# Optimization for Big Data - Mirror Descent

## Summary


*The goal of this homework is to study both from a theoretical and from a practical point of view some ingredients related to the optimization of convex and smooth function constrained on a smooth and convex set. We expect you to illustrate your homework with Python.*

*1) You are asked to answer the theoretical questions either with a handwritten report or a latex file. You are also asked to answer the practical questions with python and produce an illustrated pdf report.*
*2) You are also asked to call your file :*
*M1-NAME-SURNAME.pdf. If not, your final mark is divided by 2.*

*Deadline : 25th of april 2021*

*Individual work*


In what follows, we consider a state space $\mathbb{R}^p$ and a domain $\mathcal{D} \subset \mathbb{R}^p$ such that $\mathcal{D}$ is **closed** and **convex**. We consider a smooth function $f$ that is assumed to be $\mathcal{C}_1^L(\mathbb{R}^p, \mathbb{R}_+)$ and **convex**. We are looking for

$$x^\star = \arg\min_{x \in \mathcal{D}} f(x).$$

Below, the notation $|.|_2$ will refer to the standard Euclidean norm :

$$|x|_2 = \sqrt{\sum_{i=1}^p x_i^2},$$

whereas $|.|_1$ will refer to the $L^1$ norm :

$$|x|_1 = \sum_{i=1}^p |x_i|.$$

In what follows, $\nabla f$ will refer to the gradient of $f$.

## Part I - Elementary facts

**Question 1-a :** Prove that when $f$ is $\alpha$ strongly convex, a unique minimizer $x^\star$ exists for $f$.

**Question 1-b :** Prove that $x^\star$ satisfies

$$\forall v \in \mathcal{D} \qquad \langle \nabla f(x^\star), v - x^\star \rangle \geq 0.$$

**Question 2-a :** Recall the definition of the projection on $\mathcal{D}$ with respect to $|.|_2$. Does this projection exists ? Why (we do not ask for a proof). Below, we will denote this projection by $\Pi_{\mathcal{D}}$.

**Question 2-b :** Consider the case

$$\mathcal{D} := [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_p, b_p],$$

where $a_i < b_i$ for all $i$. Compute $\Pi_{\mathcal{D}}(x)$.

**Question 2-c :** For $p'$ an integer such that $p' \leq p$ and a radius $R > 0$, consider the set

$$\mathcal{D} := \left\{ x \in \mathbb{R}^p \ : \ x_1^2 + \ldots + x_{p'}^2 \leq R^2 \right\}.$$

Compute $\Pi_{\mathcal{D}}(x)$.

## Part II - Non-smooth domain (not so elementary)

We consider $\mathcal{S}$ the probability simplex :

$$\mathcal{S} := \{ x \in \mathbb{R}^p \, | \, x_1 + x_2 + \ldots + x_p = 1 \text{ and } \forall i \in \{1, \ldots, p\} \quad x_i \geq 0 \}$$

For a given $v \in \mathbb{R}^p$, we write $q : x \in \mathcal{S} \longmapsto \frac{1}{2}|x - v|_2^2$.

**Question 3-a :** Define the projection on $\mathcal{S}$ as a constrained minimization problem. Below, we denote by $w$ this projection.

**Question 3-b :** Prove that the Lagrangian function $\mathcal{L}$ associated to this minimization problem is :

$$\mathcal{L}(x,\xi) = \frac{1}{2}|x-v|_2^2 + \lambda\left(\sum_{i=1}^{p} x_i - 1\right) - \langle \xi, x\rangle,$$

where $\lambda \in \mathbb{R}$ and $\xi \in \mathbb{R}_+^p$. Give the relationship between the multipliers and $w$ with the help of the KKT conditions.

**Question 3-c :** Assume that for two integers $(i,j) \in \{1,\ldots,p\}$, $v_i \geq v_j$, prove that if $w_i = 0$, then $w_j = 0$.

**Question 3-d :** Prove that if $w_i > 0$, then $\xi_i = 0$. Denote by $I$ the set of "active coordinates" for the solution $w$ :

$$I = \{i \in \{1\ldots p\} : w_i > 0\},$$

and $\rho = |I|$. Prove that if we rank $w$ by decreasing values :

$$w_{(1)} \geq w_{(2)} \geq \ldots \geq w_{(\rho)} > w_{(\rho+1)} = 0,$$

then the same ranking also holds for coordinates in $v$ for integers in $I$.

