se réécrire

$$x_{n+1} = \underset{x \in \mathbb{R}^{N_1}}{\operatorname{argmin}} f(x) + \langle u_n, Ax - (d - By_n - Cz_{n+1}) \rangle + \frac{\lambda}{2} \|Ax - (d - By_n - Cz_{n+1})\|^2.$$

Donc en appliquant le Lemme B.74, on obtient que $x_{n+1} \in \partial f^*(-A^\top \hat{u}_n)$, où $\hat{u}_n := u_n + \lambda(Ax_{n+1} + By_n + Cz_{n+1} - d)$, ainsi que la relation

$$\hat{u}_n = \text{prox}_{\lambda(f^* \circ -A^\top + \langle d - By_n - Cz_{n+1}, \cdot \rangle)}(u_n) = \text{prox}_{\lambda F}(u_n + \lambda By_n + \lambda Cz_{n+1}),$$

où dans la dernière égalité nous avons posé $F = f^* \circ -A^\top + \langle d, \cdot \rangle$, et utilisé la règle de calcul Proposition III.14.iii). Enfin, la relation $y_{n+1} = \underset{y \in \mathbb{R}^{N_2}}{\operatorname{argmin}} L_\lambda(x_{n+1}, y, z_{n+1}; u_n)$ peut se réécrire

$$y_{n+1} = \underset{y \in \mathbb{R}^{N_2}}{\operatorname{argmin}} g(y) + \langle u_n, By - (d - Ax_{n+1} - Cz_{n+1}) \rangle + \frac{\lambda}{2} \|By - (d - Ax_{n+1} - Cz_{n+1})\|^2.$$

Donc en appliquant le Lemme B.74, on obtient que $y_{n+1} \in \partial g^*(-B^\top u_{n+1})$, où $u_{n+1} = u_n + \lambda(Ax_{n+1} + By_{n+1} + Cz_{n+1} - d)$, ainsi que la relation

$$\begin{aligned} u_{n+1} &= \text{prox}_{\lambda(g^* \circ -B^\top + \langle d - Ax_{n+1} - Cz_{n+1}, \cdot \rangle)}(u_n) = \text{prox}_{\lambda G}(u_n + \lambda Ax_{n+1} + \lambda Cz_{n+1} - \lambda d) \\ &= \text{prox}_{\lambda G}(\hat{u}_n - \lambda By_n), \end{aligned}$$

où nous avons posé $G = g^* \circ -B^\top$, et utilisé la règle de calcul Proposition III.14.iii). Nous avons donc bien prouvé les premières relations

$$x_{n+1} \in \partial f^*(-A^\top \hat{u}_n), \quad y_n \in \partial g^*(-B^\top u_n), \quad z_{n+1} = \nabla h^*(-C^\top u_n), \quad (\text{B.7})$$

et nous avons également les relations

$$\begin{cases} \hat{u}_n = \text{prox}_{\lambda F}(u_n + \lambda By_n + \lambda Cz_{n+1}) \\ u_{n+1} = \text{prox}_{\lambda G}(\hat{u}_n - \lambda By_n), \end{cases}$$

qu'il nous reste à travailler. Pour cela, on définit une nouvelle variable $v_n = u_n - \lambda By_n$.

- On sait que $y_n \in \partial g^*(-B^\top u_n)$, donc $-By_n \in -B\partial g^*(-B^\top u_n)$, ce qui implique que $-By_n \in \partial G(u_n)$. Donc $v_n - u_n = -\lambda By_n \in \lambda \partial G(u_n)$, ce qui est équivalent à dire que $u_n = \text{prox}_{\lambda G}(v_n)$.
- On a $u_n + \lambda By_n + \lambda Cz_{n+1} = u_n + (u_n - v_n) + \lambda Cz_{n+1} = 2u_n - v_n + \lambda Cz_{n+1}$. Or $Cz_{n+1} = C\nabla h^*(-C^\top u_n) = -\nabla H(u_n)$. Donc $u_n + \lambda By_n + \lambda Cz_{n+1} = 2u_n - v_n - \lambda \nabla H(u_n)$. Ce qui veut dire que $\hat{u}_n = \text{prox}_{\lambda F}(2u_n - v_n - \lambda \nabla H(u_n))$.
- On a $v_{n+1} = u_{n+1} - \lambda By_{n+1} = \hat{u}_n - \lambda By_n = \hat{u}_n + v_n - u_n$.

Nous avons donc prouvé que u_n réalise l'algorithme de Davis-Yin sur le dual. Nous allons maintenant prouver la convergence des itérés. Tout d'abord le pas est bien choisi puisque $\text{Lip}(\nabla H) \leq \|A\|^2/\mu$. Ensuite on a fait l'hypothèse de l'existence d'une solution x où les fonctions f, g, h sont continues. En appliquant le Théorème de Fermat puis le Théorème de Moreau-Rockafellar, on en déduit que x est une solution non dégénérée de (P). On

peut donc appliquer le Théorème de représentation primale-duale, qui nous donne l'existence d'une solution non dégénérée pour le problème dual (D). Donc on peut appliquer le Théorème III.35 pour obtenir la convergence de u_n et \hat{u}_n vers \bar{u} solution de (D). Passons maintenant aux variables primales : soit $\bar{X} := (\bar{x}, \bar{y}, \bar{z})$ une valeur d'adhérence de (x_n, y_n, z_n) . En passant à la limite dans $u_{n+1} = u_n + \lambda(Ax_{n+1} + By_{n+1} + Cz_{n+1} - d)$, on voit que $A\bar{x} + B\bar{y} + C\bar{z} = d$, donc \bar{X} appartient à la contrainte $\mathcal{C} := [\Phi x = d]$, où l'on reprend ici les notations de la Remarque B.76. De plus, en passant à la limite dans (B.7), on obtient que

$$-\Phi^\top \bar{u} = (-A^\top \bar{u}, -B^\top \bar{u}, -C^\top \bar{u}) \in \partial f(\bar{x}) \times \partial g(\bar{y}) \times \partial h(\bar{z}) = \partial \psi(\bar{X}).$$

