Lecture 4 : Lower-bound of first-order methods and Nesterov optimal algorithm

1 Introduction

In previous lectures, we already discussed about one of the most well-known optimization algorithms: *gradient descent* and its theoretical guarantee (on the convergence of iterates and of the objective functions). A natural question is: is gradient descent the *best* optimization that we can use when dealing with the class of L-smooth (and convex) functions? And what is the limit of an algorithm optimizing a function of this class? These questions will be (party) revealed in this lecture.

2 First-order methods and their fundamental limits

In this section, we consider the following question:

Question 2.1. Given an optimization algorithm $\mathcal{A}$ and a function class $\mathcal{F}$, investigate the worst objective gap at the kth iteration, i.e.,

$$\ell(\mathcal{F}, \mathcal{A}) := \sup_{f \in \mathcal{F}} f(x_k) - f^\star,$$

where x_k is the kth iteration generated by $\mathcal{A}$ and $f^\star = \inf_x f(x)$.

The sup operator in (2.1) implies that we would like find the *worst* function making the algorithm $\mathcal{A}$ sufferred the most. This gives us a lower-bound on the performance of the algorithm $\mathcal{A}$ when optimizing a function of f.

Of course, we expect that $\ell(\mathcal{F}, \mathcal{A})$ depends on the initialization x_0, the number of iterations k and the properties of the function class $\mathcal{F}$.

To this end, we consider $\mathcal{A}$ and $\mathcal{F}$ as follows:

1. $\mathcal{A}$ is a first-order algorithm, i.e., an iterative method whose sequence of iterates x_k satisfies:

$$x_k \in x_0 + \mathrm{Lin}\{\nabla f(x_0), \dots, \nabla f(x_{k-1})\}, k \geq 1.$$

In particular, gradient descent is a first-order method because:

$$\begin{aligned} x_k &= x_{k-1} - \alpha \nabla f(x_{k-1}) \\ &= x_{k-2} - \alpha \nabla f(x_{k-2}) - \alpha \nabla f(x_{k-1}) \\ &= \cdots \\ &= x_0 - \alpha(\nabla f(x_0) + \cdots + \nabla f(x_{k-1})). \end{aligned}$$

2. $\mathcal{F}$ is the set of convex and L-smooth functions, i.e., f is convex and its gradient is L-Lipschitz.

Throughout this section, we will prove the following result:

Theorem 2.2 (Lower-bound of first-order methods performance)**.** *For any $x_0 \in \mathbb{R}^n$ and $1 \leq k \leq \frac{1}{2}(n-1)$, there exists a function $f \in \mathcal{F}$ such that for any first-order algorithm $\mathcal{A}$, we have:*

$$f(x_k) - f^\star \geq \frac{3L\|x_0 - x^\star\|^2}{32(k+1)^2}$$
$$\|x_k - x^\star\|^2 \geq \frac{1}{8}\|x_0 - x^\star\|^2.$$

where $x^\star \in \mathrm{argmin} f$ and $f^\star = f(x^\star)$.

The proof consists multiple steps, which are detailed below:

Construction of the functions f We consider a family of k quadratic functions ($k = 1, \ldots, n$) given by:

$$f_k(x) = \frac{L}{4}\left\{\frac{1}{2}\left[x_1^2 + \sum_{i=1}^{k-1}(x_i - x_{i+1})^2 + x_k^2\right] - x_1\right\}$$

We prove that $f_k \in \mathcal{F}$. Given a vector $s \in \mathbb{R}^n$, we have:

$$f_k(x+s) - f_k(x) = \underbrace{\frac{L}{4}\left[s_1 x_1 + \sum_{i=1}^{k-1}(s_i - s_{i+1})^\top (x_i - x_{i+1}) + x_k s_k - s_1\right]}_{\text{first-order term}}$$
$$+ \frac{1}{2}\cdot\underbrace{\frac{L}{4}\left(s_1^2 + \sum_{i=1}^{k-1}(s_i - s_{i+1})^2 + s_k^2\right)}_{\text{second-order term}}$$

Therefore, we conclude that:

$$0 \leq s^\top \nabla^2 f(x) s = \frac{L}{4}\left(s_1^2 + \sum_{i=1}^{k-1}(s_i - s_{i+1})^2 + s_k^2\right) \leq \frac{L}{4}\left(s_1^2 + 2\sum_{i=1}^{k-1}(s_i^2 + s_{i+1}^2) + s_k^2\right) \leq L\|s\|_2^2.$$

Therefore, $0 \preceq \nabla^2 f(x) \preceq L\mathbf{I}$. Therefore, f is convex and L-smooth (prove this is an exercise).

