Problèmes Inverses et Régularisation

Arthur Leclaire

Dans tout ce cours, on considérera une image $u : \Omega \to \mathbb{R}$ en niveaux de gris définie sur

$$\Omega = [0 : M-1] \times [0 : N-1]$$

On identifiera $\mathbb{R}^\Omega$ à $\mathbb{R}^d$ où $d = MN$ est le nombre de pixels.

1 Optimisation

Dans ce cours, on va chercher à minimiser une fonction $F : \mathbb{R}^d \to \overline{\mathbb{R}}$ définie par

$$F(u) = f(u) + g(u)$$

où f et g sont deux fonctions sur $\mathbb{R}^d$. Pour la restauration d'images, on s'intéressera au cas où f est un terme d'attache aux données

$$f(u) = \frac{1}{2}\|Au - v\|_2^2 \, ,$$

où $A \in \mathbb{R}^{k\times d}, v \in \mathbb{R}^k$ et où $g(u) = R(u)$ est une fonction de régularisation.

On avait plusieurs cas (en débruitage) où il était possible de trouver explicitement le point où F atteint son minimum, et de le calculer efficacement (par exemple via la transformée de Fourier discrète). Mais lorsque la solution n'est plus explicite, on a besoin de méthodes d'optimisation pour approcher la solution.

On notera $\inf F = \inf\{ F(u) \, , \, u \in \mathbb{R}^d \}$ et $\operatorname{Argmin} F = \{ u \in \mathbb{R}^d \, , \, F(u) = \inf F \}$ (qui peut être vide).

1.1 Descente de Gradient

On considère ici la méthode de descente de gradient.

$$u_{n+1} = u_n - \tau_n \nabla F(u_n) \, ,$$

où $\tau_n > 0$ est une suite de pas. Lorsque $\tau_n = \tau$ est constante, on parle de descente de gradient à pas fixe. Lorsque l'on choisit

$$\tau_n \in \operatorname*{Argmin}_{t\in\mathbb{R}} F(u_n - t\nabla F(u_n))$$

on parle de descente de gradient à pas optimal.

Le théorème suivant donne une condition suffisante de convergence de la méthode de descente de gradient à pas fixe.

Théorème 1. *Soit $F : \mathbb{R}^d \to \mathbb{R}$ convexe différentiable telle que ∇F est L-lipschitzienne. Supposons que F atteint son infimum en un point de $\mathbb{R}^d$. Soit $\tau \in (0, \frac{2}{L})$, $u_0 \in \mathbb{R}^d$ et (u_n) la suite définie par récurrence par*

$$u_{n+1} = u_n - \tau \nabla F(u_n) \, .$$

Alors (u_n) converge vers un minimum de F.

Démonstration. Admise. Ce résultat est prouvé dans le livre [Bauschke, Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, 2011]. □

Remarque 1. Si l'on suppose que F est convexe différentiable α-fortement convexe et à gradient L-Lipschitz, alors on peut montrer que F atteint son infimum, et on peut montrer plus facilement la convergence de (u_n) vers un certain $u_* \in \operatorname{Argmin} F$ en supposant $\tau < \frac{2\alpha}{L^2}$ (ce qui est un peu plus contraignant que $\tau < \frac{2}{L}$).

La descente de gradient à pas fixe fournit une méthode très générique pour chercher les minima de F, pour peu que l'on puisse accéder simplement à son gradient. Par exemple, pour $f(u) = \frac{1}{2}\|Au - v\|_2^2$, on trouve immédiatement

$$\nabla f(u) = A^T(Au - v) ,$$

et la constante de Lipschitz de ∇f est $\|A^T A\|$ (où $\|\cdot\|$ est la norme matricielle subordonnée à $\|\cdot\|_2$). En prenant aussi une régularisation différentiable avec gradient calculable (voir par exemple Section 2.2), on peut expliciter la condition sur le pas τ pour avoir convergence de la descente de gradient à pas fixe.

