Exercise 1 (Picard's fixed point theorem)**.**
Prove the following theorem:

If $T : \mathbb{R}^d \to \mathbb{R}^d$ satisfies

$$\exists 0 < \rho < 1, \forall x \in \mathbb{R}^d, \forall y \in \mathbb{R}^d, \|T(x) - T(y)\| \leq \rho \|x - y\|$$

then T has a unique fixed point x^* such that $x^* = T(x^*)$.

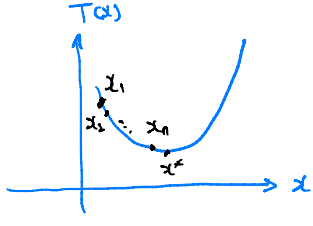
Moreover, every sequence of the form $x_{k+1} = T(x_k)$ converges to x^* with a linear convergence rate given by $\|x_k - x^*\| \leq \rho^k \|x_0 - x^*\|$.

Th: Picard's fixed point $\longrightarrow$ minimizer function

$\|T(x) - T(y)\| \leq \rho \|x - y\|, \quad 0 < \rho < 1$

sequence (x_n) satisfied $x_{n+1} = T(x_n)$

$T(x^*) = x^*$



Pf: unicité

$\exists\, x^*, y^*$, 2 points fixed.

$$\begin{cases} x^* = T(x^*) \\ y^* = T(y^*) \end{cases}$$

$\|x^* - y^*\| = \|T(x^*) - T(y^*)\| \leq \rho \|x^* - y^*\|$

$\Rightarrow 1 \leq \rho$ which is a contradiction of $0 < \rho < 1$

therefore $x^* = y^*$

existance

$(x_n)_n \xrightarrow{n \to \infty} x_\infty$

$$\|T(x_n) - T(y_n)\| \leq p\|x_n - y_n\|$$

$$\leq \dots$$

$$\leq p^n \|x_0 - y_0\| \xrightarrow{n\to\infty} 0 \quad (0 < p < 1)$$

Soit $k \in \mathbb{N}$, avec x_0 , $y_0 = T^k(x_0)$

$\downarrow$ $\downarrow$

x_n x_{n+k}

$$\|T(x_n) - T(x_{n+k})\| \xrightarrow{n\to\infty} 0$$

(x_n) is a sequence of Cauchy.

Therefore, $\exists\, x^*$ is a limite of (x_n).

$$x^* = \lim_{n\to\infty} (x_n)$$

For the continuity: $T\left(\lim_n (x_n)\right) = \lim_n T(x_n)$

$$= \lim_n x_{n+1}$$

$$= T(x^*).$$

Exercise 2 (Gradient calculus).

- Calculate the gradient of the following functions. A, M and Q are fixed matrices, b is a fixed vector. f_1 is useful for least squares and regression problems, f_2 is useful for logistic regression and binary classification, f_2 is useful for nonnegative matrix factorization.

$$f_1 : \mathbb{R}^n \to \mathbb{R}$$
$$x \mapsto \frac{1}{2}\|Ax - b\|^2 = \frac{1}{2}\sum_{i=1}^{m}\Big(\sum_{j=1}^{n} A_{ij}x_j - b_i\Big)^2$$
$$f_2 : \mathbb{R}^n \to \mathbb{R}$$
$$z \mapsto \sum_{i=1}^{n} \log(1 + \exp(z_i))$$
$$f_3 : \mathbb{R}^{m\times p} \to \mathbb{R}$$
$$P \mapsto \frac{1}{2}\|M - PQ\|_F^2 = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}(M_{ij} - \sum_{k=1}^{p} P_{ik}Q_{kj})^2$$

- Let g_1, g_2, g_3 be functions such that $g_1 : \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$, $g_2 : \mathbb{R}^{n_2} \to \mathbb{R}^{n_3}$, $g_3 : \mathbb{R}^{n_3} \to \mathbb{R}$ and let

$$f_4 = g_3 \circ g_2 \circ g_1 \,.$$

Compute the gradient of f_4 using the Jacobian matrices of g_i for $i \in \{1, 2, 3\}$.

