Aix*Marseille université

Master 2 informatique

Parcours IAAA, UE MIA

Feuille d'exercices numéro 6

Année 2023-24

Miscellaneous.

Exercice 1 Analyse de l'algorithme de descente de gradient pour une fonction de coût lisse et fortement convexe. (⋆⋆)

Définitions :

— **Fonction L-lisse.** Soit L un nombre réel positif, une fonction $f : \mathbf{R}^n \to \mathbf{R}$ est dite L-lisse si et seulement si elle est de classe C^1 et que son gradient est L-lipschitzien, c'est à dire que pour tous vecteurs x, y de $\mathbf{R}^n$: $\|f(x) - f(y)\|_2 \leq L\|x - y\|_2$.

— **Fonction μ-fortement convexe.** Soit μ un nombre réel strictement positif, une fonction différentiable $f : \mathbf{R}^n \to \mathbf{R}$ est dite μ-fortement convexe si et seulement si pour tous vecteurs x, y de $\mathbf{R}^n$, $f(x) \geq f(y) + \nabla f(y)^T(x - y) + \frac{\mu}{2}\|x - y\|_2^2$.

Soit $r : \mathbf{R}^n \to \mathbf{R}$ une fonction à minimiser. On suppose que r est L-lisse et μ-fortement convexe pour $L \geq 0$ et $\mu > 0$.

On note x^* l'unique minimum de la fonction r et on considère les valeurs successives $x_1, x_2, \ldots, x_t, \ldots$ obtenues par descente du gradient de r avec un pas fixe $\gamma = \frac{1}{L}$ en partant de $x_0 \in \mathbf{R}^n$.

1. Montrez que $1 - x \leq \exp(-x)$ pour tout réel x

   Solution : Etudions la fonction $f : x \mapsto \exp(-x) + x - 1$. Elle est continue et dérivable de dérivée $f' : x \mapsto 1 - \exp(-x)$. $\exp(-x)$ est supérieur à 1 pour x négatif et inférieur à 1 pour x positif. Donc f' est positive et f croissante pour x positif et f' est négative et f décroissante pour x négatif. On en déduit que f atteint un minimum global $f(0) = 1 + 0 - 1 = 0$ en $x = 0$ et donc que $f \geq 0$. C'est à dire que pour tout réel x, $\exp(-x) + x - 1 \geq 0$, et donc $\exp(-x) \geq 1 - x$.

2. Soit $x \in \mathbf{R}^n$. Montrez que la fonction :

$$g_x : \left| \begin{array}{l} \mathbf{R}^n \to \mathbf{R} \\ z \mapsto r(x) + \nabla r(x)^T(z - x) + \frac{\mu}{2}\|z - x\|_2^2 \end{array} \right.$$

   est μ-fortement convexe et donnez une expression explicite pour le point z^* où elle atteint son minimum.

Solution : g_x est différentiable, de gradient :

$$\nabla g_x : z \mapsto \nabla r(x) + \frac{\mu}{2}\nabla_z((z-x)^T(z-x)) = \nabla r(x) + \frac{\mu}{2}(2z - 2x + 0) = \nabla r(x) + \mu(z-x).$$

Soient $(a, b) \in (\mathbf{R}^n)^2$. On a alors :

$$g_x(b) + \nabla g_x(b)^T(a-b) + \frac{\mu}{2}\|a-b\|_2^2 - g_x(a) = \mu\left(\frac{(\|b-x\|_2^2 + \|a-b\|_2^2)}{2} + (b-x)^T(a-b) - \frac{\|a-x\|_2^2}{2}\right)$$
$$= 0.$$

Donc $g_x(a) \geq g_x(b) + \nabla g_x(b)^T(a-b) + \frac{\mu}{2}\|a-b\|_2^2$ et g_x est bien μ-fortement convexe.
g_x possède donc un unique minimum global au point z^* où son gradient s'annule, c'est à dire que $\nabla r(x) + \mu(z^* - x) = 0$, ce qui équivaut à $z^* = x - \frac{\nabla r(x)}{\mu}$.

3. En utilisant le résultat de la question précédente, montrez que pour $x \in \mathbf{R}^n$, $\|\nabla r(x)\|_2^2 \geq 2\mu\,(r(x) - r(x^*))$.

Solution : On a obtenu que pour tout $z \in \mathbf{R}^n$, $g_x(z) \geq g_x(z^*)$.
De plus, comme r est μ-fortement convexe, on a :

$$r(x^*) \geq r(x) + \nabla r(x)^T(x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2 = g_x(x^*).$$

En appliquant l'inégalité précédente en $z = x^*$, on en déduit alors que :

$$r(x^*) \geq g_x(z^*),$$

c'est à dire, en se rappelant que $z^* = x - \frac{\nabla r(x)}{\mu}$, que :

$$r(x^*) \geq r(x) - \frac{\nabla r(x)^T \nabla r(x)}{\mu} + \frac{1}{2\mu}\|\nabla r(x)\|_2^2 = r(x) - \frac{1}{2\mu}\|\nabla r(x)\|_2^2.$$

D'où :

$$\|\nabla r(x)\|_2^2 \geq 2\mu(r(x) - r(x^*)).$$

4. Donnez une interprétation verbale de la propriété que vous avez établi à la question précédente.

Solution : Intuitivement, le résultat de la question précédente garantit une taille minimale pour les gradients de r, qui croît à mesure qu'on s'éloigne du minimum de r. Cela va permettre d'établir que même si l'on se trouve « loin » du minimum de r on va obtenir des gradients suffisamment grands pour garantir une convergence rapide de la descente de gradient à pas fixe.