Autrement dit, $0 \in \partial \psi(\bar{X}) + \text{Im } \Phi^\top = \partial \psi(\bar{X}) + N_{\mathcal{C}}(\bar{X})$. Donc \bar{X} est un minimiseur de ψ sur \mathcal{C} , c'est-à-dire une solution de (P). ■

Annexe C

Annexe: Encore quelques preuves

C.I Preuves alternatives et directes de certains résultats principaux

Dans le cours nous avons énoncé certains résultats en laissant leur preuve à plus tard, notamment en utilisant la puissance des résultats sur la dualité. Ceci nous permet de gagner du temps car les preuves en question sont souvent une conséquence rapide du théorème II.99 sur la biconjuguée, lui même basé sur le théorème II.46 des minorantes affines. Néanmoins il est souvent possible de prouver ces résultats directement, avec des preuves qui ne requièrent rien d'autre que le théorème de séparation de Hahn-Banach. Dans cette annexe nous proposons une collection de telles preuves alternatives:

- Preuve de la proposition I.25 sur la *continuité de la projection* en annexe C.I.1.
- Preuve de la *séparation faible de Hahn-Banach* en annexe C.I.2.
- Preuve du théorème I.19 sur la *caractérisation des convexes comme intersection de demi-espaces* en annexe C.I.3.
- Preuve du théorème I.39 sur la *caractérisation des cônes comme intersection de demi-espaces linéaires* en annexe C.I.4.
- Preuve du théorème I.45 sur le *cône bipolaire* en annexe C.I.5.
- Preuve du théorème II.66 sur la *règle de la chaîne* en annexe C.I.6.
- Preuve du théorème II.72 sur la *règle de la somme* de Moreau-Rockafellar en annexe C.I.7.

C.I.1 Preuve de la proposition I.25 sur la continuité de la projection

Démonstration. (Voir [9, Proposition III.3.1.3]) Commençons par développer la norme au carré, en faisant apparaître les termes de projection:

$$\begin{aligned}
& \|y - x\|^2 \\
= & \|(y - x) - (\text{proj}_C(y) - \text{proj}_C(x)) + (\text{proj}_C(y) - \text{proj}_C(x))\|^2 \\
= & \|(y - x) - (\text{proj}_C(y) - \text{proj}_C(x))\|^2 + \|\text{proj}_C(y) - \text{proj}_C(x)\|^2 \\
& + 2\langle (y - x) - (\text{proj}_C(y) - \text{proj}_C(x)), \text{proj}_C(y) - \text{proj}_C(x) \rangle \\
\geq & \|\text{proj}_C(y) - \text{proj}_C(x)\|^2 \\
& + 2\langle (y - x) - (\text{proj}_C(y) - \text{proj}_C(x)), \text{proj}_C(y) - \text{proj}_C(x) \rangle.
\end{aligned}$$

On voit que l'inégalité sera prouvée pourvu qu'on arrive à monter que le produit scalaire est positif. Coupons ce terme en deux:

$$\begin{aligned}
& \langle (y - x) - (\text{proj}_C(y) - \text{proj}_C(x)), \text{proj}_C(y) - \text{proj}_C(x) \rangle \\
= & -\langle y - \text{proj}_C(y), \text{proj}_C(x) - \text{proj}_C(y) \rangle - \langle x - \text{proj}_C(x), \text{proj}_C(y) - \text{proj}_C(x) \rangle.
\end{aligned}$$

On voit alors que chacun de ces deux produits scalaires est négatif, grâce à la caractérisation de la projection par les angles, voir la proposition I.22. D'où le résultat. ■

C.I.2 Preuve de la séparation faible de Hahn-Banach

Théorème C.1 (de séparation faible de Hahn-Banach). Soit $C \subset \mathbb{R}^N$ convexe non vide, tel que $0 \notin C$. Alors il existe $\alpha \neq 0$ tel que pour tout $c \in C$, $\langle \alpha, c \rangle \geq 0$.

Démonstration. On va essayer d'utiliser le même type d'argument que pour la séparation forte en projetant 0 sur C (rappelez-vous la preuve du Théorème I.26). Sauf qu'ici C n'est pas fermé, donc la projection n'est même pas définie ! Pour contourner ce problème, on va simplement considérer un sous-ensemble fermé de C qu'on va faire grossir, et voir ce qui se passe lorsque on passe à la limite. On commence donc par regarder $C \cap \mathbb{Q}^N$. Puisque on est en dimension finie, \mathbb{Q}^N est dense dans \mathbb{R}^N , donc $C \cap \mathbb{Q}^N$ est dense dans C . De plus \mathbb{Q}^N est dénombrable, donc $C \cap \mathbb{Q}^N$ l'est aussi, donc il existe une suite $\{c_n\}_{n \in \mathbb{N}}$ telle que $C \cap \mathbb{Q}^N = \{c_n\}_{n \in \mathbb{N}}$. Maintenant nous pouvons définir les polytopes $C_n = \text{co}(c_0, \dots, c_n)$ qui vont moralement converger vers C lorsque $n \rightarrow +\infty$. Nous savons que les polytopes sont convexes par définition, ceux-ci sont non vides par construction et fermés d'après la proposition I.18, donc nous pouvons définir $p_n := \text{proj}_{C_n}(0)$. On peut raisonner comme pour la séparation forte et utiliser la caractérisation de la projection par les angles pour écrire que

