Optimization for ML: Exercise Session 1

1 Stochastic gradient descents

Exercise 1 (coordinate stochastic gradient descent)**.** Let $f : \mathbb{R}^p \to \mathbb{R}$ be a differentiable function. We assume that f can be written as a finite sum of function $f(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$, where each f_i is differentiable. (For instance, $f(\theta)$ could be the empirical risk of a machine learning algorithm on a dataset of size n.)

To minimize f, we propose a coordinate stochastic gradient descent algorithm, defined as follows. Choose an initialization $\theta_0 \in \mathbb{R}^p$ and a stepsize $\gamma \in \mathbb{R}_+$. Then for all $k \in \mathbb{N}$, sample i_{k+1} uniformly at random in $\{1, \dots, n\}$ and j_{k+1} uniformly at random in $\{1, \dots, d\}$, independently of each other and of the past. Compute θ_{k+1} such that

$$\begin{aligned} \theta_{k+1}(j_{k+1}) &= \theta_k(j_{k+1}) - \gamma_k \partial_{j_{k+1}} \ell_{i_{k+1}}(\theta_k)\,, \\ \theta_{k+1}(j) &= \theta_k(j) \qquad \text{for all } j \neq j_{k+1}\,. \end{aligned}$$

Show that the coordinate stochastic gradient descent algorithm is a stochastic gradient descent algorithm in the sense seen in the lecture. In particular, show that stochastic gradients are unbiased.

Exercise 2 (gossip problem)**.** Consider a communication network, that we model as a graph $G = (V, E)$, where V is the set of communication nodes and $E \subset \{\{i, j\} \,|\, i \neq j \in V\}$ is the set of communication links or edges. The gossip problem is an elementary problem in decentralized distributed computing where each node $v \in V$ is given a value $\theta_0(v)$ and the goal of the network is to compute the average of all these values. However, there is no central node in the network that can collect all of the information and computed the average. Instead, the nodes can only communicate along the edges of the graph when the communication link is activated.

To solve the gossip problem, we propose the following algorithm. At each iteration $k \in \mathbb{N}$, a communication link $(v_{k+1}, w_{k+1}) \in E$ is activated, uniformly at random. The nodes v_{k+1} and w_{k+1} exchange their values $\theta_k(v_{k+1})$ and $\theta_k(w_{k+1})$ and update them by averaging the two values:

$$\begin{aligned} \theta_{k+1}(v_{k+1}) &= \frac{1}{2}\theta_k(v_{k+1}) + \frac{1}{2}\theta_k(w_{k+1})\,, \\ \theta_{k+1}(w_{k+1}) &= \frac{1}{2}\theta_k(v_{k+1}) + \frac{1}{2}\theta_k(w_{k+1})\,. \end{aligned}$$

All other nodes keep their values unchanged: $\theta_{k+1}(z) = \theta_k(z)$ for $z \neq v_{k+1}, w_{k+1}$.

1. Show that the algorithm described above is a stochastic gradient descent algorithm in the sense seen in the lecture, on a function f to be determined.
2. Describe the set of minimizers of f.

2 Function structures

In this section, f denotes a function from $\mathbb{R}^p$ to $\mathbb{R}$.

2.1 Convexity

Definition. The function f is said to be *convex* if for all $\theta, \theta' \in \mathbb{R}^p$ and all $\alpha \in [0,1]$, we have

$$f((1-\alpha)\theta + \alpha\theta') \leqslant (1-\alpha)f(\theta) + \alpha f(\theta')\,.$$

Proposition 1. *We assume that f is continuously differentiable. The following conditions are equivalent:*

(i) f is convex,
(ii) for all $\theta, \theta' \in \mathbb{R}^p$, $f(\theta') \geqslant f(\theta) + \langle \nabla f(\theta), \theta' - \theta \rangle$,
(iii) for all $\theta, \theta' \in \mathbb{R}^p$, $\langle \nabla f(\theta) - \nabla f(\theta'), \theta - \theta' \rangle \geqslant 0$.

Exercise 3. The goal of this exercise is to prove the above proposition.

1. Using the function

$$g(\alpha) = (1-\alpha)f(\theta) + \alpha f(\theta') - f((1-\alpha)\theta + \alpha\theta')\,, \qquad \alpha \in [0,1]\,,$$

show that $(i) \Rightarrow (ii)$.

2. Using that f is above its tangent hyperplane at $(1-\alpha)\theta + \alpha\theta'$, show that $(ii) \Rightarrow (i)$.

3. Combining the inequality (ii) with the same inequality with θ and θ' exchanged, show that $(ii) \Rightarrow (iii)$.

4. Using the function

$$h(\alpha) = f((1-\alpha)\theta + \alpha\theta') - f(\theta) - \langle \nabla f(\theta), \alpha(\theta' - \theta) \rangle\,, \qquad \alpha \in [0,1]\,,$$

show that $(iii) \Rightarrow (ii)$.

Proposition 2. *We assume that f is twice continuously differentiable. The following conditions are equivalent:*

(i) f is convex,
(ii) for all $\theta \in \mathbb{R}^p$, $\nabla^2 f(\theta) \succcurlyeq 0$.