5. En utilisant les résultats des questions 1 et 3 montrez que pour tout entier naturel t :

$$r(x_t) - r(x^*) \leq \exp\left(-t\mu/L\right)\left(r(x_0) - r(x^*)\right).$$

Solution : Soit t un entier strictement positif. Par définition, $x_t = x_{t-1} - \frac{1}{L}\nabla r(x_{t-1})$. On va utiliser le caractère L-lisse de r pour obtenir une borne inférieure sur la descente de r entre x_t et x_{t-1}.

La première idée est de montrer que le caractère L-lisse tel que nous l'avons défini implique que le gradient de la fonction $h : x \mapsto \frac{L}{2}x^T x - r(x)$ est monotone, ce qui équivaut à montrer que cette h est convexe et à écrire la condition de convexité du premier ordre pour h.

L'inégalité de Cauchy-Schwarz donne pour tout x, y :

$$(\nabla r(x) - \nabla r(y))^T(x - y) \leq \|\nabla r(x) - \nabla r(y)\|_2 \|x - y\|_2$$

et donc pour une fonction L-lisse :

$$(\nabla r(x) - \nabla r(y))^T(x - y) \leq L\|x - y\|_2^2.$$

En terme de h, l'inégalité ci-dessus se ré-écrit plus simplement :

$$(\nabla h(x) - \nabla h(y))^T(x - y) \geq 0,$$

c'est à dire que le gradient de h est monotone.

Solution : (Continuée)

Cela implique que h est convexe. En effet, si on note $m : t \mapsto h(x+t(y-x))$, la monotonie du gradient de h nous permet de montrer que pour tout $t > 0$, $m'(t) \geq m'(0)$ et donc :

$$h(y) = m(1) = m(0) + \int_0^1 m'(t)dt \geq m(0) + m'(0) = h(x) + \nabla h(x)^T(y-x),$$

ce qui est une condition nécessaire et suffisante classique de convexité pour une fonction différentiable.

Ecrite en fonction de r, cette condition nous donne finalement :

$$\frac{L}{2}y^Ty - r(y) \geq \frac{L}{2}x^Tx - r(x) + (Lx - \nabla r(x))^T(y-x),$$

qu'on peut réarranger en :

$$r(y) \leq r(x) + \nabla r(x)^T(y-x) + \frac{L}{2}(y^Ty - x^Tx - 2x^Ty + 2x^Tx) = r(x) + \nabla r(x)^T(y-x) + \frac{L}{2}\|x-y\|_2^2.$$

Enfin, en prenant $y = x_t$ et $x = x_{t-1}$, on obtient :

$$r(x_t) \leq r(x_{t-1}) + \nabla r(x_{t-1})^T\left(-\frac{1}{L}\nabla r(x_{t-1})\right) + \frac{L}{2}\left\|\frac{1}{L}\nabla r(x_{t-1})\right\|_2^2,$$

soit :

$$r(x_t) \leq r(x_{t-1}) - \frac{1}{2L}\|\nabla r(x_{t-1})\|_2^2.$$

On peut maintenant appliquer le résultat de la question 3 pour obtenir :

$$r(x_t) \leq r(x_{t-1}) - \frac{\mu}{L}(r(x_{t-1}) - r(x^*)),$$

d'où, en utilisant le résultat de la question 1 :

$$r(x_t) - r(x^*) \leq \left(1 - \frac{\mu}{L}\right)(r(x_{t-1}) - r(x^*)) \leq \exp(-\mu/L)(r(x_{t-1}) - r(x^*)),$$

Le résultat attendu s'obtient alors par une simple récurrence sur t.

6. Que nous apprend ce résultat sur la méthode de descente de gradient à pas fixe pour une fonction de coût lisse et fortement convexe ?

Solution : Ce résultat montre que la méthode de descente de gradient à pas fixe pour une fonction de coût lisse et fortement convexe converge vers le minimum global de la fonction. De plus ce résultat montre que cette convergence se fait avec une vitesse exponentielle en le nombre d'itérations.

Exercice 2 Descente de gradient stochastique ou non et avec ou sans projection sur les contraintes. ($\star$)

1. Effectuez à la main la première étape de l'algorithme de descente de gradient avec la règle d'Armijo pour la fonction $f : x, y \mapsto 3x^2 + y^4$, en partant du point $(x_0, y_0) = (1, -2)$ et avec $s = 1$, $\sigma = 0.1$ et $\beta = .1$.
2. Effectuez à la main la première étape de l'algorithme de descente de gradient projeté avec la règle d'Armijo pour la fonction $f : x, y \mapsto x^3 + y$, avec les contraintes $x \geq 0$ et $y \geq 0$, en partant du point $(x_0, y_0) = (1, 1)$ et avec $s = 1$, $\sigma = 0.1$ et $\beta = .1$.
3. Effectuez à la main la première étape de l'algorithme de descente de gradient à pas fixe $\gamma = .01$ pour la fonction $f : x \mapsto (x - 0.1)^2 + (x - 0.2)^2 + (x - 0.15)^2 + (x - 0.3)^2$ en partant de $x = .4$.
4. Effectuez à la main la première étape de l'algorithme de descente de gradient stochastique à pas fixe $\gamma = .01$ pour la fonction de la question précédente en partant de $x = .4$.

