MASTER 2 MICAS

Machine Learning Communications and Security

Course : MICAS931 - Introduction to Optimization.

School : Institut Polytechnique de Paris

Teacher : HADI GHAUCH

Subject : Homework 1 - Submit before 22/11/2020

Etudiant : Panongbene Jean Mouhamed Sawadogo.
Email : panongbene.sawadogo@telecom-paris.fr

SOMMAIRE :

Convex Optimization

Convex Optimization

Part I : Convex Analysis

Consider a strongly convex function, f(x), (with constant $\mu > 0$), defined over a convex set X Recall that it fulfills the following :

$$f(x_2) \geqslant f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|^2, \forall (x_1, x_2) \in X \tag{1}$$

Prove all the following statements

- (1) is equivalent to a minimum positive curvature $\nabla^2 f(x) \succeq \mu I_d, \forall x \in X$

Answer : We assume that f is twice continuously differentiable.
According the Taylor-Lagrange's formula, we have :

$$\forall x, h \in X, \exists \epsilon \text{ a function define in X }, \lim_{h \to 0} \epsilon(h) = 0$$

$$f(a+h) = f(a) + \nabla f(a)^T h + \frac{1}{2} h^T \nabla^2 f(a) h + \|h\|^2 \epsilon(h)$$

Let a, h $\in$ X, using Taylor-Langrange's formula we have :

$(1) \iff f(a) + \nabla f(a)^T h + \frac{1}{2} h^T \nabla^2 f(a) h + \|h\|^2 \epsilon(h) \geqslant f(a) + \nabla f(a)^T h + \frac{\mu}{2} \|h\|^2$

$\iff \frac{1}{2} h^T \nabla^2 f(a) h + \|h\|^2 \epsilon(h) \geqslant \frac{\mu}{2} \|h\|^2$

$\iff h^T \left[\nabla^2 f(a) + 2\epsilon(h) I_d - \mu I_d\right] h \geqslant 0$

Since $\epsilon(h)$ converges to 0 when h approaches 0, the previous expression is equivalent to :

$\iff h^T \left[\nabla^2 f(a) - \mu I_d\right] h \geqslant 0$

This expression is true for all h and a in X, so we can say that :

$$(1) \iff \forall a \in X, \nabla^2 f() \succeq \mu I_d$$

- (1) is equivalent to $(\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \geqslant \mu \|x_2 - x_1\|_2^2$

Answer : Let $x_1, x_2 \in X$, we have :

$$(1) \iff \begin{cases} f(x_1) \geqslant f(x_2) + \nabla f(x_2)^T (x_1 - x_2) + \frac{\mu}{2} \|x_1 - x_2\|_2^2 \\ f(x_2) \geqslant f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|_2^2 \end{cases}$$

$\iff f(x_2)+f(x_1) \geqslant f(x_1)+f(x_2)+\nabla f(x_1)^T (x_2 - x_1)+\nabla f(x_2)^T (x_1 - x_2)+ \mu \|x_2 - x_1\|_2^2$

$\iff \nabla f(x_2)^T (x_2 - x_1) - \nabla f(x_1)^T (x_2 - x_1) \geqslant \mu \|x_2 - x_1\|_2^2$

$\iff (\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \geqslant \mu \|x_2 - x_1\|_2^2$

So we can say that :

$$(1) \iff (\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \geqslant \mu \|x_2 - x_1\|_2^2$$

- (1) implies to $f(x) - f^* \leqslant \frac{1}{2\mu} \|\nabla f(x)\|_2^2, \forall x \in X$

Answer : Let x, y $\in$ X we have :

$0 \leqslant \frac{\mu}{2} \left\| y - x + \frac{\nabla f(x)}{\mu} \right\|_2^2 = \frac{\mu}{2} \left(y - x + \frac{\nabla f(x)}{\mu} \right)^T \left(y - x + \frac{\nabla f(x)}{\mu} \right)$