Exercise 4. The goal of this exercise is to prove the above proposition.

1. Using the function

$$g(\delta) = \langle \nabla f(\theta + \delta v) - \nabla f(\theta), v \rangle\,, \qquad \delta \geqslant 0\,,$$

show that $(i) \Rightarrow (ii)$.

2. Using the function

$$h(\alpha) = \langle \nabla f((1-\alpha)\theta + \alpha\theta') - \nabla f(\theta), \theta' - \theta \rangle\,, \qquad \alpha \in [0,1]\,,$$

show that $(ii) \Rightarrow (i)$.

2.2 Strong convexity

Definition. Let $\mu > 0$. The function f is said to be *μ-strongly convex* if for all $\theta, \theta' \in \mathbb{R}^p$ and all $\alpha \in [0,1]$,

$$f((1-\alpha)\theta + \alpha\theta') \leqslant (1-\alpha)f(\theta) + \alpha f(\theta') - \frac{\mu}{2}\alpha(1-\alpha)\|\theta - \theta'\|^2 .$$

Further, the function f is said to be *strongly convex* if it is μ-strongly convex for some $\mu > 0$.

Proposition 3. *The following conditions are equivalent:*
(i) f is μ-strongly convex,
(ii) the function $g(\theta) = f(\theta) - \frac{\mu}{2}\|\theta\|^2$ is convex.

Exercise 5. Show the above proposition.

Proposition 4. *We assume that f is continuously differentiable. The following conditions are equivalent:*
(i) f is μ-strongly convex,
(ii) for all $\theta, \theta' \in \mathbb{R}^p$, $f(\theta') \geqslant f(\theta) + \langle \nabla f(\theta), \theta' - \theta \rangle + \frac{\mu}{2}\|\theta' - \theta\|^2$,
(iii) for all $\theta, \theta' \in \mathbb{R}^p$, $\langle \nabla f(\theta) - \nabla f(\theta'), \theta - \theta' \rangle \geqslant \mu\|\theta' - \theta\|^2$.

Exercise 6. Show the above proposition.

Proposition 5 (implications of strong convexity)**.** *We assume that f is continuously differentiable and μ-strongly convex. Then:*
(i) f has a unique minimizer θ_*,
(ii) we have the Polyak-Lojasiewicz condition*: for all $\theta \in \mathbb{R}^p$,*

$$\frac{1}{2}\|\nabla f(\theta)\|^2 \geqslant \mu\left(f(\theta) - f(\theta_*)\right) ,$$

(iii) for all $\theta, \theta' \in \mathbb{R}^p$, $\|\nabla f(\theta) - \nabla f(\theta')\| \geqslant \mu\|\theta - \theta'\|$.

Exercise 7. This exercise shows the above proposition.
(a) Show (i).
(b) Minimizing over θ' in both sides of Prop. 4(ii), show (ii).
(c) Show (iii).

Proposition 6. *We assume that f is twice continuously differentiable. The following conditions are equivalent:*
(i) f is μ-strongly convex,
(ii) for all $\theta \in \mathbb{R}^p$, $\nabla^2 f(\theta) \succcurlyeq \mu I_p$.

Exercise 8. Show the above proposition.

2.3 Smoothness

In all of this section, we assume that f is *convex* and *continuously differentiable.*

Definition. Let $L > 0$. The function f is said to be *L-smooth* if for all $\theta, \theta' \in \mathbb{R}^p$, for all $\alpha \in [0,1]$,

$$f((1-\alpha)\theta + \alpha\theta') \geqslant (1-\alpha)f(\theta) + \alpha f(\theta') - \frac{L}{2}\alpha(1-\alpha)\|\theta - \theta'\|^2 .$$

Further, the function f is said to be *smooth* if it is L-smooth for some $L > 0$.

We admit the following proposition, whose proof is very similar to the one for strong convexity.

Proposition 7. *The following conditions are equivalent:*
(i) f *is* L*-smooth,*
(ii) the function $g(\theta) = \frac{L}{2}\|\theta\|^2 - f(\theta)$ *is convex.*

Proposition 8. *The following conditions are equivalent:*
(i) f *is* L*-smooth,*
(ii) for all $\theta, \theta' \in \mathbb{R}^p$, $f(\theta') \leqslant f(\theta) + \langle \nabla f(\theta), \theta' - \theta \rangle + \frac{L}{2}\|\theta' - \theta\|^2$,
(iii) for all $\theta, \theta' \in \mathbb{R}^p$, $\langle \nabla f(\theta') - \nabla f(\theta), \theta' - \theta \rangle \leqslant L\|\theta' - \theta\|^2$
(iv) ∇f *is co-coercive: for all* $\theta, \theta' \in \mathbb{R}^p$,

$$\|\nabla f(\theta') - \nabla f(\theta)\|^2 \leqslant L\langle \theta' - \theta, \nabla f(\theta') - \nabla f(\theta) \rangle \,,$$

(v) ∇f *is* L*-Lipschitz: for all* $\theta, \theta' \in \mathbb{R}^p$, $\|\nabla f(\theta') - \nabla f(\theta)\| \leqslant L\|\theta' - \theta\|$.