Exercice 3 Preuve de convergence de la descente de gradient avec règle d'Armijo. ($\star\star\star$)

Dans cet exercice, on va démontrer le théorème suivant :

Theorem 1 *Stationarité des points limites de la descente de gradient avec règle d'Armijo. Soit $f : \mathbf{R}^n \to \mathbf{R}$, de classe C^1. Soit (x_k) une suite de points générée par l'application de la méthode de descente de gradient avec règle d'Armijo à f en partant de $x_0 \in \mathbf{R^n}$. Alors, tout point limite x^* de (x_k) est un point stationnaire de f, i.e. $\nabla f(x^*) = 0$.*

Rappels d'analyse réelle :

— Un point limite, aussi appelé valeur d'adhérence d'une suite (u_k), est un point l tel qu'il existe une sous-suite extraite de (u_k) qui converge vers l, c'est à dire qu'il existe une fonction $\phi : \mathbf{N} \to \mathbf{N}$ strictement croissante telle qu'on ait :

$$\lim_{k \to +\infty} u_{\phi(k)} = l.$$

— Toute suite monotone de nombres réels converge vers un nombre fini ou diverge vers l'infini (vers $+\infty$ pour une suite croissante et vers $-\infty$ pour une suite décroissante).

— Si g est continue et $\lim_{k\to+\infty} u_k = l$, alors $\lim_{k\to+\infty} g(u_k) = g(l)$.

— Théorème des accroissements finis : si a et b sont deux réels avec $a < b$ et $f : [a, b] \to \mathbf{R}$ est une fonction continue sur $[a, b]$ et dérivable sur $]a, b[$, alors il existe un réel $c \in]a, b[$, tel que :

$$\frac{f(b) - f(a)}{b - a} = f'(c).$$

— Théorème de Bolzano-Weierstrass : de toute suite bornée de vecteurs de $\mathbf{R}^n$, on peut extraire une sous-suite convergente.

On va raisonner par l'absurde. Supposons que $\hat{x}$ est un point limite de (x_k) avec $\nabla f(\hat{x}) \neq 0$.

1. Montrer que la suite $(f(x_k))$ converge vers $f(\hat{x})$.

   Solution : Par hypothèse, on peut extraire de (x_k) une suite $(x_{\phi(k)})$ qui converge vers $\hat{x}$. Comme f est continue, on en déduit que la suite $(f(x_{\phi(k)}))$ converge vers $f(\hat{x})$.
   Par ailleurs, la suite $(f(x_k))$ est décroissante par construction, elle est donc soit convergente, soit divergente vers $-\infty$. Si elle était divergente vers $-\infty$, toute suite extraite serait aussi divergente vers $-\infty$. Or on a vu que la suite extraite $(f(x_{\phi(k)}))$ converge vers $f(\hat{x})$. On en déduit que $(f(x_k))$ converge.
   Une suite convergente possède comme unique valeur d'adhérence sa limite. Or on a vu que $f(\hat{x})$ est une valeur d'adhérence de $(f(x_k))$. On en conclut que $(f(x_k))$ converge vers $f(\hat{x})$.

2. En déduire que la suite $(-\alpha_k \nabla f(x_k)^T \nabla f(x_k))$ converge vers 0, où α_k est le pas utilisé dans la descente de gradient avec règle d'Armijo à l'étape k, i.e. $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$.

   Solution : Par construction, on a pour un réel fixé $0 < \sigma < 1$ et pour tout entier positif k

   $$(f(x_k) - f(x_{k+1})) \geq \sigma \alpha_k \nabla f(x_k)^T \nabla f(x_k),$$

   avec $\alpha_k > 0$. De plus, $\nabla f(x_k)^T \nabla f(x_k) = \|\nabla f(x_k)\|_2^2 \geq 0$.
   On a donc pour tout entier positif k

   $$\frac{1}{\sigma}(f(x_k) - f(x_{k+1})) \geq \alpha_k \nabla f(x_k)^T \nabla f(x_k) \geq 0.$$

   $(f(x_k))$ étant une suite convergente, la suite $(f(x_k)) - f(x_{k+1})$ converge vers 0 et on en déduit que la suite $(\alpha_k \nabla f(x_k)^T \nabla f(x_k))$ et donc aussi la suite $(-\alpha_k \nabla f(x_k)^T \nabla f(x_k))$ convergent vers 0 (théorème des gendarmes).

3. Par hypothèse, on peut extraire de (x_k) une suite $(x_{\phi(k)})$ qui converge vers $\hat{x}$. Montrer que

$(\alpha_{\phi(k)})$ converge vers 0.

Solution : Comme $(-\alpha_k \nabla f(x_k)^T \nabla f(x_k))$ converge vers 0, la suite extraite $(-\alpha_{\phi(k)} \nabla f(x_{\phi(k)})^T \nabla f(x_{\phi(k)}))$ converge aussi vers 0.

Comme f est de classe C^1, $h : x \mapsto \nabla f(x)^T \nabla f(x))$ est continue. On en déduit que $\nabla f(x_{\phi(k)})^T \nabla f(x_{\phi(k)})) = h(x_{\phi(k)})$ converge vers $h(\hat{x}) = \nabla f(\hat{x})^T \nabla f(\hat{x}) = \|\nabla f(\hat{x})\|_2^2 \neq 0$, par hypothèse.

En combinant ces deux résultats, on obtient que $(\alpha_{\phi(k)})$ converge vers 0.

