TD 07 : Examen de l'année précédente

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

9 novembre 2023

Dauphine | PSL
UNIVERSITÉ PARIS

Note : Ce TD est basé sur l'examen du cours de l'année universitaire 2022-2023.

Remarques préliminaires

Dans cet examen, nous étudions plusieurs problèmes d'optimisation liés à l'entraînement de réseaux de neurones simples. Chaque exercice suit le même schéma en définissant son problème d'optimisation relativement à un jeu de données, un modèle/une architecture neuronale ainsi qu'un problème d'optimisation/d'entraînement.

- Les dimensions des vecteurs ou matrices seront toujours supposées supérieures ou égales à 1.
- La notation $\|\cdot\|$ désignera la norme euclidienne.
- Pour tout vecteur $\boldsymbol{u} \in \mathbb{R}^d$ avec $d \geq 1$, la i-ème coordonnée de ce vecteur sera notée $[\boldsymbol{u}]_i$.
- Pour tout entier $d \geq 1$, la notation $\mathbf{0}_{\mathbb{R}^d}$ désignera le vecteur nul de $\mathbb{R}^d$.

Exercice 1 : Problèmes en somme finie

On considère un problème de la forme

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

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

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimiser}}\, f(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{w}) \tag{2}$$

   sont-ils identiques ?

iii) Si f est fortement convexe, que peut-on dire des minima du problème (2) (et donc de ceux de (1)) ?

b) Dans la suite, on se concentre sur le problème (2).
   i) Écrire l'itération de descente de gradient avec une longueur de pas (*stepsize*) constante pour ce problème.
   ii) On suppose que f est $\mathcal{C}_L^{1,1}$. Comment peut-on alors choisir la longueur de pas constante de l'algorithme ?
   iii) Donner deux autres stratégies de choix de longueurs de pas non constantes.

c) On considère maintenant que l'on dispose d'un jeu de données contenant n exemples, et que chaque f_i dépend uniquement du i-ème exemple 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 (2)?
   ii) Écrire l'itération de l'algorithme du gradient stochastique avec une longueur de pas constante appliqué au problème (2).
   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 enfin 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 (*batch*) avec taille de fournée (*batch size*) constante égale à n_b, et longueur de pas constante.
   ii) Rappeler la définition d'une époque (*epoch*). Pour la méthode décrite en question d-i), à combien d'itérations cela correspond-il ?
   iii) Quel peut être l'intérêt de choisir $n_b = r$?
   iv) Si $r \approx n$, quel est cependant l'inconvénient de choisir $n_b = r$?

e) Supposons que les gradients ∇f_i sont parcimonieux (*sparse*). Quelle variante du gradient stochastique (par fournées) pourrait-on utiliser pour tirer profit de cette propriété, et pourquoi ?

Exercice 2 : Complétion de matrice

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

$$\underset{\boldsymbol{W}\in\mathbb{R}^{d\times d}}{\text{minimiser}}\ f(\boldsymbol{W}) := \frac{1}{2n}\sum_{(i,j)\in\mathcal{S}} ([\boldsymbol{W}]_{ij} - [\boldsymbol{X}]_{ij})^2. \tag{3}$$

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 :

$$\underset{\boldsymbol{w}\in\mathbb{R}^{d^2}}{\text{minimiser}}\ \hat{f}(\boldsymbol{w}) := \frac{1}{2n}\sum_{(i,j)\in\mathcal{S}} ([\boldsymbol{w}]_{i+(j-1)d} - [\boldsymbol{X}]_{ij})^2. \tag{4}$$

La fonction $\hat{f}$ est convexe et de classe $\mathcal{C}^1$.

i) Quelle garantie de vitesse de convergence peut-on fournir sur l'algorithme de descente de gradient lorsqu'il est appliqué au problème (4) ? Sur quelle quantité porte cette garantie ?

ii) Quelle est la garantie correspondante pour l'algorithme du gradient accéléré dû à Nesterov ? Est-elle meilleure que celle de la descente de gradient ?

iii) Lorsque $n = d^2$, la fonction $\hat{f}$ est une fonction quadratique fortement convexe. À part la méthode de Nesterov, quelle autre approche peut-on utiliser pour avoir une meilleure vitesse de convergence que la descente de gradient ?

