TD - Fermat : solution

Exercice 6 (Proximal operator of the 1-norm)
We say that a function $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is separable if there exists n functions $\phi_i : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ such that for all $x \in \mathbb{R}^n$,

$$\phi(x) = \sum_{i=1}^{n} \phi_i(x_i) \ .$$

1. *Let ϕ be a separable function. Show that*

$$\partial\phi(x) = \partial\phi_1(x_1) \times \ldots \times \partial\phi_n(x_n)$$

where $\times$ denotes the cartesian product.

▶

$$\begin{aligned} q \in \partial\phi(x) &\Rightarrow \forall y \in \mathbb{R}^n, \phi(y) \geq \phi(x) + \langle q, y - x\rangle \\ &\Rightarrow \forall i \in \{1, \ldots, n\}, \forall z \in \mathbb{R}, \phi_i(z) \geq \phi_i(x_i) + q_i(z - x_i) \Rightarrow \forall i, q_i \in \partial\phi_i(x_i) \end{aligned}$$

where the second implication comes by choosing $y = (x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n)$ and using the separability of ϕ.
For the converse

$$\begin{aligned} \forall i, q_i \in \partial\phi_i(x_i) &\Rightarrow \forall i \in \{1, \ldots, n\}, \forall y_i \in \mathbb{R}, \phi_i(y_i) \geq \phi_i(x_i) + q_i(y_i - x_i) \\ &\Rightarrow \forall y \in \mathbb{R}^n, \phi(y) \geq \phi(x) + \langle q, y - x\rangle \Rightarrow q \in \partial\phi(x) \end{aligned}$$

Where the second implication comes by summing the inequalities.

2. *Show that*

$$\inf_{x\in\mathbb{R}^n} \sum_{i=1}^{n} \phi_i(x_i) = \sum_{i=1}^{n} \inf_{x\in\mathbb{R}} \phi_i(x)$$

and

$$\arg\min_{x\in\mathbb{R}^n} \sum_{i=1}^{n} \phi_i(x_i) = \arg\min_{x\in\mathbb{R}} \phi_1(x) \times \ldots \times \arg\min_{x\in\mathbb{R}} \phi_n(x) \ .$$

▶ For all $i \in \{1, \ldots, n\}$ and for all $y_i \in \mathbb{R}$, $\phi_i(y_i) \geq \inf_{x\in\mathbb{R}} \phi_i(x)$.
Hence, $\phi(y) = \sum_{i=1}^{n} \phi_i(y_i) \geq \sum_{i=1}^{n} \inf_{x\in\mathbb{R}} \phi_i(x)$.
This implies the inequality $\inf_{y\in\mathbb{R}^n} \phi(y) \geq \sum_{i=1}^{n} \inf_{x\in\mathbb{R}} \phi_i(x)$.
For the other inequality, we have for all $y \in \mathbb{R}^n$, $\inf_{x\in\mathbb{R}^n} \phi(x) \leq \sum_{i=1}^{n} \phi_i(y_i)$. Hence, taking the infimum with respect to y_1, $\inf_{x\in\mathbb{R}^n} \phi(x) \leq \inf_{x\in\mathbb{R}} \phi_1(x) + \sum_{i=2}^{n} \phi_i(y_i)$. Taking the infimum with respect to y_2, …, y_n one after the other, we obtain the first result.
Fermat's rules and Question 1 lead to

$$x^* \in \arg\min \phi \Leftrightarrow 0 \in \partial\phi(x^*) \Leftrightarrow \forall i, 0 \in \partial\phi_i(x_i^*) \Leftrightarrow \forall i, x_i^* \in \arg\min \phi_i$$

This shows the second point.

3. *Let ϕ be a separable function. Show that*

$$\mathrm{prox}_\phi(x) = (\mathrm{prox}_{\phi_1}(x_1), \ldots, \mathrm{prox}_{\phi_n}(x_n))$$

▶ $\mathrm{prox}_\phi(x) = \arg\min_{y\in\mathbb{R}^n} \sum_{i=1}^n \phi_i(y_i) + \frac{1}{2}(x_i - y_i)^2$
Hence, $\mathrm{prox}_\phi(x)$ is the argmin of a separable function. It is the concatenation of the argmin of each summand.

