Exercise 1 (Dual of the logistic regression problem).
We consider the following logistic regression problem with ridge regularization given by

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} \log\left(1 + \exp(-y_i a_i^\top x)\right) + \frac{\lambda}{2}\|x\|^2 \tag{1}$$

where for all $i \in \{1, \dots, m\}$, $y_i \in \{-1, 1\}$ and $a_i \in \mathbb{R}^n$.

1. Let h such that $\forall u \in \mathbb{R}$, $h(u) = \log(1 + \exp(u))$. Show that for all u, $h'(u) \in [0, 1]$.

1.1 $h'(u) = \dfrac{(1+\exp(u))'}{1+\exp(u)} = \dfrac{\exp(u)}{1+\exp(u)}$

$\exp u > 0 \ \forall u \in \mathbb{R} \Rightarrow 1+\exp(u) > 0 \ \forall u \in \mathbb{R}$

$\Rightarrow \dfrac{\exp u}{1+\exp u} > \dfrac{0}{1+\exp(u)} = 0$

$\forall u \in \mathbb{R}, \ 0 < 1$

$\Rightarrow \exp u < 1 + \exp u, \ 1+\exp u > 0$, thus:

$\Rightarrow \dfrac{\exp u}{1+\exp u} < 1$

$\Rightarrow 0 < h'(u) < 1 \Rightarrow h'(u) \in]0,1[\ \forall u \in \mathbb{R}$

$\forall u \in \mathbb{R}$

$\Rightarrow h'(u) \in [0,1] \ \forall u \in \mathbb{R}$

2. Show that h is convex.

$\forall u \in \mathbb{R}, h''(x) \geq 0 \to h(u)$ is convex

$$h''(u) = \frac{e^u(e^u+1) - e^u e^u}{(e^u+1)^2}$$

$$h''(u) = \frac{e^u}{(e^u+1)^2}$$

Considering both the numerator and denominator are strictly positive, then h''(u) is also strictly positive, thus h(u) is convex.

We define the Fenchel transform of h by

$$\forall \alpha \in \mathbb{R}, h^*(\alpha) = \sup_{u \in \mathbb{R}} \alpha u - h(u) .$$

3. Let $\alpha \in]0, 1[$. Show that

$$h^*(\alpha) = C_1(\alpha) \log(\alpha) + C_2(\alpha) \log(1 - \alpha)$$

where $C_1(\alpha)$ and $C_2(\alpha)$ should be explicited.

$$\frac{\partial(\alpha u - ln(1 + e^u))}{\partial u} = 0$$

$$\alpha - \frac{e^u}{e^u + 1} = 0$$

$$u = ln(\frac{\alpha}{1 - \alpha})$$

Substituting into the initial equation to determine the supremum:

$$\alpha ln(\frac{\alpha}{1 - \alpha}) - ln(1 + \frac{\alpha}{1 - \alpha})$$

$$\alpha ln(\alpha) + (1 - \alpha) ln(1 - \alpha)$$

$$C_1 = \alpha \; C_2 = 1 - \alpha$$

4. Let $\alpha \notin [0, 1]$. Show that $h^*(\alpha) = +\infty$.

(4) $a > 1$: $\lim_{u \to \infty} au - \log(1+e^u) = \lim_{u \to \infty} au\left(1 - \frac{\log(1+e^u)}{au}\right) \overset{\text{L'Hopital}}{=}$

$\lim_{u \to \infty} au\left(1 - \frac{e^u}{1+e^u} \cdot \frac{1}{a}\right) \underset{a>1}{=} +\infty$

$a < 0$: $\lim_{u \to -\infty} au - \cancel{\log(1+e^u)}^{\to 0} = +\infty$

5. Show that $h^*(0) = h^*(1) = 0$.

When analyzing αu - h(u), for α = 0 it is possible to see that its derivative is strictly negative, so it always decreases. Therefore its maximum value is at -inf, its limit can be evaluated as 0. Same can be done for when α=1, whose derivative is strictly positive. At inf its limit is also 0. So their supremum is the same and equal to 0.

