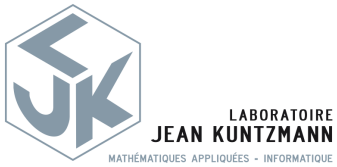

Lecture 3 : Gradient descent and theoretical properties

1 Introduction

In previous lectures, we already discussed about optimization and several theoretical properties. In this lecture, we will focus on algorithms, which are used to find global/local minima or critical points of optimization algorithm. Our spotlight is gradient descent, one of the most famous optimization algorithms and it is still very useful up to nowsaday standard.

2 Gradient descent algorithm

Consider an unconstrained optimization problem:

$$\underset{x\in\mathbb{R}^d}{\text{Minimize}} \quad f(x) \tag{OP}$$

where f is a C^1 function. Gradient descent search for a solution of (OP) by an iterative procedure, where in each iteration, we update the iterates by:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k). \tag{GD}$$

where $\alpha_k > 0$ is called step-size (also known as learning rate in Machine Learning). We can make several remark about (GD):

1. If $\|x_k\| = 0$, then $x_\ell = x_k, \forall \ell \geq k$. Thus, we already find a critical point of f. If we have in additionally that f is convex, then x_k is also a global solution of (OP).
2. If $\|x_k\| \neq 0$, for sufficiently small $\alpha_k > 0$, we have: $f(x_{k+1}) < f(x_k)$ (proof is an exercise). Thus, the objective function is monotonically decreasing along the iterations.

Useful practices concerning (GD) are:

1. *How to choose* x_0: very oftenly, we choose x_0 randomly (from a Gaussian or a uniform distribution).
2. *How to choose* α_k: The choice of α_k is a dedicated subject since the algorithm behaves differently, depending on the choice of α_k. In this lecture, we will only consider the most simple setting: $\alpha_k = \alpha$ is constant. Other criteria such as Armijo and Wolfe conditions were and will be discussed in the TP and TD.
3. *When to stop* (GD)*?* Many criteria are proposed in the literature:

- After reaching a fixed number of iterations (very common in deep learning practice).
- The gradient norm is sufficiently small: $\|\nabla f(x)\| \leq \epsilon$.
- The objective function does not decrease significantly enough $f(x_k) - f(x_{k+1}) \leq \epsilon$.

Example 2.1. Gradient descent for linear function: $f(x) = ax + b$. The update rule is given by:

$$x_{k+1} = x_k - \alpha_k a.$$

Thus, if α_k is constant, we have:

$$\lim_{k\to\infty} x_k = \begin{cases} -\infty & \alpha > 0 \\ x_0 & \alpha = 0 \\ +\infty & \alpha < 0 \end{cases},$$

which is logic since the function is not bounded below in case $\alpha \neq 0$.

Example 2.2. Gradient descent for quadratic loss function. Consider the scalar quadratic optimization problem:

$$\underset{x\in\mathbb{R}^d}{\text{Minimize}}\, f(x) := \frac{1}{2}x^\top \mathbf{A}x + b^\top x + c.$$

The gradient of x is given by:

$$\nabla f(x) = \mathbf{A}x + b.$$

Thus, the update rule for gradient descent with fixed step-size is given by:

$$x_{k+1} = x_k - \alpha_k(\mathbf{A}x_k + b) = (\mathbf{I} - \alpha_k\mathbf{A})x_k - \alpha_k b.$$

The limit of x_k will depend on the spectral of $\mathbf{A}$ and the choice of $\{\alpha_k\}_{k\in\mathbb{N}}$.

3 Theoretical guarantees of gradient descent algorithms

Well-known behavior of gradient descent is studied under the following assumption:

Definition 3.1 (L-smooth function)**.** A function is L-smooth if its gradient is L-Lipschitz, i.e., for all $x, y \in \mathbb{R}^d$, we have:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2.$$

The following lemma provides an important estimation for L-smooth functions:

Proposition 3.2 (Upperbound of L-smooth functions)**.** *If f is L-smooth, then for all $x, y \in \mathbb{R}^d$, we have:*

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

Proof. Exercise. □

Lemma 3.3 (Descent lemma)**.** *Given a L-smooth function f and an iterate given by gradient descent with constant step-size $\alpha > 0$, we have:*

$$f(x_{k+1}) \leq f(x_k) - \left(\alpha - \frac{L\alpha^2}{2}\right)\|\nabla f(x_k)\|^2.$$

Moreover, if $\alpha = \frac{1}{L}$ and $C := \inf f > -\infty$, we have:

$$\min_{k=0,\dots,K} \|\nabla f(x_k)\|^2 \leq \frac{2L}{K}(f(x_0) - C) = O\left(\frac{1}{K}\right).$$

Proof. Using Proposition 3.2, we have:

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \nabla f(x_k)^\top (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= f(x_k) - \alpha\|\nabla f(x_k)\|^2 + \frac{L\alpha^2}{2}\|\nabla f(x_k)\|^2 \\ &= f(x_k) - \left(\alpha - \frac{L\alpha^2}{2}\right)\|\nabla f(x_k)\|^2. \end{aligned} \tag{1}$$

In case $\alpha = \frac{1}{L}$, the inequality (1) become:

$$f(x_k) - f(x_{k+1}) \leq \frac{1}{2L}\|\nabla f(x_k)\|^2. \tag{2}$$

By summing these inequalities for $k = 0, \ldots, K$, we obtain:

$$f(x_0) - C \geq f(x_0) - f(x_{K+1}) = \frac{1}{2L}\sum_{i=0}^{K}\|\nabla f(x_k)\|^2 \geq \frac{K+1}{2L}\min_{k=0,\ldots,K}\|\nabla f(x_k)\|^2,$$

which concludes the proof. □

Remark 3.4. From Lemma 3.3, we conclude that to find an ϵ critical point, i.e., $\|\nabla f(x)\| \leq \epsilon$, one needs $O(\frac{1}{\epsilon^2})$ iterations.