$= \frac{\mu}{2} \|y - x\|_2^2 + \frac{1}{2} (y - x)^T \nabla f(x) + \frac{1}{2} \nabla f(x)^T (y - x) + \frac{\|\nabla f(x)\|_2^2}{2\mu}$

$= \frac{\mu}{2} \|y - x\|_2^2 + \nabla f(x)^T (y - x) + \frac{\|\nabla f(x)\|_2^2}{2\mu}$

Thus, using relation (1), we have :

$(1) \implies 0 \leqslant \frac{\mu}{2} \|y - x\|_2^2 + \nabla f(x)^T (y - x) + \frac{\|\nabla f(x)\|_2^2}{2\mu} \leqslant f(y) - f(x) + \frac{\|\nabla f(x)\|_2^2}{2\mu}$

$\implies f(x) - f(y) \leqslant \frac{\|\nabla f(x)\|_2^2}{2\mu}$

$\implies f(x) - \inf_{y \in X} f(y) \leqslant \frac{\|\nabla f(x)\|_2^2}{2\mu}$

So we have :

$$(1) \implies f(x) - f^* \leqslant \frac{\|\nabla f(x)\|_2^2}{2\mu}$$

- (1) implies to f(x)+r(x) is strongly convex for any convex f and strongly convex r

Answer : Let f be a convex function and r a strongly convexe. We assume f and r is continously differentiable, then for x, y $\in$ X we have :

$$\begin{cases} f(y) \geqslant f(x) + \nabla f(x)^T (y - x) \\ r(y) \geqslant r(x) + \nabla r(x)^T (y - x) + \frac{\mu}{2} \|y - x\|_2^2 \end{cases}$$

$\implies (f+r)(y) \geqslant (f+r)(x) + (\nabla f + \nabla r)(x)^T (y-x) + \frac{\mu}{2} \|y-x\|_2^2$

So, we can say that f + r is strongly convex

A function f : $\mathbb{R}^d \to \mathbb{R}$ is L-smooth if it is differentiable and its gradient is L-Lipschitz-continuous (usually w.r.t. norm-2) :

$$\forall x_1, x_2 \in \mathbb{R}^d, \|\nabla f(x_2) - \nabla f(x_1)\|_2 \leqslant L \|x_2 - x_1\|^2, \qquad (2)$$

Prove that (2) implies all the following statements (assume convexity if needed) :

(a) $f(x_2) \leqslant f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{L}{2} \|x_2 - x_1\|_2^2, \forall x_1, x_2 \in X$

Answer : Let x_1 and $x_2 \in X$, we have :

$f(x_2) = f(x_1) + \displaystyle\int_0^1 < \nabla f(x_1 + t(x_2 - x_1)), x_2 - x_1 > dt$

$\implies f(x_2) = f(x_1) + \displaystyle\int_0^1 < \nabla f(x_1 + t(x_2 - x_1)) - \nabla f(x_1) + \nabla f(x_1), x_2 - x_1 > dt$

$\implies f(x_2) = f(x_1) + \displaystyle\int_0^1 < \nabla f(x_1 + t(x_2 - x_1)) - \nabla f(x_1), x_2 - x_1 > dt + \displaystyle\int_0^1 < \nabla f(x_1), x_2 - x_1 > dt$

$\implies f(x_2) = f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \displaystyle\int_0^1 < \nabla f(x_1 + t(x_2 - x_1)) - \nabla f(x_1), x_2 - x_1 > dt$

Using Cauchy-Sawartz inequality, we have for all t $\in$ [0,1],

$| < \nabla f(x_1 + t(x_2 - x_1)) - \nabla f(x_1), x_2 - x_1 > | \leqslant \|\nabla f(x_1 + t(x_2 - x_1)) - \nabla f(x_1)\| \|x_2 - x_1\|$

Using inequality (2), we have :

$\|\nabla f(x_1 + t(x_2 - x_1)) - \nabla f(x_1)\| \leqslant L \|x_1 + t(x_1 + x_2) - x_1\| = tL \|x_2 - x_1\|$