Exercise 9. We admit the proof that $(i) \Leftrightarrow (ii) \Leftrightarrow (iii)$ could be done as before. This exercise completes the proof of the above proposition.

1. We assume (ii).

(a) Show that

$$f(\theta') + \langle \nabla f(\theta'), \theta + \theta'' - \theta' \rangle \leqslant f(\theta) + \langle \nabla f(\theta), \theta'' \rangle + \frac{L}{2}\|\theta''\|^2 \,.$$

(b) Isolating all terms involving θ'' on the left-hand side and optimizing over θ'', show that (iv) holds.

2. Conclude on the proof of the above proposition.

We also admit the proof of the following proposition.

Proposition 9. *We assume that* f *is twice continuously differentiable. The following conditions are equivalent:*
(i) f *is* L*-smooth,*
(ii) for all $\theta \in \mathbb{R}^p$, $\nabla^2 f(\theta) \preccurlyeq LI_p$.

1 Stochastic gradient descents

Exercise 1 (coordinate stochastic gradient descent)**.** Let $f : \mathbb{R}^p \to \mathbb{R}$ be a differentiable function. We assume that f can be written as a finite sum of function $f(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$, where each f_i is differentiable. (For instance, $f(\theta)$ could be the empirical risk of a machine learning algorithm on a dataset of size n.)

To minimize f, we propose a coordinate stochastic gradient descent algorithm, defined as follows. Choose an initialization $\theta_0 \in \mathbb{R}^p$ and a stepsize $\gamma \in \mathbb{R}_+$. Then for all $k \in \mathbb{N}$, sample i_{k+1} uniformly at random in $\{1, \dots, n\}$ and j_{k+1} uniformly at random in $\{1, \dots, d\}$, independently of each other and of the past. Compute θ_{k+1} such that

$$\begin{aligned} \theta_{k+1}(j_{k+1}) &= \theta_k(j_{k+1}) - \gamma_k \partial_{j_{k+1}} \ell_{i_{k+1}}(\theta_k)\,, \\ \theta_{k+1}(j) &= \theta_k(j) \qquad \text{for all } j \neq j_{k+1}\,. \end{aligned}$$

Show that the coordinate stochastic gradient descent algorithm is a stochastic gradient descent algorithm in the sense seen in the lecture. In particular, show that stochastic gradients are unbiased.

Ex 1

- Descent de gradient stochastique

Pour $f(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$

soit $\xi = (i,j) \sim \mathrm{Unif}(\{1,\dots,n\}) \otimes \mathrm{Unif}(\{1,\dots,p\}) = Q$

et $g(\theta,\xi) = g(\theta,(i,j)) = \partial_j f_i(\theta) \cdot e_j$

Choisit $\theta_0 \in \mathbb{R}^p$ et $\gamma \in \mathbb{R}^+$.

$\forall k \in \mathbb{N}$, on échantillons $\xi_{k+1} = (i_{k+1}, j_{k+1}) \sim Q$

$$\begin{aligned} \theta_{k+1}(j_{k+1}) &= \theta_k(j_{k+1}) - \gamma \cdot g(\theta_k, \xi_{k+1}) \\ &= \theta_k(j_{k+1}) - \gamma \cdot \partial_{j_{k+1}} f_{i_{k+1}}(\theta_k) \quad , \; j = j_{k+1} \end{aligned}$$

$$\theta_{k+1}(j) = \theta_k(j) \quad , \; j \neq j_{k+1}$$

日期: /

- Non-biaisé

$$\begin{aligned}
\mathbb{E}_{(i,j)}\left[g(\theta,(i,j))\right] &= \mathbb{E}\left[\partial_j f_i(\theta)\cdot e_j\right] \\
&= \mathbb{E}\left[e_j^T \nabla f_i(\theta)\cdot e_j\right] \\
&= \mathbb{E}\left[e_j e_j^T\right]\cdot \mathbb{E}\left[\nabla f_i(\theta)\right] \\
&= \frac{1}{p} I_d \cdot \nabla \mathbb{E}\left[f_i(\theta)\right] \\
&= \frac{1}{p}\cdot \nabla\left(\frac{1}{n}\sum_{i=1}^{n} f_i(\theta)\right) \\
&= \frac{1}{p}\cdot \nabla f(\theta)
\end{aligned}$$

Donc cet algorithme est un algorithme descent de gradient stochastique sur la fonction $\theta \longmapsto \frac{1}{p}\nabla f(\theta)$

Exercise 2 (gossip problem)**.** Consider a communication network, that we model as a graph $G = (V, E)$, where V is the set of communication nodes and $E \subset \{\{i, j\} \,|\, i \neq j \in V\}$ is the set of communication links or edges. The gossip problem is an elementary problem in decentralized distributed computing where each node $v \in V$ is given a value $\theta_0(v)$ and the goal of the network is to compute the average of all these values. However, there is no central node in the network that can collect all of the information and computed the average. Instead, the nodes can only communicate along the edges of the graph when the communication link is activated.