Suppose that computing one element of the Jacobian matrices costs C_J and that multiplying two numbers costs C_M. How much does it cost to compute $\nabla_4 f(x)$?

How to calculate the gradient ? $f: \mathbb{R}^m \to \mathbb{R}$

$f(x+h) = f(x) + h^T \nabla f(x) + o(h)$

$\langle \nabla f(x), h \rangle$

- $f(x+h) = f(x) + D[f(x)]\cdot h + o(h)$

linear function

$f: \mathbb{R}^m \to \mathbb{R}^d$

$f(x+h) = f(x) + [J_x]_{d\times n}\, h + o(h)$

$\nabla f(x) = [J_x]$

$\dfrac{\|f(x+h) - f(x)\|}{\|h\|} \xrightarrow{\|h\|\to 0} 0$

- Chain Rule

$D(f \circ g)(x) = D(f)(g(x)) \cdot D(g)(x)$

$J_x(f \circ g) = J^f_{g(x)} \cdot J^g_x$

- $f_1 : \mathbb{R}^n \to \mathbb{R}$

$$x \mapsto \frac{1}{2}\|Ax-b\|^2 = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}(A_{ij}x_j - b_i)^2$$

$$f_1(x+h) = \frac{1}{2}\|Ax-b+Ah\|^2 = \frac{1}{2}\|Ax-b\|^2 + \langle Ax-b, Ah\rangle + \frac{1}{2}\|Ah\|^2$$

$$\|a+b\|^2 = \|a\|^2 + 2\langle a, b\rangle + \|b\|^2$$

$$= f_1(x) + \langle h, A^T(Ax-b)\rangle + o(h)$$

Thus, $\nabla f_1(x) = A^T(Ax-b)$.

$$= (Ah)^T(Ax-b) = h^T\left[A^T(Ax-b)\right]$$

- $f_2 : \mathbb{R}^n \to \mathbb{R}$

$$z \mapsto \sum_{i=1}^{n}\log(1+e^{z_i})$$

This can be calculated directly.

$$\frac{\partial f_2(z_i)}{\partial z_i} = \frac{e^{z_i}}{1+e^{z_i}}, \quad \nabla f_2(z) = \left[\frac{e^{z_i}}{1+e^{z_i}}\right]_{n\times 1}$$

- $f_3 : \mathbb{R}^{m\times p} \to \mathbb{R}$

$$P \mapsto \frac{1}{2}\|M-PQ\|_F^2 = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(M_{ij} - \sum_{k=1}^{p}P_{ik}Q_{kj}\right)^2$$

$g : P \to M-PQ$

$Dg(P) = -Q$, $Dg : h \mapsto -hQ$

$D(f\circ g) = D(f)(g(x)) \circ D(g)(x)$

$\langle a, bQ\rangle = \langle aQ^T, b\rangle$

$$= \langle g(x), D(g)(x)\rangle = \langle M-xQ, -hQ\rangle$$

$$\nabla f_3(x) = \langle (xQ-M)Q^T, h\rangle$$

- $f_4 = g_3 \circ g_2 \circ g_1$, $g_1 : \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$, $g_2 : \mathbb{R}^{n_2} \to \mathbb{R}^{n_3}$, $g_3 : \mathbb{R}^{n_3} \to \mathbb{R}$

$$D(f_4)(x) = D(g_3 \circ g_2)(g_1(x)) \circ D(g_1)(x)$$

$$= D(g_3)(g_2(x))(g_1(x)) \circ D(g_2)(x)(g_1(x)) \circ D(g_1)(x)$$

$$= D(g_3)(g_2 \circ g_1)(x) \circ D(g_2 \circ g_1)(x) \circ D(g_1)(x)$$

$$= D(g_3)(g_2 \circ g_1)(x) \circ \left(D(g_2 \circ g_1) \cdot D(g_1)\right)(x)$$

$$= \langle \nabla g_3(g_2(g_1(x))),\ J_{g_2}(g_1(x))\, J_{g_1}(x)\, h \rangle$$

$$= \langle J_{g_1}^T(x)\, J_{g_2}^T(g_1(x))\, \nabla g_3(g_2(g_1(x))),\ h \rangle$$