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 (3) est alors remplacé par

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

La fonction objectif du problème (5) est de classe $\mathcal{C}^2$ et est non convexe.

i) Donner la condition nécessaire d'optimalité à l'ordre un pour le problème (5).

ii) Quelle est la vitesse de convergence de la descente de gradient sur un tel problème ? À quelle quantité cette vitesse de convergence s'applique-t-elle ?

iii) Donner la condition nécessaire d'optimalité à l'ordre deux pour le problème (5).

iv) Sous certaines hypothèses sur $\boldsymbol{X}$ et $\mathcal{S}$, on peut montrer que tous les points vérifiant la condition nécessaire d'optimalité à l'ordre deux sont des minima globaux. Est-on alors certain que la descente de gradient converge vers un minimum global ?

Exercice 3 : Gradient proximal

On considère à nouveau 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*) :

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimiser}}\ \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 + \frac{\lambda_2}{2}\|\boldsymbol{w}\|^2 + \lambda_1\|\boldsymbol{w}\|_1, \tag{6}$$

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 $\phi : \boldsymbol{w} \mapsto \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2$ est donné par

$$\nabla\phi(\boldsymbol{w}) = \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}).$$

En utilisant cette formule, écrire l'itération du gradient proximal pour le problème (6).

e) Lorsque $\lambda_2 = 0$ et $\lambda_1 > 0$, à quel algorithme le gradient proximal est-il équivalent ?

f) Lorsque $\lambda_1 = 0$, une itération du gradient proximal équivaut à résoudre un problème d'optimisation appartenant à une classe de problèmes étudiée en cours. Quelle est cette classe ?

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 4 : Optimisation décentralisée

Dans cet exercice, on considère la fonction objectif en somme finie de l'exercice 1, et on s'intéresse au problème régularisé suivant

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}} \ \sum_{i=1}^{n} f_i(\boldsymbol{w}) + \frac{\gamma}{2}\|\boldsymbol{L}\boldsymbol{w}\|^2 \tag{7}$$

où $\gamma > 0$ et $\boldsymbol{L} \in \mathbb{R}^{d\times d}$ est la matrice définie par

$$\boldsymbol{L}_{ij} = \begin{cases} 1 & \text{si } j = i+1 \text{ ou } j = i-1 \\ -2 & \text{si } j = i \\ 0 & \text{sinon.} \end{cases}$$

NB : Ce choix de matrice produit des solutions qui convergent vers une fonction lisse lorsque $d \to \infty$.

a) On modifie tout d'abord le problème en introduisant une variable auxiliaire $\boldsymbol{z} \in \mathbb{R}^d$, de sorte à obtenir

$$\begin{array}{ll} \text{minimize}_{\substack{\boldsymbol{w}\in\mathbb{R}^d \\ \boldsymbol{z}\in\mathbb{R}^d}} & \sum_{i=1}^{n} f_i(\boldsymbol{w}) + \frac{\gamma}{2}\|\boldsymbol{z}\|^2 \\ \text{s. c.} & \boldsymbol{L}\boldsymbol{w} - \boldsymbol{z} = \boldsymbol{0}. \end{array} \tag{8}$$

   i) Écrire le lagrangien associé au problème (8). Comment se reformule ce problème en utilisant ce lagrangien?

   ii) Quelle est la différence entre un lagrangien et un lagrangien augmenté ?

   iii) En quoi l'ajout de la variable $\boldsymbol{z}$ permet-il d'appliquer une approche ADMM ? Quel est l'intérêt d'une telle approche ici ?

b) On suppose maintenant que l'on travaille dans un contexte distribué, où les f_i sont réparties entre n agents. Chaque agent dispose donc d'une fonction f_i et de sa copie de $\boldsymbol{w}$, mais tous les agents connaissent le terme de régularisation.

   i) Écrire alors le problème (7) sous la forme d'un problème d'optimisation avec consensus.

   ii) En utilisant la même approche que celle de la question a), introduire une variable commune à tous les agents et reformuler le problème de la question b)i) en un problème pour lequel ADMM peut être directement appliqué.

