Révisions et annales

Optimisation pour l'apprentissage automatique

M2 IASD Apprentissage, 2024-2025

Dauphine | PSL
UNIVERSITÉ PARIS

Note : Ce TD est basé sur l'examen du cours de l'année universitaire 2023-2024. Le dernier exercice a été remplacé, et des modifications ont pu être apportées aux exercices restants afin de prendre en compte les sujets enseignés en 2024-2025.

Exercice 1 : Régression linéaire standard

Dans cet exercice, on considère un jeu de données de la forme $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ où $\boldsymbol{x}_i \in \mathbb{R}^d$ et $y_i \in \mathbb{R}$ pour $i = 1, \dots, n$. On pose $\boldsymbol{y} = [y_i] \in \mathbb{R}^n$ et $\boldsymbol{X} = [\boldsymbol{x}_i^{\mathrm{T}}] \in \mathbb{R}^{n \times d}$. On cherche un modèle linéaire qui associe les $\boldsymbol{x}_i$ aux y_i, ce que l'on formule comme le problème d'optimisation suivant :

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

a) La valeur optimale du problème (1) est-elle nécessairement 0?

b) La fonction objectif du problème (1) est convexe et de classe $\mathcal{C}^1$. Donner une caractérisation des solutions du problème (1) basée sur le gradient de f^{lin}.

c) Écrire l'itération de descente de gradient appliquée au problème (1) en utilisant une suite de longueurs de pas $\{\alpha_k\}$. Donner deux choix possibles pour cette suite.

d) Donner la vitesse de convergence de la descente de gradient sur un tel problème convexe. À quelle quantité cette vitesse s'applique-t-elle ?

e) La méthode du gradient accéléré de Nesterov possède une meilleure vitesse de convergence que celle de la descente de gradient pour un problème convexe. Que vaut cette vitesse ?

f) On suppose maintenant que les données sont telles que la fonction objectif f^{lin} du problème (1) est μ-fortement convexe et de classe $\mathcal{C}_L^{1,1}$ (son gradient est donc L-lipschitzien).

   i) Que peut-on dire des minima locaux d'une fonction fortement convexe ?

   ii) Quelle est la vitesse de convergence de la descente de gradient sur un tel problème ?

   iii) Quelle est la vitesse de convergence de l'algorithme du gradient accéléré sur un tel problème ?

Exercice 2 : Gradient stochastique et régression linéaire

Cet exercice reprend le cadre de l'exercice 1, en présentant le problème sous la forme suivante :

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

Chaque fonction f_i est de classe $\mathcal{C}^1$ et convexe.

a) Écrire l'itération du gradient stochastique appliqué au problème (2) en utilisant une longueur de pas constante.

b) On suppose que notre unité de coût correspond à un accès à un vecteur $\boldsymbol{x}_i$. Selon cette unité, quel est le coût
   i) d'une itération de descente de gradient ?
   ii) d'une itération de gradient stochastique ?

c) Rappeler la définition d'une époque (*epoch*). Si on exécute la descente de gradient et le gradient stochastique pendant le même nombre d'époques, que s'attend-on à observer en pratique ?

d) Écrire l'itération d'une méthode de gradient stochastique par fournées (*batch*) avec une taille de lot fixée à $n_b \in \{1, \ldots, n\}$ et une longueur de pas constante.

e) Justifier que la descente de gradient et le gradient stochastique sont des cas particuliers des méthodes par fournées.

f) Supposons que $n \gg 1$ et que l'on compare le gradient stochastique avec une variante par fournées où $n_b = \frac{n}{5}$ sur 10 exécutions. On observe que les 10 exécutions de la méthode par fournées sont très similaires, tandis que le comportement du gradient stochastique est très variable. Fournir une explication pour cette observation.

g) On considère enfin une itération de la méthode Adam. Expliquer en quoi cette itération diffère de celle du gradient stochastique de base.

Exercice 3 : Régression linéaire robuste

Dans cet exercice, on considère toujours un jeu de données $\boldsymbol{X} = [\boldsymbol{x}_i^{\mathrm{T}}] \in \mathbb{R}^{n\times d}$ et $\boldsymbol{y} = [y_i] \in \mathbb{R}^n$. Il est possible que ce jeu de données contienne des points aberrants (*outliers*, ou points qui ne reflètent pas le lien véritable entre les données). Ces points posent plusieurs difficultés en régression linéaire classique, en partie dues à l'utilisation d'une fonction de perte aux moindres carrés, comme dans les exercices 1 et 2.