$\implies | < \nabla f(x_1 + t(x_2 - x_1)) - \nabla f(x_1), x_2 - x_1 > | \leqslant tL \|x_2 - x_1\|^2$

$\implies \displaystyle\int_0^1 < \nabla f(x_1 + t(x_2 - x_1)) - \nabla f(x_1), x_2 - x_1 > dt \leqslant \displaystyle\int_0^1 | < \nabla f(x_1 + t(x_2 - x_1)) -$

$\nabla f(x_1), x_2 - x_1 > |dt \leqslant \int_0^1 tL\|x_2 - x_1\|^2 dt = \frac{L}{2}\|x_2 - x_1\|^2$

$\implies f(x_2) = f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \int_0^1 < \nabla f(x_1 + t(x_2 - x_1)) - \nabla f(x_1), x_2 - x_1 > dt \leqslant f(x_1) + \nabla f(x_1)^T (x_2 - x_1) \int_0^1 | < \nabla f(x_1 + t(x_2 - x_1)) - \nabla f(x_1), x_2 - x_1 > |dt \leqslant f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{L}{2}\|x_2 - x_1\|^2$

$$\implies \boxed{f(x_2) \leqslant f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{L}{2}\|x_2 - x_1\|^2}$$

(b) $f(x_2) \geqslant f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{1}{2L} \|\nabla f(x_2) - \nabla f(x_1)\|_2^2, \forall x_1, x_2 \in X$

Answer : Let $x_1, x_2 \in X$,

Let y $\in$ X,

$f(x_1) - f(x_2) = f(x_1) - f(y) + f(y) - f(x_2)$

Using (a) we have ,

$f(y) - f(x_2) \leqslant \nabla f(x_2)^T(y - x_2) + \frac{L}{2}\|y - x_2\|^2$

$\implies f(x_1) - f(x_2) \leqslant f(x_1) - f(y) + \nabla f(x_2)^T(y - x_2) + \frac{L}{2}\|y - x_2\|^2$

If we assume f is convex, then we have

$f(x_1) - f(y) \leqslant -\nabla f(x_1)^T(y - x_1)$

$\implies f(x_1) - f(x_2) \leqslant -\nabla f(x_1)^T(y - x_1) + \nabla f(x_2)^T(y - x_2) + \frac{L}{2}\|y - x_2\|^2$

$\implies f(x_1) - f(x_2) \leqslant -\nabla f(x_1)^T(y - x_2) - \nabla f(x_1)^T(x_2 - x_1) + \nabla f(x_2)^T(y - x_2) + \frac{L}{2}\|y - x_2\|^2$

$\implies f(x_1) - f(x_2) \leqslant -\nabla f(x_1)^T(x_2 - x_1) + (\nabla f(x_2) - \nabla f(x_1))^T (y - x_2) + \frac{L}{2}\|y - x_2\|^2$

Let $y = x_2 - \frac{1}{L}(\nabla f(x_2) - \nabla f(x_1))$ then $y - x_2 = -\frac{1}{L}(\nabla f(x_2) - \nabla f(x_1))$

So we have :

$f(x_1) - f(x_2) \leqslant -\nabla f(x_1)^T(x_2 - x_1) - \frac{1}{L}(\nabla f(x_2) - \nabla f(x_1))^T (\nabla f(x_2) - \nabla f(x_1)) + \frac{1}{L^2} * \frac{L}{2}\|\nabla f(x_2) - \nabla f(x_1)\|^2$

$$\Longrightarrow f(x_1) - f(x_2) \leqslant -\nabla f(x_1)^T(x_2 - x_1) - \tfrac{1}{2L}\|\nabla f(x_2) - \nabla f(x_1)\|^2$$

$$\Longrightarrow \boxed{f(x_1) + \nabla f(x_1)^T(x_2 - x_1) + \frac{1}{2L}\|\nabla f(x_2) - \nabla f(x_1)\|^2 \leqslant f(x_2)}$$

Part II : Convergence of plain GD

Consider the following optimization problem : $\underset{x\in\mathbb{R}^d}{min} f(x)$ where f is strongly convex with conqtant mu , continuous with Lipschitz constant L. In other words, f is $\mu-$strongly convex and L-smooth. Moreover, consider the Gradient descent (GD) with constant step-size

$$x_{k+1} = x_k - \alpha\nabla f(x_k), \alpha \geq 0 \qquad (3)$$

We set the (constant) step-size as $\alpha = \frac{2}{L+\mu}$. This variant is known as plain/vanilla GD.