Solutions

Solution de l'exercice 1

a) (Solutions du problème)

i) Un vecteur $\bar{\boldsymbol{w}} \in \mathbb{R}^d$ est un minimum global du problème (1) si

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

ii) Les deux problèmes sont équivalents et possèdent le même ensemble de solutions. En effet, comme $\frac{1}{n} > 0$, tout minimum global $\bar{\boldsymbol{w}} \in \mathbb{R}^d$ du problème (1) vérifie

$$\begin{aligned} & \textstyle\sum_{i=1}^{n} f_i(\bar{\boldsymbol{w}}) \quad \leq \sum_{i=1}^{n} f(\boldsymbol{w}) \quad \forall \boldsymbol{w} \in \mathbb{R}^d \\ \Leftrightarrow \quad & \tfrac{1}{n} \textstyle\sum_{i=1}^{n} f_i(\bar{\boldsymbol{w}}) \quad \leq \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{w}) \quad \forall \boldsymbol{w} \in \mathbb{R}^d, \end{aligned}$$

ce qui montre que $\bar{\boldsymbol{w}}$ est aussi un minimum global de (2). Puisque l'on a procédé par équivalence, la réciproque est vraie, et on a donc montré que

$$\operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^d} \sum_{i=1}^{n} f_i(\boldsymbol{w}).$$

iii) Si f est fortement convexe sur $\mathbb{R}^d$, alors elle possède un unique minimum global.

b) (Problème (2))

i) L'itération k de la descente de gradient appliquée au problème avec un pas constant $\alpha > 0$ est

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha \nabla f(\boldsymbol{w}_k) = \boldsymbol{w}_k - \frac{\alpha}{n} \sum_{i=1}^{n} \nabla f_i(\boldsymbol{w}_k).$$

ii) Lorsque la fonction f est $\mathcal{C}_L^{1,1}$, on peut choisir $\alpha = \frac{1}{L}$ comme taille de pas constante. *NB : Toute valeur dans* $]0, \frac{2}{L}[$ *est une réponse correcte.*

iii) Au lieu de choisir une taille de pas constante, on peut utiliser une suite décroissante de taille de pas (par exemple $\alpha_k = \frac{\alpha_0}{k+1}$ avec $\alpha_0 > 0$), ou calculer les tailles de pas de manière adaptative à chaque itération, par exemple via une recherche linéaire.

c) (Structure de somme finie)

i) Une itération de descente de gradient appliquée au problème (2) requiert n accès à un point du jeu de données.

ii) L'itération k du gradient stochastique appliqué au problème avec une taille de pas constante $\alpha > 0$ est

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

où i_k est un indice tiré aléatoirement dans $\{1, \dots, n\}$.

iii) Une itération de gradient stochastique accède uniquement à un point du jeu de données. Selon cette métrique, une itération de gradient stochastique est donc n fois moins coûteux qu'une itération de descente de gradient.

d) (Calculs parallèles sur $r \in \{1, \dots, n\}$ processeurs)

i) L'itération k du gradient stochastique par fournées appliquée au problème avec une taille de pas constante $\alpha > 0$ et une taille de fournée n_b est

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \frac{\alpha}{n_b} \sum_{i \in \mathcal{S}_k} \nabla f_i(\boldsymbol{w}_k),$$

où $\mathcal{S}_k$ est un ensemble de n_b indices tirés aléatoirement dans $\{1, \dots, n\}$ (avec ou sans remise).