To solve the gossip problem, we propose the following algorithm. At each iteration $k \in \mathbb{N}$, a communication link $(v_{k+1}, w_{k+1}) \in E$ is activated, uniformly at random. The nodes v_{k+1} and w_{k+1} exchange their values $\theta_k(v_{k+1})$ and $\theta_k(w_{k+1})$ and update them by averaging the two values:

$$\begin{aligned}
\theta_{k+1}(v_{k+1}) &= \frac{1}{2}\theta_k(v_{k+1}) + \frac{1}{2}\theta_k(w_{k+1})\,, \\
\theta_{k+1}(w_{k+1}) &= \frac{1}{2}\theta_k(v_{k+1}) + \frac{1}{2}\theta_k(w_{k+1})\,.
\end{aligned}$$

All other nodes keep their values unchanged: $\theta_{k+1}(z) = \theta_k(z)$ for $z \neq v_{k+1}, w_{k+1}$.

1. Show that the algorithm described above is a stochastic gradient descent algorithm in the sense seen in the lecture, on a function f to be determined.
2. Describe the set of minimizers of f.

日期: /

(1) Pour $\{v_{k+1}, w_{k+1}\} \sim Unif(E)$, on a

$$\begin{cases} \theta_{k+1}(v_{k+1}) = \theta_k(v_{k+1}) + \frac{1}{2}(\theta_k(w_{k+1}) - \theta_k(v_{k+1})) \\ \theta_{k+1}(w_{k+1}) = \theta_k(w_{k+1}) + \frac{1}{2}(\theta_k(v_{k+1}) - \theta_k(w_{k+1})) \\ \theta_{k+1}(z) = \theta_k(z) \qquad si \quad z \neq v_{k+1}, w_{k+1} \end{cases}$$

Soit $\xi_{k+1} = (v_{k+1}, w_{k+1}) \sim Unif(E) = Q$

et $\theta_{k+1} = \theta_k - \gamma_k \cdot g(\theta_k, \xi_{k+1})$, alors

$$\begin{cases} -\gamma_k \cdot g(\theta_k, (v_{k+1}, w_{k+1}))(v_{k+1}) = \frac{1}{2}(\theta_k(w_{k+1}) - \theta_k(v_{k+1})) \\ -\gamma_k \cdot g(\theta_k, (v_{k+1}, w_{k+1}))(w_{k+1}) = \frac{1}{2}(\theta_k(v_{k+1}) - \theta_k(w_{k+1})) \\ -\gamma_k \cdot g(\theta_k, (v_{k+1}, w_{k+1}))(z) = 0 \qquad pour \quad z \neq v_{k+1}, w_{k+1} \end{cases}$$

alors

$$-\gamma_k \cdot g(\theta_k, (v_{k+1}, w_{k+1})) = \frac{1}{2}(\theta_k(w_{k+1}) - \theta_k(v_{k+1})) \cdot (e_{v_{k+1}} - e_{w_{k+1}})$$

$$= -\frac{1}{2}\langle \theta_k, e_{v_{k+1}} - e_{w_{k+1}} \rangle \cdot (e_{v_{k+1}} - e_{w_{k+1}})$$

donc

$$\gamma_k = \frac{1}{2}, \quad g(\theta_k, (v_{k+1}, w_{k+1})) = \langle \theta_k, e_{v_{k+1}} - e_{w_{k+1}} \rangle \cdot (e_{v_{k+1}} - e_{w_{k+1}})$$

par

$$\begin{array}{ccc} f(\theta, \xi) & \xrightarrow{E} & f(\theta) \\ \int \uparrow & & \uparrow \int \\ g(\theta, \xi) & \xrightarrow{E} & \nabla f(\theta) \end{array}$$

日期: /

soit $f(\theta, \{v_{k+1}, w_{k+1}\}) = cst - \langle r, \theta \rangle + \frac{1}{2}\langle \theta, M\theta \rangle$

alors $\nabla f(\theta, \{v_{k+1}, w_{k+1}\}) = -r + \frac{1}{2}(M + M^T)\theta$

donc

$r = 0$, $\frac{1}{2}(M + M^T) = (e_{v_{k+1}} - e_{w_{k+1}})(e_{v_{k+1}} - e_{w_{k+1}})^T$

$M = (e_{v_{k+1}} - e_{w_{k+1}})(e_{v_{k+1}} - e_{w_{k+1}})^T$

alors

$f(\theta, \{v_{k+1}, w_{k+1}\}) = \frac{1}{2}\langle \theta, (e_{v_{k+1}} - e_{w_{k+1}})(e_{v_{k+1}} - e_{w_{k+1}})^T \theta \rangle$

$= \frac{1}{2}(\theta(v_{k+1}) - \theta(w_{k+1}))^2$

$$f(\theta) = \frac{1}{|E|}\sum_{\{v,w\}\in E} f(\theta, \{v, w\}) = \frac{1}{2\cdot|E|}\sum_{\{v,w\}\in E} (\theta(v_{k+1}) - \theta(w_{k+1}))^2$$

2. Describe the set of minimizers of f.

(2) $f \geq 0$ et $f(\theta) = 0$ si et seulement si θ est constant sur les composantes connexes du grape.

2 Function structures

In this section, f denotes a function from $\mathbb{R}^p$ to $\mathbb{R}$.