$n_1 \times n_2$ $\quad$ $n_2 \times n_3$ $\quad$ $n_3 \times 1$

⚠ $A_{m \times n} \times B_{n \times p} \to C_{m,n,p} = mnp\, C_M$

If we calculate from right to left:

$$C_{n_2 n_3 1} + C_{n_1 n_2 1} = (n_2 n_3 + n_1 n_2)\, C_M$$

instead of from left to right:

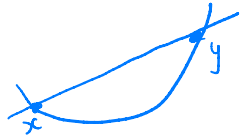
$$C_{n_1 n_2 n_3} + C_{n_1 n_3 1} = (n_1 n_2 n_3 + n_1 n_3)\, C_M$$

thus the total cost:

$$(n_1 n_2 + n_2 n_3 + n_3)\, C_J + (n_1 n_2 + n_2 n_3)\, C_M$$

What is a convex function ?

$$\forall x, y. \quad f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y)$$



- **fonction semi-continué** ssi $(x_n) \to x$

continué : $\lim_{n \to \infty} f(x_n) = f\left(\lim_{n \to \infty} x_n\right)$

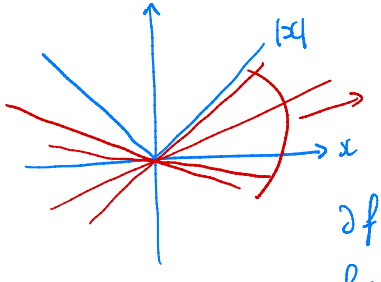
semi-continué inférieur :

$$\lim_{n \to \infty} \inf f(x_n) \geq f\left(\lim_{n \to \infty} x_n\right)$$

e.g. : $\iota_C(x) = \begin{cases} +\infty, & x \notin C \\ 0, & x \in C \end{cases}$

- **subgradient**

$x \mapsto |x|$



subgradient $\partial |\cdot|(0) = [-1, 1]$

$$\partial f(x) = \{q : \forall y, f(y) \geq f(x) + \langle q, y - x \rangle\}$$

f is differentiable : $\partial f(x) = \{\nabla f(x)\}$

Th Fermat :

x minimize $f \iff 0 \in \partial f(x)$

$\iff f(y) \geq f(x) + \langle 0, y-x \rangle = f(x)$

Exercise 3 (Convergence of gradient descent for strongly convex C^2 functions)**.**
Consider a C^2 function $f : \mathbb{R}^d \to \mathbb{R}$ such that $0 \prec \mu I \preceq \nabla^2 f(x) \preceq LI$.

1. Show that the fixed point operator $T : x \mapsto x - \gamma \nabla f(x)$ is contractant for any $0 < \gamma < \frac{2}{L}$.
2. Show that the gradient method converges linearly.
3. How many iterations are necessary to ensure that $\|x_k - x^*\| \leq \epsilon$?

Exercise 4 (Proximal operator)**.**
Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex lower-semicontinuous function such that $\operatorname{dom} f \neq \emptyset$.