On considère donc le problème de régression linéaire robuste suivant :

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

avec $\phi(t) = \frac{t^2}{1+t^2}$. La fonction ϕ s'appelle une fonction bipoids lissée. La fonction objectif f^{bw} du problème (3) est deux fois dérivable (de classe $\mathcal{C}^2$) et non convexe.

a) Supposons que $\boldsymbol{w}^* \in \mathbb{R}^d$ vérifie $\nabla f^{bw}(\boldsymbol{w}^*) = 0$. Le point $\boldsymbol{w}^*$ est-il un minimum local de f^{bw} ?

b) Le problème (3) possède généralement de nombreux points selles dits stricts, en lesquels le gradient est nul mais qui ne sont pas des minima locaux. S'attend-on à ce que la descente de gradient converge vers ces points en pratique ?

c) Donner la complexité de la descente de gradient sur un problème non convexe tel que (3). Est-ce mieux que la complexité de la même méthode sur un problème convexe ?

d) On considère maintenant une version régularisée du problème (3), donnée par

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}}\ f^{bw}(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_1, \tag{4}$$

avec $\lambda > 0$.

i) Quel est le rôle du terme de régularisation $\|\boldsymbol{w}\|_1$?

ii) Écrire l'itération du gradient proximal appliqué au problème (4) avec une longueur de pas constante.

iii) Cette instance de l'algorithme du gradient proximal correspond à ISTA (*Iterative Soft-Thresholding Algorithm*). Pourquoi cette méthode est-elle plus simple à implémenter qu'une méthode de gradient proximal générique ?

Exercice 4 : Autre régression linéaire robuste

Dans cet exercice, on considère un problème de régression linéaire robuste de la forme

$$\text{minimiser}_{\boldsymbol{w}\in\mathbb{R}^d}\ f(\boldsymbol{w}) := \frac{1}{n}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_1 = \frac{1}{n}\sum_{i=1}^{n}\left|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-y_i\right|, \tag{5}$$

où $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{x}_1^{\mathrm{T}}\\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}}\end{bmatrix} \in \mathbb{R}^{n\times d}$ et $\boldsymbol{y} = \begin{bmatrix}y_1\\ \vdots \\ y_n\end{bmatrix} \in \mathbb{R}^n$.

a) Justifier que la fonction objectif du problème (5) est convexe.

b) Soient $\boldsymbol{w}^*$ et $\bar{\boldsymbol{w}}$ deux solutions du problème (5). Justifier que le vecteur $\frac{\boldsymbol{w}^*+\bar{\boldsymbol{w}}}{2}$ est également une solution du problème.

c) Soit $i \in \{1,\dots,n\}$ et $f_i(\boldsymbol{w}) : \boldsymbol{w} \mapsto |\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i|$.

   i) Lorsque $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i \neq 0$, la fonction f_i est dérivable en $\boldsymbol{w}$. Un sous-gradient de f_i en $\boldsymbol{w}$ est alors donné par $\boldsymbol{x}_i$ si $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i > 0$, et par $-\boldsymbol{x}_i$ si $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i < 0$. Combien d'éléments possède alors le sous-différentiel de f_i en $\boldsymbol{w}$?

   ii) On suppose maintenant que $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i = 0$. Dans ce cas, on peut montrer que

$$\partial f_i(\boldsymbol{w}) = \{\gamma\,\boldsymbol{x}_i \mid \gamma \in [-1,1]\}\,.$$

   Justifier qu'un tel point $\boldsymbol{w}$ est un minimum de f_i.