2.1 Convexity

Definition. The function f is said to be *convex* if for all $\theta, \theta' \in \mathbb{R}^p$ and all $\alpha \in [0,1]$, we have

$$f((1-\alpha)\theta + \alpha\theta') \leqslant (1-\alpha)f(\theta) + \alpha f(\theta')\,.$$

Proposition 1. *We assume that f is continuously differentiable. The following conditions are equivalent:*

(i) f is convex,
(ii) for all $\theta, \theta' \in \mathbb{R}^p$, $f(\theta') \geqslant f(\theta) + \langle \nabla f(\theta), \theta' - \theta \rangle$,
(iii) for all $\theta, \theta' \in \mathbb{R}^p$, $\langle \nabla f(\theta) - \nabla f(\theta'), \theta - \theta' \rangle \geqslant 0$.

Exercise 3. The goal of this exercise is to prove the above proposition.

1. Using the function

$$g(\alpha) = (1-\alpha)f(\theta) + \alpha f(\theta') - f((1-\alpha)\theta + \alpha\theta')\,, \qquad \alpha \in [0,1]\,,$$

show that $(i) \Rightarrow (ii)$.

(i) Soit f est convexe, on a

$$g(\alpha) \geqslant (1-\alpha)f(\theta) + \alpha f(\theta') - \left((1-\alpha)f(\theta) + \alpha f(\theta')\right) \geqslant 0$$

comme f est continument différentiable

et $g(0) = f(\theta) - f(\theta) = 0 \qquad g(1) = f(\theta') - f(\theta') = 0$

lorsque $\alpha \longrightarrow 0$,

$$g(\alpha) = g(0) + g'(0)\cdot\alpha + o(\alpha)$$
$$= 0 + \left(-f(\theta) + f(\theta') - \langle \theta' - \theta, \nabla f(\theta) \rangle\right)\cdot\alpha + o(\alpha)$$

et comme $g(\alpha) \geqslant 0$, on a

$$f(\theta') \geqslant f(\theta) + \langle \theta' - \theta, \nabla f(\theta) \rangle$$

2. Using that f is above its tangent hyperplane at $(1-\alpha)\theta + \alpha\theta'$, show that $(ii) \Rightarrow (i)$.

(2) Soit $\theta, \theta' \in \mathbb{R}^p$, alors $(1-\alpha)\theta + \alpha\theta' \in \mathbb{R}^p$,

par (ii), $\forall\, \theta'' \in \mathbb{R}^p$, on a

$$f(\theta'') \geq f((1-\alpha)\theta + \alpha\theta') + \langle \nabla f((1-\alpha)\theta + \alpha\theta'),\ \theta'' - ((1-\alpha)\theta + \alpha\theta') \rangle$$

soit $\theta'' = \theta$, on a

① $$f(\theta) \geq f((1-\alpha)\theta + \alpha\theta') + \langle \nabla f((1-\alpha)\theta + \alpha\theta'),\ \alpha(\theta - \theta') \rangle$$

soit $\theta'' = \theta'$, on a

② $$f(\theta') \geq f((1-\alpha)\theta + \alpha\theta') + \langle \nabla f((1-\alpha)\theta + \alpha\theta'),\ (1-\alpha)\cdot(\theta' - \theta) \rangle$$

$(1-\alpha)\cdot$① $+\ \alpha\cdot$② alors

$$(1-\alpha) f(\theta) + \alpha f(\theta') \geq f((1-\alpha)\theta + \alpha\theta') \qquad \text{d'où (i)}$$

3. Combining the inequality (ii) with the same inequality with θ and θ' exchanged, show that $(ii) \Rightarrow (iii)$.

(3) Par (ii), $\forall\, \theta, \theta' \in \mathbb{R}^p$

$$f(\theta') \geq f(\theta) + \langle \nabla f(\theta),\ \theta' - \theta \rangle$$

$$f(\theta) \geq f(\theta') + \langle \nabla f(\theta'),\ \theta - \theta' \rangle$$

alors

$$f(\theta') + f(\theta) \geq f(\theta) + f(\theta') + \langle \nabla f(\theta) - \nabla f(\theta'),\ \theta' - \theta \rangle$$

$$\langle \nabla f(\theta) - \nabla f(\theta'),\ \theta - \theta' \rangle \geq 0$$

4. Using the function

$$h(\alpha) = f((1-\alpha)\theta + \alpha\theta') - f(\theta) - \langle \nabla f(\theta), \alpha(\theta' - \theta) \rangle , \qquad \alpha \in [0,1] ,$$

show that $(iii) \Rightarrow (ii)$.