1. Recall the definition of the domain of a convex function.
2. It is possible to prove (but we do not ask you to do it) that $\exists x_0 \in \operatorname{dom} f$ such that $\exists q_0 \in \partial f(x_0)$. Using this information, show that there exists $\alpha \in \mathbb{R}$ and $w \in \mathbb{R}^n$ such that for all x, $f(x) \geq \alpha + \langle w, x \rangle$.
3. Let us fix $x \in \mathbb{R}^n$. Let us define $g : y \mapsto f(y) + \frac{1}{2}\|x - y\|^2$. Show that g is strongly convex.
4. Show that $\lim_{\|y\| \to +\infty} g(y) = +\infty$.
5. Show that g has a minimizer and that it is unique.

We will denote this minimizer as $\operatorname{prox}_f(x)$. The function $\operatorname{prox}_f : \mathbb{R}^n \to \mathbb{R}^n$ is called the proximal operator of f.

1. $\operatorname{dom} f = \{x : f(x) < +\infty\}$.

2. $q \in \partial f(x) \iff \forall y,\ f(y) \geq f(x) + \langle q, y - x \rangle$

$\iff f(y) \geq \underbrace{f(x) + \langle q, -x \rangle}_{\alpha} + \underbrace{\langle q, y \rangle}_{\langle w, x \rangle}$

3. Strongly convex : $f - \frac{\mu}{2}\|\cdot\|^2$ is convex.

Let $h(y) = g(y) - \frac{1}{2}\|y\|^2$

$= f(y) - \langle x, y\rangle + \frac{1}{2}\|x\|^2$

is convex. (sum of the convex functions).

thus g is strongly convex.

4. $f(y) \geq \alpha + \langle w, y\rangle$

$g(y) \geq \underbrace{\boxed{\alpha + \langle w, y\rangle}}_{\text{linear}} + \frac{1}{2}\|x-y\|^2 \xrightarrow{\|y\|\to\infty} +\infty$

5. $g(y) = f(y) + \frac{1}{2}\|x-y\|^2$ which is the sum of l.s.c.

then $g(y)$ is l.s.c. and $\exists\, x^*$ s.t. g minimize.

(prop. 2.2.2).

Suppose x_1, x_2 are two minimises,

$g\left(\frac{x_1+x_2}{2}\right) \leq \frac{1}{2}g(x_1) + \frac{1}{2}g(x_2) - \frac{1}{8}\|x_1 - x_2\|^2 \quad (\lambda = \frac{1}{2})$.

$\min g \leq \frac{1}{2}\min g + \frac{1}{2}\min g - \frac{1}{8}\|x_1 - x_2\|^2$.

$\Rightarrow x_1 = x_2$.

Exercise 5. Let us denote

$$\mathrm{prox}_g(y) = \arg\min_{x \in \mathcal{X}} g(x) + \frac{1}{2}\|x - y\|^2$$

the proximal operator of g at y.
Fix $\gamma > 0$. Show that if f and g are convex, then the fixed points of the nonlinear equation

$$x = \mathrm{prox}_{\gamma g}(x - \gamma \nabla f(x))$$

are the minimizers of the function $F = f + g$.

$z = \mathrm{prox}_g(y) \iff 0 \in \partial\left(g + \frac{1}{2}\|\cdot - y\|^2\right)(z)$

Si $x \in$ intérieur relatif de $\mathrm{dom}\, g$, f, g convexes.

alors : $\partial(f+g)(x) = \partial f + \partial g$

$\partial f + \partial g = \{\nabla f + q,\ q \in \partial g\}$

$\partial\left(g + \frac{1}{2}\|\cdot - y\|^2\right)(z) \ni 0 \iff \left(\partial g + (z - y)\right) \ni 0$

$\nabla\left(\frac{1}{2}\|\cdot - y\|^2\right) = (\cdot - y) \iff (\partial g + I)(z) \ni y$

$\iff z \in (\partial g + I)^{-1}(y)$

Valeur minimisé est seule $\Rightarrow z = (\partial g + I)^{-1}(y)$

$\mathrm{prox}_g = (\partial g + I)^{-1}$.

x minimiser $f+g \Rightarrow 0 \in \partial(f+g)(x)$

$\iff 0 \in \partial f + \partial g$

$\partial(\gamma g)(x) + x - (x - \gamma \nabla f(x)) \ni 0$

$\Leftrightarrow \partial(\gamma g)(x) + \gamma \nabla f(x) \ni 0$

$\Leftrightarrow \gamma \partial g(x) + \gamma \nabla f(x) \ni 0$

$\Leftrightarrow \partial g(x) + \nabla f(x) \ni 0$ which the original function is

$F = f + g$.

