Lecture 8

8.1 Review: Lower Complexity Bound

We consider the family of quadratic functions, $f_k(x) = \frac{1}{2}\langle A_k x, x\rangle - \langle b, x\rangle$, with

$$A_k \;=\; \begin{pmatrix} \Lambda_k & 0 \\ 0 & I_{n-k} \end{pmatrix} \;\in\; \mathbb{R}^n$$

where $\Lambda_k \in \mathbb{R}^{k\times k}$ is the following tridiagonal matrix:

$$\Lambda_k \;=\; \begin{pmatrix} 1 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ \dots & & & & & \\ 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & -1 & 2 \end{pmatrix}.$$

This function satisfies the following properties:

- $0 \preceq \nabla^2 f_k(\cdot) \equiv A_k \preceq 4I$ — it is convex and has a Lipschitz gradient with constant 4
- $f_k^\star = -\frac{k}{2}$
- $R_k^2 = \|x_k^\star\|^2 \le \frac{(k+1)^3}{3}$
- We fix $b := e_1$ (the first basis vector)

We run a method for a fixed number of k iterations, on a function $\boxed{f(x) := f_{2k+1}(x)}$, starting from $0 \in \mathbb{R}^n$. Tridiagonal structure of the matrix ensures that $x_k \in \mathbb{R}^{n,k}$ (the space where only first k components are non-zero):

$$\begin{aligned} x_0 &= [0 \;\; 0 \;\; 0 \;\; \cdots \;\; 0] \\ x_1 &= [\star \;\; 0 \;\; 0 \;\; \cdots \;\; 0] \\ x_2 &= [\star \;\; \star \;\; 0 \;\; \cdots \;\; 0] \\ x_3 &= [\star \;\; \star \;\; \star \;\; \cdots \;\; 0] \\ &\dots \end{aligned}$$

Moreover, we have

- $f_{2k+1}(x) \equiv f_k(x)$ — these functions are *indistinguishable* for the method for the first k iterations.

Therefore, the following lower bound for the output of the method holds:

$$f(x_k) \;=\; f_{2k+1}(x_k) \;=\; f_k(x_k) \;\geq\; f_k^\star \;=\; -\tfrac{k}{2}.$$

At the same time,

$$f^\star \;=\; f_{2k+1}^\star \;=\; -\tfrac{2k+1}{2}$$

and

$$R^2 \;=\; R_{2k+1}^2 \;\leq\; \tfrac{2^3(k+1)^3}{3}$$

Therefore,

$$\tfrac{f(x_k)-f^\star}{R^2} \;\geq\; \tfrac{1}{R^2}\Big(\tfrac{2k+1}{2} - \tfrac{k}{2}\Big) \;=\; \tfrac{1}{2R^2}(k+1) \;\geq\; \tfrac{3}{2^4(k+1)^2},$$

and this is the lower bound!

Now, for an arbitrary Lipschitz constant $L > 0$, we can take a multiplied function:

$$\varphi(x) \;:=\; \tfrac{L}{4} f(x),$$

for which we will have:

$$\tfrac{\varphi(x_k)-\varphi^\star}{R^2} \;\geq\; \tfrac{3L}{2^6(k+1)^2}.$$

This completes the proof of the following theorem.

Theorem 8.1.1. *Let $L > 0$ and $K \geq 1$ be fixed. Then, for any first-order optimization algorithm, such that*

$$x_{k+1} \;\in\; span\Big\{\nabla f(x_0), \ldots, \nabla f(x_k)\Big\},$$

there is a convex function $f : \mathbb{R}^n \to \mathbb{R}$ with $n \geq 2K+1$ with L-Lipschitz gradient, such that

$$f(x_K) - f^\star \geq \; \tfrac{3L\|x_0 - x^\star\|^2}{2^6(K+1)^2}. \tag{8.1}$$

- $\Rightarrow$ we cannot get a rate faster than $O(\frac{1}{k^2})$.
- The gradient method: only $O(1/k)$.
- Can we achieve the optimal rate (8.1) by any method?

8.2 Review: Strongly Convex Functions

We say that a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is *strongly convex* with a constant $\mu > 0$, if

$$f(y) \;\geq\; f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{\mu}{2}\|y - x\|^2, \qquad \forall x, y \in \mathbb{R}^n. \tag{8.2}$$

if we have two functions f_1, and f_2 that satisfy (8.2) with constants $\mu_1, \mu_2 \geq 0$, then their sum $f(\cdot) = f_1(\cdot) + f_2(\cdot)$ satisfy (8.2) with constant $\mu = \mu_1 + \mu_2$. Therefore, the sum of a convex function with a strongly convex one will always give us a strongly convex function.

