Exercices : Régularisation

Optimisation pour l'apprentissage automatique

M2 IASD Apprentissage, 2024-2025

Dauphine | PSL
UNIVERSITÉ PARIS

Exercice 1 : Perte de Huber renversée

Adapté de l'examen 2019-2020.

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

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

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

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

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

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

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

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

i) Comment s'appelle un problème de cette forme ? Quel est le rôle du second terme de l'objectif ?

ii) 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 : Gradient proximal

Problème posé en session 2023-2024.

On considère un jeu de données $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ et $\boldsymbol{y} \in \mathbb{R}^n$, et le problème de régression linéaire avec régularisation "du filet élastique" (*elastic net*) :

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimiser}} \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|^2 + \frac{\lambda_2}{2}\|\boldsymbol{w}\|^2 + \lambda_1\|\boldsymbol{w}\|_1, \tag{4}$$

avec $\lambda_2 \geq 0$ et $\lambda_1 \geq 0$.

a) Quelle est l'utilité d'un terme de régularisation en général ?

b) Lorsque $\lambda_1 = 0$ et $\lambda_2 > 0$, quel est le rôle du terme de régularisation ?

c) Même question lorsque $\lambda_2 = 0$ et $\lambda_1 > 0$.

d) On rappelle que le gradient de la fonction $\varphi : \boldsymbol{w} \mapsto \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|^2$ est donné par

$$\nabla\varphi(\boldsymbol{w}) = \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}).$$

En utilisant cette formule, écrire l'itération du gradient proximal pour le problème (4).

e) Lorsque $\lambda_2 = 0$ et $\lambda_1 > 0$, à quel algorithme le gradient proximal est-il équivalent ?

f) Lorsque $\lambda_1 = 0$, donner un algorithme du cours applicable à la résolution du problème autre que le gradient proximal.

g) Lorsque $\lambda_1 > 0$ et $\lambda_2 > 0$, il n'existe pas en général de formule explicite pour les itérés du gradient proximal : en pratique, on utilise donc un algorithme d'optimisation à chaque itération du gradient proximal pour calculer les itérés (de manière approchée). Proposer un algorithme itératif parmi ceux vus en cours qui serait applicable aux itérations du gradient proximal, et justifier de son intérêt pour ce problème particulier.

Exercice 3 : Perte logistique régularisée

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

$$\ell : (h, y) \mapsto \ln\left(1 + \exp(-y\, h)\right) \tag{5}$$

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

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

avec

$$f(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + \exp(-y_i\, \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w})\right), \tag{7}$$

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

a) Le terme $\lambda\Omega(\boldsymbol{w})$ ne dépend pas des données. Comment appelle-t-on ce terme, et quel est son rôle ?

b) Lorsque $\lambda > 0$, le problème (6) est fortement convexe. Comment cela se traduit-il au niveau de la dérivée seconde de la fonction $f + \lambda\Omega$?

c) Écrire (en pseudo-code) l'itération du gradient proximal appliqué au problème (6), avec un pas générique α_k.

d) Comme Ω 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 α_k. Comparer les deux itérations.

e) 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 ?

Solutions des exercices

Solution de l'exercice 1 : Perte de Huber renversée

a) La fonction v n'est pas dérivable en tout point : l'algorithme de descente de gradient n'est donc pas défini, et pas applicable sur ce problème. En revanche, v est convexe, et il est possible de définir le sous-différentiel (l'ensemble des sous-gradients) de v en tout point : on peut alors construire des algorithmes pour résoudre (2) basés sur les sous-gradients et non le gradient.

b) i) Ce problème est un problème d'optimisation sous forme composite : la fonction objectif est donnée par la somme d'un terme lisse et d'un terme non lisse. Le but de ce dernier terme est généralement de *régulariser* la solution du problème, c'est-à-dire de pénaliser les points qui ne satisfont pas les propriétés souhaitées sur la solution.

ii) En un point $\boldsymbol{w}_k$, l'itération générique du gradient proximal (avec une longueur de pas α_k) s'écrit :

$$\boldsymbol{w}_{k+1} \in \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \left\{ f(\boldsymbol{w}_k) + \nabla f(\boldsymbol{w}_k)^{\mathrm{T}}(\boldsymbol{w} - \boldsymbol{w}_k) + \frac{1}{2\alpha_k}\|\boldsymbol{w} - \boldsymbol{w}_k\|_2^2 + \lambda \sum_{i=1}^{d} v([\boldsymbol{w}]_i) \right\}.$$

Cet algorithme n'est intéressant que dans le cas où les sous-problèmes apparaissant à chaque itération sont plus faciles à résoudre que le problème originel.

Solution de l'exercice 2 : Gradient proximal

a) Un terme de régularisation permet d'ajouter de la structure (càd des propriétés particulières) dans le problème, ce qui modifie généralement l'ensemble des solutions par rapport à une version non régularisée du problème.

b) Lorsque $\lambda_1 = 0$ et $\lambda_2 > 0$, le terme de régularisation est un terme de régularisation ℓ_2, qui permet de réduire la variance de la solution par rapport aux données.