1.2 Descente de Gradient Proximale

Définition 1. Soit $g : \mathbb{R}^d \to \overline{\mathbb{R}}$. On définit l'opérateur proximal de g par

$$\operatorname{Prox}_g(u) = \operatorname*{Argmin}_{z \in \mathbb{R}^d} \frac{1}{2}\|u - z\|_2^2 + g(z) .$$

Sans hypothèse sur g, l'argmin définissant $\operatorname{Prox}_g(u)$ est un ensemble qui peut être vide, ou contenir plusieurs points. En revanche, pour $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ convexe semi-continue inférieurement (s.c.i.) telle que $g \not\equiv +\infty$, on peut montrer que le min est atteint en un seul point, et on s'autorise alors à définir un point $p = \operatorname{Prox}_g(u) \in \mathbb{R}^d$. En particulier, lorsque $g : \mathbb{R}^d \to \mathbb{R}$ est convexe différentiable, on peut montrer que

$$p = \operatorname{Prox}_g(u) \iff u = (\mathrm{Id} + \nabla g)(p) .$$

Pour minimiser $F = f + g$, on peut alors utiliser l'algorithme de descente de gradient proximale (PGD), qui s'écrit

$$u_{n+1} = \operatorname{Prox}_{\tau g}(u_n - \tau \nabla f(u_n)) ,$$

où $\tau > 0$ est un pas fixé.

Théorème 2. *Soit $f : \mathbb{R}^d \to \mathbb{R}$ convexe différentiable à gradient L_f-Lipschitz.*

Soit $g : \mathbb{R}^d \to \mathbb{R} \setminus \{+\infty\}$ convexe s.c.i. telle que $g \not\equiv +\infty$.

On suppose que $F = f + g$ atteint son infimum en un point de $\mathbb{R}^d$.

Alors, dès que $\tau < \frac{2}{L_f}$, la suite (u_n) donnée par l'algorithme de descente de gradient proximale converge vers un certain $u_* \in \operatorname{Argmin} F$.

Démonstration. Admise. Ce résultat est prouvé dans le livre [Bauschke, Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, 2011]. □

L'intérêt de ce résultat est qu'il s'applique à des fonctions g non différentiables, et même possiblement discontinues. On notera que la condition de convergence sur le pas τ ne fait apparaître que la constante de Lipschitz de ∇f.

Exemple 1. Considérons l'indicatrice $g = \iota_C$ d'un ensemble convexe fermé $C \subset \mathbb{R}^d$. Par définition,

$$\forall u \in \mathbb{R}^d, \quad \iota_C(u) = \begin{cases} 0 & \text{si } u \in C \\ +\infty & \text{sinon} \end{cases} .$$

On peut voir dans ce cas que $\operatorname{Prox}_{\tau g}$ est bien défini, et est le projecteur orthogonal sur C. Dans ce cas, l'algorithme PGD pour minimiser $f + g$ se résume à une "descente de gradient projetée", qui permet de chercher une solution du problème d'optimisation sous contraintes $\inf_C f$.

Exemple 2. Pour $f(u) = \frac{1}{2}\|Au - v\|_2^2$, l'opérateur proximal de f est donné par

$$\forall u \in \mathbb{R}^d, \quad \mathrm{Prox}_{\tau f}(u) = \tau(\mathsf{Id} + \tau A^T A)^{-1} A^T u$$

Dans certains cas (par exemple, lorsque l'on connaît une base de diagonalisation de A), on peut calculer cet opérateur proximal, mais pas toujours. S'il est facilement accessible et si la fonction g est différentiable, alors on peut décider de renverser l'algorithme PGD pour faire l'étape de descente de gradient sur g, et l'étape proximale sur f.

1.3 Lien avec un modèle d'évolution

En prenant ici l'exemple de la régularisation de Tychonov, on va voir voir le lien entre la descente de gradient sur F et l'équation de la chaleur. En TD, on calculera l'adjoint de l'opérateur ∇ et on fera avec les discrétisations de la divergence et du laplacien, ce qui permet d'écrire

$$R(u) = \lambda\|\nabla u\|_2^2 = \lambda\langle \nabla^T \nabla u, u\rangle = -\lambda\langle \mathrm{div}(\nabla u), u\rangle = -\lambda\langle \Delta u, u\rangle \ .$$

Comme Δ est un opérateur symétrique, ceci donne que R est différentiable et que $\nabla R(u) = -2\lambda\Delta u$. Partant de u_0, faire une descente de gradient directement sur R revient donc à calculer

$$u_{n+1} - u_n = 2\lambda\tau\Delta u_n$$

ce qui revient à discrétisation en temps de l'équation de la chaleur.

