Chapitre 2

Méthodes proximales

2.1 Introduction

2.1.1 Motivation

La méthode LASSO (Least Absolute Shrinkage and Selection Operator) consiste dans sa forme la plus simple à représenter de manière linéaire une variable $Y \in \mathbb{R}^p$ à l'aide d'un ensemble de variables dites "explicatives" $\{X_1, \ldots, X_n\}$, et de manière à sélectionner parmi ces n variables X_i celles qui sont le plus pertinentes.

La première idée consiste minimiser au sens des moindres carrés l'erreur due au modèle linéaire explicatif :

$$\min_{u \in \mathbb{R}^n} \frac{1}{2} \|Y - Xu\|_2^2 \ , \tag{2.1}$$

avec $Xu = \sum_{i=1}^n u_i X_i$ et où $\|x\|_2 = \sqrt{\sum_{j=1}^p x_j^2}$ est la norme euclidienne usuelle sur l'espace $\mathbb{R}^p$. Cependant, cette méthode a l'inconvénient de distribuer l'erreur sur toutes les variables explicatives X_i, alors que certaines d'entre elles n'expliquent peut-être pas grand chose...

Une seconde idée consisterait à minimiser la somme de l'écart quadratique précédent et de la norme de comptage $\|u\|_0$ du vecteur u dans l'espace $\mathbb{R}^n$, égale au nombre de composantes non nulles du vecteur u. Ce second problème est a priori beaucoup plus difficile à résoudre que le problème quadratique car il correspond à un problème d'optimisation combinatoire de taille n.

Une troisième possibilité consiste à utiliser la norme 1 du vecteur u dans l'espace $\mathbb{R}^n$, définie par $\|u\|_1 = \sum_{i=1}^n |u_i|$, et à résoudre :

$$\min_{u \in \mathbb{R}^n} \frac{1}{2} \|Y - Xu\|_2^2 + \lambda \|u\|_1 \ , \tag{2.2}$$

dont la solution est notée $u^\sharp(\lambda)$. Cette nouvelle régression "régularisée" par la norme 1 incorpore un coefficient $\lambda \geq 0$ qui contrôle la puissance de la régularisation : l'augmentation de λ induit la diminution de certaines composantes du vecteur optimal $u^\sharp(\lambda)$ et va même rendre nulles certaines de ces composantes pour des valeurs finies de λ, ce qui correspond bien au but recherché qui est d'éliminer les variables X_i "non explicatives".

Le problème (2.2) s'inscrit dans le cadre de l'optimisation convexe non-différentiable, et on va présenter dans la suite de ce chapitre des outils et des algorithmes adaptés à la résolution de ce problème.

2.1.2 Aperçu de l'algorithme du gradient proximal

On définit l'opérateur proximal d'une fonction $J : \mathbb{R}^n \to \mathbb{R}$ par :

$$\mathcal{P}_J(u) = \underset{v\in\mathbb{R}^n}{\arg\min} \left(J(v) + \frac{1}{2} \|v - u\|_2^2 \right) ,$$

où $\|\cdot\|_2$ est la norme usuelle (euclidienne) de $\mathbb{R}^n$. On notera que calculer la fonction $\mathcal{P}_J$ en un point u est a priori un problème de complexité équivalente à celle de résoudre le problème de minimisation de la fonction J sur $\mathbb{R}^n$.

On souhaite résoudre le problème d'optimisation suivant :

$$\min_{u\in\mathbb{R}^n} \; F(u) + G(u) ,$$

en supposant que :

— la fonction F est convexe, différentiable, définie sur tout $\mathbb{R}^n$,
— la fonction G est convexe, s.c.i., propre, pas forcément différentiable,
— le calcul de $\mathcal{P}_G$ est facile.

Pour cela, la méthode du gradient proximal consiste à mettre en œuvre l'algorithme :

$$u^{(k+1)} = \mathcal{P}_{\varepsilon^{(k)}G}\Big(u^{(k)} - \varepsilon^{(k)} \nabla F\big(u^{(k)}\big) \Big) ,$$

où $\varepsilon^{(k)} > 0$ représente la longueur d'un pas, qui sera constante ou déterminée par des techniques de recherche linéaire.

Par définition de l'opérateur proximal, cette méthode s'interprète comme la minimisation de la somme de la fonction G et d'une fonction quadratique représentant F au voisinage de $u^{(k)}$:

$$u^{(k+1)} = \underset{u\in\mathbb{R}^n}{\arg\min} \left(G(u) + \left\langle \nabla F(u^{(k)}) \, , u - u^{(k)} \right\rangle + \frac{1}{2\varepsilon^{(k)}} \left\| u - u^{(k)} \right\|_2^2 \right) .$$

Remarque 14. Sous cette forme, il apparaît clairement que l'algorithme du gradient proximal se réinterprète dans le cadre du principe du problème auxiliaire (PPA) [Cohen (2004)] : linéarisant la partie F du critère, gardant telle quelle la partie G et effectuant un choix de noyau quadratique $K(u) = (1/2)\|u\|_2^2$, la k-ème itération du PPA s'écrit :

$$\min_{u\in\mathbb{R}^n} \frac{1}{2} \left\| u - u^{(k)} \right\|_2^2 + \varepsilon^{(k)} \left\langle \nabla F(u^{(k)}) \, , u - u^{(k)} \right\rangle + \varepsilon^{(k)} G(u) ,$$

qui correspond donc exactement à une itération de gradient proximal. ◇

2.1.3 Quelques exemples

Voici quelques cas où l'algorithme du gradient proximal se ramène à un algorithme connu.

— Dans le cas de la fonction nulle : $G(u) = 0$, on a :

$$\mathcal{P}_G(u) = u ,$$

et l'algorithme du gradient proximal est l'algorithme du gradient usuel :

$$u^{(k+1)} = u^{(k)} - \varepsilon^{(k)} \nabla F(u^{(k)}) .$$

— Dans le cas de la fonction caractéristique d'un ensemble : $G(u) = \chi_{U^{\text{ad}}}(u)$, on a :

$$\mathcal{P}_G(u) = \text{proj}_{U^{\text{ad}}}(u) \;\; ,$$

et l'algorithme du gradient proximal est l'algorithme du gradient projeté :

$$u^{(k+1)} = \text{proj}_{U^{\text{ad}}}\left(u^{(k)} - \varepsilon^{(k)} \nabla F(u^{(k)}) \right) \;\; .$$

2.2 Outils pour les méthodes proximales

Dans toute cette section, $\mathbb{U}$ désigne l'espace de Hilbert de dimension finie $\mathbb{R}^n$, $\|\cdot\|$ représente la norme euclidienne sur cet espace et F est une fonction définie sur $\mathbb{U}$ à valeurs dans $]-\infty,+\infty]$.

