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
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.
3. En utilisant le résultat de la question précédente, montrez que pour $x \in \mathbf{R}^n$, $\|\nabla r(x)\|_2^2 \geq 2\mu\,(r(x) - r(x^*))$.
4. Donnez une interprétation verbale de la propriété que vous avez établi à la question précédente.

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

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

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 ?

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 $f(\hat{x})$.
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. $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$.
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.
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).
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.
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.)
7. Montrer que $\lim_{k\to+\infty} \gamma_k = 0$.
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.)

9. Montrer que :

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

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

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

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

11. Conclure.

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 $A^T A$ 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 $A = U^T$ 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 $\delta > 0$, $M^TM + \delta I$ est inversible. (Vous pouvez montrer, par exemple, que la matrice est symmétrique définie positive et utiliser le théorème spectral pour conclure.)
2. Montrez que le problème d'optimisation dont $\hat{\theta}_\lambda$ est défini comme la solution possède bien

une unique solution donnée par $\hat{\theta}_\lambda = \frac{1}{n}\left(\frac{X^T X}{n} + \lambda I\right)^{-1} X^T Y$.

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 $\hat{\theta}_\lambda$.
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 $\hat{\theta}_\lambda$.
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 :

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