4. En déduire qu'il existe un entier k_0 tel que pour tout entier $k \geq k_0$, le pas initial s n'est pas satisfaisant et on le réduit au moins une fois quand on applique la règle d'Armijo à l'étape $\phi(k)$ (en le multipliant par β).

Solution : Considérons le paramètre $s > 0$ utilisé dans le calcul du pas. Par définition de $\lim_{k \to +\infty} (\alpha_{\phi(k)}) = 0$, il existe un entier positif k_0 tel que pour tout $k \geq k_0$, $|\alpha_{\phi(k)} - 0| < s$.

Comme $|\alpha_{\phi(k)} - 0| = \alpha_{\phi(k)}$, cela équivaut à avoir pour tout $k \geq k_0$, $\alpha_{\phi(k)} < s$, ce qui montre que pour tout $k \geq k_0$, le pas initial s n'a pas été satisfaisant à l'étape $\phi(k)$.

5. En déduire que pour tout $k \geq k_0$, on a

$$\frac{f\left(x_{\phi(k)}\right) - f\left(x_{\phi(k))} - \delta_k \frac{\nabla f(x_{\phi(k)})}{\|\nabla f(x_{\phi(k)})\|}\right)}{\delta_k} < \sigma \nabla f\left(x_{\phi(k)}\right)^T \frac{\nabla f(x_{\phi(k)})}{\|\nabla f(x_{\phi(k)})\|},$$

où $\delta_k = \frac{\alpha_{\phi(k)}}{\beta} \|\nabla f\left(x_{\phi(k)}\right)\|$ et σ et β sont les paramètres utilisés pour l'application de la règle d'Armijo.

Solution : Soit $k \geq k_0$. Le résultat de la question précédente nous indique que le test de la règle d'Armijo n'est pas satisfait pour l'entier m tel que $\frac{\alpha_{\phi(k)}}{\beta} = \beta^m s$, c'est à dire que

$$f\left(x_{\phi(k)}\right) - f\left(x_{\phi(k))} - \frac{\alpha_{\phi(k)}}{\beta} \nabla f(x_{\phi(k)})\right) < \sigma \frac{\alpha_{\phi(k)}}{\beta} \nabla f\left(x_{\phi(k)}\right)^T \nabla f(x_{\phi(k)})$$

De plus, $\|\nabla f(x_{\phi(k)})\|_2 \neq 0$, car si on avait $\|\nabla f(x_{\phi(k)})\|_2 = 0$ alors pour tout $q \geq \phi(k)$, on aurait $x_q = x_{\phi(k)}$ et donc on aurait $\hat{x} = x_{\phi(k)}$ et donc $\nabla f(\hat{x}) = 0$, ce qui contredirait notre hypothèse de départ.

On obtient le résultat demandé en combinant ces deux résultats.

6. En déduire que pour tout $k \geq k_0$, il existe $\gamma_k \in]0, \delta_k[$, tel que

$$\nabla f\left(x_{\phi(k)} - \gamma_k \frac{\nabla f(x_{\phi(k)})}{\|\nabla f(x_{\phi(k)})\|}\right)^T \frac{\nabla f(x_{\phi(k)})}{\|\nabla f(x_{\phi(k)})\|} < \sigma \nabla f\left(x_{\phi(k)}\right)^T \frac{\nabla f(x_{\phi(k)})}{\|\nabla f(x_{\phi(k)})\|}.$$

(Indice : utiliser le théorème des accroissements finis.)

Solution : Soit $k \geq k_0$.

Considérons la fonction h_k de $[0, \delta_k]$ vers $\mathbf{R}$ qui a γ associe

$$f\left(x_{\phi(k))} - \gamma \frac{\nabla f(x_{\phi(k)})}{\|\nabla f(x_{\phi(k)})\|}\right).$$

Le raisonnement de la question précédente nous indique que $\delta_k > 0$, et comme de plus h_k est de classe C^1 car f est de classe C^1, on peut appliquer le théorème des accroissements finis à h_k.

On obtient qu'il existe un réel γ_k dans $]0, \delta_k[$ tel que

$$\frac{f\left(x_{\phi(k))} - \delta_k \frac{\nabla f(x_{\phi(k)})}{\|\nabla f(x_{\phi(k)})\|}\right) - f(x_{\phi(k)})}{\delta_k - 0} = h_k^{'}(\gamma_k).$$

En appliquant la règle de dérivation des fonctions composées pour des fonctions de plusieurs variables (chain rule), on obtient

$$h_k^{'} : \gamma \mapsto -\nabla f\left(x_{\phi(k))} - \gamma \frac{\nabla f(x_{\phi(k)})}{\|\nabla f(x_{\phi(k)})\|}\right) \frac{\nabla f(x_{\phi(k)})}{\|\nabla f(x_{\phi(k)})\|}.$$

En remplacant $h_k^{'}(\gamma_k)$ par son expression et en combinant le résultat avec le résultat de la question 5, on obtient le résultat demandé.

7. Montrer que $\lim_{k\to+\infty} \gamma_k = 0$.

Solution : Pour tout $k \geq k_0$, on a $0 < \gamma_k < \delta_k$. Il suffit donc de montrer que $\lim_{k\to+\infty} \delta_k = 0$.

Pour tout $k \geq k_0$, on a $\delta_k = \frac{\alpha_{\phi(k)}}{\beta} \|\nabla f\left(x_{\phi(k)}\right)\|$. Comme f est de classe C^1, $x \mapsto \frac{1}{\beta}\|\nabla f(x)\|$ est continue. Et comme $(x_{\phi(k)})$ converge vers $\hat{x}$, on en déduit que la suite $(\frac{1}{\beta}\|\nabla f\left(x_{\phi(k)}\right)\|)$ converge vers $\frac{1}{\beta}\|\nabla f(\hat{x})\|$.