On considère le problème suivant :

$$\inf_{v\in\mathbb{U}} F(v) + \frac{1}{2}\|v-u\|^2 \ . \tag{2.3}$$

Pour étudier le problème (2.3), on introduit les définitions et notations suivantes.

— On appelle *régularisée de Moreau-Yosida* (aussi appelée *enveloppe de Moreau*) et on note $\mathcal{M}_F : \mathbb{U} \to \mathbb{R}$ la fonction donnant pour tout u la valeur optimale du problème (2.3) :

$$\mathcal{M}_F(u) = \inf_{v\in\mathbb{U}} F(v) + \frac{1}{2}\|v-u\|^2 \ . \tag{2.4}$$

— On appelle *opérateur proximal* et on note $\mathcal{P}_F : \mathbb{U} \rightrightarrows \mathbb{U}$ la multi-application donnant pour tout u l'ensemble (éventuellement vide) des solutions du problème (2.3) :

$$\mathcal{P}_F(u) = \arg\min_{v\in\mathbb{U}} F(v) + \frac{1}{2}\|v-u\|^2 \ . \tag{2.5}$$

Remarque 15. On fait souvent dépendre la régularisée de Moreau-Yosida d'un paramètre $c>0$, et on note alors :

$$\mathcal{M}^c_F(u) = \inf_{v\in\mathbb{U}} F(v) + \frac{1}{2c}\|v-u\|^2 \ .$$

Les propriétés de $\mathcal{M}^c_F$ se déduisent alors de celles de $\mathcal{M}_F$, puisque $\mathcal{M}^c_F = \frac{1}{c}\mathcal{M}_{cF}$. ◇

Dans le cadre de l'optimisation convexe, la fonction $\mathcal{M}_F$ et l'opérateur $\mathcal{P}_F$ présentent des propriétés particulières, qui sont énoncées dans le théorème suivant.

Théorème 13. *On suppose que la fonction $F : \mathbb{U} \to]-\infty,+\infty]$ est convexe, s.c.i., propre et sous-différentiable. Alors, le problème (2.3) admet pour tout $u \in \mathbb{U}$ une* solution unique, *et l'opérateur proximal $\mathcal{P}_F$ est une application. La régularisée de Moreau-Yosida $\mathcal{M}_F$ est une application convexe continue différentiable, avec :*

$$\nabla\mathcal{M}_F(u) = u - \mathcal{P}_F(u) \ . \tag{2.6}$$

De plus, l'opérateur proximal $\mathcal{P}_F$ est non expansif *:*

$$\|\mathcal{P}_F(v) - \mathcal{P}_F(u)\| \le \|v-u\| \ , \tag{2.7}$$

et la régularisée de Moreau-Yosida est à gradient 1*-Lipschitzien :*

$$\|\nabla\mathcal{M}_F(v) - \nabla\mathcal{M}_F(u)\| \le \|v-u\| \ . \tag{2.8}$$

Démonstration. D'après les hypothèses faites sur F, la fonction $v \mapsto F(v) + \frac{1}{2}\|v-u\|^2$ est fortement convexe, s.c.i. propre. L'inf dans le problème (2.3) est donc atteint, et la solution du problème est unique : la multi-application $\mathcal{P}_F$ est donc en fait une application de $\mathbb{U}$ dans $\mathbb{U}$.

Par le théorème 7 et le corollaire 3 sur la fonction marginale d'un opérateur dans le cas convexe-convexe, comme la fonction $u \mapsto F(v) + \frac{1}{2}\|v-u\|^2$ est différentiable, on a que la régularisée de Moreau-Yosida $\mathcal{M}_F$ de F, qui est partout finie, est une fonction convexe continue différentiable avec :

$$\nabla\mathcal{M}_F(u) = u - \mathcal{P}_F(u) \ .$$

La condition d'optimalité du problème (2.3) s'écrit :

$$\exists r_u \in \partial F\big(\mathcal{P}_F(u)\big) \ , \quad r_u + \mathcal{P}_F(u) - u = 0 \ , \tag{2.9}$$

soit encore : $u - \mathcal{P}_F(u) \in \partial F\big(\mathcal{P}_F(u)\big)$. Écrivant cette condition d'optimalité en un autre point v, et utilisant la propriété de monotonie du sous-différentiel, on en déduit l'inégalité :

$$\big\langle u - \mathcal{P}_F(u) - \big(v - \mathcal{P}_F(v)\big) \, , \mathcal{P}_F(u) - \mathcal{P}_F(v) \big\rangle \geq 0 \ , \tag{2.10}$$

qui se réécrit sous la forme :

$$\big\langle u - v \, , \mathcal{P}_F(u) - \mathcal{P}_F(v) \big\rangle \geq \big\| \mathcal{P}_F(u) - \mathcal{P}_F(v) \big\|^2 \ .$$

Par l'inégalité de Cauchy-Schwarz, on obtient :

$$\| \mathcal{P}_F(v) - \mathcal{P}_F(u) \| \leq \| v - u \| \ ,$$

ce qui prouve le caractère non expansif de l'opérateur proximal. La relation (2.10) s'écrit encore :

$$\Big\langle u - \mathcal{P}_F(u) - \big(v - \mathcal{P}_F(v)\big) \, , u - \big(u - \mathcal{P}_F(u)\big) - v + \big(v - \mathcal{P}_F(v)\big) \Big\rangle \geq 0 \ .$$

Utilisant la relation (2.6), on obtient :

$$\big\langle u - v \, , \nabla \mathcal{M}_F(u) - \nabla \mathcal{M}_F(v) \big\rangle \geq \big\| \nabla \mathcal{M}_F(u) - \nabla \mathcal{M}_F(v) \big\|^2 \ ,$$

et donc, par Cauchy-Schwarz :

$$\| \nabla \mathcal{M}_F(v) - \nabla \mathcal{M}_F(u) \| \leq \| v - u \| \ ,$$

ce qui prouve que l'application $\mathcal{M}_F$ est à gradient 1-Lipschitzien. □

On dispose du corollaire suivant.

Corollaire 5. *Soit F définie sur $\mathbb{U}$ à valeurs dans $]-\infty, +\infty]$, convexe, s.c.i. et propre. Alors, pour tout $(u, p) \in \mathbb{U} \times \mathbb{U}$, les deux affirmations suivantes sont équivalentes.*

(i) $p = \mathcal{P}_F(u)$.