$$(\forall c' \in C_n) \quad \langle p_n, c' \rangle \geq \|p_n\|^2.$$

Puisque $0 \notin C$ et que $C_n \subset C$, nous savons que $p_n \neq 0$, donc on peut définir $\alpha_n := \frac{p_n}{\|p_n\|}$ qui appartient à la sphère unité \mathbb{S} et tel que

$$(\forall c' \in C_n) \quad \langle \alpha_n, c' \rangle \geq 0.$$

Considérons maintenant un $c \in C$ quelconque. Par densité de $\{c_n\}_{n \in \mathbb{N}}$ dans C , il doit exister une sous-suite c_{n_k} qui converge vers c . De plus $\alpha_{n_k} \in \mathbb{S}$ qui est compacte, donc quitte à considérer une sous-suite, nous pouvons affirmer que α_{n_k} converge vers un certain $\alpha \in \mathbb{S}$ lorsque $k \rightarrow +\infty$. Il reste donc à passer à la limite pour écrire

$$\langle \alpha, c \rangle = \lim_{k \rightarrow +\infty} \langle \alpha_{n_k}, c_{n_k} \rangle \geq 0.$$

On pense à vérifier que $\alpha \neq 0$ qui vient du fait que $\alpha \in \mathbb{S}$. ■

Corollaire C.2 (de séparation faible de Hanh-Banach II). Soient $C, D \subset \mathbb{R}^N$ convexes non vides tels que $C \cap D = \emptyset$. Alors il existe $\alpha \neq 0$ tel que $\langle \alpha, c \rangle \geq \langle \alpha, d \rangle$ pour tout $c \in \text{cl } C, d \in \text{cl } D$.

Démonstration. C'est un corollaire direct du théorème C.1. Pour le voir, il suffit de voir que $C \cap D = \emptyset$ est équivalent à $0 \notin C - D$. Il faut donc appliquer théorème C.1 à l'ensemble $C - D$: il est non vide par construction, il est convexe car la somme de convexes est convexe (proposition I.9). Théorème C.1 nous donne alors une inégalité $\langle \alpha, c - d \rangle \geq 0$, que l'on transforme en $\langle \alpha, c \rangle \geq \langle \alpha, d \rangle$. Cette inégalité porte sur C et D , mais on peut passer à la limite pour en déduire une inégalité pour $\text{cl } C$ et $\text{cl } D$. ■

C.I.3 Preuve du théorème I.19 sur la caractérisation des convexes comme intersection de demi-espaces

Lemme C.3 (Convexe comme intersection explicite de demi-espaces). Soit $C \subset \mathbb{R}^N$ convexe fermé non vide. Alors $C = \bigcap_{x \in \mathbb{R}^N} H_x^+$ où H_x^+ est le demi-espace $H_x^+ = [\langle a_x, \cdot \rangle \leq b_x]$ défini par $a_x = x - \text{proj}_C(x)$ et $b_x = \langle a_x, \text{proj}_C(x) \rangle$.

Démonstration. Puisque C est convexe fermé non vide, on peut pour chaque $x \in \mathbb{R}^N$ définir $p_x = \text{proj}_C(x)$. D'après la caractérisation de la projection par les angles (voir Proposition I.22) nous avons

$$(\forall c \in C) \quad \langle x - p_x, c \rangle \leq \langle x - p_x, p_x \rangle.$$

Si on pose $a_x := x - p_x \in \mathbb{R}^N$, $b_x := \langle x - p_x, p_x \rangle \in \mathbb{R}$, et que l'on note le demi-espace $H_x^+ := [\langle a_x, \cdot \rangle \leq b_x]$, alors l'inégalité ci-dessus veut simplement dire que $C \subset H_x^+$. Par construction, nous avons donc que $C \subset \bigcap_{x \in \mathbb{R}^N} H_x^+$. Afin de conclure, nous allons montrer

que cette inclusion est en fait une égalité. Considérons donc $a \in \bigcap_{x \in \mathbb{R}^N} H_x^+$, et montrons

que $a \in C$. Par définition nous savons que $a \in H_a^+$, ce qui veut dire que

$$\langle a - p_a, a \rangle \leq \langle a - p_a, p_a \rangle$$

qui équivaut à

$$\langle a - p_a, a - p_a \rangle \leq 0,$$

c'est-à-dire que $\|a - p_a\|^2 \leq 0$, et donc que $p_a = a$. En conséquence $a = \text{proj}_C(a) \in C$. ■

Remarque C.4 (Face exposée). Étant donné un $x \notin C$, on a défini un demi-espace H_x^+ dans la preuve ci-dessus. Considérons H_x l'hyperplan associé, et regardons l'ensemble $F_C(x) := H_x \cap C$. On peut voir sur des exemples simples (faire un dessin !) que $F_C(x)$ fait partie du bord de C , et si C est simplement un polygone ou un polyèdre, on se convainc que $F_C(x)$ est toujours une *face* de C . En général l'ensemble $F_C(x)$ est appelé une **face exposée** de C , on en parle également dans la remarque II.107.

