Optimisation pour l'apprentissage automatique

M2 IASD Apprentissage

Examen 2024/2025

Dauphine | PSL
UNIVERSITÉ PARIS

Durée : 2 heures.
Contenu : Trois exercices.
Les exercices ont vocation à être indépendants. Il n'est pas nécessaire de les rédiger sur des copies séparées.
Documents autorisés : Une feuille A4 recto-verso de notes manuscrites ou imprimées.
Si les étudiants pensent constater une erreur dans l'une des questions, ils sont invités à l'indiquer explicitement sur leur copie, et à poursuivre leur composition en tenant compte de cette erreur.

Remarques préliminaires

- Les dimensions des vecteurs ou matrices sont toujours supposées supérieures ou égales à 1.
- Pour tout vecteur $\boldsymbol{u} \in \mathbb{R}^m$, la i-ème coordonnée de ce vecteur sera notée $[\boldsymbol{u}]_i$.
- Pour tout entier $d \geq 1$, la notation $\mathbf{0}_d$ désignera le vecteur nul de $\mathbb{R}^d$.
- Pour tout vecteur $\boldsymbol{u} \in \mathbb{R}^m$, on notera $\|\boldsymbol{u}\|_2 = \sum_{j=1}^m [\boldsymbol{u}]_j^2$ la norme euclidienne, ou ℓ_2, de ce vecteur, et $\|\boldsymbol{u}\|_1 = \sum_{j=1}^m |[\boldsymbol{u}]_j|$ sa norme ℓ_1.

Exercice 1 : Pseudo-perte de Huber

Le but de cet exercice est de considérer un problème de régression basé sur la *pseudo-perte de Huber* (dite Huber lissée) suivante :

$$\begin{array}{rccl} p: & \mathbb{R} & \to & \mathbb{R} \\ & t & \mapsto & p(t) := \sqrt{1+t^2} - 1. \end{array}$$

On peut montrer que p est de classe $\mathcal{C}^1$ et qu'il s'agit d'une fonction fortement convexe.

Partant de la perte p et d'un jeu de données de régression $\{(\boldsymbol{x}_i, y_i)\}_{i=1,\dots,n}$ avec $\boldsymbol{x}_i \in \mathbb{R}^d$ et $y_i \in \mathbb{R}^n$, on considère le problème

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

où $f_i(\boldsymbol{w}) = p(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i)$ est de classe $\mathcal{C}^1$ pour tout $i = 1, \dots, n$.

a) Donner la définition d'un minimum global du problème $\text{minimiser}_{t\in\mathbb{R}}\, p(t)$.

b) Montrer que $\text{argmin}_{t\in\mathbb{R}}\, p(t) = \{0\}$.

c) Pourquoi l'ensemble des solutions du problème (1) et l'ensemble des solutions du problème

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

sont-ils identiques ?

d) On peut montrer que la fonction f n'est pas fortement convexe en général.

i) À partir des hypothèses sur p, justifier en revanche que la fonction f est convexe.

ii) D'après la question d)i), comment peut-on alors caractériser les solutions du problème (1) ?

e) Écrire l'itération de descente de gradient appliquée au problème (1) avec une longueur de pas constante pour ce problème.

f) Donner deux autres stratégies de choix de longueurs de pas non constantes.

g) Quelle est la vitesse de convergence de la descente de gradient appliquée au problème (1) ? À quelle quantité cette vitesse de convergence s'applique-t-elle ?

h) Quel est le coût en termes d'accès aux données d'une itération de descente de gradient appliquée au problème (1)?

i) Écrire l'itération de l'algorithme du gradient stochastique avec une longueur de pas constante appliqué au problème (1).

j) 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.

k) Sous les bonnes hypothèses, la méthode du gradient stochastique possède une vitesse de convergence (en espérance) en $\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$ pour la même quantité que celle considérée dans la question c), et avec K représentant le nombre d'itérations.

i) En termes d'itérations, justifier que cette borne n'est pas meilleure que celle de la descente de gradient.

ii) Rappeler la définition d'une époque (*epoch*) vue en cours.

iii) Exprimer les vitesses de convergence du gradient stochastique et de la descente de gradient en termes d'époques. Quelle méthode semble alors la meilleure ?

l) On suppose enfin que les différents exemples du jeu de données sont répartis sur r processeurs, avec $1 < r < n$.

i) Écrire une itération de l'algorithme de gradient stochastique par fournées (*batch*) avec taille de fournée (*batch size*) constante égale à n_b, et longueur de pas constante.

ii) Justifier que l'algorithme de descente de gradient est un cas particulier du gradient stochastique par fournées.

iii) Quel peut être l'intérêt de choisir $n_b = r$?

iv) Donner un autre avantage de choisir $n_b = r$ plutôt que $n_b = 1$.

m) On considère finalement le problème

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimiser}}\, f(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2. \tag{2}$$

i) Justifier que la fonction objectif du problème (2) est λ-fortement convexe.

ii) En utilisant $\nabla\left(\frac{\lambda}{2}\|\cdot\|_2^2\right)(\boldsymbol{w}) = \lambda\boldsymbol{w}$, écrire une itération du gradient stochastique appliquée au problème (2).

iii) Justifier alors que l'on parle parfois de *weight decay* (décroissance des poids) lorsque l'itération de la question m)ii) est utilisée en apprentissage profond.