The optimal solution of the function f_k We compute the minimum of the function f_k. Since f is convex, it is sufficient to find x such that $\nabla f(x) = 0$. The gradient of f (up to a constant $\frac{L}{4}$) is given by:

$$\nabla f(x) = \begin{pmatrix} 2x_1 - x_2 \\ -x_1 + 2x_2 - x_3 \\ \ldots \\ -x_{k-2} + 2x_{k-1} - x_k \\ -x_{k-1} + 2x_k \end{pmatrix} - e_1$$

Therefore, to have $\nabla f(x) = 0$, one deduces (by induction) that:

$$x_i = (k - i + 1)x_k,$$

and hence, we obtain:

$$[x_k^\star]_i = \begin{cases} 1 - \frac{i}{k+1}, & i = 1, \dots, k \\ \mathbb{R}, & k+1 \le i \le n. \end{cases}$$

The optimal value of f_k is given by:

$$f_k^\star = f_k(x_k^\star) = -\frac{L}{8} \cdot \frac{k}{k+1} = -\frac{L}{8}\left(1 - \frac{1}{k+1}\right).$$

The iterates generated by first-order methods $\mathcal{A}$ Let's fix a value $1 \le p \le n$.

Proposition 2.3 (Properties of first-order methods)**.** *Let $x_0 = 0$. For any sequence $\{x_k\}_{k=0}^p$ generated by a first-order method $\mathcal{A}$, we have:*

$$x_k \in \mathbb{R}^{k,n} := \{x \in \mathbb{R}^n \mid x_i = 0, k+1 \le i \le n\}.$$

As a consequence, we have: $f_p(x_k) = f_k(x_k) \ge f_k^\star, \forall p = k, \dots, n.$

Proof. The proof is conducted by induction and the form of the gradient of f.

Since $f_p - f_k$ contains monomial of the form $x_j x_l$ where $j, l \ge k$, we have:

$$f_p(x_k) = f_k(x_k) \ge f_k^\star.$$

□

The actual proof of Theorem 2.2 WLOG, we consider first-order methods $\mathcal{A}$ whose initilization is $x_0 = 0$. Otherwise, we can use the following remark:

Remark 2.4. One can choose different $x_0 \neq 0$. However, the result remains the same by considering $g_k(\cdot) = f_k(\cdot + x_0)$.

We prove two claims one by one: since we assume $k \le \frac{1}{2}(n-1)$, we consider the function $f = f_{2k+1}$ and $x^\star = x_{2k+1}^\star$: We have:

$$\begin{aligned}\|x^\star - x_0\| &= \sum_{i=1}^{2k+1}\left(\frac{i}{2k+2}\right)^2 = \frac{1}{(2k+2)^2}\sum_{i=1}^{2k+1} i^2 \\ &= \frac{1}{(2k+2)^2}\frac{(2k+1)(2k+2)(4k+3)}{6} \le \frac{2k+2}{3}.\end{aligned}$$

1. **First claim**:

$$\frac{f(x_k) - f^\star}{\|x_0 - x^\star\|^2} \ge \frac{L}{8} \cdot \frac{\frac{1}{k+1} - \frac{1}{2k+2}}{\frac{2k+2}{3}} = \frac{3L}{32} \cdot \frac{1}{(k+1)^2}.$$

2. **Second claim**: Consider the quantity $\|x_k - x^\star\|^2$:

$$\begin{aligned}\|x_k - x^\star\|_2^2 &\ge \sum_{i=k+1}^{2k+1}([x_{2k+1}^\star]_i)^2 = \sum_{i=k+1}^{2k+1}\left(1 - \frac{i}{2k+2}\right)^2 \\ &= k+1 - \frac{1}{k+1}\sum_{i=k+1}^{2k+1} i + \frac{1}{4(k+1)^2}\sum_{i=k+1}^{2k+1} i^2 \\ &= k+1 - \frac{3k+2}{2} + \frac{(2k+1)(7k+6)}{24(k+1)} \\ &\ge \frac{2k^2+7k+6}{24(k+1)} \ge \frac{2k^2+7k+6}{16(k+1)^2}\|x_0 - x^\star\|^2 \ge \frac{1}{8}\|x_0 - x^\star\|^2.\end{aligned}$$