4. *Let F be the 1-norm, that is $F(x) = \sum_{i=1}^n |x_i|$.*
Show that F is convex and separable.
▶ It is clear that F is separable. To show that F is convex, we compute for $t \in [0, 1]$,

$$F(tx + (1-t)y) \leq F(tx) + F((1-t)y) \leq tF(x) + (1-t)F(y)$$

where the second inequality is the triangle inequality.

5. *Recall the proximal operator of the absolute value and give the formula for the proximal operator of the 1-norm.*
▶ The proximal operator of the absolute value is the soft thresholding

$$\mathrm{prox}_{|.|}(x) = S_1(x) = \begin{cases} x-1 & \text{if } x > 1 \\ 0 & \text{if } x \in [-1,1] \\ x+1 & \text{if } x < -1 \end{cases}$$

The proximal operator of the 1-norm is thus

$$\mathrm{prox}_{\|.\|}(x) = (\mathrm{prox}_{|.|}(x_1), \ldots, \mathrm{prox}_{|.|}(x_2),).$$

Exercise 7(LASSO)
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.*
▶ Denote $f(x) = \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1$. By Fermat's rule 0 is solution if and only if

$$0 \in \partial f(0).$$

$(x \mapsto \frac{1}{2}\|Ax - b\|_2^2)$ is differentiable and $\lambda > 0$, so $\partial f(x) = \{A^\top(Ax - b)\} + \lambda\partial\|\cdot\|_1(x)$.
Moreover, $\partial\|\cdot\|_1(0) = [-1, 1] \times \ldots \times [-1, 1] = B_\infty$ so

$$0 \in \partial f(0) \Leftrightarrow 0 \in \{-A^\top b\} + \lambda\partial\|\cdot\|_1(0) \Leftrightarrow A^\top b \in \lambda B_\infty \Leftrightarrow \left\|A^\top b\right\|_\infty \leq \lambda$$

2. *For an arbitrary λ, provide the expression of the proximal gradient algorithm, using the step size suggested in Exercise 4.*

   ▶ We will take as stepsize $\gamma = 1/L$ where L is the Lipschitz constant of the gradient of $(x \mapsto \frac{1}{2}\|Ax - b\|_2^2)$, that is $L = \|A\|^2$.

   The proximal gradient algorithm starts at $x_0 \in \mathbb{R}^n$ and consists in the recurrence

$$x_{k+1} = \mathrm{prox}_{\gamma\lambda\|.\|_1}(x_k - \gamma \nabla f(x_k)) = S_{\lambda/\|A\|^2}\Big(x_k - \frac{1}{\|A\|^2} A^\top (Ax_k - b)\Big)$$

   where S_λ is the soft thresholding operator.

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 ?*

   ▶ The iterates of the proximal gradient algorithm are guaranteed to satisfy

$$f(x_k) - f(x^*) \leq \frac{\|A\|^2}{2k} \|x_0 - x^*\|^2 = \frac{\|A\|^2 D^2}{2k}$$

   Hence, if $k \geq \frac{\|A\|^2 D^2}{2\epsilon}$, $f(x_k) - f(x^*) \leq \epsilon$.

Exercice 9 (Proximal stochastic gradient for logistic regression)
We consider a classification problem defined by observations $(x_i, y_i)_{1\leq i\leq n}$ 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 $\forall i \in \{1, \ldots, n\}$, $\mathbb{P}(Y_i = y_i|x_i) = \dfrac{1}{1 + \exp(-y_i(x_i^\top w + w_0))}$.*

   ▶ $\mathbb{P}(Y_i = 1|x_i) = \frac{\exp(x_i^\top w + w_0)}{1 + \exp(x_i^\top w + w_0)} = \frac{1}{1 + \exp(-(x_i^\top w + w_0))} = \frac{1}{1 + \exp(-y_i(x_i^\top w + w_0))}$

   $\mathbb{P}(Y_i = -1|x_i) = 1 - \mathbb{P}(Y_i = 1|x_i) = \frac{\exp(-(x_i^\top w + w_0))}{1 + \exp(-(x_i^\top w + w_0))} = \frac{1}{1 + \exp(x_i^\top w + w_0)} = \frac{1}{1 + \exp(-y_i(x_i^\top w + w_0))}$