Si maintenant, on applique la descente de gradient à

$$F(u) = \frac{1}{2}\|u - v\|_2^2 + \lambda\|\nabla u\|_2^2$$

on obtient la suite

$$u_{n+1} - u_n = -\tau(u_n - v) + 2\lambda\tau\Delta u_n \ .$$

On retrouve bien un schéma que vous avez étudié au premier semestre, où l'on a le terme $-\tau(u_n - v)$ lié à l'attache aux données, et un terme de diffusion en Δu_n. Et l'on voit que cette suite (u_n) converge effectivement vers l'arg-minimum de F que l'on avait pu calculer explicitement dans le cours précédent.

2 Régularisation et Opérateurs Proximaux

2.1 Régularisation de Tychonov

Pour $R(u) = \frac{\lambda}{2}\|Bu\|_2^2$, on peut calculer

$$\nabla R(u) = \lambda B^T B u$$

et en déduire Prox_R. La constante de Lipschitz de ∇R est liée au calcul de norme matricielle

$$\|B^T B\|_2 = \sup_{\|u\|_2 \leqslant 1} \langle B^T B u, u\rangle = \sup_{\|u\| \leqslant 1} \|Bu\|_2^2.$$

Par exemple, si B est l'opérateur ∇, on a

$$\|\nabla^T \nabla u\|_2 = \sup_{\|u\|_2 \leqslant 1} \|\nabla u\|_2^2 = \sup_{\|u\|_2 \leqslant 1} \|\partial_1 u\|_2^2 + \|\partial_2 u\|_2^2 \ .$$

De plus, la formule de Parseval donne que

$$\|\partial_1 u\|_2^2 + \|\partial_2 u\|_2^2 = \frac{1}{MN} \sum_{(\xi,\zeta)\in\Omega} 4\left(\sin^2\left(\frac{\pi\xi}{M}\right) + \sin^2\left(\frac{\pi\zeta}{N}\right)\right) |\hat{u}(\xi,\zeta)|^2 \leqslant 8\|u\|_2^2 \ .$$

2.2 Régularisation parcimonieuse

Soit $B \in \mathbb{R}^{d\times d}$ une matrice orthogonale.

Pour $R(u) = \lambda\|Bu\|_q$ avec $q \in \{0,1\}$, on a calculé dans le cours/td précédent une expresion de l'opérateur proximal associé à R. On a vu que cela revenait à appliquer une fonction de seuillage aux coefficients dans la base B : pour $q = 0$, on applique le seuillage dur avec la fonction

$$\forall t \in \mathbb{R}, \quad h_{0,\lambda}(t) = t\mathbf{1}_{|t|>\sqrt{2\lambda}} ,$$

et un seuillage doux avec la fonction

$$\forall t \in \mathbb{R}, \quad h_{1,\lambda}(t) = \mathrm{sgn}(t)(|t| - \lambda)_+ .$$

On rappelle le théorème qu'on a démontré là-dessus.

Théorème 3. *Soit $q \in \{0,1\}$. Soit $R(u) = \lambda\|Bu\|_q$ avec $\lambda > 0$ et $B \in \mathbb{R}^{d\times d}$ une matrice orthogonale.*