Or on a vu à la question 3 que $\lim_{k\to+\infty} \alpha_k = 0$. On en déduit que (δ_k) converge et que $\lim_{k\to+\infty} \delta_k = \frac{1}{\beta}\|\nabla f(\hat{x})\| \times 0 = 0$.

8. Montre qu'il existe une sous-suite extraite de $\left(\frac{\nabla f(x_{\phi(k)})}{\|\nabla f(x_{\phi(k)})\|}\right)$ qui converge vers un vecteur $v \in \mathbf{R}^n$. (Indice : utiliser le théorème de Bolzano-Weierstrass.)

Solution : On a déjà justifié que pour tout k, $\|\nabla f\left(x_{\phi(k)}\right)\| \neq 0$, il suffit donc de montrer que $\left(\frac{\nabla f\left(x_{\phi(k)}\right)}{\|\nabla f\left(x_{\phi(k)}\right)\|}\right)$ est une suite bornée pour pouvoir appliquer le théorème et obtenir le résultat souhaité.

C'est à dire qu'il suffit de montrer qu'il existe un réel M tel que, pour tout entier positif k

$$\left\| \frac{\nabla f\left(x_{\phi(k)}\right)}{\|\nabla f\left(x_{\phi(k)}\right)\|} \right\| \leq M.$$

Or, pour tout entier positif k

$$\left\| \frac{\nabla f\left(x_{\phi(k)}\right)}{\|\nabla f\left(x_{\phi(k)}\right)\|} \right\| = \frac{1}{\|\nabla f\left(x_{\phi(k)}\right)\|} \left\|\nabla f\left(x_{\phi(k)}\right)\right\| = 1.$$

On peut donc prendre $M = 1$.

9. Montrer que :

$$\nabla f(\hat{x})^T v \leq 0.$$

(Indice : utiliser les résultats des question 6, 7 et 8.)

Solution : Le résultat de la question précédente nous indique qu'il existe un vecteur $v \in \mathbf{R}^n$ une fonction strictement croissante $\psi : \mathbf{N} \to \mathbf{N}$ telle que

$$\lim_{s \to +\infty} \frac{\nabla f\left(x_{\phi\circ\psi(s)}\right)}{\|\nabla f\left(x_{\phi\circ\psi(s)}\right)\|} = v.$$

Par ailleurs, $\lim_{k\to+\infty} x_{\phi\circ\psi}(k) = \hat{x}$ puisque $(x_{\phi\circ\psi(k)})$ est une suite extraite de $(x_{\phi(k)})$. f étant de classe C^1, nous pouvons appliquer le résultat de la question 7, pour obtenir

$$\lim_{s \to +\infty} \nabla f\left(x_{\phi\circ\psi(s)} - \gamma_k \frac{\nabla f(x_{\phi\circ\psi(s)})}{\|\nabla f(x_{\phi\circ\psi(s)})\|}\right) = \nabla f(\hat{x}).$$

En prenant $k = \psi(s)$ et en passant à la limite quand $s \to +\infty$ des deux côtés de l'équation obtenue à la question 6, on obtient donc

$$\nabla f(\hat{x})^T v \leq \sigma \nabla f(\hat{x})^T v$$

Comme $0 < \sigma < 1$, on en déduit le résultat demandé.

10. Montrer qu'il existe une fonction strictement croissante $\psi : \mathbf{N} \to \mathbf{N}$, telle que :

$$\nabla f(\hat{x})^T v = \lim_{k \to +\infty} \|\nabla f\left(x_{\psi\circ\phi(k)}\right)\|_2.$$

Solution : *Il y avait une erreur d'énoncé ici : il faut prendre $\phi \circ \psi$ et non $\psi \circ \phi$.*

En prenant la fonction ψ de la question précédente et en utilisant que toute suite extraite d'une suite convergente converge vers la même limite, on obtient par un argument de continuité

$$\lim_{k \to +\infty} \nabla f\left(x_{\phi\circ\psi(k)}\right)^T \frac{\nabla f\left(x_{\phi\circ\psi(k)}\right)}{\|\nabla f\left(x_{\phi\circ\psi(k)}\right)\|} = \nabla f(\hat{x})^T v.$$

Or

$$\begin{aligned}
\nabla f\left(x_{\phi\circ\psi(k)}\right)^T \frac{\nabla f\left(x_{\phi\circ\psi(k)}\right)}{\|\nabla f\left(x_{\phi\circ\psi(k)}\right)\|} &= \frac{\nabla f\left(x_{\phi\circ\psi(k)}\right)^T \nabla f\left(x_{\phi\circ\psi(k)}\right)}{\|\nabla f\left(x_{\phi\circ\psi(k)}\right)\|} \\
&= \frac{\|\nabla f\left(x_{\phi\circ\psi(k)}\right)\|^2}{\|\nabla f\left(x_{\phi\circ\psi(k)}\right)\|} \\
&= \|\nabla f\left(x_{\phi\circ\psi(k)}\right)\|
\end{aligned}$$

On vient de voir que le terme de gauche de cette égalité converge vers $\nabla f(\hat{x})^T v$ quand $k \to +\infty$. On en déduit que le terme de droite converge vers $\nabla f(\hat{x})^T v$ quand $k \to +\infty$.