Exercise 6 (Taylor-Lagrange inequality)**.** The goal of this exercise is to prove Taylor-Lagrange inequality. This is a fundamental inequality for the study of gradient descent and related methods.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function whose gradient is L-Lipschitz continuous *i.e.* $\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$ for all x, y.

1. Prove that for all x, y, $\langle \nabla f(y) - \nabla f(x), y - x\rangle \leq L\|y - x\|^2$.

2. Set $\varphi(t) = f(x + t(y - x))$ for all $t \in [0, 1]$. Prove that
$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle = \varphi(1) - \varphi(0) - \varphi'(0)\,.$$

3. Deduce that
$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle = \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x\rangle dt$$

4. Using the first question, conclude that
$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2\,.$$

1. $|\langle \nabla f(y) - \nabla f(x),\ y - x \rangle| \leq \|\nabla f(y) - \nabla f(x)\| \cdot \|y - x\|$

$\leq L \|y - x\|^2$

2. $\varphi(t) = f(x + t(y - x))$. $\varphi(0) = f(x)$. $\varphi(1) = f(y)$

$g : t \mapsto x + t(y - x)$. $\varphi(t) = (f \circ g)(t)$

$D\varphi(t) = D(f)g(t) \circ D\,g(t)$

$= \langle \nabla f(g(t)),\ y-x \rangle$, $\varphi'(0) = \langle \nabla f(x),\ y-x \rangle$.

thus: $f(y) - f(x) - \langle \nabla f(x),\ y-x \rangle = \varphi(1) - \varphi(0) - \varphi'(0)$

3. $\varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)\,dt = \int_0^1 \langle \nabla f(g(t)),\ y-x \rangle dt$

$= \int_0^1 \langle \nabla f(x + t(y-x)),\ y-x \rangle dt$

$\varphi'(0) = \langle \nabla f(x),\ y-x \rangle$.

thus, $\varphi(1) - \varphi(0) - \varphi'(0) = \int_0^1 \langle \nabla f(x+t(y-x)),\ y-x \rangle - \langle \nabla f(x),\ y-x \rangle dt$

$= \int_0^1 \langle \nabla f(x+t(y-x)) - \nabla f(x),\ y-x \rangle dt$

4. $f(y) - f(x) - \langle \nabla f(x),\ y-x \rangle$

$= \int_0^1 \langle \nabla f(x+t(y-x)) - \nabla f(x),\ y-x \rangle dt$

$= \int_0^1 \frac{1}{t} \langle \nabla f(x+t(y-x)) - \nabla f(x),\ t(y-x) \rangle dt$

$\le \int_0^1 \frac{1}{t} \cdot L\, t^2 \|y-x\|^2 dt$

$= \frac{L}{2} \|y-x\|^2$

Exercise 15 (Proximal gradient for logistic regression)**.**
We consider a classification problem defined by observations $(x_i, y_i)_{1\le i\le n}$ where for all i, $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$. We propose the following linear model for the generation of the data. Each observation is supposed to be independent and there exists a vector $w \in \mathbb{R}^p$ and $w_0 \in \mathbb{R}$ such that for all i, (y_i, x_i) is a realization of the random variable (Y, X) whose law satisfies

$$\mathbb{P}_{w,w_0}(Y = 1|X) = \frac{\exp(X^\top w + w_0)}{1 + \exp(X^\top w + w_0)}\,.$$

