Tutorial 3 : Gradient descent and theoretical properties

Exercise 1 (Gradient descent update)**.** Given a sequence of pair $(x_i, y_i), y_i \in \{\pm 1\}$, we consider the following optimization problem (also known as *logistic regression*).

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-y_i x_i^\top \theta)). \tag{1}$$

Prove that the update of gradient descent for $\mathcal{L}$ is given by:

$$\theta_{k+1} = \theta_k - \frac{\alpha}{n}\sum_{i=1}^{n} \frac{-y_i}{1 + \exp(y_i x_i^\top \theta_k)} x_i.$$

Exercise 2 (Unproven proposition 1)**.** If f is L-smooth, then for all $x, y \in \mathbb{R}^d$, we have:

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|x - y\|_2^2.$$

Exercise 3 (Unproven proposition 2)**.** Consider $f : \mathbb{R}^d \to \mathbb{R}$ a μ-strongly convex function, we have:

1. $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{1}{2}\mu t(1-t)\|x - y\|^2, \forall x, y \in \mathbb{R}^d, \forall t \in [0, 1]$.
2. If f is C^1, then $f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|x - y\|^2, \forall x, y \in \mathbb{R}^d$.
3. $(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \mu \|x - y\|^2, \forall x, y \in \mathbb{R}^d$.
4. If f is C^2, then $\nabla^2 f(x) \succeq \mu \mathbf{I}$ (i.e., $\nabla^2 f(x) - \mu \mathbf{I}$ est positive semidefinite).

Exercise 4 (Gradient descent on quadratic optimization)**.** Consider a simple quadratic optimization of the form:

$$f(x) = \frac{1}{2} x^\top \mathbf{A} x$$

where $\mathbf{A} \in \mathbb{S}^{d\times d}$ is a symmetric matrix. Remind that if $\mathbf{A} \in \mathbb{S}^{d\times d}$ is a symmetric matrix, there exists an orthogonal $\mathbf{Q} \in \mathbb{R}^{d\times d}$ and a diagonal matrix $\mathbf{D} \in \mathbb{R}^{d\times d}$ such that $\mathbf{A} = \mathbf{Q}^\top \mathbf{D}\mathbf{Q}$. Answer the following question:

1. Prove that f is convex if and only if $\mathbf{D}$ has nonnegative entries. Deduce that f is μ-strongly convex if and only if all the coefficients in the diagonal of $\mathbf{D}$ are at least μ.
2. If f is convex, prove that $f^\star = \min_x f(x) = 0$.

3. If f is μ-strongly convex, what can we deduce about the convergence speed $f(x_k)$ to 0 if x_k is generated by gradient descent (with a proper choice of the learning rate).

4. If f is only convex, can we say the same thing about the convergence speed of $f(x_k)$ obtained by gradient descent (with a proper choice of the learning rate).

Exercise 5 (Armijo and Wolfe conditions)**.** Armijo (A) and Wolfe (W) conditions provides criteria to choose the step-size $\alpha_k > 0$ when minimizing a C^1 function f. Assume that at the kth iteration, one considers update the $(k+1)$th iterate as:

$$x_{k+1} = x_k + \alpha p_k,$$

these two conditions requires:

$$f(x_{k+1}) \leq f(x_k) + \alpha_k c_1 \nabla f(x_k)^\top p_k, \qquad c_1 \in (0,1) \tag{A}$$

$$\nabla f(x_{k+1})^\top p_k \geq c_2 \nabla f(x_k)^\top p_k, \qquad c_2 \in (c_1, 1) \tag{W}$$

Our goal is to prove that if f is C^1 and bounded below, $p_k^\top \nabla f(x_k) < 0$, then there exists $\alpha > 0$ satisfying (A) and (W). Consider two functions:

$$\varphi(\alpha) = f(x_k + \alpha p_k) \quad , \quad \ell(\alpha) = f(x_k) + \alpha c_1 \nabla f(x_k)^\top p_k.$$

1. Define $g(\alpha) = \varphi(\alpha) - \ell(\alpha)$. Prove that there is an interval $(0, \bar{\alpha})$ such that $g(\alpha) > 0, \forall \alpha \in (0, \bar{\alpha})$ and $g(0) = g(\bar{\alpha}) = 0$.

2. By mean value theorem, prove that there exists $\tilde{\alpha}$ such that:

$$\frac{\varphi(\bar{\alpha}) - \varphi(0)}{\bar{\alpha}} = \varphi'(\tilde{\alpha}).$$

3. Conclude that $\tilde{\alpha}$ satisfies both (A) and (W).

Exercise 6 (More about line search)**.** Consider the line search conditions (A) and (W).

1. Show that if $0 < c_2 < c_1 < 1$, there may not exists α_k satisfying both (A) and (W).

2. Prove that if α_k satisfies (W) for some $c_2 \in (0,1)$ and $\nabla f(x_k)^\top p_k < 0$, then the following equation (a.k.a curvature condition) is satisfied:

$$(x_{k+1} - x_k)^\top (\nabla f(x_{k+1}) - \nabla f(x_k)) > 0.$$

Exercise 7 (Linear convergence of iterates)**.** Consider a μ-strongly convex, L-smooth function f. Prove that the iterates generated by gradient descent with step-size $1/L$ satisfying that:

$$\|x_k - x^\star\|_2^2 = O\left(\left(1 - \frac{\mu}{L}\right)^k\right).$$

where $x^\star$ is the unique global optimal solution of f.