(ii) $u - p \in \partial F(p)$.

Démonstration. Cette équivalence a été vue dans la preuve du théorème précédent. □

Une autre propriété importante de l'opérateur proximal est sa décomposition, qui permet de relier l'opérateur proximal d'une fonction et celui de sa conjuguée de Fenchel.

Théorème 14. *Soit F définie sur $\mathbb{U}$ à valeurs dans $]-\infty, +\infty]$, convexe, s.c.i. et propre. Alors, pout tout $u \in \mathbb{U}$,*

$$\mathcal{P}_F(u) + \mathcal{P}_{F^\star}(u) = u \ .$$

Démonstration. Soit $u \in \mathbb{U}$. On sait que $u - \mathcal{P}_F(u) \in \partial F\big(\mathcal{P}_F(u)\big)$, et donc par le théorème 3 que $\mathcal{P}_F(u) \in \partial F^\star\big(u - \mathcal{P}_F(u)\big)$. Par le corollaire 5, on en déduit que $u - \mathcal{P}_F(u) = \mathcal{P}_{F^\star}(u)$, d'où le résultat. □

2.2.1 Régularisée de Moreau-Yosida

L'intérêt de la régularisée de Moreau-Yosida pour l'optimisation vient du résultat suivant.

Théorème 15. *Soit F définie sur $\mathbb{U}$ à valeurs dans $]-\infty,+\infty]$, convexe, s.c.i. et propre.*

1. *La régularisée $\mathcal{M}_F$ est convexe, s.c.i., propre, différentiable à gradient Lipschitzien.*
2. *Pour tout $u \in \mathbb{U}$, on a $\mathcal{M}_F(u) \leq F(u)$.*
3. $\inf_{u\in\mathbb{U}} F(u) = \inf_{u\in\mathbb{U}} \mathcal{M}_F(u)$.
4. $\arg\min_{u\in\mathbb{U}} F(u) = \arg\min_{u\in\mathbb{U}} \mathcal{M}_F(u) = \left\{u^\sharp \in \mathbb{U}\ ,\ \ u^\sharp = \mathcal{P}_F(u^\sharp)\right\}$.

Démonstration. Le point *1* a déjà été montré. Le point *2* est évident et le point *3* résulte du fait que la condition d'optimalité du problème de minimisation de la régularisée de Moreau-Yosida $\mathcal{M}_F$ s'écrit $u^\sharp = \mathcal{P}_F(u^\sharp)$, de telle sorte que la condition d'optimalité du problème (2.3) prend la forme : $0 \in \partial F(u^\sharp)$. Pour le point *4*, voir [Gilbert (2018), §3.7.2]. □

On considère alors le problème :

$$\min_{u\in\mathbb{U}} F(u)\ . \tag{2.11}$$

Le fait de disposer d'une fonction différentiable à gradient Lipschitzien ayant le même minimum et le même ensemble d'arg min que la fonction F suggère d'effectuer la minimisation suivante :

$$\min_{u\in\mathbb{U}} \mathcal{M}_F(u)\ , \tag{2.12}$$

et donc de trouver une solution du problème (2.11) par des méthodes de type gradient (avec un pas scalaire $\rho > 0$) sur la fonction $\mathcal{M}_F$ plutôt que sur la fonction F :

$$u^{(k+1)} = u^{(k)} - \rho\nabla\mathcal{M}_F(u^{(k)})\ .$$

Un tel algorithme ne peut bien sûr en général pas être mis en œuvre car la simple évaluation de $\mathcal{M}_F$ au point $u^{(k)}$ est de même complexité que la résolution du problème initial (2.11). On va voir dans la suite que l'on peut utiliser les méthodes proximales pour certains types de problème d'optimisation. Plus tard dans le cours, on verra aussi comment utiliser la transformée de Moreau-Yosida dans les problèmes de recherche de point-selle.

Exercice 2. *Fonction de Huber.* La fonction de Huber de paramètre λ, définie sur $\mathbb{U}$ à valeurs dans $\mathbb{R}$, est donnée par :

$$H_\lambda(u) = \begin{cases} \frac{1}{2\lambda}\|u\|^2 & \text{si } \|u\| \leq \lambda \\ \|u\| - \frac{\lambda}{2} & \text{sinon} \end{cases}\ ,$$

où $\|\cdot\|$ est la norme euclidienne de $\mathbb{U}$.

1. Calculer l'opérateur proximal de la fonction $F = \lambda\|\cdot\|$.
2. Calculer la régularisée de Moreau-Yosida de la fonction $F = \lambda\|\cdot\|$.
3. Faire le lien avec la fonction H_λ.

2.2.2 Opérateur proximal

Soit F définie sur $\mathbb{U}$ à valeurs dans $]-\infty,+\infty]$, convexe, s.c.i., propre et sous-différentiable. On s'intéresse au problème (2.11), que l'on rappelle :

$$\min_{u\in\mathbb{U}} F(u)\ . \tag{2.13}$$

On rappelle que l'opérateur proximal est caractérisé par la relation :

$$u-\mathcal{P}_F(u)\in\partial F\big(\mathcal{P}_F(u)\big)\ , \tag{2.14}$$

(voir le corollaire 5). On a le résultat suivant.

Théorème 16. *Soit F définie sur $\mathbb{U}$ à valeurs dans $]-\infty,+\infty]$, convexe, s.c.i. et propre. Alors, les deux propositions suivantes sont équivalentes.*

1. *Le point $u^\sharp$ est solution du problème (2.13) de minimisation de F :*

$$u^\sharp\in\arg\min_{u\in\mathbb{U}} F(u)\ .$$

2. *Le point $u^\sharp$ est un point fixe de l'opérateur proximal :*

$$u^\sharp=\mathcal{P}_F(u^\sharp)\ .$$

Démonstration. Utilisant la caractérisation (2.14) de l'opérateur proximal, on a :

$$u^\sharp=\mathcal{P}_F(u^\sharp)\iff 0\in\partial F(u^\sharp)\iff u^\sharp\in\arg\min_{u\in\mathbb{U}} F(u)\ ,$$

d'où le résultat. □

On en déduit l'algorithme dit du *point proximal*, consistant à résoudre le problème (2.13) par une méthode itérative de recherche d'un point fixe de $\mathcal{P}_F$:

$$u^{(k+1)}=\mathcal{P}_F(u^{(k)})\ .$$

L'*algorithme* du point proximal qui en découle se formule de la manière suivante. On notera que cet algorithme n'est a priori pas intéressant en pratique car l'évaluation de $\mathcal{P}_{cF}$ au point $u^{(k)}$ est de même complexité que la résolution du problème initial (2.13).