1. Show that $\forall i \in \{1, \dots, n\}$, $\mathbb{P}(Y_i = y_i|x_i) = \dfrac{1}{1 + \exp(-y_i(x_i^\top w + w_0))}$.
2. Show that the maximum likelihood estimator is

$$(\hat{w}, \hat{w}_0) = \arg\min_{w,w_0} \sum_{i=1}^{n} \log(1 + \exp(-y_i(x_i^\top w + w_0)))$$

3. Denote $f(w, w_0) = \sum_{i=1}^n \log(1 + \exp(-y_i(x_i^\top w + w_0)))$. Compute $\nabla f(w, w_0)$.
4. Compute the proximal operator of $(x \mapsto \frac{\lambda}{2}\|x\|^2)$ for $\lambda > 0$.
5. Write the proximal gradient method for the logistic regression problem with ridge regularizer

arg min f + g

$$(\hat{w}^{(\lambda)}, \hat{w}_0^{(\lambda)}) = \arg\min_{w,w_0} \underbrace{\sum_{i=1}^{n} \log(1 + \exp(-y_i(x_i^\top w + w_0)))}_{f} + \underbrace{\frac{\lambda}{2}\|w\|^2}_{g}\,.$$

6. Compute the Hessian matrix and write Newton's method for the same problem.

1. $y_i = \{-1, 1\}$.

if $y_i = 1$, $\mathbb{P}(Y = y_i = 1 | X) = \dfrac{1}{1 + \exp(-(X^T w + w_0))}$

if $y_i = -1$, $\mathbb{P}(Y = y_i = -1 | X) = 1 - \mathbb{P}(Y = y_i = 1 | x_i) = \dfrac{1}{1 + \exp(X^T w + w_0)}$

thus, $\mathbb{P}(Y = y_i | X) = \dfrac{1}{1 + \exp(-y_i(X^T w + w_0))}$

2. $\prod_{i=1}^{n} \mathbb{P}(Y = y_i | X) = \prod_{i=1}^{n} \dfrac{1}{1 + \exp(-y_i(X^T w + w_0))}$

$\log \prod_{i=1}^{n} \mathbb{P}(Y = y_i | X) = -\sum_{i=1}^{n} \log\left[1 + \exp(-y_i(X^T w + w_0))\right] = \ell(\hat{w}, w_0, X)$

$(\hat{w}, \hat{w}_0) = \arg\min \sum_{i=1}^{n} \log(1 + \exp(-y_i(X^T w + w_0)))$

3. $\nabla f(w_0, w) = \begin{bmatrix} \frac{\partial f}{\partial w_0} \\ \frac{\partial f}{\partial w_i} \end{bmatrix} = \begin{bmatrix} -\sum_{i=1}^{n} y_i \cdot \frac{1}{1+\exp\{-y_i(x^T w + w_0)\}} \\ -\sum_{i=1}^{n} y_i x_i^T \frac{1}{1+\exp(-y_i(x_i^T w + w_0))} \end{bmatrix}_{p+1}$

4. $g: x \mapsto \frac{\lambda}{2}\|x\|^2$

$$\text{prox}_g = \arg\min_x \ \frac{\lambda}{2}\|x\|^2 + \frac{1}{2}\|x-y\|^2$$

$$= \arg\min_x \ \frac{\lambda}{2}\|x\|^2 + \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 - \langle x, y\rangle.$$

$$= \arg\min_x \ \frac{\lambda+1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 - \langle x, y\rangle.$$

Let $h(x) = \frac{\lambda+1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 - \langle x, y\rangle$.

$\nabla h(x) = (\lambda+1)x - y \overset{\Delta}{=} 0$, $x = \frac{y}{\lambda+1}$

thus, $x = \text{prox}_g(y) = \frac{y}{\lambda+1}$.

