

## TD 04 : Régularisation

Outils d'optimisation pour les sciences des données et de la décision, M2 MIAGE

8 novembre 2024



### Exercice 1 : Perte de Huber renversée

On considère un modèle linéaire  $\mathbf{x} \mapsto \mathbf{w}^T \mathbf{x}$  et un jeu de données  $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ , où  $\mathbf{x}_i \in \mathbb{R}^d$  et  $y_i \in \mathbb{R}$ .

Dans cet exercice, on renverse l'idée de la perte de Huber, en proposant une fonction de perte qui ressemble à la valeur absolue sur  $[-1, 1]$  et à une quadratique partout ailleurs. On définit donc la "perte de Huber renversée" comme

$$v: \mathbb{R} \rightarrow \mathbb{R} \\ t \mapsto v(t) := \begin{cases} |t| & \text{if } |t| < 1 \\ \frac{t^2+1}{2} & \text{sinon.} \end{cases} \quad (1)$$

Cette fonction est convexe mais non lisse (car non dérivable en 0).

a) On s'intéresse tout d'abord au problème non lisse suivant :

$$\underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimiser}} \frac{1}{n} \sum_{i=1}^n v(\mathbf{x}_i^T \mathbf{w} - y_i). \quad (2)$$

Peut-on appliquer l'algorithme de descente de gradient ? Si non, quel outil peut-on utiliser pour construire des algorithmes pour résoudre (2) ?

b) On considère maintenant le problème

$$\underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimiser}} f(\mathbf{w}) + \lambda \sum_{i=1}^d v([\mathbf{w}]_i), \quad (3)$$

où  $f$  est une fonction de perte de classe  $\mathcal{C}^1$ , et  $\lambda > 0$ .

- Comment s'appelle un problème de cette forme ? Quel est le rôle du second terme de l'objectif ?
- Écrire l'itération générique de la méthode du gradient proximal pour ce problème. À quelle condition est-il envisageable d'utiliser cette méthode en pratique ?

## Exercice 2 : Autour du problème de LASSO

On considère le problème dit de *LASSO généralisé* donné par

$$\underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimiser}} \frac{1}{2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{A}\mathbf{w}\|_1, \quad (4)$$

où  $\mathbf{X} \in \mathbb{R}^{n \times d}$  et  $\mathbf{y} \in \mathbb{R}^n$  représentent un jeu de données,  $\lambda > 0$  et  $\mathbf{A} \in \mathbb{R}^{m \times d}$  avec  $m$  un entier supérieur ou égal à 1. On rappelle que

$$\forall \mathbf{v} \in \mathbb{R}^m, \|\mathbf{v}\|_1 = \sum_{i=1}^m |v_i|,$$

où les  $v_i$  sont les coefficients du vecteur  $\mathbf{v}$ .

- a) Pourquoi ne peut-on pas appliquer la méthode de descente de gradient à ce problème ?
- b) Écrire l'itération de l'algorithme du gradient proximal appliqué à ce problème avec une taille de pas constante. On donne  $\nabla \psi(\mathbf{w}) = \mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{y})$  pour  $\psi : \mathbf{w} \mapsto \frac{1}{2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$ .
- c) On considère le cas particulier  $m = d$  et  $\mathbf{A} = \mathbf{I} \in \mathbb{R}^{d \times d}$ .
  - i) Que représente alors le terme en  $\lambda \|\mathbf{A}\mathbf{w}\|_1$  dans la fonction objectif du problème (4) ? Quelle est son utilité ?
  - ii) À quel algorithme vu en cours correspond alors la méthode de gradient proximal décrite en question b) ?

### Exercice 3 : Perte logistique régularisée

On se donne un problème de régression logistique construit à partir d'un jeu de données  $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$  pour lequel on considère une perte logistique

$$\ell : (h, y) \mapsto \ln(1 + \exp(-y h)) \quad (5)$$

et un modèle linéaire  $\mathbf{x} \mapsto \mathbf{x}^T \mathbf{w}$  paramétré par  $\mathbf{w} \in \mathbb{R}^d$ . Le problème d'optimisation associé est ainsi

$$\underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimiser}} f(\mathbf{w}) + \lambda \Omega(\mathbf{w}), \quad (6)$$

avec

$$f(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^n \ln(1 + \exp(-y_i \mathbf{x}_i^T \mathbf{w})), \quad (7)$$

$\lambda > 0$  et  $\Omega(\mathbf{w}) := \frac{1}{2} \|\mathbf{w}\|^2$ . La fonction  $\Omega$  est de classe  $\mathcal{C}^2$ , et on a  $\nabla \Omega(\mathbf{w}) = \mathbf{w}$ .

- Le terme  $\lambda \Omega(\mathbf{w})$  ne dépend pas des données. Comment appelle-t-on ce terme, et quel est son rôle ?
- Lorsque  $\lambda > 0$ , le problème (6) est fortement convexe. Comment cela se traduit-il au niveau de la dérivée seconde de la fonction  $f + \lambda \Omega$  ?
- Écrire (en pseudo-code) l'itération du gradient proximal appliqué au problème (6), avec un pas générique  $\alpha_k$ .
- Comme  $\Omega$  est dérivable, on peut aussi appliquer l'algorithme de descente de gradient à (6). Écrire (en pseudo-code) l'itération de la descente de gradient pour ce problème, avec un pas générique  $\alpha_k$ .
- Les solutions de (6) sont généralement de norme euclidienne plus faible que celles du problème non régularisé ( $\lambda = 0$ ). Comment la régularisation influe-t-elle sur cette norme, et quel est le but derrière la réduction de cette norme ?