ii) Une époque (*epoch*) est une unité de coût correspondant à n accès à un point d'un jeu de données à n éléments. Par conséquent, une époque représente le coût de $\frac{n}{n_b}$ itérations d'une méthode de gradient stochastique par fournées de taille n_b.

iii) Choisir $n_b = r$ permet de calculer tous les gradients de la fournée en parallèle, en distribuant les calculs sur les r processeurs disponibles.

iv) Si $r \approx n$, choisir $n_b = r$ correspondra à une méthode avec une grande taille de fournée. Comme discuté en cours, cela signifie que son comportement et son coût seront proches de celui d'une descente de gradient. Chaque itération sera ainsi sensiblement plus coûteuse qu'une itération de gradient stochastique, et par conséquent la convergence sera plus lente en termes d'époques.

e) La variante `Adagrad` est particulièrement adaptées aux problèmes avec gradients parcimonieux car elle adapte progressivement la taille de pas à chacune des coordonnées. *NB : Avec la même justification, la variante `RMSProp` serait acceptée comme réponse valide.*

Solution de l'exercice 2

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) (Formulation convexe)

i) Comme le problème est convexe, on sait qu'après $K \geq 1$ itérations de descente de gradient, l'itéré $\boldsymbol{w}_K$ vérifie

$$\hat{f}(\boldsymbol{w}_K) - \min_{\boldsymbol{w} \in \mathbb{R}^{d^2}} \hat{f}(\boldsymbol{w}) \leq \mathcal{O}\left(\frac{1}{K}\right).$$

On dit alors que la descente de gradient converge en vitesse $\frac{1}{K}$.

ii) La vitesse de convergence pour l'algorithme du gradient accéléré sur un tel problème convexe est $\frac{1}{K^2}$, qui est plus rapide que celle de la descente de gradient (car elle converge plus vite vers 0).

iii) Lorsque $\hat{f}$ est une quadratique fortement convexe, la méthode de la boule lestée (ou *heavy ball*, due à Polyak) possède la même vitesse de convergence que l'algorithme accéléré, qui est meilleure que celle de la descente de gradient (*NB: La valeur de cette vitesse, en* $\left(1-\sqrt{\frac{\mu}{L}}\right)^K$, *n'est pas requise pour répondre à la question*).

c) (Cas non convexe)

i) Si $\bar{\boldsymbol{u}} \in \mathbb{R}^d$ est un minimum local du problème (5), alors $\nabla \tilde{f}(\bar{\boldsymbol{u}}) = \mathbf{0}$.

ii) Pour une telle fonction non convexe, après $K \geq 1$ itérations de descente de gradient, on a

$$\min_{0 \leq k \leq K-1} \|\nabla f(\boldsymbol{w}_k)\| \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right),$$

et on dit alors que la descente de gradient converge en vitesse $\frac{1}{\sqrt{K}}$.

iii) Si $\bar{\boldsymbol{u}} \in \mathbb{R}^d$ est un minimum local du problème (5), alors $\nabla \tilde{f}(\bar{\boldsymbol{u}}) = \mathbf{0}$ et $\nabla^2 \tilde{f}(\bar{\boldsymbol{u}}) \succeq \mathbf{0}$.

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. *NB: En pratique, si on choisit le point initial au hasard, on peut montrer que l'on converge presque sûrement vers un minimum global.*

Solution de l'exercice 3

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.

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).

d) La $k$ième itération de l'algorithme du gradient proximal appliqué au problème (6) s'écrit

$$\boldsymbol{w}_{k+1} \in \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^d} \left\{ \phi(\boldsymbol{w}_k) + \frac{1}{n}(\boldsymbol{X}\boldsymbol{w}_k - \boldsymbol{y})^{\mathrm{T}} \boldsymbol{X}(\boldsymbol{w} - \boldsymbol{w}_k) + \frac{1}{2\alpha_k}\|\boldsymbol{w} - \boldsymbol{w}_k\|^2 + \frac{\lambda_2}{2}\|\boldsymbol{w}\|^2 + \lambda_1 \|\boldsymbol{w}\|_1 \right\}$$

où $\alpha_k > 0$.

e) Lorsque $\lambda_1 = 0$, le sous-problème proximal est un problème aux moindres carrés linéaires.

f) *Il y a plusieurs réponses valides.* On peut utiliser une méthode de sous-gradient pour résoudre le sous-problème, car cet algorithme est applicable même en présence d'une fonction non lisse comme la norme ℓ_1.