5. prox gradient method: $x_{k+1} = \text{prox}_{\gamma g}(x_k - \gamma_k \nabla f(x_k))$

$$(w_0^{k+1}, w^{k+1}) = \text{prox}_{\gamma g}\left((w_0^k, w^k) - \gamma_k \nabla f(w_0^k, w^k)\right)$$

$$= \frac{1}{\lambda+1}\left((w_0^k, w^k) - \gamma_k \nabla f(w_0^k, w^k)\right)$$

6. Newton method: $x_{k+1} = x_k - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$

$$\nabla f(w_0, w) = \begin{bmatrix} \frac{\partial f}{\partial w_0} \\ \frac{\partial f}{\partial w_i} \end{bmatrix} = \begin{bmatrix} -\sum_{i=1}^{n} y_i \cdot \frac{1}{1+\exp\{-y_i(x_i^T w + w_0)\}} \\ -\sum_{i=1}^{n} y_i x_i^T \frac{1}{1+\exp(-y_i(x_i^T w + w_0))} \end{bmatrix}_{p+1}$$

$$\nabla^2 f(w_0, w) = \begin{bmatrix} \frac{\partial^2 f}{\partial w_0^2} & \frac{\partial^2 f}{\partial w_0 \partial w_i} \\ \frac{\partial^2 f}{\partial w_i \partial w_0} & \frac{\partial^2 f}{\partial w_i^2} \end{bmatrix}$$

Exercise 13 (LASSO). Let $\lambda > 0$, A be some $m \times n$ matrix and b be a vector of size m. We consider the problem

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \lambda \|x\|_1 .$$

1. Prove that the solution is $\{0\}$ for large λ.
2. For an arbitrary λ, provide the expression of the proximal gradient algorithm, using the step size suggested in Exercise 4.
3. Assume that the initial point is at distance D from a minimizer. How many iterations are needed (at most) to achieve an ε-minimizer?

1. $f(\lambda, x) = \frac{1}{2}\|Ax-b\|^2 + \lambda\|x\|$

$$\lim_{\lambda \to \infty} f(\lambda, x) = \begin{cases} \frac{1}{2}\|b\|^2, & x = 0 \\ \infty, & x \neq 0 \end{cases}$$

2. $f : x \mapsto \|Ax-b\|$. $g : x \mapsto \|x\|$

both f and g are convex and f is differentiable.

$\nabla f(x) = A^T(Ax-b)$. take A^TA as the Lipschitz const.

$$\|\nabla f(x) - \nabla f(y)\| = \|A^TAx - A^TAy\|$$

$$\leq \sup_{z \neq 0} \frac{\|A^TAz\|}{\|z\|} \|x-y\| \quad (?)$$

$$= \|A^TA\| \, \|x-y\|$$

g is not differentiable, but $g: x \mapsto \|x\| + \frac{1}{2}\|x-y\|^2$ is easy to find its minimum.

$$\partial g(x) = \begin{cases} \lambda \operatorname{sgn}(x) + x - y & x \neq 0 \\ [-\lambda, \lambda] - y & x = 0 \end{cases}$$

- if $|y| > \lambda$, then $0 \notin \partial g(0)$.

  if $x \in \mathbb{R}_+^*$, $\partial g(x) = 0 \Rightarrow x = y - \lambda$

  if $x \in \mathbb{R}_-^*$, $\partial g(x) = 0 \Rightarrow x = y + \lambda$

  $x = y + \lambda \operatorname{sgn}(y)$

- if $|y| \leq \lambda$, then $0 \in \partial g(0)$. 0 is a minimizer of g.

Therefore, $x = \operatorname{sgn}(y)\left(|y| - \lambda\right)_+$ $\quad (\cdot)_+ = \max(\cdot, 0)$

$$\operatorname{prox}_{\gamma g}(y) = \left(\operatorname{sgn}(y_1)\left(|y_1| - \lambda\right)_+, \cdots, \operatorname{sgn}(y_n)\left(|y_n| - \lambda\right)_+\right)$$