2. *Show that the maximum likelihood estimator is*

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

   ▶ As the observations are independent, the likelihood is

$$p(x, y; w) = \prod_{i=1}^n \frac{1}{1 + \exp(-y_i x_i^\top w)}.$$

The log-likelihood is thus

$$\log(p(x,y;w)) = \sum_{i=1}^{n} -\log(1+\exp(-y_i x_i^\top w))$$

and the maximum likelihood estimator is

$$\hat{w} = \arg\max_w p(x,y;w) = \arg\min_w \sum_{i=1}^{n} \log(1+\exp(-y_i x_i^\top w))$$

3. *Denote $f(w) = \sum_{i=1}^n \log(1+\exp(-y_i x_i^\top w))$. Compute $\nabla f(w)$.*
   ▶ We denote $f_i(w) = \log(1+\exp(-y_i(x_i^\top w + w_0)))$. $\nabla f(w,w_0) = \sum_{i=1}^n \nabla f_i(w,w_0)$, where
   $$\nabla_{w_0} f_i(w,w_0) = \frac{-y_i \exp(-y_i(x_i^\top w + w_0))}{1+\exp(-y_i(x_i^\top w + w_0))}$$
   $$\nabla_{w} f_i(w,w_0) = \frac{-y_i \exp(-y_i(x_i^\top w + w_0))}{1+\exp(-y_i(x_i^\top w + w_0))} x_i$$
4. *Compute the proximal operator of $(x \mapsto \frac{\lambda}{2}\|x\|^2)$.*
   ▶ $p = \operatorname{prox}_{\frac{\lambda}{2}\|\cdot\|_2^2}(y) = \arg\min_x \frac{\lambda}{2}\|x\|_2^2 + \frac{1}{2}\|y-x\|_2^2$ so p is solution to $\lambda p + p - y = 0$ which gives $p = \frac{1}{1+\lambda} y$.
5. *Write the proximal stochastic gradient method for the logistic regression problem with ridge regularizer*
   $$(\hat{w}^{(\lambda)}, \hat{w}_0^{(\lambda)}) = \arg\min_{w,w_0} \sum_{i=1}^{n} \log(1+\exp(-y_i(x_i^\top w + w_0))) + \frac{\lambda}{2}\|w\|^2 \ .$$
   ▶ Note that $\nabla f(w,w_0) = \frac{1}{n}\sum_{i=1}^n n\nabla f_i(w,w_0)$ so if $i_{k+1} \sim U(\{1,\dots,n\})$, then $\mathbb{E}(n\nabla f_{i_{k+1}}(w,w_0)) = \nabla f(w,w_0)$.
   At iteration k :
   Generate i_{k+1} uniformly at random
   Set $\gamma_k = \frac{\gamma_0}{k+1}$. $w_{k+1} = \operatorname{prox}_{\frac{\gamma_k\lambda}{2}\|\cdot\|_2^2}(w_k - \gamma_k n \nabla f_{i_{k+1}}(w_k)) = \frac{1}{1+\lambda\gamma_k}\Big(w_k - \gamma_k n\nabla f_{i_{k+1}}(w_k)\Big)$

Exercice 10 (Optimisation with explicit constraints)
We consider the following optimization problem

$$\min_{x\in C} f(x) \tag{1}$$

where $C \subset \mathbb{R}^d$ is a convex set and $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable.

1. *We define the convex indicator function of the set C as*

$$\iota_C(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{if } x \notin C \end{cases}$$

Show that (1) *is equivalent to*

$$\min_{x \in \mathbb{R}^d} f(x) + \iota_C(x) \tag{2}$$

► *Clearly, the solution of* (8) *is in C (elsewhere $\iota_C(x) = +\infty$), and $\forall x \in C$, $\iota_C(x) = 0$, therefore* (8) *is equivalent to (7).*

2. *Show that for all $x \in C$, $\partial\iota_C(x) = \{q \in \mathbb{R}^d \ : \ \forall y \in C, \langle q, y - x\rangle \leq 0\}$ and that $\partial\iota_C(x)$ is a cone (it is called the normal cone to C at x). Show that for all $x \notin C$, $\partial\iota_C(x) = \emptyset$.*