C.I.4 Preuve du théorème I.39 sur la caractérisation des cônes comme intersection de demi-espaces linéaires

Démonstration. Si K est une intersection de demi-espaces linéaires (qui sont des cônes) alors K est un cône d'après la proposition I.31. Supposons maintenant que K est un cône fermé. Puisque K est convexe fermé, on peut reprendre les résultats du lemme C.3, qui nous donne que $K = \cap_x H_x^+$, où $H_x^+ = [\langle a_x, \cdot \rangle \leq b_x]$. Pour conclure, nous allons tout simplement montrer que $b_x = 0$ pour tout x . Fixons donc $x \in \mathbb{R}^N$, et rappelons que dans le lemme C.3 nous avons $b_x = \langle x - p_x, p_x \rangle$ où $p_x = \text{proj}_K(x)$.

La caractérisation de la projection par les angles (proposition I.22) nous indique que pour tout $c \in K$, $\langle x - p_x, c - p_x \rangle \leq 0$. Nous allons utiliser cette inégalité deux fois. Premièrement, puisque $p_x \in K$ et que K est positivement homogène, nous pouvons prendre $c = 2p_x \in K$ et voir que $\langle x - p_x, p_x \rangle \leq 0$. Deuxièmement, nous pouvons aussi prendre $c = \frac{1}{2}p_x \in K$ et voir que $-\frac{1}{2}\langle x - p_x, p_x \rangle \leq 0$. En combinant ces deux résultats nous voyons bien que $\langle x - p_x, p_x \rangle = 0$. ■

C.I.5 Preuve du théorème I.45 sur le cône bipolaire

Démonstration. On procède par double inclusion.

\subset : D'après le théorème I.39, on sait qu'on peut écrire K comme une intersection de demi-espaces linéaires, autrement dit $K = \cap_{i \in I} [\langle a_i, x \rangle \leq 0]$. Pour tout i nous avons que $K \subset [\langle a_i, x \rangle \leq 0]$, donc en appliquant deux fois la décroissance du passage au polaire (proposition I.44) on obtient $K^* \supset [\langle a_i, x \rangle \leq 0]^*$ puis $K^{**} \subset [\langle a_i, x \rangle \leq 0]^{**}$. Or on sait calculer que $[\langle a_i, x \rangle \leq 0]^{**} = (\mathbb{R}_+ a_i)^* = [\langle a_i, x \rangle \leq 0]$. Ceci étant vrai pour tout $i \in I$, nous en déduisons donc que $K^{**} \subset \cap_{i \in I} [\langle a_i, x \rangle \leq 0]^{**} = K$.

\supset : Soit $x \in K$, montrons que $x \in (K^*)^*$. Pour cela, prenons $x^* \in K^*$ quelconque et montrons que $\langle x^*, x \rangle \leq 0$. Or ceci est évidemment vrai puisque $x \in K$ et $x^* \in K^*$. ■

C.I.6 Preuve du théorème II.66 sur la règle de la chaîne

Démonstration. On procède par double inclusion.

\supset : Soit $x^* \in A^\top \partial g(Ax)$, c'est-à-dire que $x^* = A^\top y^*$ où $y^* \in \partial g(Ax)$. Alors

$$(\forall y \in \mathbb{R}^M) \quad g(y) - g(Ax) - \langle y^*, y - Ax \rangle \geq 0$$

Si on prend $y = Ax'$ pour $x' \in \mathbb{R}^N$ quelconque on a

$$(\forall x' \in \mathbb{R}^N) \quad g(Ax') - g(Ax) - \langle y^*, Ax' - Ax \rangle \geq 0,$$

et on conclut en voyant que $\langle y^*, Ax' - Ax \rangle = \langle x^*, x' - x \rangle$.

\subset : Soit $x^* \in \partial f(x)$ et montrons que $x^* = A^\top \alpha$ où $\alpha \in \partial g(Ax)$. On va utiliser un argument de séparation, donc on va introduire quelques ensembles. Tout d'abord

$$\mathcal{F} := \{(y, r) \in \mathbb{R}^M \times \mathbb{R} \mid y = Ax', r = f(x) + \langle x^*, x' - x \rangle, x' \in \mathbb{R}^N\}.$$

On voit que \mathcal{F} est un espace affine, donc il est convexe et fermé. De plus $\mathcal{F} \neq \emptyset$ car il contient $(Ax, f(x))$, et on a bien $f(x) \in \mathbb{R}$ puisque $x \in \text{dom } \partial f \subset \text{dom } f$ (voir proposition II.63). On introduit ensuite $C := \text{int epi } g$. Par hypothèse on sait que $\text{epi } g$ est convexe fermé non vide (voir les propositions II.13, II.21 et II.38). Puisque l'intérieur d'un convexe est convexe, on en déduit que C est convexe (proposition I.12). De plus $\text{cont } g \neq \emptyset$ donc d'après la proposition II.15 on sait que l'intérieur de l'épigraphe est non vide, c'est-à-dire $C \neq \emptyset$. On va vouloir séparer C et \mathcal{F} , donc il nous reste à montrer que leur intersection est vide. Si il existait $(y, r) \in C \cap \mathcal{F}$, alors on aurait d'une part que $y = Ax'$ et $r = f(x) + \langle x^*, x' - x \rangle$. D'autre part puisque $(y, r) \in \text{int epi } g$ on aurait que $g(y) < r$ (voir proposition II.15). Cette dernière inégalité s'écrit

$$g(y) - r = f(x') - f(x) - \langle x^*, x' - x \rangle < 0$$

et contredit $x^* \in \partial f(x)$. Nous sommes donc prêts à utiliser le théorème C.2 de séparation faible de Hahn-Banach. Il existe $(\alpha, \beta) \in \mathbb{R}^M \times \mathbb{R}$ non nul tel que (on utilise le fait que $\text{epi } g = \text{adh } C$):

$$(\forall (y, r) \in \text{epi } g)(\forall (y', r') \in \mathcal{F}) \quad \langle \alpha, y \rangle + \beta r \geq \langle \alpha, y' \rangle + \beta r'. \quad (\text{C.1})$$

Nous allons maintenant distinguer trois cas en fonction du signe de β .