What happen if the function f is also convex In case f is convex, we can show guarantee on the objective function. It is given by the following result.

Proposition 3.5 (Gradient descent for convex, L-smooth functions)**.** *Suppose that $\alpha_k = \alpha = \frac{1}{L}$ and* (OP) *admit at least a global solution $x^\star$. The iterates x_k generated by gradient descent satisfy that:*

$$f(x_k) - f(x^\star) \leq \frac{L}{2k}\|\theta_0 - \theta^\star\|^2 = O\left(\frac{1}{k}\right).$$

Proof. Since f is convex, we have:

$$f(x^\star) \geq f(x) - \nabla f(x)^\top (x - x^\star).$$

Applying this inequality to (2) gives us:

$$\begin{aligned} f(x_k) &\leq f(x_{k-1}) - \frac{1}{2L}\|\nabla f(x_{k-1})\|^2 \\ &\leq f(x^\star) + \nabla f(x_{k-1})^\top (x_{k-1} - x^\star) - \frac{1}{2L}\|\nabla f(x_{k-1})\|^2 \\ &\leq f(x^\star) + \frac{L}{2}\left(\|x_{k-1} - x^\star\|^2 - \|\underbrace{x_{k-1} - \frac{1}{L}\nabla f(x_{k-1})}_{x_k} - x^\star\|^2\right) \end{aligned}$$

By summing these in inequality from $i = 0$ to k, we obtain :

$$\begin{aligned} kf(x^\star) + \frac{L}{2}\left(\|x_0 - x^\star\|^2 - \|x_k - x^\star\|^2\right) &\geq \sum_{i=1}^{k} f(x_i) \\ &\geq kf(x_k) \quad \text{(since } f(x_k) \text{ is monotone)}, \end{aligned}$$

which conclude the proof. □

Remark 3.6. We see that with the convexity of f, we get a stronger result: we proved that the sequence $f(x_k) \to f(x^\star)$ when $k \to \infty$ with rate $O\left(\frac{1}{k}\right)$.

However, we are not unable to prove that $x_k \to x^\star$ (the convergence of iterates). Showing iterate convergence usually requires additional assumption such as Kurdyka-Łojasiewicz inequality. Without this type of hypothesis, the iterates convergence of gradient descent remains an open question.

What happen if the function f is μ-strongly convex

Definition 3.7 (μ-strongly convex function)**.** A function f is called μ-strongly convex if $f(\cdot) - \frac{\mu}{2}\|\cdot\|_2^2$ is convex.

When $\mu = 0$, strong convexity is equivalent to convexity.

Lemma 3.8 (Properties of strongly convex functions)**.** *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).*

Proof. Exercise. □

Proposition 3.9 (Gradient descent for μ-strongly convex, L-smooth functions)**.** *Assume the same hypothesis (on f and $\alpha_k = \alpha = \frac{1}{L}$) as in Proposition 3.5. In addition, we also suppose that f is mu-strongly convex ($\mu > 0$). Then, the iterates generated by* (GD) *verify:*

$$f(x_k) - f(x^\star) \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f(x^\star)), \forall k \in \mathbb{N}.$$

Proof. By applying the second Lemma 3.8 with $y = x^\star$, the optimal solution of f (whose existence and unicity are guaranteed by Lemma 3.8) gives :

$$\begin{aligned} f(x^\star) = f(y) &\geq f(x) + \nabla f(x)^\top (y-x) + \frac{\mu}{2}\|y-x\|^2 \\ &\geq \inf_{y \in \mathbb{R}} f(x) + \nabla f(x)^\top (y-x) + \frac{\mu}{2}\|y-x\|^2 \\ &= f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2. \end{aligned} \tag{3}$$

Combining this inequality with (1) gives:

$$\begin{aligned} f(x_{k+1}) - f(x^\star) &\leq f(x_k) - f(x^\star) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \\ &\leq f(x_k) - f(x^\star) - \frac{\mu}{L}(f(x_k) - f(x^\star)) \\ &= \left(1 - \frac{\mu}{L}\right)(f(x_k) - f(x^\star))). \end{aligned}$$

The result is obtained by applying the previous inequality in a recursive manner. □

Remark 3.10. If we get $f(x_k) - f(x^\star) \leq c^k (f(x_0) - f(x^\star))$, we say that the objective function converge linearly.

Remark 3.11. Under the hypothesis of Proposition 3.9, we can also prove that $x_k \to x^\star$ linearly.

4 Summary on the theoretical guarantees of gradient descent

Hypothesis	Results	Convergence rate
• f is L-smooth. • $\inf_\theta f(\theta) > 0$. • Step-size $\alpha = \frac{1}{L}$	$\min_{k\in\mathbb{N}} \|\nabla f(\theta_k)\|$ tend vers zero	$O\left(\sqrt{\frac{1}{k}}\right)$
• f is L-smooth. • An optimal solution $\theta^\star$ exist. • Step-size $\alpha = \frac{1}{L}$ • f is convex.	$\lim_{k\to\infty} f(\theta_k) = f(\theta^\star)$	$O\left(\frac{1}{k}\right)$
• f is L-smooth. • Step-size $\alpha = \frac{1}{L}$ • f is μ-strongly convex.	• $\lim_{k\to\infty} \theta_k = \theta^\star$ • $\lim_{k\to\infty} f(\theta_k) = f(\theta^\star)$	$O\left(\left(1 - \frac{\mu}{L}\right)^k\right)$

Table 1: Résumé des résultats concernant la descente de gradient.