lim de xlog(x) = > log(x)/(1/x), l'hôpital = -x =>0

6. Give a convex function f and a matrix M such that for all $x \in \mathbb{R}^n$,

$$\sum_{i=1}^{m} \log\left(1 + \exp(-y_i a_i^\top x)\right) = f(Mx)$$

Define the matrix multiplication to have the minus signal too, so we can use the fact that log(1 + exp(u)) is convex:

$M = [-y_1a_1^T, -y_2a_2^T, \dots, -y_ma_m^T]^T$

$\implies Mx = [-y_1a_1^Tx, -y_2a_2^Tx, \dots, -y_ma_m^Tx]^T$

Given x = [$x_1, \dots, x_m$]^T, f(x) = $\sum_{i=1}^{m} log(1 + exp(x_i))$ which is convex since log(1 + exp(u)) is convex and the sum of convex functions is also convex.

Thus we have f(Mx) = $\sum_{i=1}^{m} log(1 + exp(-y_i a_i^T x))$

7. By introducing a new variable $z \in \mathbb{R}^m$ and a linear constraint, define a problem equivalent to (1) of the form

$$\min_{x\in\mathbb{R}^n, z\in\mathbb{R}^m} f(z) + g(x) \qquad (2)$$
$$\text{st} : Mx = z$$

Please explicit the functions f and g and the matrix M.

f, M are the same as item 1.6 and $g(x) = \frac{\lambda}{2}||x||^2$

8. Write the Lagrangian associated to Problem (2).

L(x, z, ϕ) = f(z) + g(x) + <ϕ, Mx-z> = g(x) + <ϕ, Mx> + f(z) - <ϕ, z>

9. Calculate the dual function.

$$D(\Phi) = \inf_{x,z} L(x, z, \Phi)$$
$$D(\Phi) = \inf_{z} f(z) - \langle \Phi, z\rangle + \inf_{x} g(x) + \langle \Phi, Mx\rangle$$

$\Phi_M'(z) = M^T(Ax - B)$

$$F(x,z) = f(x) + g(z)$$

For $A \to A\binom{x}{z} = \underbrace{Mx}_{z} - z = 0 \to A = (M, -1)$

$\min f(x) + g(Mx) = \min F(x,z)$ et $A\binom{x}{z} = 0$ or $F(\alpha)$ and $A\alpha = 0$

$$\mathcal{L}(\alpha, \lambda) = F(\alpha) + \langle \lambda, A\alpha \rangle = f(x) + g(z) + \langle \lambda, Mx - z \rangle$$

Dual $\Phi(\lambda) = \inf \mathcal{L}(\alpha, \lambda)$

$$= -\sup_x \langle -A^T\lambda, \alpha \rangle - f(x) - g(z) = \sup_x \langle -A^T\lambda, \alpha \rangle - F(\alpha)$$

$$= -F^*(-A^T\lambda)$$

Exercise 2 (Subgradient method)**.**
Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. We assume that f has a minimizer x^*.
The subgradient method is the following algorithm that starts at $x^0 \in \mathbb{R}^n$ and for all $k \in \mathbb{N}$:

$$\text{Take } g_k \in \partial f(x_k)$$
$$x_{k+1} = x_k - \gamma_k g_k \,.$$

The sequence $(\gamma_k)_k$ is such that $\gamma_l > 0$ for all l, $\sum_{k=0}^{+\infty} \gamma_k = +\infty$ and $\lim_{N\to\infty} \frac{\sum_{k=0}^N \gamma_k^2}{\sum_{k=0}^N \gamma_k} = 0$.