- Cas $\beta < 0$: Ceci est impossible. En effet on voit que $(Ax, f(x)) \in \mathcal{F}$ tandis que $(Ax, f(x) + 1) \in \text{epi } g$, donc si on injecte cela dans (C.1) on obtient

$$\langle \alpha, Ax \rangle + \beta f(x) + \beta \geq \langle \alpha, Ax \rangle + \beta f(x)$$

qui veut bien dire que $\beta \geq 0$.

- Cas $\beta = 0$: Ceci va également être impossible ! En effet commençons par voir que cela implique

$$(\forall y \in \text{dom } g)(\forall y' \in \text{Im } A) \quad \langle \alpha, y - y' \rangle \geq 0. \quad (\text{C.2})$$

Si on utilise notre hypothèse que $\text{cont } g \cap \text{Im } A \neq \emptyset$ on obtient un $\hat{y} = A\hat{x} \in \text{cont } g = \text{int dom } g$ d'après le théorème II.44. On dispose donc d'une boule $\mathbb{B}(\hat{y}, \delta) \subset \text{dom } g$. D'autre part on a $\hat{y} \in \text{Im } A$ donc on déduit de (C.2) que

$$(\forall y \in \mathbb{B}(\hat{y}, \delta)) \quad \langle \alpha, y - \hat{y} \rangle \geq 0.$$

Ceci implique que $\alpha = 0$, or on a déjà $\beta = 0$ et le théorème de séparation nous a dit que $(\alpha, \beta) \neq 0$: contradiction.

- Cas $\beta > 0$: quitte à diviser par β et renommer α , on peut supposer que $\beta = 1$ dans (C.1), ce qui nous donne

$$(\forall (y, r) \in \text{epi } g)(\forall (y', r') \in \mathcal{F}) \quad \langle \alpha, y - y' \rangle + r - r' \geq 0.$$

En particulier on a pour tout $y \in \text{dom } g$ que $(y, g(y)) \in \text{epi } g$; et pour tout $x' \in \mathbb{R}^N$ que $(Ax', f(x) + \langle x^*, x' - x \rangle) \in \mathcal{F}$; donc on peut écrire

$$(\forall y \in \text{dom } g)(\forall x' \in \mathbb{R}^N) \quad \langle \alpha, y - Ax' \rangle + g(y) - f(x) - \langle x^*, x' - x \rangle \geq 0. \quad (\text{C.3})$$

Si on prend (C.3) avec y quelconque et $x' = x$ alors

$$(\forall y \in \text{dom } g) \quad \langle \alpha, y - Ax \rangle + g(y) - g(Ax) \geq 0,$$

ce qui veut dire que $-\alpha \in \partial g(Ax)$. Si on prend (C.3) avec x' quelconque et $y = Ax$ alors

$$(\forall x' \in \mathbb{R}^N) \quad \langle \alpha, Ax - Ax' \rangle - \langle x^*, x' - x \rangle \geq 0,$$

ce qui équivaut à

$$(\forall x' \in \mathbb{R}^N) \quad \langle -A^\top \alpha - x^*, x' - x \rangle \geq 0.$$

Ceci implique que $-A^\top \alpha - x^* = 0$, autrement dit $x^* = -A^\top \alpha \in A^\top \partial g(Ax)$.



C.I.7 Preuve du théorème II.72 sur la règle de la somme de Moreau-Rockafellar

Démonstration. Soit $x \in \mathbb{R}^N$, soit $z^* \in \partial(f + g)(x)$, et montrons que $z^* \in \partial f(x) + \partial g(x)$. La preuve va reposer sur la séparation faible entre deux convexes définis de manière astucieuse. On introduit les ensembles suivants:

$$\begin{aligned} B &= \{(y, r) \in \mathbb{R}^N \times \mathbb{R} \mid g(y) \leq g(x) - r\}, \\ C &= \{(y, r) \in \mathbb{R}^N \times \mathbb{R} \mid f(y) \leq f(x) + r + \langle z^*, y - x \rangle\}. \end{aligned}$$

D'une part, on voit que $C = \text{epi } h$, où $h(y) := f(y) - f(x) - \langle z^*, y - x \rangle$. Clairement on a $h \in \Gamma_0(\mathbb{R}^N)$, donc C est convexe fermé non vide. D'autre part, on peut voir que $B = -\text{epi } \hat{h}$, où $\hat{h}(y) = g(-y) - g(x)$. De même on voit que $\hat{h} \in \Gamma_0(\mathbb{R}^N)$, donc B est convexe fermé non vide. On considère maintenant $\text{int } C = \text{int epi } h$. On sait que cet intérieur est non vide grâce à la proposition II.15 et le fait que $\text{cont } h = \text{cont } f$, ce dernier étant non vide par hypothèse. Enfin, on se convainc que $\text{int } C \cap B = \emptyset$. En effet, si il existait $(y, r) \in \text{int } C \cap B$, on aurait à la fois

$$f(y) < f(x) + r + \langle z^*, y - x \rangle \quad \text{et} \quad g(y) \leq g(x) - r.$$

Si on faisait la somme de ces inégalités, on obtiendrait

$$(f + g)(y) < (f + g)(x) + \langle z^*, y - x \rangle,$$

ce qui contredirait le postulat de départ que $z^* \in \partial(f + g)(x)$.