(1) Prove the convergence of GD with constant step size. Specifically, show that the iterations in (3) are such that :

$$\|x_k - x^*\|_2^2 \leqslant \left(1 - \frac{2}{1+\frac{L}{2}}\right)^{2k} \|x_0 - x^*\|_2^2$$

Hints : you may use these observations.

From smoothness and vanishing gradient of the optimal point, conclude

$f(x_k) - f(x^*) \leqslant \frac{L}{2}\|x_k - x^*\|_2^2 \qquad (*)$

Use the coercivity of the gradient :

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geqslant \frac{\mu L}{\mu + L}\|x - y\|_2^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

Iterate over k and use (*) to obtain :

$$f(x_k) - f(x^*) \leqslant \frac{L}{2}\prod_{i\in[k]}\left(1 - 2\alpha\frac{\mu L}{\mu + L}\right)\|x_0 - x^*\|_2^2$$

Answer : We have

$$\|x_{k+1} - x^*\|^2 = \|x_k - \alpha \nabla f(x_k) - x^*\|^2 = \| (x_k - x^*) - \alpha \nabla f(x_k)\|^2$$

$\Longrightarrow \|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\alpha < \nabla f(x_k), x_k - x^* > +\alpha^2 \|\nabla f(x_k)\|^2$

We have $\nabla f(x^*) = 0$ so, we can write :

$\Longrightarrow \|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\alpha < \nabla f(x_k) - \nabla f(x^*), x_k - x^* > +\alpha^2 \|\nabla f(x_k) - \nabla f(x^*)\|^2$

Using the coercivity of the gradient of f, we have :

$(\nabla f(x) - \nabla f(y))^T (x - y) \geqslant \frac{\mu L}{\mu + L} \|x - y\|^2 + \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|^2 \; \forall x, y \in X$

we have :

$\|x_{k+1} - x^*\|^2 \leqslant \|x_k - x^*\|^2 - 2\alpha \left(\frac{\mu L}{\mu + L} \|x_k - x^*\|^2 + \frac{1}{\mu + L} \|\nabla f(x_k) - \nabla f(x^*)\|^2 \right) + \alpha^2 \|\nabla f(x_k) - \nabla f(x^*)\|^2$

$\Longrightarrow \|x_{k+1} - x^*\|^2 \leqslant \left(1 - 2\alpha \frac{\mu L}{\mu + L}\right) \|x_k - x^*\|^2 + \left(\alpha^2 - \frac{2\alpha}{\mu + L}\right) \|\nabla f(x_k) - \nabla f(x^*)\|^2$

We have :

$$\alpha^2 - \frac{2\alpha}{\mu + L} = 0$$

so, we have :

$\|x_{k+1} - x^*\|^2 \leqslant \left(1 - 2\alpha \frac{\mu L}{\mu + L}\right) \|x_k - x^*\|^2$

$\|x_{k+1} - x^*\|^2 \leqslant \left(\frac{L - \mu}{\mu + L}\right)^2 \|x_k - x^*\|^2$

$\Longrightarrow \|x_{k+1} - x^*\|^2 \leqslant (1 - \alpha\mu)^2 \|x_k - x^*\|^2$

When we iterate over k on the right side, we get :

$\Longrightarrow \|x_{k+1} - x^*\|^2 \leqslant (1 - \alpha\mu)^{2(k+1)} \|x_0 - x^*\|^2$

we can conclude that :

$$\boxed{\|x_k - x^*\|^2 \leqslant \left(1 - \frac{2}{1+\frac{L}{\mu}}\right)^{2k} \|x_0 - x^*\|^2}$$

2) What is the convergence rate of this variant, in the $\mathcal{O}$(-) sense ?

