Exercises : Descente de gradient

Optimisation pour l'apprentissage automatique

M2 IASD Apprentissage, 2024-2025

Dauphine | PSL
UNIVERSITÉ PARIS

Exercice 1: Réseaux linéaires à une couche

(Adapté d'un exercice d'examen 2021-2022.)

Dans cet exercice, on considère un jeu de données à labels scalaires, à savoir $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ où $\boldsymbol{x}_i \in \mathbb{R}^{d_x}$ et $y_i \in \mathbb{R}$ pour tout $i = 1, \dots, n$. On construit une architecture neuronale très basique avec une seule couche linéaire homogène et pas d'activation, afin de prédire la valeur y_i à partir du vecteur $\boldsymbol{x}_i$: le modèle obtenu est ainsi

$$\begin{array}{rccc} h^{lin}(\cdot; \boldsymbol{w}) : & \mathbb{R}^{d_x} & \longrightarrow & \mathbb{R}^{d_y} \\ & \boldsymbol{x} & \longmapsto & \boldsymbol{W}_1 \boldsymbol{x}, \end{array} \tag{1}$$

avec $\boldsymbol{W}_1 \in \mathbb{R}^{1 \times d_x}$. En posant $d = d_x$ et $\boldsymbol{w} = \boldsymbol{W}_1^{\mathrm{T}} \in \mathbb{R}^d$, on formule le problème de déterminer le meilleur modèle comme suit :

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

a) La formulation (2) correspond à un problème bien connu en apprentissage. Quel est ce problème ?

b) La fonction objectif f^{lin} est $\mathcal{C}_L^{1,1}$ (son gradient est L-lipschitzien). Lorsque la valeur de L est connue, comment celle-ci peut-elle être employée dans un algorithme tel que la descente de gradient ?

c) Le problème (2) est convexe avec une fonction objectif de classe $\mathcal{C}^1$.

   i) Que peut-on dire d'un point $\bar{\boldsymbol{w}}$ tel que $\nabla f^{lin}(\bar{\boldsymbol{w}}) = \mathbf{0}_{\mathbb{R}^d}$?

   ii) On peut obtenir une vitesse de convergence pour la descente de gradient sur ce problème. De quoi s'agit-il, et à quelle(s) quantité(s) s'applique la vitesse de convergence dans le contexte de cette question ?

d) On suppose dans cette question que les données sont telles que f^{lin} soit μ-fortement convexe, en plus des propriétés mentionées plus haut.

   i) Soient $\boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^d$ deux vecteurs tels que $\nabla f^{lin}(\boldsymbol{w}) = \nabla f^{lin}(\boldsymbol{v}) = \mathbf{0}_{\mathbb{R}^d}$. Que peut-on dire de $\boldsymbol{v}$ et $\boldsymbol{w}$?

   ii) La complexité de la descente de gradient dans ce cas est-elle meilleure que celle de la question c)ii) ?

Exercice 2: Réseaux linéaires à deux couches

(Adapté d'un exercice d'examen de 2021-2022.)

Soit un jeu de données $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ avec $\boldsymbol{x}_i \in \mathbb{R}^{d_x}$ et $\boldsymbol{y}_i \in \mathbb{R}^{d_y}$. Notre objectif est d'apprendre une fonction de $\mathbb{R}^{d_x}$ dans $\mathbb{R}^{d_y}$ qui renvoie $\boldsymbol{y}_i$ lorsque $\boldsymbol{x}_i$ est passé en entrée. Le modèle que l'on choisit ici est celui d'un réseau de neurones à deux couches linéaires :

$$h(\cdot; \boldsymbol{w}) : \begin{array}{ccc} \mathbb{R}^{d_x} & \longrightarrow & \mathbb{R}^{d_y} \\ \boldsymbol{x} & \longmapsto & \boldsymbol{W}_2(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2, \end{array} \tag{3}$$

