TD 07 : Révisions et annales

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

29 novembre 2024

Dauphine | PSL
UNIVERSITÉ PARIS

Note : Ce TD est basé sur l'examen du cours de l'année universitaire 2023-2024. Certains exercices ont été modifiés afin de prendre en compte les sujets enseignés en 2024-2025.

Exercice 1 : Un problème non convexe

Problème posé à l'examen 2023-2024 du cours "Optimization for Machine Learning" en MIAGE ID apprentissage.

Soit un jeu de données $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ où l'on suppose que $y_i \in (0,1)$ pour tout i. Partant de la fonction de perte

$$\ell(h, y) := \left(y - \frac{1}{1+\exp(-h)}\right)^2, \tag{1}$$

le problème d'expliquer les données par un modèle linéaire peut s'écrire :

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimiser}}\ \phi(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^n \ell\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}, y_i\right) = \frac{1}{n}\sum_{i=1}^n \left(y_i - \frac{1}{1+\exp(-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w})}\right)^2. \tag{2}$$

La fonction ϕ de ce problème est de classe $\mathcal{C}^2$ et est non convexe.

a) Justifier que 0 est un minorant de la fonction ϕ. Est-ce nécessairement sa valeur optimale ?

b) Supposons que l'on applique l'algorithme de descente de gradient appliqué à (2).

   i) Écrire l'itération de cet algorithme appliqué avec une taille de pas quelconque.

   ii) Donner deux choix possibles pour la taille de pas.

   iii) Sous les bonnes hypothèses, quelle est la complexité de cette méthode sur un tel problème ? À quelle quantité ce résultat s'applique-t-il ?

c) Supposons que l'algorithme de descente de gradient renvoie un point en lequel le gradient est nul. Est-ce nécessairement un minimum ?

d) Rappeler la condition nécessaire d'optimalité à l'ordre deux. Un point qui vérifie cette condition est-il un minimum ?

e) Supposons que l'on choisisse un point initial aléatoire pour la descente de gradient, et que l'algorithme converge vers un point vérifiant la condition nécessaire d'optimalité à l'ordre deux. Comment expliquer ce phénomène ?

Exercice 2 : Gradient proximal

Problème posé à l'examen 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*) :

$$\underset{\boldsymbol{w} \in \mathbb{R}^d}{\text{minimiser}} \frac{1}{2n} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 + \frac{\lambda_2}{2} \|\boldsymbol{w}\|^2 + \lambda_1 \|\boldsymbol{w}\|_1, \tag{3}$$

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 (3).

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 : Somme finie et optimisation distribuée

On considère un problème de la forme

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimiser}} \sum_{i=1}^{n} f_i(\boldsymbol{w}), \tag{4}$$

où chaque $f_i : \mathbb{R}^d \to \mathbb{R}$ est supposée être de classe $\mathcal{C}^1$.

a) On s'intéresse tout d'abord à la caractérisation des solutions de ce problème.

i) Donner la définition d'un minimum global du problème (4).

ii) Pourquoi l'ensemble des solutions du problème (4) et l'ensemble des solutions du problème

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimiser}}\, f(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{w}) \tag{5}$$

sont-ils identiques ?

iii) Si f est fortement convexe, que peut-on dire des minima du problème (5) (et donc de ceux de (4)) ?

b) Dans la suite, on se concentre sur le problème (5).

i) Écrire l'itération de descente de gradient avec une taille de pas constante pour ce problème.

ii) On suppose que les fonctions $\{f_i\}_{i=1}^n$ sont convexes. Quelle est alors la complexité de l'algorithme de descente de gradient sur le problème (5) ? À quelle quantité cette vitesse s'applique-t-elle ?

iii) Donner le nom d'un algorithme qui possède une meilleure complexité que la descente de gradient sur ce problème sous l'hypothèse que toutes les fonctions f_i sont convexes. Quelle est cette vitesse de convergence ?

c) On considère maintenant que l'on dispose d'un jeu de n données et que chaque f_i dépend d'un exemple distinct du jeu de données.

i) Quel est le coût en termes d'accès aux données d'une itération de descente de gradient appliquée au problème (5) ?

ii) Écrire l'itération de l'algorithme du gradient stochastique avec une taille de pas constante appliqué au problème (5).

iii) Comparer le coût de l'algorithme du gradient stochastique en termes d'accès aux données avec le coût de l'algorithme de descente de gradient établi en question c-i).

d) On suppose maintenant que les différents exemples du jeu de données sont répartis sur r processeurs, avec r compris entre 1 et n.

i) Écrire une itération de l'algorithme de gradient stochastique par fournées de taille n_b en utilisant une taille de pas constante.

ii) Compte tenu du contexte de la question, quel peut être l'intérêt pratique de choisir $n_b = r$?

iii) Si $r \approx n$, que risque-t-on d'observer en termes de performance si l'on choisit $n_b = r$?

Solutions des exercices

Solution de l'exercice 1 : Un problème non convexe

a) Un carré est toujours positif ou nul, et une moyenne de termes positifs et nuls est toujours positive ou nulle. Par conséquent, on a $\phi(\boldsymbol{w}) \geq 0$ pour tout $\boldsymbol{w} \in \mathbb{R}^d$, ce qui justifie que 0 est un minorant de la fonction ϕ.