Autres réponses possibles : réduire la norme ℓ_2 de la solution, garantir l'unicité de la solution lorsque le terme d'attache aux données est convexe.

c) Lorsque $\lambda_2 = 0$ et $\lambda_1 > 0$, le terme de régularisation est un terme de régularisation ℓ_1, qui vise à favoriser les solutions parcimonieuses (ayant un grand nombre de coefficients nuls).

Autres réponses possibles : réduire la norme ℓ_1 de la solution, identifier les composantes de $\boldsymbol{w}$ les plus utiles dans l'attache aux données.

d) La $k$ième itération de l'algorithme du gradient proximal appliqué au problème (4) s'écrit

$$\boldsymbol{w}_{k+1} \in \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \left\{ \varphi(\boldsymbol{w}_k) + \frac{1}{n}(\boldsymbol{X}\boldsymbol{w}_k - \boldsymbol{y})^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{w} - \boldsymbol{w}_k) + \frac{1}{2\alpha_k}\|\boldsymbol{w} - \boldsymbol{w}_k\|^2 + \frac{\lambda_2}{2}\|\boldsymbol{w}\|^2 + \lambda_1\|\boldsymbol{w}\|_1 \right\}$$

où $\alpha_k > 0$.

e) Quand $\lambda_2 = 0$, l'algorithme du gradient proximal est équivalent à l'algorithme ISTA.

f) Lorsque $\lambda_1 = 0$, le problème est un problème quadratique fortement convexe. On peut donc lui appliquer l'algorithme de descente de gradient, mais aussi les variantes avec momentum (algorithme de Nesterov, *Heavy ball*). En écrivant la fonction objectif sous la forme

$$\varphi(\boldsymbol{w}) + \frac{\lambda_2}{2}\|\boldsymbol{w}\|^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\left((\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i)^2 + \lambda_2\|\boldsymbol{w}\|^2\right),$$

on voit que chacun des termes de la somme dépend d'un unique point du jeu de données. Cela justifie qu'on puisse appliquer l'algorithme du gradient stochastique et ses variantes à ce problème.

g) On peut considérer la résolution du sous-problème par l'algorithme du gradient proximal lui-même ! En effet, on peut considérer la norme ℓ_1 comme le terme de régularisation de la fonction objectif du sous-problème, et définir alors une méthode proximale. Celle-ci correspondrait alors à appliquer l'algorithme ISTA, dont les itérations ont l'avantage d'être définies de manière explicite et sont donc aisées à calculer.

On peut également envisager d'appliquer l'algorithme de sous-gradient à ce problème non lisse, même s'il exploite moins la structure du problème que le gradient proximal.

Solution de l'exercice 3 : Perte logistique régularisée

a) Le terme $\lambda\Omega(\boldsymbol{w})$ est un terme de régularisation : son rôle est de pénaliser les vecteurs $\boldsymbol{w}$ qui ne satisfont pas une certaine structure ou, de manière équivalente, de favoriser certaines formes de solutions.

b) Puisque le problème est fortement convexe et que $f + \lambda\Omega$ est de classe $\mathcal{C}^2$, on sait qu'il existe $\mu > 0$ tel que

$$\forall \boldsymbol{w} \in \mathbb{R}^d, \quad \nabla^2 f(\boldsymbol{w}) + \lambda\nabla^2\Omega(\boldsymbol{w}) \succeq \mu \boldsymbol{I}_d.$$

c) L'itération du gradient proximal sur ce problème s'écrit :

$$\boldsymbol{w}_{k+1} \in \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \left\{ f(\boldsymbol{w}_k) + \nabla f(\boldsymbol{w}_k)^{\mathrm{T}}(\boldsymbol{w} - \boldsymbol{w}_k) + \frac{1}{2\alpha_k}\|\boldsymbol{w} - \boldsymbol{w}_k\|^2 + \lambda\Omega(\boldsymbol{w}) \right\}.$$

d) Comme f et Ω sont de classe $\mathcal{C}^2$, elles sont dérivables et le gradient de la fonction objectif de (6) est donc donné par $\nabla f(\boldsymbol{w}) + \lambda\Omega(\boldsymbol{w})$. Par conséquent, l'itération de la descente de gradient sur ce problème s'écrit :

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k\left(\nabla f(\boldsymbol{w}_k) + \lambda\nabla\Omega(\boldsymbol{w}_k)\right).$$

Cette itération est en fait identique à celle du gradient proximal sur ce problème.

e) L'un des buts de la régularisation en norme ℓ_2 ou norme euclidienne est réduire la variance de la solution du problème par rapport aux données : pour cela, la régularisation ℓ_2 impose une contrainte implicite sur la norme euclidienne de la solution. Une solution de norme plus faible sera ainsi moins sensible à une variation dans les données.