3 Optimal algorithm with Nesterov acceleration

Knowing that the lower-bound is $O\left(\frac{1}{k^2}\right)$, we wonder if there exists an optimization attaining this limit. The answer is affirmative, which is the famous Nesterov acceleration method. It is defined as follows:

$$\begin{aligned} \lambda_k &= \frac{1+\sqrt{1+4\lambda_{k-1}^2}}{2} \\ \gamma_k &= \frac{1-\lambda_k}{\lambda_{k+1}} \\ x_{k+1} &= y_k - \frac{1}{L}\nabla f(y_k) \\ y_{k+1} &= (1-\gamma_k)x_{k+1} + \gamma_k x_k \end{aligned} \tag{Nes-Acc}$$

where $\lambda_0 = 0$ and $x_0 = y_0$ is a (random) initialization. The main result of this section is the following:

Theorem 3.1 (Theoretical guarantee of (Nes-Acc))**.** *Given a convex and L-smooth function f, the iterates generated by* (Nes-Acc) *satisfy:*

$$f(x_k) - f^\star \le \frac{2L\|y_1 - x^\star\|^2}{(k+1)^2}.$$

where we assume $x^\star \in \operatorname{argmin} f$ is a non-empty set and $f^\star = f(x^\star)$.

Proof. Since f is convex and L-smooth, we have:

$$\begin{aligned} f\left(x - \frac{1}{L}\nabla f(x)\right) - f(y) &\le f\left(x - \frac{1}{L}\nabla f(x)\right) - f(x) + \nabla f(x)^\top (x-y) \\ &\le -\frac{1}{2L}\|\nabla f(x)\|^2 + \nabla f(x)^\top (x-y). \end{aligned}$$

Applying the previous results for $x = y_k$ and $y = x_k$, we have:

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

Similarly, we apply the result for $x = y_k$ and $y = x^\star$:

$$f(x_{k+1}) - f(x^\star) \le -\frac{L}{2}\|x_{k+1} - y_k\|^2 - L(x_{k+1}-y_k)^\top (y_k - x^\star) \tag{2}$$

Multiplying (1) by $(\lambda_k - 1)$ and adding to (2) yields:

$$\lambda_k \underbrace{\left(f(x_{k+1}) - f^\star\right)}_{\delta_{k+1}} - (\lambda_k - 1)\underbrace{\left(f(x_k) - f^\star\right)}_{\delta_k} \le -\frac{L\lambda_k}{2}\|x_{k+1} - y_k\|^2 - L(x_{k+1}-y_k)^\top (\lambda_k y_k - (\lambda_k - 1)x_k - x^\star)$$

By definition, we have: $\lambda_{k-1}^2 = \lambda_k^2 - \lambda_k$. Therefore, multiplying the previous inequality by λ_k, we have:

$$\begin{aligned} \lambda_k^2 \delta_{k+1} - \lambda_{k-1}^2 \delta_k &\le -\frac{L}{2}\left(\|\lambda_k(x_{k+1}-y_k)\|^2 + 2\lambda_k(x_{k+1}-y_k)^\top(\lambda_k y_k - (\lambda_k - 1)x_k - x^\star)\right) \\ &= -\frac{L}{2}\left(\|\lambda_k x_{k+1} - (\lambda_k - 1)x_k - x^\star\|^2 - \|\lambda_k y_k - (\lambda_k - 1)x_k - x^\star\|^2\right) \\ &= -\frac{L}{2}\left(\|\lambda_{k+1} y_{k+1} - (\lambda_{k+1} - 1)x_{k+1} - x^\star\|^2 - \|\lambda_k y_k - (\lambda_k - 1)x_k - x^\star\|^2\right) \end{aligned}$$

Telescoping, we get:

$$\lambda_{k-1}^2 \delta_k \leq \frac{L}{2} \|x_1 - x^\star\|^2.$$

The proof is finished by noticing that $\lambda_{k-1} \geq \frac{k}{2}, \forall k \geq 1$. □

4 Summary on the theoretical guarantees of first-order methods

Hypothesis	Gradient descent	Nesterov acceleration
• f is L-smooth. • An optimal solution $\theta^\star$ exist. • Step-size $\alpha = \frac{1}{L}$ • f is convex.	$f(x_k) - f^\star = O\left(\frac{1}{k}\right)$ $x_k \to x^\star \in \text{argmin} f$?	$f(x_k) - f^\star = O\left(\frac{1}{k^2}\right)$ $x_k \to x^\star$

Table 1: Comparison between gradient descent and Nesterov acceleration