Algorithme 1.

1. *Choisir un point initial $u^{(0)}\in\mathbb{U}$ et un coefficient $c>0$. Poser $k=0$.*
2. *Calculer : $u^{(k+1)}=\mathcal{P}_{cF}(u^{(k)})$ et faire $k\leftarrow k+1$.*

La question de la convergence de cet algorithme est réglée par le théorème suivant.

Théorème 17. *Soit F définie sur $\mathbb{U}$ à valeurs dans $]-\infty,+\infty]$, convexe, s.c.i. propre. On suppose que le problème (2.13) admet un ensemble de solutions $U^\sharp$ non vide, et on note $F^\sharp$ la valeur optimale du problème. Alors, la suite $\big\{u^{(k)}\big\}_{k\in\mathbb{N}}$ engendrée par l'algorithme 1 converge vers un point $u^\sharp\in U^\sharp$, et on a la majoration :*

$$F(u^{(k)})-F^\sharp\le\frac{1}{2ck}\mathrm{dist}(u^{(0)},U^\sharp)^2\qquad\forall k>0\ .$$

Pour la preuve on se reportera à [Beck (2017), Theorem 10.28]. On remarquera que l'algorithme converge pour *toute* valeur positive du coefficient c. Dans le cas où la fonction F est *fortement convexe*, on peut montrer que la vitesse de convergence de l'algorithme du point proximal est *linéaire*, et non en $1/k$ comme dans le théorème 17 ci-dessus (voir [Beck (2017), Theorem 10.29] pour plus de détails).

Remarque 16. Choisissant un paramètre $\rho \in [0, 1[$, la version relaxée de l'algorithme du point proximal (avec $c = 1$) s'écrit :

$$\begin{aligned} u^{(k+\frac{1}{2})} &= \mathcal{P}_F(u^{(k)}) \ , \\ u^{(k+1)} &= (1-\rho)u^{(k)} + \rho u^{(k+\frac{1}{2})} \ , \end{aligned}$$

ou de manière équivalente :

$$\begin{aligned} u^{(k+1)} &= u^{(k)} - \rho\big(u^{(k)} - \mathcal{P}_F(u^{(k)})\big) \ , \\ &= u^{(k)} - \rho\nabla\mathcal{M}_F(u^{(k)}) \ . \end{aligned}$$

On en déduit que l'algorithme du point proximal relaxé est l'algorithme du gradient appliquée à la régularisée de Moreau-Yosida de F. Une fois de plus, cet algorithme relaxé a le défaut de nécessiter à chaque itération un calcul (ici le gradient de la régularisée de Moreau-Yosida) de même complexité que la résolution du problème initial. ◇

Règles de calcul pour l'opérateur proximal. On donne quelques opérations élémentaires utiles pour le calcul des fonctions prox.

Une propriété très utile de l'opérateur proximal est donnée par la proposition suivante.

Proposition 12. *On suppose que l'application F définie sur $\mathbb{R}^n$ à valeurs dans $]-\infty, +\infty]$, est* additive *:*

$$F(u) = \sum_{i=1}^{n} F_i(u_i) \ .$$

Alors l'opérateur proximal de F est le produit cartésien des opérateurs proximaux des F_i :

$$\mathcal{P}_F(u) = \prod_{i=1}^{n} \mathcal{P}_{F_i}(u_i) \ .$$

Démonstration. Comme la fonction F est additive en $(u_1, \ldots, u_n)$ et qu'il en est de même pour la fonction norme au carré, la minimisation dans le problème (2.3) se fait indépendamment selon chaque u_i, d'où le résultat. □

Remarque 17. Dans le cas où chaque fonction F_i est convexe s.c.i. propre, l'ensemble $\mathcal{P}_{F_i}(u_i)$ est réduit à un point, de telle sorte que l'opérateur proximal de F s'écrit :

$$\mathcal{P}_F(u) = \big(\mathcal{P}_{F_1}(u_1), \ldots, \mathcal{P}_{F_n}(u_n)\big) \ .$$

Autrement dit, l'opérateur proximal $\mathcal{P}_F$ se calcule composante par composante. ◇

On donne quelques transformations classiques d'une fonction F dont l'opérateur proximal se calcule facilement à partir de celui de F. Les calculs sont laissés au soin du lecteur.

Ajout d'un terme linéaire. Soit F une fonction propre et soit $a \in \mathbb{U}$. On définit la fonction G par : Soit $G(u) = F(u) + \langle a \, , u\rangle$. Alors,

$$\mathcal{P}_G(u) = \mathcal{P}_F(u - a) \ .$$

Ajout d'un terme quadratique. Soit F une fonction propre, soit $\lambda > 0$ et soit $a \in \mathbb{U}$. On définit la fonction G par : $G(u) = F(u) + \frac{\lambda}{2}\|u - a\|^2$. Alors,

$$\mathcal{P}_G(u) = \mathcal{P}_{\theta F}\big(\theta u + (1-\theta)a\big) \ ,$$

avec $\theta = 1/(1+\lambda)$.

Mise à l'échelle et translation. Soit F une fonction propre, soit $\lambda > 0$ et soit $a \in \mathbb{U}$. On définit la fonction G se déduisant de F par une transformation linéaire : $G(u) = F(\lambda u + a)$. Alors,

$$\mathcal{P}_G(u) = \frac{1}{\lambda}\big(\mathcal{P}_{\lambda^2 F}(\lambda u + a) - a\big) \ .$$

Fonction perspective. Soit F une fonction propre et soit $\lambda > 0$. On définit la fonction *perspective* G par : $G(u) = \lambda F(u/\lambda)$. Alors,

$$\mathcal{P}_G(u) = \lambda \mathcal{P}_{\lambda^{-1}F}(u/\lambda) \ .$$

Opérateurs proximaux usuels. On donne enfin quelques fonctions dont l'opérateur proximal se calcule de manière analytique. On trouvera dans [Combettes and Pesquet (2011)] l'expression de l'opérateur proximal d'un grand nombre de fonctions. On trouvera aussi dans le chapitre 6 de l'ouvrage [Beck (2017)] de nombreux exemples concernant la régularisée de Moreau-Yosida et l'opérateur proximal.