1. Using the definition of the subgradient, show that for all $x \in \mathbb{R}^n$ and $g \in \partial f(x)$, we have
$$f(x+g) \geq f(x) + \langle g, g \rangle$$

2.1 $x \in \mathbb{R}^n$ and $x+g \in \mathbb{R}^n \Rightarrow x \in \mathrm{dom}(f)$ and $x+g \in \mathrm{dom}(f)$
$g \in \partial f(x)$, thus by definition:
$\forall y \in \mathbb{R}^n$, $f(y) - f(x) \geq \langle g, y - x \rangle$
in particular, for $y = x + g$ we have:
$$f(x+g) - f(x) \geq \langle g, x + g - x \rangle$$
$$f(x+g) \geq f(x) + \langle g, g \rangle \quad \blacksquare$$

2. Suppose that f is L-Lipschitz ($|f(x) - f(y)| \leq L \|x - y\|$). Show that there exists $M > 0$ such that for all $x \in \mathbb{R}^n$ and for all $g \in \partial f(x)$, we have $\|g\|_2 \leq M$.

$f(x+g) \geq f(x) + \langle g, g \rangle \Rightarrow f(x+g) - f(x) \geq \langle g, g \rangle$

$|f(x+g) - f(x)| \leq L \|g\| \Rightarrow \|g\|_2^2 \leq L \|g\|$

$\|g\| \leq L$

How do you go from the first line to the second one?

L||g|| >= |f(x+g)-f(x)| >= ||g||^2

f is L-Lipschitz

3. Conversely, suppose that there exists $M \geq 0$ such that for all $x \in \mathbb{R}^n$ and for all $g \in \partial f(x)$, $\|g\|_2 \leq M$. Show that f is Lipschitz continuous.

(just need to mention that for the same x,y in the second part)

$\forall x \in \mathbb{R}^n \wedge \forall g \in \partial f(x), ||g||_2 \leq M$

$\Longrightarrow$ Given $x \in \mathbb{R}^n, y \in \mathbb{R}^n, g_1 \in f(x)$,

$f(y) - f(x) \geq \langle g_1, y - x \rangle \implies f(x) - f(y) \leq \langle g_1, x - y \rangle \leq ||g_1||_2 ||x - y||_2 \leq M||x - y||_2$

$\Longrightarrow f(x) - f(y) \leq M||x - y||_2$

Also, by choosing the opposite order (ie: x = y and y = x)

Given $y \in \mathbb{R}^n, x \in \mathbb{R}^n, g_2 \in \partial f(y)$,

$f(x) - f(y) \geq \langle g_2, x - y \rangle \implies f(y) - f(x) <= \langle g_2, y - x \rangle \leq ||g_2||_2 ||y - x||_2 \leq M||y - x||_2 = M||x - y||_2$

$\Longrightarrow f(y) - f(x) \leq M||x - y||_2$

Finally,

$max(f(x) - f(y), f(y) - f(x)) = |f(x) - f(y)| \leq M||x - y||_2$

Which implies f is Lipschitz continuous

For the rest of the exercise, we assume that f is L-Lipschitz.

4. Find $\beta(k)$ such that for all $k \in \mathbb{N}$,

$$\frac{1}{2} \|x_{k+1} - x^*\|_2^2 = \frac{1}{2} \|x_k - x^*\|_2^2 + \gamma_k \langle g_k, x_* - x_k \rangle + \beta(k) \|g_k\|_2^2 .$$

$x_{k+1} = x_k - \gamma \nabla f(x_k)$

$$\frac{1}{2}\|\underbrace{x_k - \gamma\nabla f(x_k) - x^*}_{\alpha-\beta}\|^2 = \frac{1}{2}\left(\|\gamma\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - 2\langle \gamma\nabla f(x_k), x_k - x^*\rangle\right)$$