où $\boldsymbol{W}_1 \in \mathbb{R}^{d_x \times m}$, $\boldsymbol{b}_1 \in \mathbb{R}^m$, $\boldsymbol{W}_2 \in \mathbb{R}^{m \times d_y}$ et $\boldsymbol{b}_2 \in \mathbb{R}^{d_y}$. On considère que le modèle h est paramétré par $\boldsymbol{w} \in \mathbb{R}^d$, où $d = d_x m + m + m d_y + d_y$ et $\boldsymbol{w}$ représente tous les coefficients de $\boldsymbol{W}_1, \boldsymbol{b}_1, \boldsymbol{W}_2, \boldsymbol{b}_2$ mis sous forme vectorielle. Notre but est de déterminer une valeur de $\boldsymbol{w}$ telle que $h(\boldsymbol{x}_i; \boldsymbol{w}) \approx \boldsymbol{y}_i$ pour tout i, ce que l'on quantifie au moyen d'une fonction de perte aux moindres carrés.

Au final, on obtient le problème suivant

$$\underset{\boldsymbol{w} \in \mathbb{R}^d}{\text{minimiser}}\, f(\boldsymbol{w}) := \frac{1}{2n} \sum_{i=1}^{n} \|h(\boldsymbol{x}_i; \boldsymbol{w}) - y_i\|^2. \tag{4}$$

a) Donner un minorant de la fonction objectif du problème (4).

b) En général, le problème (4) est non convexe. Que cela implique-t-il concernant l'ensemble de ses minima locaux et celui de ses minima globaux?

c) La fonction f est de classe $\mathcal{C}^1$.

   i) Supposons que $\boldsymbol{w}^*$ soit une solution de (4). Que peut-on dire de la dérivée de f en $\boldsymbol{w}^*$?

   ii) Écrire l'itération de la descente de gradient appliquée au problème (4) avec une longueur de pas non spécifique.

   iii) Donner deux stratégies possibles pour choisir cette longueur de pas.

   iv) Ce problème est non convexe : sur quelle quantité peut-on alors donner des garanties de complexité dans ce cadre ? Ces garanties sont-elles meilleures que pour un problème convexe ?

Exercice 3: Complétion de matrice

(Adapté d'un exercice d'examen de 2022-2023.)

Soit une matrice de données $\boldsymbol{X} \in \mathbb{R}^{d\times d}$ dont on ne connaît qu'un ensemble d'entrées $\mathcal{S} \subset \{1,\dots,d\}^2$ de taille $n \leq d^2$. On se donne alors le problème

$$\underset{\boldsymbol{W}\in\mathbb{R}^{d\times d}}{\text{minimiser}}\ f(\boldsymbol{W}) := \frac{1}{2n}\sum_{(i,j)\in\mathcal{S}} ([\boldsymbol{W}]_{ij} - [\boldsymbol{X}]_{ij})^2. \tag{5}$$

a) Si $n = d^2$, justifier que $\boldsymbol{W}^* = \boldsymbol{X}$ est l'unique solution du problème.

b) Le problème ci-dessus est convexe en les coefficients de $\boldsymbol{W}$. En notant $\boldsymbol{w} \in \mathbb{R}^{d^2}$ le vecteur colonne formé en mettant bout à bout les colonnes de $\boldsymbol{W}$ dans l'ordre, le problème se reformule comme suit :

$$\underset{\boldsymbol{w}\in\mathbb{R}^{d^2}}{\text{minimiser}}\ \hat{f}(\boldsymbol{w}) := \frac{1}{2n}\sum_{(i,j)\in\mathcal{S}} ([\boldsymbol{w}]_{i+(j-1)d} - [\boldsymbol{X}]_{ij})^2. \tag{6}$$

La fonction $\hat{f}$ est convexe et de classe $\mathcal{C}^1$. Quelle garantie de complexité peut-on fournir sur l'algorithme de descente de gradient lorsqu'il est appliqué au problème (6) ? Sur quel critère porte cette garantie ?