d) Pour tout réel $t \in \mathbb{R}$, on pose $\mathrm{sgn}(t) = 1$ si $t > 0$, $\mathrm{sgn}(t) = 0$ si $t = 0$ et $\mathrm{sgn}(t) = -1$ si $t < 0$. On montre alors que le vecteur

$$\boldsymbol{g}(\boldsymbol{w}) = \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\mathrm{sign}(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}), \quad \text{avec} \quad \mathrm{sign}(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}) = \begin{bmatrix}\mathrm{sgn}(\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{w}-y_i)\\ \vdots \\ \mathrm{sgn}(\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{w}-y_n)\end{bmatrix} \tag{6}$$

   est un sous-gradient de f en $\boldsymbol{w}$.
   Écrire l'itération de la méthode de sous-gradient avec une longueur de pas constante et le choix de $g(\boldsymbol{w})$ comme sous-gradient.

e) D'après la question c), quel est l'intérêt de choisir $g(\boldsymbol{w})$ comme sous-gradient du point de vue de l'optimisation ?

Solutions des exercices

Solution de l'exercice 1 : Régression linéaire standard

a) 0 est un minorant de la fonction objectif, mais il n'en est pas nécessairement le minimum (il faut pour cela qu'il existe $\boldsymbol{w}^* \in \mathbb{R}^d$ tel que $\boldsymbol{X}\boldsymbol{w}^* = \boldsymbol{y}$).

b) Pour une telle fonction convexe et $\mathcal{C}^1$, on sait que $\boldsymbol{w}^* \in \mathbb{R}^d$ est une solution du problème si et seulement si $\nabla f^{lin}(\boldsymbol{w}^*) = \boldsymbol{0}$.

c) L'itération de descente de gradient s'écrit

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

avec $\alpha_k > 0$. Cette suite de longueurs de pas peut être choisie constante, décroissante ou de manière adaptative en fonction de l'itération.

d) Sur ce problème, et sous les bonnes hypothèses sur α_k, on montre qu'après $K \geq 1$ itérations de la descente de gradient, on a $f^{lin}(\boldsymbol{w}_K) - \min_{\boldsymbol{w}\in\mathbb{R}^d} f^{lin}(\boldsymbol{w}) \leq \mathcal{O}\left(\frac{1}{K}\right)$.

e) Sur la même classe de problèmes, l'algorithme du gradient accéléré possède une vitesse en $\mathcal{O}\left(\frac{1}{K^2}\right)$.

f) i) Une fonction fortement convexe possède un unique minimum global (tout minimum local est global par convexité, et la convexité forte garantit l'unicité).

ii) Pour une fonction μ-fortement convexe et $\mathcal{C}_L^{1,1}$, les valeurs de fonction convergent vers la valeur optimale en $\mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^K\right)$.

iii) Dans le même cadre que la question précédente, la vitesse de convergence du gradient accéléré est $\mathcal{O}\left(\left(1 - \sqrt{\frac{\mu}{L}}\right)^K\right)$.

Solution de l'exercice 2 : Gradient stochastique et régression linéaire

a) Si $\alpha > 0$ désigne la longueur de pas, l'itération du gradient stochastique s'écrit

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla f_{i_k}(\boldsymbol{w}_k),$$

où i_k est un indice tiré aléatoirement entre 1 et n.

b) Avec cette unité de coût,

i) Une itération de descente de gradient correspond à n accès à un vecteur $\boldsymbol{x}_i$.

ii) Une itération de gradient stochastic correspond à 1 accès à un vecteur $\boldsymbol{x}_i$.

c) Telle que définie dans le cours, une époque est une unité de coût équivalente à n accès à un point du jeu de données dans un jeu de données à n éléments.
Si on exécute les deux algorithmes pendant le même nombre d'époques, on s'attend à ce que le gradient stochastique converge très rapidement durant les premières époques, puis que l'algorithme atteigne une phase de stagnation/d'oscillation autour d'une valeur. A contrario, on s'attend à ce que l'algorithme de descente de gradient converge lentement mais de manière monotone vers la solution.

d) $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \frac{\alpha}{n_b} \sum_{i \in \mathcal{S}_k} \nabla f_i(\boldsymbol{w}_k)$, où $\alpha > 0$ est la longueur de pas constante et $\mathcal{S}_k$ est un ensemble de taille n_b d'indices tirés au hasard (avec ou sans remise) dans $\{1, \dots, n\}$.

e) En prenant $|\mathcal{S}_k| = 1$, on retrouve l'algorithme du gradient stochastique. En prenant $|\mathcal{S}_k| = n$ et en tirant les indices sans remise, on obtient la descente de gradient.

f) La méthode par fournées calcule des pas de plus petite variance que celle du gradient stochastique de base. Il est donc attendu que la différence d'une exécution à l'autre varie moins pour une grande taille de fournée.

g) La méthode Adam utilise à la fois une normalisation de la longueur de pas différente pour chaque coordonnée (*diagonal scaling*) et un terme de momentum pour définir le pas en combinant un pas de gradient avec le pas de l'itération précédente.