(4) $h(0) = f(\theta) - f(\theta) - \langle \nabla f(\theta), 0 \rangle = 0$

$h'(\alpha) = \langle \nabla f((1-\alpha)\theta + \alpha\theta'), \theta' - \theta \rangle - \langle \nabla f(\theta), \theta' - \theta \rangle$

$= \langle \nabla f((1-\alpha)\theta + \alpha\theta') - \nabla f(\theta), \theta' - \theta \rangle$

Par (iii), on a

$\langle \nabla f((1-\alpha)\theta + \alpha\theta') - \nabla f(\theta), (1-\alpha)\theta + \alpha\theta' - \theta \rangle \geq 0$

$\langle \nabla f((1-\alpha)\theta + \alpha\theta') - \nabla f(\theta), \alpha(\theta' - \theta) \rangle \geq 0$

donc comme $\alpha \geq 0$, on a

$h'(\alpha) \geq 0$

et puisque $h(0) = 0$ alors

$0 = h(0) \leq h(1) = f(\theta') - f(\theta) - \langle \nabla f(\theta), \theta' - \theta \rangle$

donc

$f(\theta') \geq f(\theta) + \langle \nabla f(\theta), \theta' - \theta \rangle$

Proposition 2. *We assume that f is twice continuously differentiable. The following conditions are equivalent:*

(i) f is convex,

(ii) for all $\theta \in \mathbb{R}^p$, $\nabla^2 f(\theta) \succcurlyeq 0$.

Exercise 4. The goal of this exercise is to prove the above proposition.

1. Using the function

$$g(\delta) = \langle \nabla f(\theta + \delta v) - \nabla f(\theta), v \rangle , \qquad \delta \geqslant 0 ,$$

show that $(i) \Rightarrow (ii)$.

(i) D'après Prop 1 (iii) et puisque f est convexe,

$$\langle \nabla f(\theta+\delta v) - \nabla f(\theta) , \theta+\delta v-\theta \rangle \geqslant 0$$

$$\delta \cdot \langle \nabla f(\theta+\delta v) - \nabla f(\theta) , v \rangle \geqslant 0$$

$$\delta \cdot g(\delta) \geqslant 0$$

$$g(\delta) \geqslant 0$$

et on a

$$g(0) = \langle \nabla f(\theta) - \nabla f(\theta) , v \rangle = 0$$

$$g'(\delta) = \langle v \cdot \nabla^2 f(\theta+\delta v) , v \rangle$$

lorsque $\delta \longrightarrow 0$,

$$g(\delta) = g(0) + g'(0) \cdot \delta + o(\delta)$$

$$= 0 + \langle v \cdot \nabla^2 f(\theta) , v \rangle \cdot \delta + o(\delta)$$

alors

$$\langle v \cdot \nabla^2 f(\theta) , v \rangle \cdot \delta + o(\delta) \geqslant 0$$

donc

$$\nabla^2 f(\theta) \geqslant 0$$

2. Using the function

$$h(\alpha) = \langle \nabla f((1-\alpha)\theta + \alpha\theta') - \nabla f(\theta), \theta' - \theta \rangle , \qquad \alpha \in [0,1] ,$$

show that $(ii) \Rightarrow (i)$.

(2) $h(0) = \langle \nabla f(\theta) - \nabla f(\theta), \theta' - \theta \rangle = 0$

$h'(\alpha) = \langle (\theta' - \theta) \cdot \nabla^2 f((1-\alpha)\theta + \alpha\theta') , \theta' - \theta \rangle$

D'après Prop 2 (ii)

$$\nabla^2 f((1-\alpha)\theta + \alpha\theta') \geq 0$$

donc

$$h'(\alpha) \propto \langle \theta' - \theta, \theta' - \theta \rangle = \|\theta' - \theta\|^2 \geq 0$$

alors

$$h(1) = \langle \nabla f(\theta') - \nabla f(\theta) , \theta' - \theta \rangle \geq 0$$

d'après Prop 1 , f est convexe.

2.2 Strong convexity

Definition. Let $\mu > 0$. The function f is said to be *μ-strongly convex* if for all $\theta, \theta' \in \mathbb{R}^p$ and all $\alpha \in [0,1]$,

$$f((1-\alpha)\theta + \alpha\theta') \leqslant (1-\alpha)f(\theta) + \alpha f(\theta') - \frac{\mu}{2}\alpha(1-\alpha)\|\theta - \theta'\|^2 .$$

Further, the function f is said to be *strongly convex* if it is μ-strongly convex for some $\mu > 0$.

Proposition 3. *The following conditions are equivalent:*

(i) f is μ-strongly convex,

(ii) the function $g(\theta) = f(\theta) - \frac{\mu}{2}\|\theta\|^2$ is convex.

Exercise 5. Show the above proposition.

Proposition 4. *We assume that f is continuously differentiable. The following conditions are equivalent:*

(i) f is μ-strongly convex,

(ii) for all $\theta, \theta' \in \mathbb{R}^p$, $f(\theta') \geqslant f(\theta) + \langle \nabla f(\theta), \theta' - \theta \rangle + \frac{\mu}{2}\|\theta' - \theta\|^2$,

(iii) for all $\theta, \theta' \in \mathbb{R}^p$, $\langle \nabla f(\theta) - \nabla f(\theta'), \theta - \theta' \rangle \geqslant \mu \|\theta' - \theta\|^2$.

Exercise 6. Show the above proposition.