Il ne s'agit pas pour autant de la valeur optimale. Pour que 0 soit la valeur optimale de ϕ, il faut qu'il existe $\bar{\boldsymbol{w}} \in \mathbb{R}^d$ tel que $\phi(\bar{\boldsymbol{w}}) = 0$, ce qui n'est pas garanti par l'énoncé.

b) (Application de la descente de gradient)

i) $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla\phi(\boldsymbol{w}_k)$ avec $\alpha_k > 0$.

ii) On peut choisir une taille de pas constante ($\alpha_k = \alpha > 0$ pour tout k), une taille de pas décroissante ($\alpha_k \to 0$) ou encore une taille de pas adaptative, choisie à chaque itération en fonction de l'itéré courant $\boldsymbol{w}_k$ (et des valeurs de ϕ et de son gradient en ce point), par exemple par une recherche linéaire.

iii) Sous les bonnes hypothèses, on garantit que la descente de gradient vérifie $\min_{0\leq\ell\leq k-1} \|\nabla\phi(\boldsymbol{x}_\ell)\| \leq \epsilon$ en au plus $\mathcal{O}(\epsilon^{-2})$ itérations.

c) La fonction est non convexe, donc un point en lequel le gradient est nul peut ne pas être un minimum (ce peut être un maximum, local ou global, ou encore un point selle).

d) Pour la fonction ϕ de classe $\mathcal{C}^2$, la condition nécessaire d'optimalité à l'ordre deux s'énonce comme suit :

$$\bar{\boldsymbol{w}} \in \mathbb{R}^d \text{ minimum local de } \phi \quad \Rightarrow \quad \nabla\phi(\bar{\boldsymbol{w}}) = \mathbf{0} \quad \text{and} \quad \nabla^2\phi(\bar{\boldsymbol{w}}) \succeq \mathbf{0}.$$

e) Un théorème énoncé en cours établit que la descente de gradient converge presque sûrement (avec probabilité 1) vers un point vérifiant la condition nécessaire d'optimalité à l'ordre deux. Il est donc logique que ce soit ce que l'on observe avec un tirage aléatoire de $\boldsymbol{w}_0$.

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 (3) 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.

Solution de l'exercice 3 : Somme finie et optimisation distribuée

a) (Caractérisation des solutions)

i) Un point $\boldsymbol{w}^* \in \mathbb{R}^d$ est un minimum global du problème (4) si

$$\sum_{i=1}^{n} f_i(\boldsymbol{w}^*) \leq \sum_{i=1}^{n} f_i(\boldsymbol{w}) \forall \boldsymbol{w} \in \mathbb{R}^d.$$

ii) La multiplication de la fonction objectif d'un problème par un scalaire positif ne change pas l'ensemble des solutions. Dans le cas présent, on a

$$\begin{aligned}
\boldsymbol{w}^* \in \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{i=1}^{n} f_i(\boldsymbol{w}) \quad &\iff \quad \sum_{i=1}^{n} f_i(\boldsymbol{w}^*) \leq \sum_{i=1}^{n} f_i(\boldsymbol{w}) \forall \boldsymbol{w} \in \mathbb{R}^d \\
&\iff \quad \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{w}^*) \leq \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{w}) \forall \boldsymbol{w} \in \mathbb{R}^d \\
&\iff \quad \boldsymbol{w}^* \in \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w}).
\end{aligned}$$

iii) Si f est fortement convexe, alors elle possède un unique minimum local qui est global (l'ensemble des minima, qu'ils soient globaux ou locaux, consiste donc en un seul élément).

b) Problème (5).

i) $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha \nabla f(\boldsymbol{w}_k)$, où $\alpha > 0$

ii) Lorsque les fonctions f_i sont convexes, alors f l'est aussi. Par conséquent, on peut garantir que $f(\boldsymbol{w}_k) - \min_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) \leq \epsilon$ en au plus $\mathcal{O}(\epsilon^{-1})$ itérations.

iii) L'algorithme du gradient accéléré, aussi appelé algorithme de Nesterov, possède une meilleure vitesse de convergence que la descente de gradient dans le cadre convexe. Cette complexité est en $\mathcal{O}(\epsilon^{-1/2})$.

c) (Chaque f_i dépend d'un exemple distinct du jeu de données.)

i) Une itération de descente de gradient coûte n accès aux données.

ii) $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha \nabla f_{i_k}(\boldsymbol{w}_k$, où $\alpha > 0$ et $i_k \in \{1, \ldots, n\}$ est un indice tiré aléatoirement.

iii) L'algorithme ne requiert qu'un seul accès aux données par itération. Son coût par itération est donc n fois moins élevé que celui de la descente de gradient.

d) (Données réparties sur r processeurs, $1 \leq r \leq n$)

i) $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \frac{\alpha}{n_b} \sum_{i \in S_k} \nabla f_i(\boldsymbol{w}_k)$, où $\alpha > 0$ et S_k est un ensemble de n_b indices tirés aléatoirement (avec ou sans remise) dans $\{1, \ldots, n\}$.

ii) En choisissant $n_b = r$, on pourra effectuer les calculs de gradient en parallèle sur les r processeurs.

iii) Si $r \approx n$, choisir $n_b = r$ placerait la méthode dans le régime de "grande fournée" (*large batch*), ce qui signifierait que son comportement pratique serait plus proche de celui de la descente de gradient que du gradient stochastique.