Using $\|x-y\|^2 = \|x-z\|^2 + \|y-z\|^2 - 2\langle x-z, z-y\rangle$

$\alpha - z = -\gamma\nabla f(x_k)$

$\alpha - \beta = x_k - \gamma\nabla f(x_k) - x^*$

$\beta - z = -x_k + x^*$

$$= \frac{1}{2}\left(\|x_k - x^*\|^2 + \gamma^2\|g_k\|_2^2 + 2\gamma\langle g_k, x^* - x_k\rangle\right)$$

$\beta(k)\|g_k\|_2^2 = \gamma^2\|g_k\|_2^2 \to \beta(k) = \gamma_k^2$

I think there is ½ missing : beta(k)=½ (gamma_k)^2

could have done || (x_k-x^*) - \gamma_k g_k||²

5. Show that for all $k \in \mathbb{N}$, $f(x^*) \geq f(x_k) + \langle g_k, x_* - x_k\rangle$
6. Deduce from this inequality that

$$\gamma_k\big(f(x_k) - f(x_*)\big) \leq \frac{1}{2}\|x_k - x^*\|_2^2 - \frac{1}{2}\|x_{k+1} - x^*\|_2^2 + \beta(k)M^2$$

5-6:

$\beta(k)\|g_k\|_2^2 = \gamma^2\|g_k\|_2^2 \to \beta(k) = \gamma_k^2$

$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$

$f(x^*) \geq f(x) + \langle g_k, x^* - x\rangle$

$\frac{1}{2}\left(\|x_{k+1} - x^*\|^2 - \|x_k - x^*\|^2\right) - \beta(k)\|g_k\|^2 = \gamma\langle g_k, x^* - x_k\rangle$

$\gamma(f(x^*) - f(x_k)) \geq \gamma\langle g_k, x^* - x_k\rangle$ and $\|g_k\| \leq M$

$\gamma(f(x_k) - f(x^*)) \leq \frac{1}{2}\|x_k - x^*\|^2 - \frac{1}{2}\|x_{k+1} - x^*\|^2 + \beta(k)M^2$

7. Show that

$$\sum_{l=0}^{k} \gamma_l \big(f(x_l) - f(x_*)\big) \leq \frac{1}{\sum_{l=0}^{k} \gamma_l} \Big(\frac{1}{2} \|x_0 - x_*\|_2^2 + \sum_{l=0}^{k} \beta(l) M^2\Big)$$

8. Denote $\bar{x}_k = \frac{1}{\sum_{l=0}^{k} \gamma_l} \sum_{j=0} \gamma_j x_j$. Using the fact that $\bar{x}_k$ is a convex combination of the previous iterates, find a bound on $f(\bar{x}_k) - f(x_*)$.

9. Using the properties of the sequence (γ_k), show that $f(\bar{x}_k)$ converges to $f(x_*)$.

Exercise 3 (ℓ_1 regression)**.**
In this exercise we study the following ℓ_1 regression problem

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_1 \tag{3}$$

where $A = \begin{bmatrix} a_1^\top \\ \vdots \\ a_m^\top \end{bmatrix}$ is a $m \times n$ matrix and $b \in \mathbb{R}^m$.
As the objective is affine by parts and lower bounded, it has at least one minimizer x_* (you do not have to show this fact).
The objective function is affine by parts, so it could be dealt with by linear optimization. However, we are going to show that algorithms based on subgradients can also solve this problem.

1. We define $f : \mathbb{R}^n \to \mathbb{R}$ such that $f(x) = \|Ax - b\|_1 = \sum_{j=1}^{n} |a_j^\top x - b_j|$. Is f convex? differentiable? separable?

It is convex as it is the norm 1 of a linear function, it is not differentiable as it is the norm1
Une Fonction est séparable s'il s'écrit sous la forme

$z < 1$
$z = 1$
$z > 1$

$_1, \ldots, h_p$ tels que $H(v) = \sum_{i=1}^{p} h_i(v_i)$.

$\partial h_p(v_p)$.