c) On suppose maintenant que la matrice de données $\boldsymbol{X}$ est symétrique semi-définie positive et de rang $1 \ll d$. Dans ce cas, au lieu de chercher une matrice $\boldsymbol{W}$ arbitraire, on peut chercher à calculer une matrice de rang 1 par construction, que l'on note $\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$ avec $\boldsymbol{u} \in \mathbb{R}^d$. Le problème (5) est alors remplacé par

$$\underset{\boldsymbol{u}\in\mathbb{R}^{d}}{\text{minimiser}}\ \tilde{f}(\boldsymbol{u}) := \frac{1}{2n}\sum_{(i,j)\in\mathcal{S}} ([\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}]_{ij} - [\boldsymbol{X}]_{ij})^2. \tag{7}$$

La fonction objectif du problème (7) est de classe $\mathcal{C}^2$ et est non convexe.

i) Donner la condition nécessaire d'optimalité à l'ordre un pour le problème (7).

ii) Quelle est la complexité de la descente de gradient sur un tel problème ? À quel critère ce résultat s'applique-t-il ?

iii) Donner la condition nécessaire d'optimalité à l'ordre deux pour le problème (7).

iv) Sous certaines hypothèses sur $\boldsymbol{X}$ et $\mathcal{S}$, on peut montrer que tout vecteur vérifiant la condition nécessaire d'optimalité à l'ordre deux est un minimum global. Dans ce cas, comment garantir que l'algorithme de descente de gradient converge vers un tel vecteur, et non vers un point selle ?

Correction

Correction de l'exercice 1: Réseau à une couche

a) La formulation (2) correspond à un problème aux moindres carrés linéaires.

b) Si l'on connaît la constante de Lipschitz L associée au gradient, alors la valeur constante $\alpha = \frac{1}{L}$ peut être utilisée comme longueur de pas. *NB: Toute valeur positive strictement inférieure à* $\frac{2}{L}$ *conduit à la convergence théorique de l'algorithme.*

c)

i) Comme le problème est convexe, tout point $\bar{\boldsymbol{w}}$ tel que $\nabla f^{lin}(\bar{\boldsymbol{w}}) = \mathbf{0}_{\mathbb{R}^d}$ est un minimum global.

ii) En règle générale, une vitesse de convergence correspond à la valeur d'un critère garantie pour un budget donné. Dans le cadre de cette question, on peut établir des vitesses de convergence pour un budget d'itérations sur la quantité $f(\boldsymbol{w}_k) - \min_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w})$. Il est également possible d'obtenir une vitesse de convergence sur la quantité $\min_{0\leq k\leq K-1} \|\nabla f(\boldsymbol{w})\|$ dans ce cas.

d)

i) Comme la fonction est fortement convexe et de classe $\mathcal{C}^1$, elle possède un unique minimum global, qui correspond à l'unique point $\boldsymbol{w}$ tel que $\nabla f^{lin}(\boldsymbol{w}) = \mathbf{0}_{\mathbb{R}^d}$. Par conséquent, si $\boldsymbol{w}$ et $\boldsymbol{v}$ sont tels que $\nabla f^{lin}(\boldsymbol{w}) = \nabla f^{lin}(\boldsymbol{v}) = \mathbf{0}_{\mathbb{R}^d}$, alors nécéssairement $\boldsymbol{v} = \boldsymbol{w}$.

ii) Lorsque le problème est fortement convexe, la complexité de la descente de gradient est en $\mathcal{O}(\ln(\epsilon^{-1}))$, ce qui est meilleur que $\mathcal{O}(\epsilon^{-1})$, au sens où la borne dans le cas fortement convexe croît moins lentement que celle dans le cas convexe lorsque ϵ diminue. *NB : Ce n'est pas demandé dans la question, mais cette borne s'applique aussi à la quantité* $f(\boldsymbol{w}_k) - \min_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w})$ *ainsi qu'à* $\|\boldsymbol{w}_k - \boldsymbol{w}^*\|$ *où* $\boldsymbol{w}^*$ *est l'unique minimum du problème.*

Correction de l'exercice 2: Réseaux linéaires à deux couches

a) La fonction objectif étant toujours positive ou nulle, toute valeur négative ou nulle en est un minorant.