Answer : The rate of convergence of this variant is $(c)^k = \mathcal{O}((c)^{2k})$ where $c = 1 - \frac{2}{1+\frac{L}{\mu}}$

3) Recall : μ and L are upper and lower bounds on the largest and smallest eigenvalues of the Hessian of a μ-strongly convex and L-smooth, respectively

a) Wath happens to the convergence rate (question 1) when L/$\mu \to 1$? Explain to which scenario does this correspond. Discuss the pratical implications of this scenario for DG.

Answer : When $\frac{L}{\mu} \to 1$, the convergence constant c $= 1 - \frac{2}{1+\frac{L}{\mu}} \to 0$ so the convergence rate $(c)^k$ decreases quickly : We have a fastest convergence for GD.

b) What happens to the convergence rate (question 1) when L/$\mu \to \infty$? Explain to which scenario does this correspond. Discuss the practical implications of this scenario for GD.

Answer : When $\frac{L}{\mu} \to \infty$, the convergence constant c $= 1 - \frac{2}{1+\frac{L}{\mu}} \to 1$ so the convergence rate $(c)^k$ decreases slowly : We have a slowest convergence for GD.

Part III : Finding L and μ

Consider a linear ridge regression : $\underset{w}{min}$ f(w)=$\frac{1}{N}\sum_{i\in[N]} f_i(w) + \lambda\|w\|_2^2$ where the loss for sample i is given by : $f_i(w) = \left(y_i - w^T x_i\right)^2$. Use the Bodyfat dataset (available in the ML toolbox in MATLAB)

a) Is f Lipschitz continuous ? If so, find a small Lipschitz constant L ?

Answer : We have for all w :

$$f(w) = \frac{1}{N} \sum_{i \in [N]} f_i(w) + \lambda \|w\|_2^2$$

$\implies f(w) = \frac{1}{N} \sum_{i \in [N]} \left(f_i(w) + \lambda \|w\|_2^2\right)$

Let $g_i(w) = f_i(w) + \lambda \|w\|_2^2$

then we have :

$f(w) = \frac{1}{N} \sum_{i \in [N]} g_i(w)$

NB : we consider the functions gi thus defined in the other questions

Let B $>$ 0 and $\mathcal{H} = \left\{w \in \mathbb{R}^d : \|w\|_2 \leqslant B\right\}$

Let w $\in \mathcal{H}$ then we have :

$g_i(w) = \left(y_i - w^T x_i\right)^2 + \lambda \|w\|_2^2$

$\implies \nabla g_i(w) = -2x_i \left(y_i - w^T x_i\right) + 2\lambda w$

$\implies \nabla g_i(w) = -2x_i y_i + 2\left(x_i x_i^T + \lambda I_d\right) w$

$\implies \|\nabla g_i(w)\| = \|-2x_i y_i + 2\left(x_i x_i^T + \lambda I_d\right) w\| \leqslant 2\|x_i y_i\| + 2\| \left(x_i x_i^T + \lambda I_d\right) \| B$

$\implies \|\nabla f_i(w)\| = \|\frac{1}{N} \sum_{i \in [N]} \nabla g_i(w)\| \leqslant \frac{1}{N} \sum_{i \in [N]} \|\nabla g_i(w)\| \leqslant \frac{2}{N} \sum_{i \in [N]} \|x_i y_i\| + \frac{2B}{N} \sum_{i \in [N]} \| \left(x_i x_i^T + \lambda I_d\right) \|$

$\implies \|\nabla f_i(w)\| \leqslant \frac{2}{N} \sum_{i \in [N]} \|x_i y_i\| + \frac{2B}{N} \sum_{i \in [N]} \| \left(x_i x_i^T + \lambda I_d\right) \|$

So we can say that f is L-lipschitz and we can takes

$$L = \frac{2}{N} \sum_{i \in [N]} \|x_i y_i\| + \frac{2B}{N} \sum_{i \in [N]} \| \left(x_i x_i^T + \lambda I_d \right) \|$$

as lipschitz constant

b) Is f strongly convex ? If so, find a large μ ?