Norme ℓ_1 ("soft thresholding"). Soit F définie sur $\mathbb{U}$ à valeurs dans $]-\infty, +\infty]$, donnée par $F(u) = \|u\|_1$, et soit $\varepsilon > 0$. La proposition 12 s'applique à cette fonction, et la i-ème composante de l'opérateur proximal de F est :

$$\mathcal{P}_{\epsilon F}(u)_i = \begin{cases} u_i + \varepsilon & \text{si} \quad u_i \leq -\varepsilon \\ 0 & \text{si} \quad |u_i| \leq \varepsilon \\ u_i - \varepsilon & \text{si} \quad u_i \geq \varepsilon \end{cases} \ .$$

Norme euclidienne. Soit $F(u) = \|u\|_2$. Alors,

$$\mathcal{P}_{\varepsilon F}(u) = \begin{cases} \big(1 - \varepsilon/\|u\|_2\big)u & \text{si } \|u\|_2 \geq \varepsilon \\ 0 & \text{sinon} \end{cases} \ .$$

Fonction quadratique. Soit $F(u) = \frac{1}{2}u^\top A u + b^\top u + c$, avec A semi-définie positive. Alors,

$$\mathcal{P}_{\varepsilon F}(u) = \big(\mathrm{I} + \varepsilon A\big)^{-1}(u - \varepsilon b) \ .$$

Fonction support d'un ensemble convexe $U^{\mathrm{ad}} \subset \mathbb{U}$. Soit $F(u) = \sup_{v \in U^{\mathrm{ad}}} \langle u\,, v\rangle$. Alors,

$$\mathcal{P}_{\varepsilon F}(u) = u - \varepsilon\, \mathrm{proj}_{U^{\mathrm{ad}}}(u/\varepsilon) \ .$$

Exercice 3. *Norme ℓ_0 ("hard thresholding").* Soit F définie sur $\mathbb{U}$ à valeurs dans $]-\infty, +\infty]$, donnée par $F(u) = \|u\|_0$, et soit $\varepsilon > 0$. On rappelle que $\|u\|_0$ est la *norme de comptage* du vecteur u, qui est égale au nombre de composantes non nulles du vecteur u.

1. Donner l'expression de la fonction $\mathbb{I} : \mathbb{R} \to \mathbb{R}$, telle que $F(u) = \sum_{i=1}^n \mathbb{I}(u_i)$.
2. Calculer l'opérateur proximal de $\varepsilon\mathbb{I}$.
3. En déduire l'opérateur proximal de εF.

2.3 Algorithme du gradient proximal

L'une des grandes réussites des méthodes proximales est de pouvoir traiter des problèmes d'optimisation dans lesquels la fonction à minimiser est la somme de deux fonctions dont l'une est« lisse » (différentiable) et l'autre simplement sous-différentiable, mais d'opérateur proximal connu. On présente ici l'approche correspondant à ce type de problème.

Dans toute cette section, $\mathbb{U}$ désigne l'espace de Hilbert de dimension finie $\mathbb{R}^n$ et $\|\cdot\|$ représente la norme euclidienne sur cet espace.

2.3.1 Présentation

On pose le problème d'optimisation :

$$\min_{u\in\mathbb{U}} \; F(u) + G(u) \;, \tag{2.15}$$

où les fonctions F et G sont définies sur l'espace $\mathbb{U}$ à valeurs dans $]-\infty, +\infty]$. On utilisera dans la suite la notation :

$$J(u) = F(u) + G(u) \;.$$

On fait sur les fonctions F et G les hypothèses suivantes.

Hypothèse 3.

1. La fonction F est convexe propre, son domaine est l'espace $\mathbb{U}$ tout entier et elle est différentiable à gradient Lipschitzien de constante $L > 0$:

$$\|\nabla F(u) - \nabla F(v)\| \leq L\,\|u - v\| \qquad \forall u,\; v \in \mathbb{U} \;.$$

2. La fonction G est convexe s.c.i. propre.
3. Le minimum de la fonction $J = F + G$ est fini et atteint en (au moins) un point $u^\sharp \in \mathbb{U}$.

Remarque 18. On remarquera que le problème (2.11) étudié dans le cadre de l'algorithme du point proximal au §2.2.2 entre dans ce cadre, puisqu'il suffit de considérer dans (2.15) que la partie différentiable du critère est identiquement nulle. ◇

Le théorème suivant est à la base de l'algorithme du gradient proximal.

Théorème 18. *Sous l'hypothèse 3, une solution quelconque $u^\sharp$ du problème (2.15) est telle que, pour tout $\varepsilon > 0$, on a :*

$$u^\sharp = \mathcal{P}_{\varepsilon G}\big(u^\sharp - \varepsilon\nabla F(u^\sharp)\big) \;.$$

Démonstration. Soit $u^\sharp$ une solution du problème (2.15) :

$$\exists r^\sharp \in \partial G(u^\sharp),\; r^\sharp + \nabla F(u^\sharp) = 0 \;.$$

Pour tout $\varepsilon > 0$, cette dernière relation se met sous la forme équivalente :

$$\varepsilon r^\sharp + u^\sharp - \big(u^\sharp - \varepsilon\nabla F(u^\sharp)\big) = 0 \;,$$

qui exprime le fait que le point $u^\sharp$ satisfait la condition d'optimalité du problème :

$$\min_{v\in\mathbb{U}} \; \varepsilon G(v) + \frac{1}{2}\left\|v - \big(u^\sharp - \varepsilon\nabla F(u^\sharp)\big)\right\|^2 \;.$$

Par définition de l'opérateur proximal, on en déduit que $u^\sharp = \mathcal{P}_{\varepsilon G}\big(u^\sharp - \varepsilon\nabla F(u^\sharp)\big)$. □

Le théorème 18 conduit alors, dans l'optique de résoudre le problème 2.15, à proposer l'algorithme suivant, dit *algorithme du gradient proximal*.

Algorithme 2.