Du coup f(x) n'est pas séparable par rapport aux variable x = (x1 , x2 , x3 , ………. , xn)

"

2. For $z \in \mathbb{R}$, calculate $\phi(z) = \max_{u \in [-1,1]} uz$.

the max = Z, when u = sign(z)

3. Express $\max_{u \in [-1,1]^m} \langle u, Ax - b \rangle$ as a function of x.

max = abs(Ax-b) ?how? You can write the sum, it is max if each term is positive

$= -F^*(-A'\lambda)$

$\max \langle u, Ax-b \rangle = \max \sum_{i}^{m} u_i (Ax-b)_i$

If z equal to 4, we fall on ex 3,2

So $u_i = \text{sign}(Ax-b)_i$

4. Let $u(x)$ be any element of $\arg\max_{u \in [-1,1]^m} \langle u, Ax - b \rangle$.
Show that this $\arg\max$ is never empty and that $f(x) = \langle u(x), Ax - b \rangle$.

As arg max = sign(Ax-b), donc <u(x),Ax-b> = <sign(Ax-b),Ax-b> = abs(Ax-b)

5. Show that $A^\top u(x) \in \partial f(x)$.

As we know, subgradient de f(x) = subgradient(abs) * subgradient(Ax-b), subgradient (abs) = u(x), and subgradient(Ax-b) = At => A.T*u(x)

6. Show that $\left\|A^\top u(x)\right\|_2 \leq \sqrt{m}\,\|A\|$ where $\|A\|$ is the operator norm of A.

Using the definition of subgradients, f(y) - f(x) >= <At*u(x), y-x>
(i) with y = x + At*u(x) : ||At*u(x)||^2 <= f(x + At*u(x)) - f(x)

(ii) f(x + At*u(x)) <= ||A*At*u(x)|| + f(x), using the "inégalité triangulaire"

(iii) ||A*At*u(x)|| = ||A||*||A||*||u(x)|| <= m||A||^2

above proof they less than m||A||^2 -> proof less than(m)^2||A||^2 below
By definition of Matrix Norm, we will have ||A.x|| <= ||A||.||x||
→ ||A^T.u(x)|| <= ||u(x)||.||A^T|| = ||u(x)||.||A|| (norm 2 of matrix, ||A|| = ||A^T||). Because uj(x) = sign(...) → ||u(x)|| = sqrt(m)
this proves ||At.x||<=m.||A||
(the definition of Matrix norm here, image below):

We can very easily build matrix norms from vectorial norms: they are then called *subordinate matrix norms*. For this, we can define $||A||$ by the following equivalent formulas:

$$||A|| = \sup_{x\neq 0}\frac{||Ax||}{||x||} = \sup_{||x||=1}||Ax|| = \sup_{0<||x||\leqslant 1}\frac{||Ax||}{||x||}.$$

We have: $||Ax|| \leqslant ||A||\ ||x||$.

The matrix norms subordinate to the most usual norms that we have described above are therefore, for $A = (a_{i,j})_{\substack{1\leqslant i\leqslant n\\ 1\leqslant j\leqslant n}}$:

- $||A||_1 = \sup_{x\neq 0}\frac{||Ax||_1}{||x||_1} = \max_{1\leqslant j\leqslant n}\sum_{i=1}^{n}|a_{ij}|$
- $||A||_2 = \sup_{x\neq 0}\frac{||Ax||_2}{||x||_2} = \sqrt{\rho(A^*A)} = ||A^*||_2$ where $\rho(A^*A)$ represents the largest absolute value of A^*A (spectral radius of A^*A)
- $||A||_\infty = \sup_{x\neq 0}\frac{||Ax||_\infty}{||x||_\infty} = \max_{1\leqslant i\leqslant n}\sum_{j=1}^{n}|a_{ij}|.$

The norm $||\ ||_2$ is invariant by unitary transformation: if U is a unitary matrix, i.e., if U fulfils the relation $U^*U = I$, then we have

$$||A||_2 = ||AU||_2 = ||UA||_2 = ||U^*AU||_2.$$