**Introduction to Continuous optimization**

Assessment

(6th January 2021)

# Exercise I

We denote $\mathbb{R}^{n\times n}_{\text{sym}}$ the space of dimension $n(n+1)/2$ of symmetric $n\times n$ matrices. We consider the scalar product $X : Y = \sum_{i,j} X_{i,j}Y_{i,j} = \mathrm{Tr}\,(XY)$ (or $\mathrm{Tr}\,(X^TY)$ but it is the same here since $X, Y$ are symmetric).

Let $\mathcal{S}_+ \subset \mathbb{R}^{n\times n}_{\text{sym}}$ be the set of $n\times n$ symmetric, positive semidefinite matrices: $X = X^T$, $(X\xi)\cdot\xi \geq 0$ for any $\xi \in \mathbb{R}^n$. Let $\mathcal{S}_{++}$ be the interior of $\mathcal{S}_+$, that is, the set of positive definite matrices: $(X\xi)\cdot\xi > 0$ for all $\xi \neq 0$.

We let, for $X \in \mathbb{R}^{n\times n}_{\text{sym}}$:

$$h(X) := \begin{cases} -\ln\det X & \text{if } X \in \mathcal{S}_{++}, \\ +\infty & \text{else.} \end{cases}$$

**1.** Let $X \in \mathcal{S}_{++}$, $H$ a symmetric matrix. Show that for $t \in \mathbb{R}$ with $|t|$ small enough, $X + tH \in \mathcal{S}_{++}$.

**2.** Using $X + tH = X(I + tX^{-1}H)$, show that

$$\nabla h(X) = -X^{-1}.$$

We recall that $\det(I + A) = 1 + \mathrm{Tr}\,A + o(\|A\|)$.

**3.** One now wants to compute the conjugate $h^*(Y) = \sup_X X : Y - h(X)$.

Let $Y \in \mathbb{R}^{n\times n}_{\text{sym}}$ and assume $e$ is an eigenvector of $Y$ with eigenvalue $\lambda \in \mathbb{R}$ (and $|e| = 1$).

Considering first $X$ of the form $te\otimes e + \varepsilon I$ (where for $e \in \mathbb{R}^n \setminus \{0\}$ with $|e| = 1$, $e \otimes e$ is the matrix $e_ie_j$ which has eigenvector $e$ with eigenvalue 1), $\varepsilon > 0$, $t \to +\infty$, show that $h^*(Y) = +\infty$ if $\lambda \geq 0$.

Deduce that $\mathrm{dom}\,h^* \subset \{Y : -Y \in \mathcal{S}_{++}\}$.

**4.** Now, assuming $-Y > 0$ we admit (even if it is quite easy to show) that $\sup_X X : Y - h(X)$ is reached at some positive matrix $X$.

Show that $X = -Y^{-1}$. Deduce the expression of $h^*$. Deduce also that $h$ is convex.

**5.** We consider the problem $\min_{X\in\mathcal{S}_+} C : X$ and the Bregman distance

$$D_h(X, Y) = h(X) - h(Y) - \nabla h(Y) : (X - Y)$$

induced by $h$, defined for $X, Y \in \mathcal{S}_{++}$. Write the expression of an iteration of non-linear gradient descent for the problem, with step $\tau > 0$, relative to the Bregman distance $D_h$. Why can we always assume that $C$ is symmetric? What assumption is needed on $C$ in order for the problem to have a solution (and the algorithm to be well defined for all $k$)?

## Exercice II - prox

Compute the proximity operator (for some parameter $\tau > 0$):

$$\operatorname{prox}_{\tau g}(x) = \arg\min_z g(z) + \frac{1}{2\tau}|z - x|^2$$

for the convex functions:

**1.** $g_1(x) = -\ln x$ for $x > 0$, $+\infty$ else;

**2.** $g_2(x) = \sum_{i=1}^n \frac{1}{3}|x_i|^3$, $(x \in \mathbb{R}^n)$;

**3.** $g_3(x) = \sum_{i=1}^n \frac{2}{3}|x_i|^{3/2}$, $(x \in \mathbb{R}^n)$;

**4.** $g_4(x) = \frac{1}{2}\sum_i x_i^2$ if $x_i \geq 0$, $i = 1, \dots, n$, and $+\infty$ else, defined for $x \in \mathbb{R}^n$ (and with domain $\operatorname{dom} g_4 = [0, +\infty)^n$).

## Exercise III - rate for the proximal point algorithm

We consider $M$ a maximal-monotone operator, defined in a Hilbert space $\mathcal{X}$. Given $x^0 \in \mathcal{X}$, we let for $k \geq 0$:

$$x^{k+1} = (I + M)^{-1} x^k$$

that is, the iterations of the proximal-point algorithm.

**1.** Let $x^*$ be a zero, that is, a point such that $Mx^* \ni 0$ (we assume the set $M^{-1}(0)$ is not empty). Show that $x^* = (I + M)^{-1}(x^*)$ and that

$$|x^{k+1} - x^*|^2 + |x^k - x^{k+1}|^2 \leq |x^k - x^*|^2.$$

**2.** Show that $|x^{k+1} - x^k|$ is a decreasing function of $k \geq 0$.

**3.** Deduce that

$$|x^{k+1} - x^k| \leq \frac{|x^0 - x^*|}{\sqrt{k+1}}.$$

**4.** Let $x^{k_l}$ be a (weakly) converging subsequence, to some point $\bar{x}$. Show that for any $x' \in \mathcal{X}$ and $y' \in Mx'$,

$$\langle x' - \bar{x}, y' \rangle \geq 0.$$

Deduce that $0 \in M\bar{x}$.