1. *Choisir un point initial* $u^{(0)} \in \mathbb{U}$. *Poser* $k = 0$.
2. *À l'itération* k *de l'algorithme,*
   - *choisir un pas* $\varepsilon^{(k)} > 0$,
   - *mettre à jour la variable* u *par :*

$$u^{(k+1)} = \mathcal{P}_{\varepsilon^{(k)}G}\Big(u^{(k)} - \varepsilon^{(k)}\nabla F\big(u^{(k)}\big)\Big) \,, \tag{2.16}$$

3. *Faire* $k \leftarrow k+1$ *et retourner à l'étape 2 jusqu'à la convergence.*

Exemple 1. Considérons le cas où la fonction F est une fonction quadratique en u et où la fonction G correspond à la norme ℓ_1. On retrouve la méthode LASSO introduite en introduction de ce chapitre. Dans ce cas, l'algorithme du gradient proximal est désigné par l'acronyme ISTA "Iterative Shrinkage-Thresholding Algorithm". △

Afin d'alléger les notations, on introduit l'*opérateur gradient* $\mathcal{G}_\varepsilon$ défini par :

$$\mathcal{G}_\varepsilon(u) = \frac{1}{\varepsilon}\Big(u - \mathcal{P}_{\varepsilon G}\big(u - \varepsilon\nabla F(u)\big)\Big) \,.$$

Cet opérateur permet d'écrire l'itération (2.16) du gradient proximal sous la forme :

$$u^{(k+1)} = u^{(k)} - \varepsilon^{(k)}\mathcal{G}_{\varepsilon^{(k)}}(u^{(k)}) \,. \tag{2.17}$$

Théorème 19. *Sous l'hypothèse 3, un point* $u^\sharp \in \mathbb{U}$ *est une solution du problème* (2.15) *si et seulement si il est tel que :*

$$\mathcal{G}_\varepsilon(u^\sharp) = 0 \quad \forall \epsilon > 0 \,.$$

Démonstration. Par le théorème 18 et la définition de la fonction $\mathcal{G}_\varepsilon$, le sens direct de la proposition est vérifiée.

Réciproquement, utilisant la caractérisation donnée par le corollaire 5 de l'opérateur proximal de la fonction εG au point $u - \varepsilon\nabla F(u)$, on a que :

$$\big(u - \varepsilon\nabla F(u)\big) - \mathcal{P}_{\varepsilon G}\big(u - \varepsilon\nabla F(u)\big) \in \varepsilon\partial G\big(\mathcal{P}_{\varepsilon G}\big(u - \varepsilon\nabla F(u)\big)\big) \,.$$

Par définition de $\mathcal{G}_\varepsilon$, on a $\mathcal{P}_{\varepsilon G}\big(u - \varepsilon\nabla F(u)\big) = u - \varepsilon\mathcal{G}_\varepsilon(u)$, et cette condition s'écrit donc :

$$\mathcal{G}_\varepsilon(u) \in \nabla F(u) + \partial G\big(u - \varepsilon\mathcal{G}_\varepsilon(u)\big) \,. \tag{2.18}$$

et donc :

$$\mathcal{G}_\varepsilon(u^\sharp) = 0 \quad \Rightarrow \quad u^\sharp \in \operatorname*{arg\,min}_{u\in\mathbb{U}} \Big(F(u) + G(u)\Big) \,,$$

ce qui achève la démonstration. □

2.3.2 Convergence

On suppose que la constante de Lipschitz L du gradient de la fonction F est connue. On fait alors le choix d'un pas *constant* dans l'algorithme du gradient proximal :

$$\varepsilon^{(k)} = \varepsilon \qquad \forall k \in \mathbb{N} \,,$$

vérifiant l'inégalité :

$$\varepsilon \leq \frac{1}{L} \,. \tag{2.19}$$

Théorème 20. *Sous l'hypothèse 3 et avec le choix de pas (2.19), la suite des itérées $u^{(k)}$ de l'algorithme 2 du gradient proximal engendre une suite monotone décroissante $\big\{J(u^{(k)})\big\}_{k\in\mathbb{N}}$ qui converge vers la valeur optimale $J^\sharp = J(u^\sharp)$ du critère. De plus, l'écart $\big(J(u^{(k)}) - J^\sharp\big)$ est majorée par une constante arbitrairement petite σ en un nombre d'itérations d'ordre $1/\sigma$.*

Démonstration. On suit le schéma classique de preuve des algorithmes d'optimisation.

1. Inégalité fondamentale. On rappelle que l'hypothèse de gradient Lipschitzien de F implique :

$$F(v) \leq F(u) + \big\langle \nabla F(u) \,, v - u \big\rangle + \frac{L}{2} \, \|v - u\|^2 \quad \forall u \,, \;\; v \in \mathbb{U} \,.$$

Avec $v = u - \varepsilon \mathcal{G}_\varepsilon(u)$, cette inégalité s'écrit :

$$F\big(u - \varepsilon \mathcal{G}_\varepsilon(u)\big) \leq F(u) - \varepsilon \big\langle \nabla F(u) \,, \mathcal{G}_\varepsilon(u) \big\rangle + \frac{\varepsilon^2 L}{2} \, \|\mathcal{G}_\varepsilon(u)\|^2 \quad \forall u \in \mathbb{U} \,.$$

Ajoutant de part et d'autre de cette inégalité le terme $G\big(u - \varepsilon \mathcal{G}_\varepsilon(u)\big)$, on obtient :

$$J\big(u - \varepsilon \mathcal{G}_\varepsilon(u)\big) \leq F(u) - \varepsilon \big\langle \nabla F(u) \,, \mathcal{G}_\varepsilon(u) \big\rangle + \frac{\varepsilon^2 L}{2} \, \|\mathcal{G}_\varepsilon(u)\|^2 + G\big(u - \varepsilon \mathcal{G}_\varepsilon(u)\big) \,.$$

Utilisant d'une part la convexité de F :

$$F(u) \leq F(w) + \big\langle \nabla F(u) \,, u - w \big\rangle \,,$$

d'autre part celle de G pour tout $r \in \partial G\big(u - \varepsilon \mathcal{G}_\varepsilon(u)\big)$:

$$G\big(u - \varepsilon \mathcal{G}_\varepsilon(u)\big) \leq G(w) + \big\langle r \,, u - \varepsilon \mathcal{G}_\varepsilon(u) - w \big\rangle \,,$$

et choisissant $r = \mathcal{G}_\varepsilon(u) - \nabla F(u)$ qui appartient à $\partial G\big(u - \varepsilon \mathcal{G}_\varepsilon(u)\big)$ par la relation (2.18), on obtient que, pour tout $w \in \mathbb{U}$:

$$\begin{aligned} J\big(u - \varepsilon \mathcal{G}_\varepsilon(u)\big) &\leq F(w) + \big\langle \nabla F(u) \,, u - w \big\rangle \\ &\qquad - \varepsilon \big\langle \nabla F(u) \,, \mathcal{G}_\varepsilon(u) \big\rangle + \frac{\varepsilon^2 L}{2} \, \|\mathcal{G}_\varepsilon(u)\|^2 \\ &\qquad + G(w) + \big\langle \mathcal{G}_\varepsilon(u) - \nabla F(u) \,, u - \varepsilon \mathcal{G}_\varepsilon(u) - w \big\rangle \\ &\leq J(w) + \big\langle \mathcal{G}_\varepsilon(u) \,, u - w \big\rangle + \Big(\frac{\varepsilon^2 L}{2} - \varepsilon\Big) \, \|\mathcal{G}_\varepsilon(u)\|^2 \,. \end{aligned}$$