Deduce that :

$$\lambda = \frac{\sum_{i=1}^{\rho} v_{(i)} - 1}{\rho}$$

**Question 3-e :** Assume that the integer $\rho$ is known, prove that

$$w_i = \max\{v_i - \lambda, 0\}$$

**Question 3-f :** Prove that the following algorithm computes $w$.

ALGORITHM 1 (PROJECTION ON $\mathcal{S}$) *Input :* $v \in \mathbb{R}^p$

- *Sort* $v_{(1)} \geq v_{(2)} \geq \ldots \geq v_{(p)}$.
- *Compute* $\rho^\star$ *defined by :*

$$\rho^\star = \max\left\{j \leq p : v_{(j)} - \frac{1}{j}\left(\sum_{k=1}^{j} v_{(k)} - 1\right) \geq 0\right\}$$

- *Compute* $\lambda^\star$ *defined by :* $\lambda^\star = \frac{1}{\rho^\star}\left(\sum_{k=1}^{\rho^\star} v_{(k)} - 1\right)$
- *Return :*

$$w_i = \max\{v_i - \lambda^\star, 0\}$$

**Question 3-g :** Implement this projection in Python with a program from you.

**Question 3-h :** What is the complexity cost of a such algorithm ?

## Part III - Projected Gradient Descent

Below, we consider that $x^\star = \arg\min_{x\in\mathbb{R}^p} f(x) \in \mathcal{D}$. We also assume that $f$ is strongly convex of parameter $\alpha$.

**Question 4-a :** We introduce the *projected gradient descent algorithm* as :

ALGORITHM 2 (PGD) *Initialization :* $x_0 \in \mathbb{R}^p$

- *Choose a step-size* $\rho > 0$
- *Iterate :*
  - *Compute* $d_k = \nabla f(x_k)$ *and*

$$\tilde{x}_{k+1} = x_k - \rho\nabla f(x_k).$$

  - *Upgrade the new position of the algorithm :*

$$x_{k+1} = \Pi_{\mathcal{D}}(\tilde{x}_{k+1}).$$

Prove that the algorithm always belongs to $\mathcal{D}$.

**Question 4-b :** Show that when $\rho \in (0, \frac{2\alpha}{L^2})$, the algorithm converges exponentially fast towards $x^\star$.

**Question 4-c :** What is the numerical "cost" of the algorithm to achieve an $\epsilon$ solution ?

**Question 4-d :** Discuss on the effect of the dimension when looking at the simplex constraint of Question 3.

## Part IV - Projected stochastic strongly convex case

**Question 5-a :** Assume that we only have access to a noisy gradient within a framework of stochastic optimization :

$$x_{k+1} = \Pi_{\mathcal{D}}\left[x_k - \gamma_{k+1}[\nabla f(x_k) + \xi_{k+1}]\right],$$

where $(\xi_{k+1})_{k\geq 1}$ is a sequence of i.i.d. centered random noises with

$$\sigma^2 = \sup_{k\geq 1} \mathbb{E}[\|\xi_{k+1}\|^2] < +\infty.$$

The purpose of the next questions is to derive a mathematical study of the projected stochastic gradient descent algorithm. Prove that :

$$2\gamma_{k+1}\left[f(x_k) - f(x^\star) + \frac{\alpha}{2}|x_{k-1} - x^\star|_2^2\right] \leq |x_{k-1} - x^\star|_2^2 \\ - \mathbb{E}[|x_k - x^\star|_2^2 \,|\mathcal{F}_{k-1}] + \sigma^2\gamma_{k+1}^2 L^2$$

**Question 5-b :** Conclude that for a fixed horizon $N > 0$ and a constant step-size $\gamma$, if we define $\bar{x}_N = \frac{1}{N}\sum_{k=1}^N x_k$, one has :

$$\mathbb{E}[2(f(\bar{x}_N) - f(x^\star)) + \alpha|\bar{x}_n - x^\star|^2] \leq \sigma^2 L^2\gamma + \frac{D^2}{N\gamma}$$

Conclude an optimal tuning of the parameter $\gamma$.

**Question 5-c :** Coming back to 5.a and choosing $\gamma_k = \frac{1}{\alpha k}$, prove that

$$\mathbb{E}[f(\bar{x}_N)] - f(x^\star) \leq \frac{D^2 \log n}{\alpha n}$$

where $D$ refers to the diameter of $\mathcal{D}$.

**Question 5-d :** Compare the rates obtained by the two step-size strategies.

## Part V - Projected stochastic convex case

We are now interested in the weaker situation of convex function $f$.

**Question 6-a :** Repeating the arguments of Question 5.a, prove that :

$$\gamma_{k+1}\mathbb{E}[f(x_k) - f(x^\star)] \leq \frac{|x_1 - x^\star|_2^2 + \sigma^2\sum_{j=1}^k \gamma_j^2}{2\sum_{j=1}^k \gamma_j}.$$

**Question 6-b :** Define now

$$\bar{x}_N = \sum_{k=1}^N \left(\frac{\gamma_{k+1}}{\sum_{j=1}^k \gamma_{j+1}} x_k\right),$$

prove that a suitable constant step-size yields a $\mathcal{O}(N^{-1/2})$ convergence rate. Discuss on the "not-anytime" feature of a such strategy.