Les conditions sont donc réunies pour appliquer le théorème C.2 de séparation faible de Hahn-Banach. Il existe $(\alpha, \beta) \in \mathbb{R}^N \times \mathbb{R}$ non nul tel que (on utilise le fait que $\text{adh int } C = C$):

$$(\forall (y, r) \in C)(\forall (y', r') \in B) \quad \langle (\alpha, \beta), (y, r) \rangle \geq \langle (\alpha, \beta), (y', r') \rangle.$$

Autrement dit

$$(\forall (y, r) \in C)(\forall (y', r') \in B) \quad \langle \alpha, y \rangle + \beta r \geq \langle \alpha, y' \rangle + \beta r'. \quad (\text{C.4})$$

On fait maintenant une distinction de cas sur le signe de $\beta \in \mathbb{R}$.

- Cas $\beta < 0$: C'est impossible, car on voit que $(x, 0) \in B$ et $(x, 1) \in C$, ce qui implique via (C.4) que $\beta \geq 0$.
- Cas $\beta = 0$: Ici (C.4) devient

$$(\forall (y, r) \in C)(\forall (y', r') \in B) \quad \langle \alpha, y - y' \rangle \geq 0.$$

Soit $y' \in \text{cont } f \cap \text{dom } g$. Premièrement on peut définir $(y', g(x) - g(y')) \in B$. Deuxièmement on dispose d'un voisinage $\mathbb{B}(y', \varepsilon) \subset \text{cont } f \subset \text{dom } f$ sur lequel $(y, h(y)) \in C$. On a alors

$$(\forall y \in \mathbb{B}(y', \varepsilon)) \quad \langle \alpha, y - y' \rangle \geq 0,$$

ce qui implique que $\alpha = 0$. Or $(\alpha, \beta) = (0, 0)$ est une contradiction.

- Cas $\beta > 0$: quitte à diviser (C.4) par β et renommer α , on peut supposer que $\beta = 1$ ce qui donne

$$(\forall (y, r) \in C)(\forall (y', r') \in B) \quad \langle \alpha, y - y' \rangle + r - r' \geq 0.$$

Premièrement, on utilise le fait que $(x, 0) \in C$ et que pour tout $y' \in \text{dom } g$ on a $(y', g(x) - g(y')) \in B$, ce qui donne

$$(\forall y' \in \text{dom } g) \quad g(y') - g(x) - \langle \alpha, y' - x \rangle \geq 0,$$

ce qui veut dire que $\alpha \in \partial g(x)$. Deuxièmement, on utilise le fait que $(x, 0) \in B$ et que pour tout $y \in \text{dom } f$ on a $(y, h(y)) \in C$ ce qui donne

$$(\forall y \in \text{dom } f) \quad h(y) + \langle \alpha, y - x \rangle = f(y) - f(x) - \langle z^* - \alpha, y - x \rangle \geq 0,$$

ce qui veut dire que $z^* - \alpha \in \partial f(x)$. Ceci conclut la preuve.



C.II Preuves de petits résultats laissés en exercice

Certains des résultats du cours sont laissés à titre d'exercice: d'une part parce que nous n'avons pas le temps de tout faire en cours, d'autre part car ils sont selon moi de bon exercices pour vérifier votre maîtrise des outils. Cette section en réunit des preuves, à titre d'exhaustivité.

C.II.1 Cônes tangent et normal à un polyèdre

Lemme C.5 (Cône tangent est croissant). Soient $C, D \subset \mathbb{R}^N$ convexes fermés non vides, et $x \in C$. Si $C \subset D$ alors $T_C(x) \subset T_D(x)$.

Démonstration. Soit $d \in T_C(x)$, alors $d = \lim d_n$ où $d_n = \lambda_n(c_n - x)$ avec $c_n \in C$. Par hypothèse, $c_n \in D$ donc $d \in T_D(x)$. ■

Lemme C.6 (Cône tangent est local). Soient $C \subset \mathbb{R}^N$ convexe fermé non vide, et $x \in C$. Alors

$$(\forall \varepsilon > 0) \quad T_{C \cap \mathbb{B}(x, \varepsilon)}(x) = T_C(x).$$

Démonstration. Notons $U = \mathbb{B}(x, \varepsilon)$.

\subset : immédiat car $C \cap U \subset C$ et le cône tangent est croissant (voir Lemme C.5).

\supset : Soit $d \in T_C(x)$, alors $d = \lim d_n$ avec $d_n = \lambda_n(c_n - x)$ où $c_n \in C$. Définissons $t_n = \min\{1; \frac{\varepsilon}{\|c_n - x\|}\}$ et posons $\hat{c}_n = x + t_n(c_n - x)$. D'une part, nous avons que $t_n \in]0, 1]$ donc par convexité il est clair que $\hat{c}_n \in C$. D'autre part, nous avons également $\hat{c}_n \in U$ puisque

$$\|\hat{c}_n - x\| = t_n \|c_n - x\| \leq \frac{\varepsilon}{\|c_n - x\|} \|c_n - x\| = \varepsilon.$$

Donc $\hat{c}_n \in C \cap U$, et on peut écrire $d_n = \hat{\lambda}_n(\hat{c}_n - x)$ avec $\hat{\lambda}_n = \frac{\lambda_n}{t_n} \geq 0$. ■

Lemme C.7. Le Théorème I.57 sur les cônes tangent et normal à un polyèdre est vrai.