Answer : Consider this function : $h(w) = \lambda \|w\|_2^2$ then :

$\nabla h(w) = 2\lambda w$

$\implies \nabla^2 h(w) = 2\lambda I_d$

$\implies \nabla^2 h(w) \succ 2\lambda I_d$

So, the h function is 2λ-strongly convex (**)

For i $\in [N]$, we have

$f_i(w) = \left(y_i - w^T x_i \right)^2$

$\implies \nabla f_i(w) = -2x_i \left(y_i - w^T x_i \right)$

$\implies \nabla^2 f_i(w) = 2x_i x_i^T$

$\implies \nabla^2 f_i(w) \succcurlyeq 0$

so we can say that the function f_i is convex (***)

(**) and (***) $\implies g_i(w) = f_i(w) + h(w)$ is 2λ-strongly convex

So, $f(w) = \dfrac{1}{N} \displaystyle\sum_{i \in [N]} g_i(w)$ is 2λ-strongly convex and we can choose

$$\boxed{\mu = 2\lambda}$$

c) There is a simple way (trick) to find L and μ, for the optimization problem considered here.
Express, mathematically, the steps for doing so, and derive the expressions for L, μ, and their ratio L/μ

What inherent properties of the dataset impact (and determine) the ratio L/μ ?

Answer : We can find the values of L and μ by calculating the eigenvalues of the Hessian matrix of f. Indeed μ and L are respectively the min and the max of the eigenvalues of the Hessian matrix of f.

step 1 : We compute the matrix $\nabla^2 f(w) = \dfrac{2}{N} \sum_{i \in [N]} x_i x_i^T + 2\lambda I_d$

step 2 : Compute the eigenvalues $\{\lambda_i\}_{i \in [N]}$ of $\nabla^2 f(w)$

step 3 : compute the values of L and μ

L= max $\{\lambda_i\}_{i \in [N]}$

μ= min $\{\lambda_i\}_{i \in [N]}$

To calculate μ and L, we need to determine the eigenvalues of $\nabla^2 f(w) = \dfrac{2}{N} \sum_{i \in [N]} x_i x_i^T + 2\lambda I_d$.
So, the larger our dataset, i.e. N is big , more we need times to calculate $\dfrac{2}{N} \sum_{i \in [N]} x_i x_i^T$.
Thus, we can say that the size of our data set (N) impacts the calculation of the ratio $\frac{L}{\mu}$

d) What can you say about the ratio L/μ for the Bodyfat dataset ? is it a 'good' or 'bad' setup for a plain GD method

Answer : The bodyfat dataset contains 252 items, so the number of items is relatively small. So this is a good dataset for GD method

Part IV : Duality and Optimality for Equality Constrained Quadratic Program

Consider the following Equality Constrained Quadratic Program (ECQP).

$$ECQP : x^* := \begin{cases} \underset{x \in \mathbb{R}^d}{arg\ min} f_0(x) = x^T S x \\ \qquad s.t. \qquad\quad Ax = b \end{cases}$$

where $\mathbf{S} \in \mathbb{R}^{d \times d} \succ 0$ is a positive definite matrix, A $\in \mathbb{R}^{d \times d}$ and b $\in \mathbb{R}^d$

1) Derive the Lagrangian $\mathcal{L}()$, the Lagrange function g(), and the dual problem, that correspond to the above ECQP

Answer : We have

$$\mathcal{L}(x, \upsilon) = f_0(x) + \upsilon^T (Ax - b)$$

$$\Longrightarrow \boxed{\mathcal{L}(x, \upsilon) = x^T S x + \upsilon^T (Ax - b)}$$

Nothing that for all υ, the function $x \to \mathcal{L}(x, \upsilon)$ is strongly convex and continuously differentiable. So, we have :

$\hat{x} = \underset{x \in \mathbb{R}^d}{arg\ min}\ \mathcal{L}(x, \upsilon) \in \{x \in \nabla \mathcal{L}_x(x, \upsilon) = 0\}$