**5.** Let $T : \mathcal{X} \to \mathcal{X}$ be a 1-Lipschitz operator and, for $\theta \in (0,1)$, let $T_\theta = (1-\theta)I + \theta T$. Let $x^*$ be a fixed point of $T$ (and therefore also of $T_\theta$ for any $\theta$). We now consider the algorithm given by

$$x^{k+1} = T_\theta x^k.$$

Use the parallelogram identity to show that:

$$|x^{k+1} - x^*|^2 \leq |x^k - x^*|^2 - \theta(1-\theta)|Tx^k - x^k|^2$$

**6.** As before, deduce that:

$$|Tx^k - x^k| \leq \frac{|x^0 - x^*|}{\sqrt{\theta(1-\theta)}\sqrt{k+1}}.$$

(Remark: in this framework, one can show [Baillon-Bruck 1996] that a similar estimate holds in any metric space, but it is much harder).

**7.** Application: show that the over-relaxed proximal point algorithm:

$$\begin{aligned} x^{k+\frac{1}{2}} &= (I+M)^{-1}x^k \\ x^{k+1} &= x^k + \lambda(x^{k+\frac{1}{2}} - x^k) \end{aligned}$$

for $1 < \lambda < 2$ is a converging method.

## Exercise IV - Yosida approximation

Let $A$ be a maximal monotone operator in a Hilbert space, and defined the Yosida approximation, for $\lambda > 0$, as

$$A_\lambda x = \frac{x - J_{\lambda A}x}{\lambda}$$

where $J_{\lambda A} = (I + \lambda A)^{-1}$ is the resolvent.

**1.** Show that $A_\lambda$ is a monotone operator.

**2.** Show that $A_\lambda x = J_{A^{-1}/\lambda}(x/\lambda)$. Deduce that it is $(1/\lambda)$-Lipschitz. Bonus: show that it is $\lambda$-co-coercive.

**3.** Let $x \in \operatorname{dom} A$ (that is, $Ax \neq \emptyset$). Show that

$$\lim_{\lambda \to 0} A_\lambda x = A_0 x := \arg\min_{p \in Ax} |p|.$$

*Hint:* first, show that if $p_\lambda = A_\lambda x$ then $p_\lambda \in A(x - \lambda p_\lambda)$. Using the monotonicity of $A$, deduce that for any $p \in Ax$, $|p_\lambda|^2 \leq \langle p_\lambda, p\rangle$, hence that $|p_\lambda| \leq |p|$. Conclude by using that $A$ is maximal.

**4.** Contraction semigroup: since $A_\lambda$ is Lipschitz, by the Cauchy-Lipschitz theorem, one can solve for all $x \in \mathcal{X}$:

$$\begin{cases} \dot{X}^\lambda(t,x) = -A_\lambda X^\lambda(t,x) & t > 0, \\ X^\lambda(0,x) = x \end{cases}$$

and the solution, which is at least $C^1$ in time, satisfies:

$$X^\lambda(t,x) = x - \int_0^t A_\lambda X^\lambda(s,x)ds = X^\lambda(t',x) - \int_0^{t-t'} A_\lambda X^\lambda(s, X^\lambda(t',x))ds$$

for any $t' < t$. In particular, $X^\lambda(t,x) = X^\lambda(t-t', X^\lambda(t',x))$.

Show that for any $x, y \in \mathcal{X}$, $t \mapsto |X^\lambda(t,x) - X^\lambda(t,y)|^2$ is non-increasing. Deduce that $|X^\lambda(t,x) - X^\lambda(t,y)| \le |x-y|$ for all $t \ge 0$.

**5.** Show that $t \mapsto |A_\lambda X^\lambda(t,x)|$ is nonincreasing. If $x \in \operatorname{dom} A$, show that $|A_\lambda X(t,x)| \le |A_0 x|$ for all $\lambda > 0$ and $t \ge 0$.
*Hint:* use that

$$X^\lambda(t+h,x) - X^\lambda(t,x) = X^\lambda(t-t', X^\lambda(t'+h,x)) - X^\lambda(t-t', X^\lambda(t',x))$$

for any $h > 0$, $t$, $t' < t$, and the previous question.

**6.** Using question **2.**, show that for $\lambda, \mu > 0$ and for any $x \in \mathcal{X}$, $A_\lambda x = A_\mu(x + (\mu-\lambda)A_\lambda x)$. Deduce that:

$$\frac{\partial}{\partial t}|X^\lambda(t,x) - X^\mu(t,x)|^2 \le (\mu-\lambda)(|A_\lambda X^\lambda(t,x)|^2 - |A_\mu X^\mu(t,x)|^2)$$

(or any similar estimate) and in particular that if $x \in \operatorname{dom} A$, $|X^\lambda(t,x) - X^\mu(t,x)| \le C|A_0 x|\sqrt{|\mu-\lambda|t}$ for some constant $C > 0$.

What can you conclude? (Without justifying everything, unless there is still time.)