Démonstration. Dans la preuve on notera I l'ensemble des contraintes actives, et J son complémentaire. On note que, par définition, nous avons $A_I x = b_I$ et $A_J x < b_J$.

Commençons par vérifier que l'inégalité $A_Jx < b_J$ reste vraie au voisinage de x . Pour tout $j \in J$, on a $\langle a_j, x \rangle < b_j$. Donc par continuité de la forme linéaire, il existe un ε_j tel que sur $\mathbb{B}(x, \varepsilon)$ on ait $\langle a_j, x' \rangle < b_j$. Puisque J est fini, on donc bien peut prendre $\varepsilon = \min \varepsilon_j$. Nous avons donc montré que $\mathbb{B}(x, \varepsilon) \subset [A_Jx' \leq b_J]$. Or $x' \in C$ si et seulement si $A_Ix' \leq b_I$ et $A_Jx' \leq b_J$. Donc nous pouvons en déduire que $C \cap \mathbb{B}(x, \varepsilon) = [A_Ix \leq b_I] \cap \mathbb{B}(x, \varepsilon)$, qui est une description locale de C . Or le cône tangent est local (voir Lemme C.6) donc

$$T_C(x) = T_{C \cap \mathbb{B}(x, \varepsilon)}(x) = T_{[A_Ix \leq b_I] \cap \mathbb{B}(x, \varepsilon)}(x) = T_{[A_Ix \leq b_I]}(x).$$

Nous allons maintenant conclure en montrant que $T_{[A_Ix \leq b_I]}(x) = [A_Ix \leq 0]$.

\subset : Soit $d \in T_{[A_Ix \leq b_I]}(x)$; alors $d = \lim d_n = \lambda_n(c_n - x)$ avec $c_n \in [A_Ix \leq b_I]$ et $\lambda_n \geq 0$. On doit montrer que $d \in [A_Ix \leq 0]$, or cet ensemble est fermé donc il suffit de montrer que $d_n \in [A_Ix \leq 0]$. On calcule alors

$$A_Id_n = \lambda_n(A_Ic_n - A_Ix) = \lambda_n(A_Ic_n - b_I) \leq \lambda_n(b_I - b_I) = 0,$$

où nous avons utilisé le fait que $A_Ix = b_I$, par définition de I .

\supset : Soit $d \in [A_Ix \leq 0]$, et montrons que $d \in T_{[A_Ix \leq b_I]}(x)$. Il suffit de prendre $d_n = d$, et d'écrire $d_n = \lambda_n(c_n - x)$, avec $\lambda_n = 1$ et $c_n = x + d$. Il faut vérifier que $c_n \in [A_Ix \leq b_I]$, et effectivement

$$A_Ic_n = A_Ix + A_Id = b_I + A_Id \leq b_I.$$

On a donc bien calculé le cône tangent, et le cône normal s'obtient en prenant le polaire (cf. Proposition I.47). ■

Lemme C.8. *La proposition II.15 sur l'intérieur de l'épigraphie est vraie.*

Démonstration.

- 1) On peut faire un raisonnement par l'absurde : supposons que $f(x) \geq r$. Si $f(x) > r$ alors clairement $(x, r) \notin \text{epi } f$ ce qui contredit notre hypothèse. Donc on a $f(x) = r$. On peut alors considérer la suite $(x, r - \frac{1}{n})$ qui converge vers (x, r) mais qui n'appartient pas à $\text{epi } f$. Ceci contredit le fait que $(x, r) \in \text{int epi } f$.
- 2) Soit $\delta > 0$, on va prendre $\delta = (1/2)(r - f(x))$. f étant continue en x , on sait qu'il existe $\varepsilon > 0$ tel que $f(\mathbb{B}(x, \varepsilon)) \subset]f(x) - \delta, f(x) + \delta[$. Quitte à prendre ε plus petit, on peut supposer que $\varepsilon < \delta$. Soit $(x', r') \in \mathbb{B}(x, \varepsilon)$. Alors en particulier on a $x' \in \mathbb{B}(x, \varepsilon)$ et $r' \in \mathbb{B}(r, \varepsilon)$. On a alors que $r' \geq r - \varepsilon \geq r - \delta$

$$f(x') < f(x) + \delta = (1/2)(f(x) + r) = r - \delta \leq r'.$$

Lemme C.9. *Le Lemme II.43 sur la topologie des fonctions convexes est vrai.* ■

Démonstration. Il ya deux implications triviales ici. Si f est localement Lipschitzienne, alors elle est localement continue. Si elle est localement continue, alors on peut choisir un voisinage compact sur lequel elle est continue; elle y atteint forcément ses bornes, donc en particulier elle y est majorée. Il reste donc à montrer l'implication difficile: localement borné implique localement Lipschitz.

Supposons que f est localement bornée: il existe $\delta > 0$ et $M \in \mathbb{R}$ tels que $f(x) \leq M$ pour tout $x \in \mathbb{B}(\bar{x}, 2\delta)$. Nous voulons montrer que f est localement Lipschitz, donc donnons nous deux points $x, y \in \mathbb{B}(\bar{x}, \delta/2)$. Dans la suite on notera $N := \|y - x\|$ pour simplifier. Pour commencer, on pose $\hat{y} = y + \frac{\delta}{\|y - x\|}(y - x)$ et on observe que

$$\|\hat{y} - \bar{x}\| = \|y - x + \frac{\delta}{N}(y - x)\| \leq \|y - x\| + \delta \leq \|y - \bar{x}\| + \|\bar{x} - x\| + \delta \leq 2\delta,$$

ce qui veut dire que $\hat{y} \in \mathbb{B}(\bar{x}, 2\delta)$. Ainsi nous savons que $f(\hat{y}) \leq M$.