Proposition 5 (implications of strong convexity). *We assume that f is continuously differentiable and μ-strongly convex. Then:*

(i) f has a unique minimizer θ_*,

(ii) we have the Polyak-Lojasiewicz condition*: for all $\theta \in \mathbb{R}^p$,*

$$\frac{1}{2}\|\nabla f(\theta)\|^2 \geqslant \mu\,(f(\theta) - f(\theta_*))\ ,$$

(iii) for all $\theta, \theta' \in \mathbb{R}^p$, $\|\nabla f(\theta) - \nabla f(\theta')\| \geqslant \mu\|\theta - \theta'\|$.

Exercise 7. This exercise shows the above proposition.
(a) Show (i).
(b) Minimizing over θ' in both sides of Prop. 4(ii), show (ii).
(c) Show (iii).

Exercice 7

(a) Soient $\theta_*, \theta_*' \in \mathbb{R}^p$ deux minimiseurs de f,

par Prop 4 (iii),

$$\langle \underbrace{\nabla f(\theta_*)}_{=0} - \underbrace{\nabla f(\theta_*')}_{=0}, \theta_* - \theta_*' \rangle \geqslant \mu \cdot \|\theta_* - \theta_*'\|^2$$

$$0 \geqslant \mu \cdot \|\theta_* - \theta_*'\|^2$$

donc on a $\theta_* = \theta_*'$, d'où l'unicité.

Existence

Par Prop 4 (ii), $f(\theta) \geqslant f(0) + \langle \nabla f(0), \theta \rangle + \frac{\mu}{2}\|\theta\|^2$

$$\geqslant f(0) - \|\nabla f(0)\| \cdot \|\theta\| + \frac{\mu}{2} \cdot \|\theta\|^2$$

$$\xrightarrow{\|\theta\| \to +\infty} +\infty$$

Il existe $R > 0$, tel que si $\|\theta\| \geqslant R$, alors $f(\theta) \geqslant f(0) + 1$

Comme f est continue sur $\overline{B(0,R)}$ compact,

donc f atteint son minimum sur $\overline{B(0,R)}$.

(b) Minimizing over θ' in both sides of Prop. 4(ii), show (ii).

(b) D'après Prop 4 (ii),

$\forall\ \theta, \theta' \in \mathbb{R}^p$, $f(\theta') \geq f(\theta) + \langle \nabla f(\theta), \theta' - \theta \rangle + \frac{\mu}{2} \cdot \|\theta' - \theta\|^2$

soit $g(\theta') = f(\theta) + \langle \nabla f(\theta), \theta' - \theta \rangle + \frac{\mu}{2} \cdot \|\theta' - \theta\|^2$

$\nabla g(\theta') = \nabla f(\theta) + \mu \cdot (\theta' - \theta)$

par la règle de Fermat, on a que g est optimal dans $\theta' = \theta - \frac{1}{\mu} \nabla f(\theta)$.

Et soit θ^* est le minimiseur de f, alors on a

$$f(\theta^*) \geq f(\theta) + \langle \nabla f(\theta), \theta - \tfrac{1}{\mu} \nabla f(\theta) - \theta \rangle + \tfrac{\mu}{2} \|\theta - \tfrac{1}{\mu} \nabla f(\theta) - \theta\|^2$$
$$= f(\theta) - \tfrac{1}{\mu} \|\nabla f(\theta)\|^2 + \tfrac{1}{2\mu} \|\nabla f(\theta)\|^2$$

donc

$$\tfrac{1}{2} \|\nabla f(\theta)\|^2 \geq \mu \cdot (f(\theta) - f(\theta^*))$$

(c) Show (iii).

(c) Par Prop 4 (iii),

$\forall\ \theta, \theta' \in \mathbb{R}^p$, $\mu \cdot \|\theta - \theta'\|^2 \leq \langle \nabla f(\theta) - \nabla f(\theta'), \theta - \theta' \rangle$
$\leq \|\nabla f(\theta) - \nabla f(\theta')\| \cdot \|\theta - \theta'\|$

alors

$$\mu \cdot \|\theta - \theta'\| \leq \|\nabla f(\theta) - \nabla f(\theta')\|$$

Proposition 6. *We assume that f is twice continuously differentiable. The following conditions are equivalent:*

(i) f is μ-strongly convex,

(ii) for all $\theta \in \mathbb{R}^p$, $\nabla^2 f(\theta) \succcurlyeq \mu I_p$.

Exercise 8. Show the above proposition.

Exercice 8

Soit $g(\theta) = f(\theta) - \frac{\mu}{2}\|\theta\|^2$, alors

f est fortement convexe $\iff$ g est convexe

$\iff \forall \theta \in \mathbb{R}^p,\ \nabla^2 g(\theta) \geq 0$

$\iff \forall \theta \in \mathbb{R}^p,\ \nabla^2 f(\theta) - \mu \cdot I_p \geq 0$

2.3 Smoothness

In all of this section, we assume that f is *convex* and *continuously differentiable.*

Definition. Let $L > 0$. The function f is said to be *L-smooth* if for all $\theta, \theta' \in \mathbb{R}^p$, for all $\alpha \in [0,1]$,

$$f((1-\alpha)\theta + \alpha\theta') \geqslant (1-\alpha)f(\theta) + \alpha f(\theta') - \frac{L}{2}\alpha(1-\alpha)\|\theta - \theta'\|^2 .$$

Further, the function f is said to be *smooth* if it is L-smooth for some $L > 0$.