On peut sinon 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.

Note: On peut également combiner les deux idées précédentes en considérant la norme ℓ_2 comme le terme de régularisation.

Solution de l'exercice 4

a) (Problème (8))

i) Le lagrangien associé au problème est

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{\lambda}) = \sum_{i=1}^{n} f_i(\boldsymbol{w}) + \frac{\gamma}{2}\|\boldsymbol{z}\|^2 + \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{L}\boldsymbol{w} - \boldsymbol{z}).$$

On sait alors que le problème (8) se reformule en le problème sans contraintes

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}} \max_{\boldsymbol{\lambda}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\lambda});$$

ii) Un lagrangien augmenté possède un terme de régularisation supplémentaire qui pénalise les points ne vérifiant pas les contraintes :

$$\mathcal{L}^a(\boldsymbol{w}, \boldsymbol{\lambda}, \mu) = \mathcal{L}(\boldsymbol{w}, \boldsymbol{\lambda}) + \frac{\mu}{2}\|\boldsymbol{L}\boldsymbol{x} - \boldsymbol{z}\|^2.$$

iii) L'ajout de la variable $\boldsymbol{z}$ conduit à un problème séparable, ce qui permet donc d'envisager l'application de la méthode ADMM. Pour ce problème particulier, l'intérêt est que l'on découple le terme dépendant des données du terme de régularisation, ce qui peut conduire à des problèmes en $\boldsymbol{w}$ et $\boldsymbol{z}$ plus simples à résoudre.

b) (Contexte distribué)

i) Sous les hypothèses de la question, le problème d'optimisation avec consensus s'écrit

$$\begin{aligned} \text{minimize}_{\substack{\boldsymbol{w}\in\mathbb{R}^d \\ \boldsymbol{w}^{(1)}\in\mathbb{R}^d \\ \vdots \\ \boldsymbol{w}^{(n)}\in\mathbb{R}^d}} \quad & \textstyle\sum_{i=1}^{n} f_i(\boldsymbol{w}^{(i)}) + \frac{\gamma}{2}\|\boldsymbol{L}\boldsymbol{w}^{(i)}\|^2 \\ \text{s. c.} \quad & \boldsymbol{w}^{(i)} - \boldsymbol{w} = \boldsymbol{0} \quad \forall (i,j) \in \{1,\dots,n\}. \end{aligned}$$

ii) En introduisant une variable $\boldsymbol{z}$ commune à tous les agents, on obtient le problème

$$\begin{array}{lll} \text{minimize}_{\substack{\boldsymbol{w}^{(1)}\in\mathbb{R}^d \\ \vdots \\ \boldsymbol{w}^{(n)}\in\mathbb{R}^d \\ \boldsymbol{z}\in\mathbb{R}^d}} & \sum_{i=1}^n f_i(\boldsymbol{w}^{(i)}) + \frac{\gamma}{2}\|\boldsymbol{z}\|^2 & \\ \text{s. c.} & \boldsymbol{w}^{(i)} - \boldsymbol{w} = \boldsymbol{0} & \forall (i,j) \in \{1,\dots,n\} \\ & \boldsymbol{L}\boldsymbol{w} - \boldsymbol{z} = \boldsymbol{0}, & \end{array}$$

où l'on voit qu'ADMM peut être appliqué à ce problème en posant $\boldsymbol{u} = \begin{bmatrix} \boldsymbol{w}^{(1)} \\ \vdots \\ \boldsymbol{w}^{(n)} \\ \boldsymbol{w} \end{bmatrix}$ et $\boldsymbol{v} = \boldsymbol{z}$.