

## TD 02 : Autour de la descente de gradient

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

11 octobre 2024



### Exercice 2.1: Réseaux linéaires à une couche

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

Dans cet exercice, on considère un jeu de données à labels scalaires, à savoir  $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$  où  $\mathbf{x}_i \in \mathbb{R}^{d_x}$  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  $\mathbf{x}_i$  : le modèle obtenu est ainsi

$$\begin{aligned} h^{lin}(\cdot; \mathbf{w}) : \mathbb{R}^{d_x} &\longrightarrow \mathbb{R}^{d_y} \\ \mathbf{x} &\longmapsto \mathbf{W}_1 \mathbf{x}, \end{aligned} \quad (1)$$

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

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

- La formulation (2) correspond à un problème bien connu en apprentissage. Quel est ce problème ?
- 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 ?
- Le problème (2) est convexe avec une fonction objectif de classe  $\mathcal{C}^1$ .
  - Que peut-on dire d'un point  $\bar{\mathbf{w}}$  tel que  $\nabla f^{lin}(\bar{\mathbf{w}}) = \mathbf{0}_{\mathbb{R}^d}$  ?
  - Quelle est la complexité de la descente de gradient sur ce problème ?
- On suppose dans cette question que les données sont telles que  $f^{lin}$  soit  $\mu$ -strongly convexe, en plus des propriétés mentionnées plus haut.
  - Soient  $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d$  deux vecteurs tels que  $\nabla f^{lin}(\mathbf{w}) = \nabla f^{lin}(\mathbf{v}) = \mathbf{0}_{\mathbb{R}^d}$ . Que peut-on dire de  $\mathbf{v}$  et  $\mathbf{w}$  ?
  - La complexité de la descente de gradient dans ce cas est-elle meilleure que celle de la question c)ii) ?

## Exercice 2.2: Réseaux linéaires à deux couches

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

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

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

où  $\mathbf{W}_1 \in \mathbb{R}^{d_x \times m}$ ,  $\mathbf{b}_1 \in \mathbb{R}^m$ ,  $\mathbf{W}_2 \in \mathbb{R}^{m \times d_y}$  et  $\mathbf{b}_2 \in \mathbb{R}^{d_y}$ . On considère que le modèle  $h$  est paramétré par  $\mathbf{w} \in \mathbb{R}^d$ , où  $d = d_x m + m + m d_y + d_y$  et  $\mathbf{w}$  représente tous les coefficients de  $\mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2$  mis sous forme vectorielle. Notre but est de déterminer une valeur de  $\mathbf{w}$  telle que  $h(\mathbf{x}_i; \mathbf{w}) \approx \mathbf{y}_i$  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

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

- 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 continûment différentiable, ou de classe  $\mathcal{C}^1$ .
  - i) Supposons que  $\mathbf{w}^*$  soit une solution de (4). Que peut-on dire de la dérivée de  $f$  en  $\mathbf{w}^*$  ?
  - 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 : quelle est donc la complexité de la descente de gradient dans ce cas ? À quelle critère cette complexité est-elle associée ?

### Exercice 2.3: Complétion de matrice

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

Soit une matrice de données  $\mathbf{X} \in \mathbb{R}^{d \times d}$  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

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

- a) Si  $n = d^2$ , justifier que  $\mathbf{W}^* = \mathbf{X}$  est l'unique solution du problème.
- b) Le problème ci-dessus est convexe en les coefficients de  $\mathbf{W}$ . En notant  $\mathbf{w} \in \mathbb{R}^{d^2}$  le vecteur colonne formé en mettant bout à bout les colonnes de  $\mathbf{W}$  dans l'ordre, le problème se reformule comme suit :

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

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 quelle critère porte cette garantie ?

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

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

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).

## Correction

### Correction de l'exercice 2.1: 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  $\frac{1}{L}$  conduit à la convergence théorique de l'algorithme.*
- c)
- i) Comme le problème est convexe, tout point  $\bar{\mathbf{w}}$  tel que  $\nabla f^{lin}(\bar{\mathbf{w}}) = \mathbf{0}_{\mathbb{R}^d}$  est un minimum global.
  - ii) Pour un tel problème convexe, étant donnée une précision  $\epsilon > 0$ , l'algorithme a une complexité en  $\mathcal{O}(\epsilon^{-1})$ . *NB : Ce n'est pas explicitement demandé dans la question, mais cette borne signifie que l'algorithme calcule un itéré tel que  $f(\mathbf{w}_k) - \min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}) < \epsilon$  en au plus  $\mathcal{O}(\epsilon^{-1})$  itérations.*
- 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  $\mathbf{w}$  tel que  $\nabla f^{lin}(\mathbf{w}) = \mathbf{0}_{\mathbb{R}^d}$ . Par conséquent, si  $\mathbf{w}$  et  $\mathbf{v}$  sont tels que  $\nabla f^{lin}(\mathbf{w}) = \nabla f^{lin}(\mathbf{v}) = \mathbf{0}_{\mathbb{R}^d}$ , alors nécessairement  $\mathbf{v} = \mathbf{w}$ .
  - 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  $\epsilon$  diminue. *NB : Ce n'est pas demandé dans la question, mais cette borne s'applique à la quantité  $\|\mathbf{w}_k - \mathbf{w}^*\|$  où  $\mathbf{w}^*$  est l'unique minimum du problème.*

### Correction de l'exercice 2.2: 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  $\mathbf{w}^*$  est une solution du problème (4), on sait alors que son gradient est nul, alors  $\nabla f(\mathbf{w}^*) = \mathbf{0}$ .
- d) Si la longueur de pas est notée  $\alpha_k > 0$ , la  $k$ ième itération de la descente de gradient s'écrit

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

*NB : Ne pas oublier la positivité de la valeur  $\alpha_k$  !*

- 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 montrer (notamment avec le choix de longueur de pas  $\alpha_k = \frac{1}{L}$ ) que l'algorithme calcule  $\mathbf{w}_k$  tel que  $\|\nabla f(\mathbf{w}_k)\| < \epsilon$  (avec  $\epsilon > 0$ ) en au plus  $\mathcal{O}(\epsilon^{-2})$  itérations.

### Correction de l'exercice 2.3: Complétion de matrice

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

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

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

- 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  $f(\mathbf{w}_k) - \min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w})$ , c'est-à-dire à l'écart à la valeur optimale.

- c) (Cas non convexe)

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

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

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