► *By definition, $\partial\iota_C(x) = \{q \in \mathbb{R}^d \ : \ \forall y \in R^d, \iota_C(y) \geq \iota_C(x) + \langle q, y - x\rangle\}$. If $x \in C$, then $\iota_C(x) = 0$, thus $\partial\iota_C(x) = \{q \in \mathbb{R}^d \ : \ \forall y \in C, \langle q, y - x\rangle \leq 0\}$. Clearly, $\partial\iota_C(x)$ is a cone because if $q' = \lambda q$ with $q \in \partial\iota_C(x)$ and $\lambda \geq 0$, then $q' \in \partial\iota_C(x)$. If $x \notin C$, then $\iota_C(x) = +\infty$, and no vector q can fulfill the condition, therefore $\partial\iota_C(x) = \emptyset$.*

3. *Show that x^* is a solution to* (2) *if and only if*

$$-\nabla f(x^*) \in \partial\iota_c(x^*) \ .$$

► *We have $\partial(f + \iota_c)(x^*) = \nabla f(x^*) + \partial\iota_c(x^*)$ because ι_c is convex, f is differentiable, and $0 \in \mathrm{relint}(\mathrm{dom}(\iota_c) - \mathrm{dom}(f)) = \mathrm{relint}(C - R^d) = R^d$. Therefore x^* is a solution to* (2) *if and only if $0 \in \nabla f(x^*) + \partial\iota_c(x^*)$, i.e. $-\nabla f(x^*) \in \partial\iota_c(x^*)$.*

4. *Denote*

$$\mathcal{H}_{w,b} = \{x \in \mathcal{X} : \ \langle w, x\rangle + b = 0\}$$

Compute $\partial\iota_{\mathcal{H}_{w,b}}(x)$ for all $x \in \mathbb{R}^d$.

► *Since $\mathcal{H}_{w,b}$ is an hyperplane, it is convex. Therefore we can use the result of question 2 : if $x \notin \mathcal{H}_{w,b}$, $\partial\iota_{\mathcal{H}_{w,b}}(x) = \emptyset$. Otherwise, if $x \in \mathcal{H}_{w,b}$, $\langle w, x\rangle + b = 0$ and $\partial\iota_{\mathcal{H}_{w,b}}(x) = \{q \in \mathbb{R}^d \ : \ \forall y \in \mathbb{R}^d : \langle w, y\rangle + b = 0, \langle q, y - x\rangle \leq 0\}$. If there is a y such that $\langle q, y - x\rangle < 0$, then $y' = 2x - y$ is such that $\langle w, y'\rangle + b = 0$ and $\langle q, y' - x\rangle > 0$, which makes a contradiction. Therefore $\partial\iota_{\mathcal{H}_{w,b}}(x) = \{q \in \mathbb{R}^d \ : \ \forall y \in \mathbb{R}^d : \langle w, y\rangle + b = 0, \langle q, y - x\rangle = 0\}$. Therefore $\partial\iota_{\mathcal{H}_{w,b}}(x)$ is a 1-dimensional vector space, and we note that $w \in \iota_{\mathcal{H}_{w,b}}(x)$. We conclude that $\partial\iota_{\mathcal{H}_{w,b}}(x) = \mathrm{span}(w)$.*

5. *Prove that the distance of a point z to $\mathcal{H}$ is equal to*

$$d(z, \mathcal{H}_{w,b}) = \min_{x \in \mathcal{H}_{w,b}} \|x - z\|_2 = \frac{|\langle w, z\rangle + b|}{\|w\|_2} \ .$$

► *Let $f(x) = \frac{1}{2}\|x - z\|^2$ and $C = \mathcal{H}_{w,b}$, and let us use the result of questions 3 and 4 : $-\nabla f(x^*) \in \partial\iota_c(x^*) \Leftrightarrow \exists \nu \in \mathbb{R} : -(x^* - z) = \nu w$, i.e. $x^* = z - \nu w$. However, we have $\langle w, x^*\rangle + b = 0$, thus $\nu = \frac{\langle w, z\rangle + b}{\|w\|^2}$. Finally, we get $\|x^* - z\|_2 = \|\nu w\|_2 = \frac{\langle w, z\rangle + b}{\|w\|}$.*