11. Conclure.

Solution : Comme $\|\nabla f\left(x_{\phi\circ\psi(k)}\right)\|_2$ est une suite extraite de $\|\nabla f\left(x_{\phi(k)}\right)\|_2$ et que, par un argument de continuité, $\|\nabla f\left(x_{\phi(k)}\right)\|_2$ converge vers $\|\nabla f\left(\hat{x}\right)\|_2$, on a que $\lim_{k\to+\infty} \|\nabla f\left(x_{\phi\circ\psi(k)}\right)\|_2 = \|\nabla f(\hat{x})\|_2 \geq 0$.

D'aprés la question précédente, cette limite est aussi égale à $\nabla f(\hat{x})^T v$, qui d'après la question 9 est négatif. On en déduit que $\|\nabla f(\hat{x})\|_2 = 0$ et donc que $\nabla f(\hat{x}) = 0$, ce qui contredit notre hypothèse de départ.

Exercice 4 Optimisation, normes, valeurs singulières et éléments propres. ($\star$)

1. Prouvez que $\|x\|_2 = \sqrt{x^T x}$ et que $\|Qx\|_2 = \|x\|_2$ pour toute matrice orthogonale Q.
2. Prouvez que $\|A\|_F = \|(\sigma_1, \sigma_2, \ldots, \sigma_r)\|_2$.
3. Prouvez que $\|A\|_2 = \sigma_1$.
4. Soit M une matrice carrée à coefficients réels. Prouvez que $\sigma_r(M) \leq \lambda \leq \sigma_1(M)$ pour toute valeur propre λ de M. Commencez par le cas d'une matrice de rang 1.
5. Montrez que toutes les valeurs propres d'une matrice orthogonale ont un module de 1.

Exercice 5 Moindres carrés linéaires et pseudo-inverse. ($\star\star$)

Etant donné A une matrice à coefficients réels de taille m par n et y une matrice colonne de taille m, on cherche à déterminer si $l : x \mapsto \|Ax - y\|_2$ possède une minimum global et à identifier le ou les $x \in \mathbf{R}^n$ où cet éventuel minimum global est atteint.

1. Montre qu'on peut répondre à notre question en étudiants $f : x \mapsto (Ax - y)^T(Ax - y)$.
2. Justifier que f est de classe C^1 et calculer le gradient de f.
3. En déduire une condition nécessaire d'optimum local pour f.
4. Calculez la Hessienne de f.
5. En déduire que f admet (au moins) un minimum global.
6. Donnez une condition nécessaire d'optimum global pour f.
7. On suppose que A^TA est inversible. Prouvez que l admet un unique minimum global et donnez une expression explicite pour ce minimum global.
8. On définit la pseudo-inverse A^+ d'une matrice A à coefficients réels de taille m par n par le biais de sa décomposition en valeurs singulières $A = U\Sigma V^T$ comme $A^+ := V\Sigma^+U^T$ où Σ^+ est la matrice diagonale contenant sur la case i de sa diagonale l'inverse de la i-ième valeurs singulière de A si celle-ci est strictement positive et 0 sinon. Montrez que si A est carrée et inversible, $A^+ = A$.
9. Montrez que $(A^+)^+ = A$, que $(A^+)^T = (A^T)^+$, et que $A(A^T)(A^+)^T = A = (A^+)^T(A^T)A$.
10. Montrez que AA^+ est la projection orthogonale sur l'image de A, que A^+A est la projection orthogonale sur l'espace des lignes de A (co-image), que $I - AA^+$ est la projection orthogonale sur le co-noyau de A et que $I - A^+A$ est la projection orthogonale sur sur le noyau de A.
11. Soit X une matrice à coefficients réels de taille n par d. Montrez que $\theta^* = X^+y$ est la solution de norme minimale du problème des moindres carrés linéaires consistant à minimiser $\|y - X\theta\|_2$. Précisez l'ensemble des autres solutions possibles s'il y en a.

Exercice 6 Régression linéaire et régularisation L_2. ($\star\star$)

On observe $(x_1, y_1), \dots (x_n, y_n) \in (\mathbf{R}^d \times \mathbf{R})^n$. On modélise $x_1, \dots, x_n$ comme les valeurs observées de variables aléatoires $X_1, \dots, X_n$ i.i.d. de loi P et $y_1, \dots, y_n$ comme les valeurs observées de variables aléatoires $Y_1, \dots Y_n$ telles qu'il existe $\theta^* \in \mathbf{R}^d$ tel que pour tout i dans $\{1, \dots n\}$, $Y_i = X_i^T\theta^* + \epsilon_i$, avec $\epsilon_1, \dots, \epsilon_n$ i.i.d. de loi N telle que $\mathbf{E}(N) = 0$ et $\mathbf{Var}(N) = \sigma^2$.

On note X la matrice de taille n par d dont les lignes sont $x_1, \dots, x_n$ et Y la matrice colonne contenant $y_1, \dots, y_n$. On cherche à estimer θ^* et on considère deux estimateurs :

— L'estimateur de la régression « ridge » :

$$\hat{\theta}_\lambda := \arg\min_{\theta \in \mathbf{R}^d} \frac{1}{n}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_2^2,$$

défini pour $\lambda > 0$;

— L'estimateur de norme minimale parmi les estimateurs au moindre carrés :

$$\hat{\theta} := \arg \min_{\theta \in \Theta_{LS}} \|\theta\|_2$$

où $\Theta_{LS} = \arg\min_{\theta \in \mathbf{R}^d} \|Y - X\theta\|_2^2$.