We have :

$\nabla \mathcal{L}_x(x, \upsilon) = \nabla_x \left(x^T S x + \upsilon^T (Ax - b) \right)$

$\Longrightarrow \nabla \mathcal{L}_x(x, \upsilon) = 2Sx + A^T \upsilon$

$\Longrightarrow \nabla \mathcal{L}_x(x, \upsilon) = 0 \Longleftrightarrow 2Sx + A^T \upsilon = 0$

$\Longrightarrow x = -(2S)^{-1} A^T \upsilon$ So we have

$$\boxed{\hat{x} = -(2S)^{-1} A^T \upsilon}$$

We have $g(\upsilon) = \underset{x \in \mathbb{R}^d}{inf}\, \mathcal{L}(x, \upsilon) = \mathcal{L}(\hat{x}, \upsilon)$

$\Longrightarrow g(\upsilon) = \mathcal{L}(-(2S)^{-1} A^T \upsilon, \upsilon)$

$\Longrightarrow g(\upsilon) = \left[-(2S)^{-1} A^T \upsilon\right]^T S \left[-(2S)^{-1} A^T \upsilon\right] + \upsilon^T \left(A \left[-(2S)^{-1} A^T \upsilon\right] - b\right)$

$\Longrightarrow g(\upsilon) = \frac{1}{4} \left[A^T \upsilon\right]^T S^{-1} \left[A^T \upsilon\right] - \frac{1}{2} \left[A^T \upsilon\right]^T S^{-1} \left[A^T \upsilon\right] - \upsilon^T b$

$$\boxed{g(\upsilon) = -\frac{1}{4}\left[A^T\upsilon\right]^T S^{-1}\left[A^T\upsilon\right] - \upsilon^T b}$$

The Lagragian dual problem D is define by :

$$\underset{\upsilon}{max}\ g(\upsilon) = \underset{\upsilon}{min}\ -g(\upsilon) = \underset{\upsilon}{min}\ \left\{\frac{1}{4}\left[A^T\upsilon\right]^T S^{-1}\left[A^T\upsilon\right] + \upsilon^T b\right\}$$

As S $\succ$ 0 then $S^{-1} \succ 0, \Longrightarrow$ the dual problem is trongly convex QP in υ. So we have :

$$-\nabla g(\upsilon) = \frac{1}{2}AS^{-1}A^T\upsilon + b$$

$\Longrightarrow -\nabla g(\upsilon) = 0 \Longleftrightarrow \frac{1}{2}AS^{-1}A^T\upsilon + b = 0$

$\Longrightarrow -\nabla g(\upsilon) = 0 \Longleftrightarrow \upsilon = -2\left(AS^{-1}A^T\right)b$

So we have :

$$\boxed{\hat{\upsilon} = arg\,\underset{\upsilon}{max}\ g(\upsilon) = -2\left(AS^{-1}A^T\right)b}$$

$$\Longrightarrow \hat{x} = S^{-1}A^T\left(AS^{-1}A^T\right)b$$

$$\Longrightarrow \boxed{\hat{x} = \left(S^{-1}A^T\right)A\left(S^{-1}A^T\right)b}$$

2) Derive the KKT conditions that correspond to the above ECQP. Use these KKT condition to derive a closed-form analytical solution, x^* ? , as a function of optimal dual variables, μ^* ?

Answer : We have :

1- The ECQP problem is strongly convex

2- $\nabla\mathcal{L}(\hat{x}, \hat{\upsilon}) = 0$

3- The contrain function (Ax-b) is affine

So using the KKT condition, we can say : the close form analytical solution of ECQP problem is given by :

$$\boxed{\hat{x} = \left(S^{-1}A^T\right) A \left(S^{-1}A^T\right) b}$$

we have also :

$$\boxed{\hat{\upsilon} = arg \underset{\upsilon}{max}\ g(\upsilon) = -2\left(AS^{-1}A^T\right) b}$$