b) Comme la fonction est non convexe, les minima locaux de la fonction ne sont pas nécessairement globaux (mais peuvent l'être).

c) Si $\boldsymbol{w}^*$ est une solution du problème (4), on sait alors que son gradient est nul, alors $\nabla f(\boldsymbol{w}^*) = \mathbf{0}$.

d) Si la longueur de pas est notée $\alpha_k > 0$, la $k$ième itération de la descente de gradient s'écrit

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

NB : Ne pas oublier la positivité de la valeur α_k *!*

e) On peut choisir une longueur de pas constante, une longueur de pas décroissante (par exemple en $\frac{1}{k+1}$), ou encore une longueur de pas adaptative calculée via une recherche linéaire.

f) Pour un problème non convexe comme (4), on peut obtenir une borne sur le nombre d'itérations nécessaires pour calculer $\boldsymbol{w}_k$ tel que $\|\nabla f(\boldsymbol{w}_k)\| < \epsilon$ (avec $\epsilon > 0$). Cette borne (en $\mathcal{O}(\epsilon^{-2})$, ce qui n'est pas explicitement demandé dans la question) est moins bonne que celle du cas convexe (en $\mathcal{O}(\epsilon^{-1})$, ce qui n'est pas explicitement demandé dans la question).

Correction de l'exercice 3: Complétion de matrice

a) Les valeurs de la fonction $f(\boldsymbol{W})$ sont toujours positives ou nulles. De plus, lorsque $n = d^2$, on a

$$f(\boldsymbol{W}) = 0 \quad \Leftrightarrow \quad ([\boldsymbol{W}]_{ij} - [\boldsymbol{X}]_{ij})^2 = 0 \ \forall (i,j) \in \{1,\dots,d\}^2 \quad \Leftrightarrow \quad \boldsymbol{W} = \boldsymbol{X}.$$

Par conséquent, $f(\boldsymbol{X}) \leq f(\boldsymbol{W})$ pour tout $\boldsymbol{W} \in \mathbb{R}^{d^2 \times d^2}$ et $f(\boldsymbol{X}) < f(\boldsymbol{W})$ si $\boldsymbol{X} \neq \boldsymbol{W}$, ce qui prouve que le problème possède un unique minimum global donné par $\boldsymbol{W}^* = \boldsymbol{X}$.

b) Comme le problème est convexe, on sait que la complexité de la descente de gradient est en $\mathcal{O}(\epsilon^{-1})$ pour une tolérance $\epsilon > 0$. Cette borne de complexité s'applique à la quantité $f(\boldsymbol{w}_k) - \min_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w})$, c'est-à-dire à l'écart à la valeur optimale.

c) (Cas non convexe)

i) Si $\bar{\boldsymbol{u}} \in \mathbb{R}^d$ est un minimum local du problème (7), alors $\nabla \tilde{f}(\bar{\boldsymbol{u}}) = \boldsymbol{0}$.

ii) Pour une telle fonction non convexe, étant donné $\epsilon > 0$, on sait que la méthode calcule un itéré tel que $\|\nabla f(\boldsymbol{w}_k)\| < \epsilon$ en au plus $\mathcal{O}(\epsilon^{-2})$ itérations. On dit alors que la complexité de la descente de gradient est en ϵ^{-2}.

iii) Si $\bar{\boldsymbol{u}} \in \mathbb{R}^d$ est un minimum local du problème (7), alors $\nabla \tilde{f}(\bar{\boldsymbol{u}}) = \boldsymbol{0}$ et $\nabla^2 \tilde{f}(\bar{\boldsymbol{u}}) \succeq \boldsymbol{0}$.

iv) Même dans ce cas, il n'est pas certain que la descente de gradient converge vers un minimum global. En effet, si on initialise la descente de gradient en un maximum local ou un point selle en lequel le gradient est nul, l'algorithme ne pourra jamais bouger de ce point, et n'atteindra donc jamais un minimum global. En pratique, si on choisit le point initial au hasard, on sait cependant que l'on convergera presque sûrement vers un point stationnaire à l'ordre deux, c'est-à-dire un minimum global.