1. Montrez que pour toute matrice M de taille n par m et tout réel $\delta > 0$, $M^T M + \delta I$ est inversible. (Vous pouvez montrer, par exemple, que la matrice est symmétrique définie positive et utiliser le théorème spectral pour conclure.)

Solution : $(M^T M + \delta I)^T = (M^T M)^T + \delta I^T = M^T (M^T)^T + \delta I = M^T M + \delta I$, donc $M^T M + \delta I$ est symmétrique.

Soit $v \in \mathbf{R}^m \setminus \{0\}$. On a vu dans le DM 1 que $M^T M$ est positive (car elle est symmétrique et ses valeurs propres sont positives), donc $v^T M^T M v \geq 0$. De plus $v^T \delta I v = \delta \|v\|_2^2 > 0$ car $v \neq 0$ et $\delta > 0$. Donc $v^T (M^T M + \delta I) v = v^T M^T M v + v^T \delta I v > 0$. Donc $M^T M + \delta I$ est définie positive.

Comme $M^T M + \delta I$ est symmétrique on peut appliquer le théorème spectral et l'écrire $U \mathrm{diag}(\lambda_1, \ldots, \lambda_m) U^T$ pour une matrice orthogonale (réelle) U et des réels $\lambda_1, \ldots, \lambda_m$. Ces réels sont strictement positifs car $M^T M + \delta I$ est définie positive. On peut donc définir $A := U^T \mathrm{diag}(1/\lambda_1, \ldots, 1/\lambda_m) U$. On vérifie facilement que $(M^T M + \delta I)^T A = A (M^T M + \delta I)^T = I$. $M^T M + \delta I$ est donc inversible.

2. Montrez que le problème d'optimisation dont $\hat{\theta}_\lambda$ est défini comme la solution possède bien une unique solution donnée par $\hat{\theta}_\lambda = \frac{1}{n} \left(\frac{X^T X}{n} + \lambda I \right)^{-1} X^T Y$.

Solution : Considérons la fonction $f : \theta \mapsto \frac{1}{n}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_2^2$, définie sur $\mathbf{R}^d$.

f est clairement coercive, elle admet donc un minimum global.

De plus f est de classe C^2. Son gradient est :

$$\begin{aligned}\nabla f(\theta) &= \frac{1}{n}\nabla((Y - X\theta)^T(Y - X\theta)) + \lambda\nabla(\theta^T\theta) \\ &= \frac{1}{n}(-2\nabla(\theta^T X^T Y) + \nabla(\theta^T X^T X\theta)) + \lambda\nabla(\theta^T\theta) \\ &= \frac{1}{n}(-2X^T Y + 2X^T X\theta) + 2\lambda\theta.\end{aligned}$$

Sa Hessienne est :

$$\begin{aligned}\nabla^2 f(\theta) &= \frac{1}{n}\nabla(2X^T X\theta) + 2\lambda\nabla(\theta) \\ \nabla^2 f(\theta) &= 2\left(\frac{X^T X}{n} + \lambda I\right)\end{aligned}$$

D'après la question 1, sa Hessienne est toujours définie positive et donc f est strictement convexe. Il existe donc bien une unique solution globale $\hat{\theta}_\lambda$ au problème d'optimisation considéré.

De plus, d'après la condition nécessaire d'existence d'optimum local pour les fonction de classe C^1, on doit avoir $\nabla f(\hat{\theta}_\lambda) = 0$, c'est à dire $\frac{X^T Y}{n} = \left(\frac{X^T X}{n} + \lambda I\right)\hat{\theta}_\lambda$. D'après la question 1, $\frac{X^T X}{n} + \lambda I$ est inversible et donc $\hat{\theta}_\lambda = \left(\frac{X^T X}{n} + \lambda I\right)^{-1}\frac{X^T Y}{n}$.

3. Montrez que le problème d'optimisation dont $\hat{\theta}$ est défini comme la solution possède bien une unique solution donnée par $\hat{\theta} = X^\dagger Y$, où $X^\dagger$ est la pseudo-inverse (de Moore-Penrose) de X.

4. Montrez que $\hat{\theta}_\lambda$ tend vers $\hat{\theta}$ quand λ tend vers 0 par valeurs positives. (Indice : commencer par le cas où $d = n = 1$, puis le cas où X est diagonale, puis le cas général.)

5. Montrez qu'une simple descente de gradient à pas fixe sur $f(\theta) := \|Y - X\theta\|^2$ converge vers $\hat{\theta}$. Précisez comment choisir le pas. (Vous pouvez vous inspirer de l'exercice sur la convergence de la descente de gradient dans le cas lisse et fortement convexe.)