Avec un choix de pas conforme à (2.19), on en déduit :

$$J\big(u - \varepsilon \mathcal{G}_\varepsilon(u)\big) \leq J(w) + \big\langle \mathcal{G}_\varepsilon(u) \,, u - w \big\rangle - \frac{\varepsilon}{2} \, \|\mathcal{G}_\varepsilon(u)\|^2 \,. \tag{2.20}$$

2. Variation de la fonction J. Écrivant l'inégalité (2.20) pour $u = w = u^{(k)}$, on obtient :

$$J(u^{(k+1)}) - J(u^{(k)}) \leq -\frac{\varepsilon}{2} \left\| \mathcal{G}_\varepsilon(u^{(k)}) \right\|^2 .$$

On en déduit que la suite $\left\{ J(u^{(k)}) \right\}_{n\in\mathbb{N}}$ est décroissante, minorée par $J^\sharp$ et donc convergente.

3. Distance à l'optimum. Écrivant l'inégalité (2.20) pour $u = u^{(k)}$ et $w = u^\sharp$, on obtient :

$$\begin{aligned} J(u^{(k+1)}) &\leq J(u^\sharp) + \left\langle \mathcal{G}_\varepsilon(u^{(k)}) \, , u^{(k)} - u^\sharp \right\rangle - \frac{\varepsilon}{2} \left\| \mathcal{G}_\varepsilon(u^{(k)}) \right\|^2 \\ &= J(u^\sharp) + \frac{1}{2\varepsilon} \Big(\left\| u^{(k)} - u^\sharp \right\|^2 - \left\| u^{(k)} - u^\sharp - \varepsilon \mathcal{G}_\varepsilon(u^{(k)}) \right\|^2 \Big) , \end{aligned}$$

et donc :

$$J(u^{(k+1)}) - J^\sharp \leq \frac{1}{2\varepsilon} \Big(\left\| u^{(k)} - u^\sharp \right\|^2 - \left\| u^{(k+1)} - u^\sharp \right\|^2 \Big) . \tag{2.21}$$

On en déduit en particulier que $\left\| u^{(k+1)} - u^\sharp \right\| \leq \left\| u^{(k)} - u^\sharp \right\|$: la distance de $u^{(k)}$ à toute solution optimale $u^\sharp$ décroît au cours des itérations de l'algorithme.

4. Vitesse de convergence. Sommant les inégalités (2.21) entre les indices 1 et k, on obtient :

$$\begin{aligned} \sum_{\ell=1}^{k} \Big(J(u^{(\ell)}) - J^\sharp \Big) &\leq \frac{1}{2\varepsilon} \Big(\left\| u^{(0)} - u^\sharp \right\|^2 - \left\| u^{(k)} - u^\sharp \right\|^2 \Big) \\ &\leq \frac{1}{2\varepsilon} \left\| u^{(0)} - u^\sharp \right\|^2 . \end{aligned}$$

La suite $\left\{ J(u^{(k)}) \right\}_{n\in\mathbb{N}}$ étant décroissante, $J(u^{(k)}) - J^\sharp \leq J(u^{(\ell)}) - J^\sharp$ pour tout $\ell \leq k$, d'où :

$$J(u^{(k)}) - J^\sharp \leq \frac{1}{2\varepsilon k} \left\| u^{(0)} - u^\sharp \right\|^2 .$$

On en déduit que la suite $\left\{ J(u^{(k)}) \right\}_{n\in\mathbb{N}}$ converge vers $J^\sharp$, et que l'écart $\big(J(u^{(k)}) - J^\sharp \big)$ est majorée par une constante arbitrairement petite σ en un nombre d'itérations k d'ordre $1/\sigma$. □

Remarque 19. Dans le cas où la constante de Lipschitz du gradient de la fonction F n'est pas connue, on peut quand même mettre en œuvre l'algorithme du gradient proximal en ajustant le pas $\varepsilon^{(k)}$ par une recherche linéaire de type Armijo. Plus précisément, initialisant l'itération k avec un pas $\varepsilon^{(k)}$ « assez grand », on effectue un pas de l'algorithme :

$$u^+ = u^{(k)} - \varepsilon^{(k)} \mathcal{G}_{\varepsilon^{(k)}}(u^{(k)}) .$$

On valide ce résultat, c'est-à-dire on passe à l'itération suivante en fixant $u^{(k+1)} = u^+$, si la condition de décroissance :

$$F\big(u^+\big) \leq F(u^{(k)}) - \varepsilon^{(k)} \big\langle \mathcal{G}_{\varepsilon^{(k)}}(u) \, , \nabla F(u^{(k)}) \big\rangle + \frac{\varepsilon^{(k)}}{2} \left\| \mathcal{G}_{\varepsilon^{(k)}}(u^{(k)}) \right\|^2 ,$$

est vérifiée. Dans le cas contraire, on fait décroître le pas $\varepsilon^{(k)}$ en le multipliant par un coefficient $\beta \in]0,1[$ et on recalcule u^+. Cet algorithme qui incorpore la recherche linéaire dans la méthode du gradient proximal converge dans les mêmes conditions que l'algorithme à pas fixe. On pourra consulter les détails de cet algorithme dans [Beck and Teboulle (2010)]. ◇

Remarque 20. On retrouve l'algorithme du point proximal à partir de l'algorithme du gradient proximal en l'appliquant au cas où la partie différentiable F du critère est identiquement nulle et donc où $J = G$. Alors, l'algorithme du point proximal :

$$u^{(k+1)} = \mathcal{P}_{\varepsilon^{(k)} J}\big(u^{(k)}\big) ,$$

converge pour le choix de pas $\varepsilon^{(k)} = \varepsilon$, pourvu que l'on ait $\varepsilon > 0$. ◇

2.3.3 Accélération de l'algorithme du gradient proximal

Il existe une version accélérée de l'algorithme du gradient proximal, appelé FISTA ("Fast Iterative Shrinkage-Thresholding Algorithm"). Cet algorithme, dont on trouve l'analyse complète dans [Beck and Teboulle (2009)] et [Beck and Teboulle (2010)], est présenté brièvement. Il s'applique au même problème et avec les mêmes hypothèses que l'algorithme standard du gradient proximal.