Maintenant on pose $\alpha = \frac{\|y - x\|}{\|y - x\| + \delta}$. Il est clair que $\alpha \in [0, 1]$ et $\alpha \leq \frac{\|y - x\|}{\delta}$ par définition. De plus,

$$(1 - \alpha)x + \alpha\hat{y} = (1 - \alpha)x + \alpha y + \alpha \frac{\delta}{N}y - \alpha \frac{\delta}{N}x = y\alpha(1 + \frac{\delta}{N}) + x - x\alpha(1 + \frac{\delta}{N}).$$

Or $\alpha(1 + \frac{\delta}{N}) = \frac{N}{\delta+N} \frac{\delta+N}{N} = 1$, d'où

$$y = (1 - \alpha)x + \alpha\hat{y}.$$

Puisque on a une combinaison convexe de x et y , on peut utiliser la convexité de f pour écrire

$$f(y) - f(x) \leq (1 - \alpha)f(x) + \alpha f(\hat{y}) - f(x) = \alpha(f(\hat{y}) - f(x)) \leq \alpha(M - f(x)).$$

De plus, en utilisant encore la convexité de f :

$$f(\bar{x}) \leq (1/2)f(x) + (1/2)f(2\bar{x} - x).$$

Or $f(2\bar{x} - x) \leq M$ puisque $2\bar{x} - x \in \mathbb{B}(\bar{x}, 2\delta)$ puisque $\|2\bar{x} - x - \bar{x}\| = \|x - \bar{x}\| \leq \delta/2$. Donc $2f(\bar{x}) \leq f(x) + M$, autrement dit $-f(x) \leq M - 2f(\bar{x})$. Si on réinjecte ceci plus haut, on obtient que

$$f(y) - f(x) \leq \alpha(M - f(x)) \leq 2\alpha(M - f(\bar{x}))$$

et on conclut en utilisant le fait que $\alpha \leq \frac{N}{\delta}$ que l'on peut prendre $L = \frac{N}{\delta}(M - f(\bar{x}))$. ■

Lemme C.10. La Proposition II.50 sur la caractérisation des fonctions fortement convexes par la norme est vraie.

Démonstration. Puisque $\frac{\mu}{2} \|\cdot\|^2$ est continue et à valeurs finies, on voit que f est s.c.i.propre ssi g l'est aussi. Donc il reste à montrer que f est μ -fortement convexe ssi g est convexe. f est μ -convexe si et seulement si

$$f(tu + (1-t)v) \leq tf(u) + (1-t)f(v) - \frac{\mu}{2}t(1-t)\|u-v\|^2, \quad \forall u, v \in \mathbb{R}^N, \forall t \in [0, 1].$$

On peut invoquer l'identité du parallélogramme:

$$(\forall x, y \in \mathbb{R}^N)(\forall t \in [0, 1]) \quad \|(1-t)x + ty\|^2 = (1-t)\|x\|^2 + t\|y\|^2 - t(1-t)\|x-y\|^2,$$

et voir que la forte convexité est équivalente à

$$f(tu + (1-t)v) \leq tf(u) + (1-t)f(v) - \frac{\mu}{2} \left(-\|tu + (1-t)v\|^2 + t\|u\|^2 + (1-t)\|v\|^2 \right)$$

c'est à dire

$$f(tu + (1-t)v) - \frac{\mu}{2}\|tu + (1-t)v\|^2 \leq t(f(u) - \frac{\mu}{2}\|u\|^2) + (1-t)(f(v) - \frac{\mu}{2}\|v\|^2)$$

ce qui est équivalent à dire que g est convexe. ■

Bibliographie

- [1] Kazunori Akiyama, Antxon Alberdi, Walter Alef, Keiichi Asada, Rebecca Azulay, Anne-Kathrin Bacsko, David Ball, Mislav Baloković, John Barrett, and Dan Bintley. First M87 event horizon telescope results. IV. Imaging the central supermassive black hole. *The Astrophysical Journal Letters*, 875(1):L4, 2019.
- [2] Aliprantis and Border. *Infinite Dimensional Analysis - A Hitchhiker's Guide*. 2006.
- [3] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2nd edition, 2017.
- [4] Jonathan M. Borwein and Adrian S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, 2006.
- [5] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge university press, 2004.
- [6] Guillaume Garrigos. Optimisation pour la licence 3. <https://guillaume-garrigos.com/L3optimisation>, 2022.
- [7] Guillaume Garrigos. Optimization for machine learning. <https://guillaume-garrigos.com/M2optimization>, 2022.
- [8] Jean-Baptiste Hiriart-Urruty. *Optimisation et analyse convexe : Exercices et problèmes corrigés, avec rappels de cours*. EDP Sciences, Ulis, France, 2009.
- [9] Jean-Baptiste Hiriart-Urruty and Claude Lemarechal. *Convex Analysis and Minimization Algorithms I: Part 1: Fundamentals*. Springer Science & Business Media, 1996.
- [10] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10051–10060, 2019.
- [11] Melvyn B. Nathanson. Polytopes, polyhedra, and the Farkas lemma, 2023.
- [12] Juan Peypouquet. *Convex Optimization in Normed Spaces*. SpringerBriefs in Optimization. Springer International Publishing, Cham, 2015.