Pour $u \in \mathbb{R}^d$, on définit $T_q^\lambda u \in \mathbb{R}^d$ par ses coordonnées dans la nouvelle base : pour tout $j = 1,\dots,d$,

$$(BT_q^\lambda u)_j = h_{q,\lambda}((Bu)_j) .$$

Avec $f(u) = \frac{1}{2}\|Au - v\|_2^2$ et $g(u) = \lambda\|Bu\|_1$, la fonction

$$F(u) = \frac{1}{2}\|Au - v\|_2^2 + \lambda\|Bu\|_1$$

est strictement convexe et on peut de plus montrer que $F(u) \to +\infty$ quand $\|u\|_2 \to \infty$. On peut donc montrer qu'il existe un unique $u_* \in \mathbb{R}^d$ tel que $F(u_*) = \inf F$. La norme ℓ^1 n'étant pas différentiable, on ne peut pas utiliser l'algorithme de descente de gradient pour minimiser F. En revanche, puisqu'on sait calculer l'opérateur proximal de g, on peut utiliser l'algorithme PGD, qui s'écrit donc ici

$$u_{n+1} = T_1^\lambda(u_n - \tau\nabla f(u_n)) .$$

L'analyse ci-dessus montre que (u_n) converge alors vers u_* dès que $\tau < \frac{2}{\|A^TA\|}$. Cet algorithme est connu dans la littérature sous le nom *Iterative Soft-Tresholding*, et a été introduit dans [Daubechies, Defrise, De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, 2004].

2.3 Régularisation par variation totale

On peut aussi considérer d'autres fonctions de régularisation plus compliquées, comme par exemple, la variation totale définie par

$$\mathrm{TV}(u) = \sum_{(x,y)\in\Omega} \|\nabla u(x,y)\| .$$

On notera qu'il existe plusieurs variantes selon la norme utilisée sur $\mathbb{R}^2$ pour calculer $\|\nabla u(x,y)\|$. Par exemple la norme euclidienne $\|(a,b)\|_2 = \sqrt{a^2+b^2}$ mène à la TV dite isotrope, alors que la norme $\|(a,b)\|_1 = |a| + |b|$ mène à la TV dite non-isotrope.

La résolution d'un problème inverse avec régularisation TV mène au problème de minimisation

$$\operatorname*{Argmin}_{u\in\mathbb{R}^d} \frac{1}{2}\|Au - v\|_2^2 + \lambda\mathrm{TV}(u) .$$

La régularisation TV favorise des images dites *cartoon*, c'est-à-dire lisses par morceaux, avec des discontinuités nettes. Mais le dernier problème d'optimisation n'a en général pas de solution explicite, même lorsque $A = \mathrm{Id}$ (cas du débruitage). Autrement dit, on n'a pas une formule close pour calculer $\mathrm{Prox}_{\lambda\mathrm{TV}}$. De plus, la variation totale n'est pas lisse : la fonction $u \mapsto \mathrm{TV}(u)$ n'est pas différentiable. La résolution des problèmes d'optimisation correspondants nécessitent donc des outils d'optimisation avancés, notamment des outils de dualité convexe.

2.4 Régularisation par TV lissée

Soit $\varepsilon > 0$. On pose

$$\mathsf{TV}_\varepsilon(u) = \sum_{(x,y)\in\Omega} \sqrt{\varepsilon^2 + \partial_1 u(x,y)^2 + \partial_2 u(x,y)^2}\ .$$

On définit aussi la fonction appelée Huber-TV par

$$\mathsf{HTV}_\varepsilon(u) = \sum_{(x,y)\in\Omega} \psi_\varepsilon(\|\nabla u(x,y)\|_2)\ .$$

où

$$\forall t \in \mathbb{R},\quad \psi_\varepsilon(t) = \begin{cases} \frac{t^2}{2\varepsilon} & \text{si } |t| < \varepsilon \\ |t| - \frac{\varepsilon}{2} & \text{sinon} \end{cases}\ .$$

On peut voir que ces deux fonctions sont bien différentiables. Par exemple, comme ψ_ε est $\mathscr{C}^1$, on peut voir que

$$\nabla\mathsf{TV}_\varepsilon(u) = \nabla^T\left(\frac{\nabla u}{\sqrt{\varepsilon^2 + \|\nabla u\|_2^2}}\right)\ .$$

On peut voir que $\nabla\mathsf{TV}_\varepsilon$ est $\frac{8}{\varepsilon}$-Lipschitz, ce qui permet de donner des conditions de convergence de la descente de gradient lorsqu'on utilise TV_ε pour fonction de régularisation.