Algorithme 3.

1. *Choisir un point initial* $u^{(0)} = v^{(0)} \in \mathrm{dom} G$.
2. *À l'itération* k *de l'algorithme,*
   — *choisir un pas* $\varepsilon^{(k)} > 0$,
   — *mettre à jour le couple* (u, v) *par :*

$$u^{(k+1)} = \mathcal{P}_{\varepsilon^{(k)} G}\Big(v^{(k)} - \varepsilon^{(k)} \nabla F\big(v^{(k)}\big)\Big) \,, \tag{2.22a}$$

$$v^{(k+1)} = u^{(k+1)} + \frac{k-1}{k+2}\big(u^{(k+1)} - u^{(k)}\big) \,. \tag{2.22b}$$

3. *Faire* $k \leftarrow k+1$ *et retourner à l'étape 2 jusqu'à la convergence.*

On notera que la seconde étape (2.22b) correspond à une phase d'extrapolation de la variable u. Sur ce nouvel algorithme, on peut faire les remarques suivantes.

— Il a la même complexité de mise en œuvre que l'algorithme 2.
— Il converge sous les mêmes hypothèses que l'algorithme 2, avec des pas $\varepsilon^{(k)}$ fixes ($\leq 1/L$), ou encore avec des pas obtenus par recherche linéaire.

L'intérêt de l'algorithme FISTA vient de sa vitesse de convergence. On montre en effet que l'écart $\big(J(u^{(k)}) - J^\sharp\big)$ est majorée par une constante arbitrairement petite σ en un nombre d'itérations d'ordre $1/\sqrt{\sigma}$, alors qu'il était d'ordre $1/\sigma$ pour l'algorithme du gradient proximal standard. Cette amélioration justifie le qualificatif "accéléré" pour cet algorithme.

Remarque 21. La suite $\big\{J(u^{(k)})\big\}_{k\in\mathbb{N}}$ engendrée par l'algorithme 3 n'est a priori pas monotone décroissante. On sait cependant en faire une version modifiée garantissant la décroissance du critère au cours des itérations. ◇

2.3.4 Exemple

On considère l'exemple suivant :

$$\min_{u\in\mathbb{R}^{1000}} \frac{1}{2}\,\|Au - b\|_2^2 + \lambda\,\|u\|_1 \,, \tag{2.23}$$

où A est une matrice $(2000, 1000)$ et où b est un vecteur de dimension 2000, tous deux obtenus par tirage aléatoire de leurs coefficients suivant la loi normale centrée réduite. On rappelle que la constante de Lipschitz associé au gradient de la partie quadratique du critère est la valeur propre maximale de la matrice $A^\top A$. Le coefficient $\lambda > 0$ permet de faire varier le poids de la partie en norme L_1 du critère.

On applique alors à cet exemple (de type LASSO) les algorithmes du gradient proximal standard et accéléré. La convergence de ces deux méthodes est illustrée sur la figure 2.1 (on notera que l'ordonnée de ce graphe est en échelle logarithmique). L'ensemble des propriétés vues précédemment se trouvent vérifiées sur cet exemple : convergence du gradient proximal standard, amélioration dû à l'algorithme FISTA et décroissance non monotone de ce dernier.

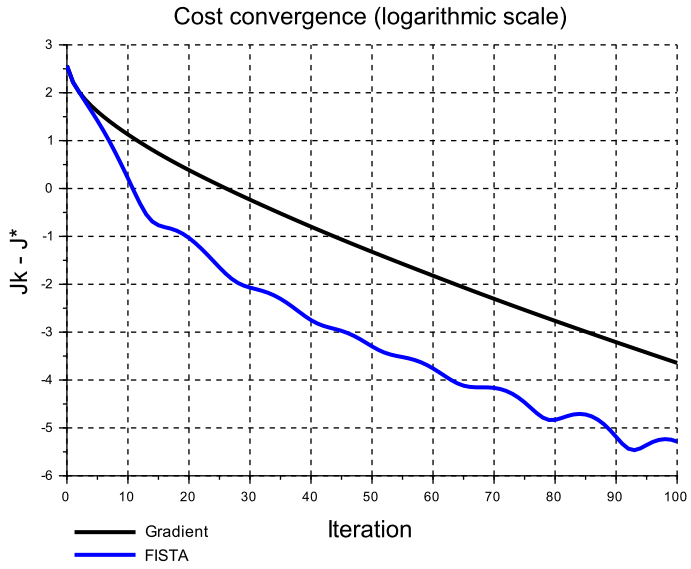

FIGURE 2.1 – Convergence des algorithmes de type gradient proximal

2.4 Extensions

On présente ici quelques développements possibles dans le cadre des méthodes proximales.

2.4.1 Résolvante d'un opérateur ⊖

Soit M un opérateur multi-valué définie sur l'espace $\mathbb{U} = \mathbb{R}^n$ à valeurs dans $\mathbb{U}$. On appelle ε-résolvante de M l'opérateur R_ε (a priori multi-valué) défini par :

$$R_\varepsilon = \big(\mathrm{I}_\mathbb{U} + \varepsilon M\big)^{-1} ,$$

où $\mathrm{I}_\mathbb{U}$ est l'application identité sur $\mathbb{U}$. On a les propriétés suivantes.

— Si l'opérateur M est maximal, alors le domaine de R_ε est l'espace $\mathbb{U}$ tout entier.

— Si M est monotone, alors R_ε est mono-valué, Lipschitz de constante 1 sur son domaine.

On peut montrer que l'opérateur prox $\mathcal{P}_{\varepsilon F}$ est la résolvante de l'opérateur sous-différentiel ∂F :

$$\mathcal{P}_{\varepsilon F}(u) = \big(\mathrm{I}_\mathbb{U} + \varepsilon \partial F\big)^{-1}(u) .$$

Comme l'opérateur sous-différentiel d'une fonction convexe s.c.i. est maximal monotone, on retrouve le fait que l'opérateur prox est mono-valué défini sur tout l'espace $\mathbb{U}$. On retrouve aussi la relation :

$$w = \mathcal{P}_{\varepsilon F}(u) \quad \Leftrightarrow \quad u \in \big(\mathrm{I}_\mathbb{U} + \varepsilon \partial F\big)(w) .$$

2.4.2 Méthodes algorithmiques avancées ▲

1. Nesterov's second and third methods.
2. Douglas-Rachford splitting algorithm.
3. Alternating Direction Method of Multiplier (ADMM).