**Question 6-c :** Choosing now $\gamma_{k+1} \propto (k+1)^{-1/2}$, what convergence rate is obtained ?

## Part VI - Mirror Descent - convex case

The objective of the rest of the theoretical part is to avoid the projection, as it may be a real additional cost for large dimensional problems. In this view, we introduce $\varphi$ a smooth strongly convex function on $\mathcal{D}$ and the Bregman divergence

$$\forall (x,z) \in \mathcal{D}^2 \qquad D_\varphi(x,z) = \varphi(x) - \varphi(z) - \langle \nabla\varphi(z), x - z\rangle.$$

We assume that $\varphi$ is $\rho$ strongly convex.

**Question 7-a :** Prove that $D_\varphi \geq 0$ and is a convex function of the first coordinate. Compute $\nabla_x D_\varphi(x,z)$.

**Question 7-b :** Show that $D_\varphi$ satisfies the three points lemma :

$$D_\varphi(x,z) = D_\varphi(x,y) + D_\varphi(y,z) - \langle \nabla\varphi(z) - \nabla\varphi(y), x - y\rangle.$$

**Question 7-c :** Assume that $\mathcal{D} = \mathbb{R}^p$ (no constraints) and $\varphi$ is the square function $\varphi(x) = |x|_2^2$, prove that :

$$D_\varphi(x,z) = |x - z|_2^2.$$

**Question 7-d :** Assume that $\mathcal{D} = \mathcal{S}$ (simplex) and $\varphi$ is the negative entropy $\varphi(x) = \sum_{i=1}^p x_i \log(x_i)$, prove that

$$D_\varphi(x,z) = \sum_{i=1}^p x_i \log\left(\frac{x_i}{z_i}\right)$$

What is the name of a such divergence ?

We introduce now the Mirror Descent algorithm :

ALGORITHM 3 (MIRROR DESCENT ON $\mathcal{D}$) *Initialization :* $x_0 \in \mathcal{D}$

- *Input : step-size sequence* $(\gamma_{k+1})_{k\geq 0}$

- *Iterate :*
  - *Compute the gradient of $f$ : $g_k = \nabla f(x_k)$*
  - *Upgrade the new position of the algorithm :*

$$x_{k+1} = \arg\min_{x \in \mathcal{D}} \left\{ \langle g_k, x - x_k \rangle + \frac{1}{\gamma_{k+1}} D_\varphi(x, x_k) \right\}$$

**Question 8-a :** Write an explicit upgrade when $\mathcal{D} = \mathbb{R}^p$ and $\varphi(x) = |x|_2^2$.

**Question 8-b :** Prove that when $\mathcal{D} = \mathcal{S}$ and $\varphi(x) = \sum_{i=1}^p x_i \log(x_i)$ :

$$\forall j \in \{1, \ldots, p\} \qquad x_{k+1,j} = \frac{x_{k,j} e^{-\gamma_{k+1} g_{k,j}}}{\sum_{i=1}^p x_{k,i} e^{-\gamma_{k+1} g_{k,i}}}.$$

**Question 8-c :** Using the definition of the algorithm and the three points lemma, prove that for any $x \in \mathcal{D}$, we have :

$$\gamma_{k+1} \langle g_k, x_{k+1} - x \rangle \leq D_\varphi(x, X_k) - D_\varphi(x, X_{k+1}) - D_\varphi(X_{k+1}, X_k).$$

**Question 8-d :** Show that

$$\gamma_{k+1} \langle g_k, x_{k+1} - x_k \rangle \leq \frac{\gamma_{k+1}^2 |g_{k+1}|^2}{2\rho} + \frac{\rho}{2} |x_{k+1} - x_k|^2$$

**Question 8-e :** Assume that $|\nabla f|$ is bounded over $\mathcal{D}$ by $M$, using the convexity of $f$ and a telescopic sum argument, prove that if we define $\bar{x}_N$ as in Question 6-b, then :

$$f(\bar{x}_N) - f(x^\star) \leq \frac{\sup_{x \in \mathcal{D}} D_\varphi(x, x_0) + \frac{M^2}{2\rho} \sum_{k=0}^N \gamma_{k+1}^2}{\sum_{k=0}^N \gamma_{k+1}}$$

**Question 9 :** Present the Markowitz portfolio problem. To do this, you are allowed (and even asked) to find the needed documentation by yourself on www.

**Question 10 :** Compare the mirror descent and the projected gradient descent over the simplex from a numerical point of view with a large number $p$ of assets in a porfolio with the Markowitz model with correlated and uncorrelated framework.

**Question 11 :** Would it be possible to handle the mirror descent with a stochastic optimzation algorithm ? If yes, try it on the Markowitz model !