6. Calculez le biais de $\hat{\theta}_\lambda$.

Solution : Notons ϵ le vecteur colonne contenant $\epsilon_1, \dots, \epsilon_n$. On a :

$$\begin{aligned}
\mathbf{E}[\hat{\theta}_\lambda] &= \mathbf{E}\left[\frac{1}{n}\left(\frac{X^TX}{n} + \lambda I\right)^{-1} X^TY\right] \\
&= \mathbf{E}\left[\frac{1}{n}\left(\frac{X^TX}{n} + \lambda I\right)^{-1} X^T(X\theta^* + \epsilon)\right] \\
&= \mathbf{E}_X\left[\frac{1}{n}\left(\frac{X^TX}{n} + \lambda I\right)^{-1} X^TX\theta^*\right] + \mathbf{E}_{X,\epsilon}\left[\frac{1}{n}\left(\frac{X^TX}{n} + \lambda I\right)^{-1} X^T\epsilon\right].
\end{aligned}$$

Or :

$$\begin{aligned}
\mathbf{E}_{X,\epsilon}\left[\frac{1}{n}\left(\frac{X^TX}{n} + \lambda I\right)^{-1} X^T\epsilon\right] &= \mathbf{E}_X\left[\mathbf{E}_\epsilon\left[\frac{1}{n}\left(\frac{X^TX}{n} + \lambda I\right)^{-1} X^T\epsilon\right]\right] \\
&= \mathbf{E}_X\left[\frac{1}{n}\left(\frac{X^TX}{n} + \lambda I\right)^{-1} X^T\mathbf{E}_\epsilon(\epsilon)\right] \\
&= 0.
\end{aligned}$$

Donc :

$$\begin{aligned}
\mathbf{E}[\hat{\theta}_\lambda] &= \mathbf{E}_X\left[\left(\frac{X^TX}{n} + \lambda I\right)^{-1} \frac{X^TX}{n}\theta^*\right] \\
&= \mathbf{E}\left[\left(\hat{\Sigma} + \lambda I\right)^{-1} \hat{\Sigma}\right]\theta^*,
\end{aligned}$$

où on a noté $\hat{\Sigma} := \frac{X^TX}{n}$ la covariance empirique.
Le biais de $\hat{\theta}_\lambda$ est donc $\left(\mathbf{E}\left[\left(\hat{\Sigma} + \lambda I\right)^{-1} \hat{\Sigma}\right] - I\right)\theta^*$.

7. Calculez $v(\hat{\theta}_\lambda) := \mathbf{E}\left[\left\|\hat{\theta}_\lambda - \mathbf{E}[\hat{\theta}_\lambda]\right\|_2^2\right]$, la variance totale de $\hat{\theta}_\lambda$.

Solution :

$$\begin{aligned}
v(\hat{\theta}_\lambda) &= \mathbf{E}\left[\hat{\theta}_\lambda^T\hat{\theta}_\lambda\right] - \mathbf{E}\left[\hat{\theta}_\lambda\right]^T \mathbf{E}\left[\hat{\theta}_\lambda\right] \\
&= \mathbf{E}\left[\frac{(X\theta^* + \epsilon)^TX}{n}\left(\frac{X^TX}{n} + \lambda I\right)^{-2} \frac{X^T(X\theta^* + \epsilon)}{n}\right] - \left\|\mathbf{E}\left[\left(\hat{\Sigma} + \lambda I\right)^{-1} \hat{\Sigma}\right]\right\|_2^2 \\
&= 0 + 0 + 0 + \mathbf{E}\left[(\theta^*)^T\hat{\Sigma}\left(\hat{\Sigma} + \lambda I\right)^{-2} \hat{\Sigma}\theta^*\right] - \left\|\mathbf{E}\left[\left(\hat{\Sigma} + \lambda I\right)^{-1} \hat{\Sigma}\right]\right\|_2^2 \\
&= \mathbf{E}\left[\left\|\left(\hat{\Sigma} + \lambda I\right)^{-1} \hat{\Sigma}\right\|_2^2\right] - \left\|\mathbf{E}\left[\left(\hat{\Sigma} + \lambda I\right)^{-1} \hat{\Sigma}\right]\right\|_2^2.
\end{aligned}$$

8. Calculez le biais et la variance (totale) de $\hat{\theta}$ et commentez.

Exercice 7 Weyl's inequality. ($\star\star$)

Nous allons prouver un résultat qui peut être utile, par exemple, pour analyser les propriétés statistiques de l'estimateur usuel utilisé dans le cadre d'une analyse en composantes principales.

Considérons une matrice symmétrique M à coefficients réels et de taille d par d à laquelle on ajoute une perturbation, représentée par une autre matrice symétrique P à coefficients réels et de taille d par d. On cherche à trouver des conditions sous lesquelles on peut garantir que les valeurs propres de M et celles de $M' = M + P$ sont proches.

Pour toute matrice symmétrique A à coefficients réels et de taille d par d, on note $\lambda_1(A), \lambda_2, \ldots, \lambda_d(A)$ les valeurs propres de A triées par ordre décroissant.

1. Montrer que :

$$|\lambda_1(M') - \lambda_1(M)| \leq \|P\|.$$

   (Indice : nous avons défini et prouvé certaines propriétés de la norme matricielle $\|.\|_2$ en cours.)

2. Soit $i \in \{2, 3, \ldots, d\}$. Notons $\mathcal{E}_i^d$ l'ensemble de tous les sous-espaces de dimension i de $\mathbf{R}^d$. Montrer que pour toute matrice symmétrique A à coefficients réels et de taille d par d, la i-ième valeur propre de A est donnée par :

$$\lambda_i(A) = \min_{E \in \mathcal{E}_i^d} \max_{v \in E^\perp \cap \mathcal{S}^{d-1}} v^T A v,$$

   où $E^\perp$ est l'ensemble des vecteurs de $\mathbf{R}^d$ orthogonaux à tous les éléments de E et $\mathcal{S}^{d-1}$ est l'ensemble des vecteurs de $\mathbf{R}^d$ de norme euclidienne 1.

3. Montrer que :

$$\max_{i \in \{1,2,\ldots,d\}} |\lambda_i(M') - \lambda_i(M)| \leq \|P\|_2.$$