Solution de l'exercice 3 : Régression linéaire robuste

a) Comme la fonction f^{bw} est non convexe, un point en lequel le gradient est nul n'est pas forcément un minimum local. Ce peut être un maximum local ou un point selle.

b) Avec une initialisation aléatoire, on sait que la descente de gradient a une probabilité nulle de converger vers un maximum ou un point selle. On ne s'attend donc pas à ce que l'algorithme converge vers ces points en pratique.

c) Pour un problème non convexe, et sous les bonnes hypothèses, la descente de gradient atteint un point tel que $\|\nabla f(\boldsymbol{w}_k)\| \leq \varepsilon$ en au plus $\mathcal{O}(\varepsilon^{-2})$ itérations. Cette borne de complexité (et la garantie associée) est moins bonne que celle sur un problème convexe, en $\mathcal{O}(\varepsilon^{-1})$.

   i) Le terme $\|\boldsymbol{w}\|_1$ permet de favoriser les solutions parcimonieuses, avec de nombreux coefficients nuls.

   ii) Si $\alpha > 0$ désigne la longueur de pas constante, l'algorithme du gradient proximal s'écrit

$$\boldsymbol{w}_{k+1} \in \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^d} \left\{ f^{bw}(\boldsymbol{w}_k) + \nabla f^{bw}(\boldsymbol{w}_k)^{\mathrm{T}}(\boldsymbol{w} - \boldsymbol{w}_k) + \frac{1}{2\alpha_k} \|\boldsymbol{w} - \boldsymbol{w}_k\|_2^2 + \lambda \|\boldsymbol{w}\|_1 \right\}$$

   iii) L'itération de ISTA est définie de manière explicite, alors que l'itération du gradient proximal est en général définie de manière implicite (par une résolution de sous-problème).

Solution de l'exercice 4 : Autre régression linéaire robuste

a) La norme $\|\cdot\|_1$ est une fonction convexe (car c'est une norme), et la composition d'une fonction affine par une fonction convexe donne une fonction convexe. On en déduit donc que la fonction objectif du problème (5) est convexe.

b) On peut directement utiliser le fait que l'ensemble des minima d'un problème convexe est un ensemble convexe, soit le redémontrer. Si l'on note $f^* = f(\boldsymbol{w}^*) = f(\bar{\boldsymbol{w}})$ la valeur optimale du problème, on a

$$f^* \leq f\left(\tfrac{\boldsymbol{w}^* + \bar{\boldsymbol{w}}}{2}\right) \leq \frac{1}{2} f(\boldsymbol{w}^*) + \frac{1}{2} f(\bar{\boldsymbol{w}}) \leq \frac{1}{2} f^* + \frac{1}{2} f^* = f^*,$$

d'où $f\left(\tfrac{\boldsymbol{w}^* + \bar{\boldsymbol{w}}}{2}\right) = f^*$.

c) i) Comme f_i est dérivable en $\boldsymbol{w}$, le sous-différentiel ne possède qu'un élément, qui est $\nabla f_i(\boldsymbol{w})$.

ii) Le sous-différentiel $\partial f_i(\boldsymbol{w})$ contient le vecteur nul (prendre $\gamma = 0$ dans la définition). Les conditions d'optimalité en optimisation convex non lisse garantissent alors que $\boldsymbol{w}$ est un minimum de f_i.

d) $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha g(\boldsymbol{w}_k)$, avec $\alpha > 0$ la longueur de pas constante.

e) Si il existe un point $\boldsymbol{w}$ tel que $\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y} = \boldsymbol{0}$, alors on aura $\boldsymbol{g}(\boldsymbol{w}) = \boldsymbol{0}$, ce qui garantira que la méthode s'arrête en la solution.