We admit the following proposition, whose proof is very similar to the one for strong convexity.

Proposition 7. *The following conditions are equivalent:*

(i) f is L-smooth,

(ii) the function $g(\theta) = \frac{L}{2}\|\theta\|^2 - f(\theta)$ is convex.

Proposition 8. *The following conditions are equivalent:*

(i) f is L-smooth,

(ii) for all $\theta, \theta' \in \mathbb{R}^p$, $f(\theta') \leqslant f(\theta) + \langle \nabla f(\theta), \theta' - \theta \rangle + \frac{L}{2}\|\theta' - \theta\|^2$,

(iii) for all $\theta, \theta' \in \mathbb{R}^p$, $\langle \nabla f(\theta') - \nabla f(\theta), \theta' - \theta \rangle \leqslant L\|\theta' - \theta\|^2$

(iv) ∇f is co-coercive: for all $\theta, \theta' \in \mathbb{R}^p$,

$$\|\nabla f(\theta') - \nabla f(\theta)\|^2 \leqslant L\langle \theta' - \theta, \nabla f(\theta') - \nabla f(\theta) \rangle ,$$

(v) ∇f is L-Lipschitz: for all $\theta, \theta' \in \mathbb{R}^p$, $\|\nabla f(\theta') - \nabla f(\theta)\| \leqslant L\|\theta' - \theta\|$.

Exercise 9. We admit the proof that $(i) \Leftrightarrow (ii) \Leftrightarrow (iii)$ could be done as before. This exercise completes the proof of the above proposition.

1. We assume (ii).

(a) Show that

$$f(\theta') + \langle \nabla f(\theta'), \theta + \theta'' - \theta' \rangle \leqslant f(\theta) + \langle \nabla f(\theta), \theta'' \rangle + \frac{L}{2}\|\theta''\|^2 .$$

(b) Isolating all terms involving θ'' on the left-hand side and optimizing over θ'', show that (iv) holds.

2. Conclude on the proof of the above proposition.

Exercice 9

(1) (a) D'après Prop 8 (ii), soit $\theta, \theta'' \in \mathbb{R}^p$,

$$f(\theta''+\theta) \leqslant f(\theta) + \langle \nabla f(\theta), \theta''+\theta-\theta \rangle + \frac{L}{2}\|\theta''+\theta-\theta\|^2$$

$$= f(\theta) + \langle \nabla f(\theta), \theta'' \rangle + \frac{L}{2}\|\theta''\|^2$$

et par la convexité de f,

$\forall \theta' \in \mathbb{R}^p$, $f(\theta''+\theta) \geqslant f(\theta') + \langle \nabla f(\theta'), \theta+\theta''-\theta' \rangle$

alors

$$f(\theta') + \langle \nabla f(\theta'), \theta+\theta''-\theta' \rangle \leqslant f(\theta) + \langle \nabla f(\theta), \theta'' \rangle + \frac{L}{2}\|\theta''\|^2$$

(b)

$$\langle \nabla f(\theta') - \nabla f(\theta), \theta'' \rangle - \frac{L}{2}\|\theta''\|^2 \leqslant f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta-\theta' \rangle$$

Par la règle de Fermat, le terme gauche est optimal à $\theta'' = \frac{1}{L}(\nabla f(\theta') - \nabla f(\theta))$.

Alors

$$\langle \nabla f(\theta') - \nabla f(\theta), \frac{1}{L}(\nabla f(\theta') - \nabla f(\theta)) \rangle - \frac{L}{2}\|\frac{1}{L}(\nabla f(\theta') - \nabla f(\theta))\|^2$$

$$\leqslant f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta-\theta' \rangle$$

日期: /

$$\frac{1}{2L}\|\nabla f(\theta')-\nabla f(\theta)\|^2 \le f(\theta)-f(\theta')-\langle \nabla f(\theta'),\, \theta-\theta'\rangle$$

on échange θ' et θ, alors

$$\frac{1}{2L}\|\nabla f(\theta)-\nabla f(\theta')\|^2 \le f(\theta')-f(\theta)-\langle \nabla f(\theta),\, \theta'-\theta\rangle$$

en sommant les deux,

$$\frac{1}{L}\|\nabla f(\theta')-\nabla f(\theta)\|^2 \le \langle \nabla f(\theta')-\nabla f(\theta),\, \theta'-\theta\rangle$$

$$\|\nabla f(\theta')-\nabla f(\theta)\|^2 \le L\cdot\langle \nabla f(\theta')-\nabla f(\theta),\, \theta'-\theta\rangle$$

(iv) $\Rightarrow$ (v) Par le C-S

$$\|\nabla f(\theta')-\nabla f(\theta)\|^2 \le L\cdot\langle \nabla f(\theta')-\nabla f(\theta),\, \theta'-\theta\rangle$$

$$\le L\cdot\|\nabla f(\theta')-\nabla f(\theta)\|\cdot\|\theta'-\theta\|$$

alors

$$\|\nabla f(\theta')-\nabla f(\theta)\| \le L\cdot\|\theta'-\theta\|$$

日期: /