MICAS901 - Introduction to Optimization: Homework 1

Please submit your solution by email before November 22nd, 2020

Hadi Ghauch, Telecom Paris,
email: hadi.ghauch@telecom-paristech.fr

All the steps and derivations are needed for full credit. The solution for each homework should be submitted individually by each student (one student per homework). Regarding the questions that need programming or coding, feel free to use any programing language you are comfortable with. The homework is worth 20% of total course grade.

Convex Optimization

Part 1: Convex Analysis

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

$$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T(\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{\mu}{2}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2 \text{ }, \forall \text{ } (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{X} \tag{1}$$

Prove all the following statements

- (1) is equivalent to a minimum positive curvature $\nabla^2 f(\boldsymbol{x}) \succeq \mu \boldsymbol{I}_d, \forall \boldsymbol{x} \in \mathcal{X}$
- (1) is equivalent to $\left(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\right)^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq \mu \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2$
- (1) implies $f(\boldsymbol{x}) - f^\star \leq \dfrac{1}{2\mu}\|\nabla f(\boldsymbol{x})\|_2^2, \forall \boldsymbol{x}$
- (1) implies $f(\boldsymbol{x}) + r(\boldsymbol{x})$ is strongly convex for any convex f and strongly convex r

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

$$\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d, \|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2 \leq L\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 \tag{2}$$

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

(a) $f(\boldsymbol{x}_2) \leq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T(\boldsymbol{x}_2 - \boldsymbol{x}_1) + \dfrac{L}{2}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2, \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$

(b) $f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T(\boldsymbol{x}_2 - \boldsymbol{x}_1) + \dfrac{1}{2L}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2, \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$

Part II: Convergence of plain GD

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

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k) \ , \quad \alpha > 0 \tag{3}$$

We set the (constant) step-size as $\alpha = 2/(L+\mu)$. 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:

$$\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2 \leq \left(1 - \frac{2}{1 + L/\mu}\right)^{2k} \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2$$

Hints: you may use these observations.

From smoothness and vanishing gradient of the optimal point, conclude
$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \leq \frac{L}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2 \qquad (*)$

Use the coercivity of the gradient:

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

Iterate over k and use $(*)$ to obtain

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \leq \frac{L}{2} \prod_{i\in[k]} \left(1 - 2\alpha\frac{\mu L}{\mu + L}\right) \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2$$

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

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

Part III: Finding L and μ

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

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

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

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?

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 ?

Part IV: Duality and Optimality for Equality Constrained Quadratic Program

Consider the following Equality Constrained Quadratic Program (ECQP).

$$\textbf{ECQP:} \quad \boldsymbol{x}^\star := \begin{cases} \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} & f_0(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{S}\boldsymbol{x} \\ \text{s.t.} & \boldsymbol{A}\boldsymbol{x} = \mathbf{b} \end{cases} \tag{4}$$

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

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

2) Derive the KKT conditions that correspond to the above ECQP. Use these KKT condition to derive a closed-form analytical solution, $\boldsymbol{x}^\star$, as a function of optimal dual variables, $\mu^\star$