Exercice 2 : Perte non convexe

Soit un jeu de données $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ où l'on suppose que $\boldsymbol{x}_i \in \mathbb{R}^d$ et $y_i \in (0,1)$ pour tout i. On considère le problème

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimiser}}\, \phi(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^n \left(y_i - \frac{1}{1+\exp(-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w})}\right)^2. \tag{3}$$

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

a) Donner la complexité de la descente de gradient sur un problème non convexe tel que le problème (3). À quel critère cette borne de complexité s'applique-t-elle ?

b) Supposons que $\boldsymbol{w}^* \in \mathbb{R}^d$ vérifie $\nabla\phi(\boldsymbol{w}^*) = \mathbf{0}_d$. Le point $\boldsymbol{w}^*$ est-il un minimum local de ϕ ?

c) Quelle condition impliquant à la fois la dérivée d'ordre 1 et la dérivée d'ordre 2 de ϕ est toujours vérifiée par un minimum local de ϕ ?

d) Réciproquement, un point vérifiant la condition de la question c) est-il un minimum local de ϕ ?

e) Soit $\bar{\boldsymbol{w}} \in \mathbb{R}^d$ tel que $\nabla\phi(\bar{\boldsymbol{w}}) = \mathbf{0}_d$. Si $\bar{\boldsymbol{w}}$ ne vérifie pas la condition de la question c), s'attend-on à ce que la descente de gradient converge vers $\bar{\boldsymbol{w}}$ en pratique ? Justifier votre réponse.

Exercice 3 : Approximation parcimonieuse de matrice

Dans cet exercice, on considère que l'on observe un sous-ensemble $\mathcal{S} \subset \{1, \dots, d_1\} \times \{1, \dots, d_2\}$ de taille n des entrées d'une matrice $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$. On cherche à calculer une approximation parcimonieuse de $\boldsymbol{X}$ qui colle également aux données observées, ce que l'on modélise par le problème suivant :

$$\underset{\boldsymbol{W} \in \mathbb{R}^{d_1 \times d_2}}{\text{minimiser}} \frac{1}{2n} \sum_{(i,j) \in \mathcal{S}} (\boldsymbol{W}_{ij} - \boldsymbol{X}_{ij})^2 + \lambda \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} |\boldsymbol{W}_{ij}|, \tag{4}$$

avec $\lambda > 0$. Le problème (4) peut se reformuler comme un problème d'optimisation sur un vecteur de variables. En effet, si l'on définit un vecteur $\boldsymbol{w} \in \mathbb{R}^d$ comme la concaténation des colonnes de $\boldsymbol{W} \in \mathbb{R}^{d_1 \times d_2}$ (avec $d = d_1 d_2$), le problème (4) peut se ré-écrire comme

$$\underset{\boldsymbol{w} \in \mathbb{R}^d}{\text{minimiser}} f^{\boldsymbol{X}}(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1, \qquad f^{\boldsymbol{X}}(\boldsymbol{w}) := \frac{1}{2n} \sum_{(i,j) \in \mathcal{S}} \left([\boldsymbol{w}]_{i+(j-1)d_1} - \boldsymbol{X}_{ij} \right)^2. \tag{5}$$

où $\|\boldsymbol{w}\|_1 = \sum_{i=1}^{d} |[\boldsymbol{w}]_i|$.

a) Les deux termes de la fonction objectif du problème (4) (ou de celle de (5)) ont des rôles différents dans le problème : lesquels ?

b) Écrire l'itération de l'algorithme du gradient proximal appliqué au problème (5) avec une longueur de pas constante.

c) La fonction $f^{\boldsymbol{X}}$ est de classe $\mathcal{C}_L^{1,1}$ avec $L = 1$. Proposer alors un choix de longueur de pas constante, et expliquer l'intérêt de ce choix.

d) Quel est l'intérêt pratique de la méthode du gradient proximal dans le cas de problèmes tels que (5) avec un terme ℓ_1 ? À quelle méthode le gradient proximal est-il équivalent dans ce cas ?

e) Dans cette question, on souhaite appliquer une méthode de sous-gradient pour résoudre le problème (5) (et ainsi le problème (4)).

   i) Écrire une itération de la méthode du sous-gradient appliquée au problème (5) avec une taille de pas constante.

   ii) Donner une condition d'optimalité vérifiée par la solution de (5).

   iii) Supposons que l'on dispose d'un code pour évaluer la fonction objectif du problème (4). Expliquer comment la différentiation automatique permet de calculer des sous-gradients du problème (5).

   iv) Supposons que le paramètre $\lambda > 0$ soit suffisamment large numériquement pour que la première partie de la fonction objectif puisse être négligée dans l'optimisation. Quelle sera alors la solution du problème (4)?