The most important example of strongly convex function is the squared Euclidean norm:

$$d(x) \;=\; \tfrac{1}{2}\|x\|^2 \;=\; \tfrac{1}{2}\langle x, x\rangle.$$

Let us check (8.2) directly. We have

$$\begin{aligned} d(y) - d(x) - \langle \nabla d(x), y - x \rangle &= \tfrac{1}{2}\|y\|^2 - \tfrac{1}{2}\|x\|^2 - \langle x, y - x \rangle \\ &= \tfrac{1}{2}\|y - x + x\|^2 - \tfrac{1}{2}\|x\|^2 - \langle x, y - x \rangle \\ &= \tfrac{1}{2}\|y - x\|^2 + \tfrac{1}{2}\|x\|^2 + \langle x, y - x \rangle - \tfrac{1}{2}\|x\|^2 - \langle x, y - x \rangle \\ &= \tfrac{1}{2}\|y - x\|^2. \end{aligned}$$

Therefore, for the squared Euclidean norm, inequality (8.2) is satisfied as equation, with $\mu = 1$.

As a direct consequence of these observations, let us consider a regularized objective

$$g(y) \;:=\; f(y) + \tfrac{1}{2}\|y - x_0\|^2,$$

where f is an arbitrary differentiable convex function. Hence, g is strongly convex with constant $\mu = 1$, and by (8.2), for the optimum $x_g^\star := \arg\min_{y \in \mathbb{R}^n} g(y)$ we have

$$g(y) \;\geq\; g(x_g^\star) + \langle \nabla g(x_\star), y - x_\star \rangle + \tfrac{\mu}{2}\|y - x^\star\|^2 \;=\; g^\star + \tfrac{\mu}{2}\|y - x_g^\star\|^2. \tag{8.3}$$

Therefore, by strong convexity, we obtain a strengthening (8.3) of a trivial inequality: $g(y) \geq g^\star$ that is the definition of $g^\star$.

8.3 Nesterov's Fast Gradient Method

8.3.1 Analysis

Before, we analyzed methods built on some "physical" or "geometrical" intuition. Such methods are easy to describe by analogy to some known phenomena. However it is difficult to analyze them, after the method is already rigidly fixed. Now, we try a different approach. We will immediately start to look into the essence of what we want to achieve, and that would lead us to the development of a method.

We are interested in solving an unconstrained optimization problem,

$$\min_{x \in \mathbb{R}^n} f(x), \tag{8.4}$$

where f is convex and it has a Lipschitz gradient. These are the main two inequalities that characterize our problem class and that we will employ:

$$0 \;\leq\; f(y) - f(x) - \langle \nabla f(x), y - x \rangle \;\leq\; \tfrac{L}{2}\|y - x\|^2, \qquad x, y \in \mathbb{R}^n.$$

We fix a sequence $A_k > 0$ of growing coefficients that will give us the "rate" of the method. The idea is to prove the following inequality:

$$f(x_k) - f^\star \;\leq\; \tfrac{\|x_0 - x^\star\|^2}{2A_k}, \tag{8.5}$$

where A_k specifies exactly the *rate of convergence*. To achieve rate (8.1), we hope to have $A_k \approx \frac{k^2}{L}$.

Instead of (8.5), let us try to ensure a bit more general goal, the same inequality but for an arbitrary $x \in \mathbb{R}^n$:

$$f(x_k) - f(x) \;\leq\; \tfrac{\|x_0 - x\|^2}{2A_k},$$

which is equivalent to

$$\varphi_k(x) \;:=\; \tfrac{1}{2}\|x_0 - x\|^2 + A_k f(x) \;\geq\; A_k f(x_k), \qquad x \in \mathbb{R}^n. \tag{8.6}$$

Notice that in the left hand side of (8.6) we have a value of a strongly convex function at arbitrary point x (as a sum of a convex function $A_k f(\cdot)$ and strongly convex $\frac{1}{2}\|x_0 - \cdot\|^2$), and in the right hand side we have a constant term. Thus, from (8.6) we want to achieve that

$$\varphi_k^\star \;=\; \min_{x \in \mathbb{R}^n} \varphi_k(x) \;\geq\; A_k f(x_k). \tag{8.7}$$

However, by strong convexity, we know an improved inequality:

$$\varphi_k(x) \;\geq\; \varphi_k^\star + \tfrac{1}{2}\|x^\star_{\varphi_k} - x\|^2 \overset{(8.7)}{\geq} A_k f(x_k) + \tfrac{1}{2}\|x^\star_{\varphi_k} - x\|^2.$$

Therefore, we refine our goal. Now we want to construct a sequence of growing coefficients A_k (growing as fast as possible), and two sequences of points $\{x_k\}_{k\geq 0}$ and $\{v_k\}_{k\geq 0}$ such that, for any $k \geq 0$.

$$\tfrac{1}{2}\|x_0 - x\|^2 + A_k f(x) \;\geq\; \tfrac{1}{2}\|v_k - x\|^2 + A_k f(x_k), \qquad x \in \mathbb{R}^n. \tag{8.8}$$

Clearly, if (8.8) is satisfied, than we achieve our initial goal (8.5), by plugging $x := x^\star$.

First, let us check how to start. We can assume that $A_0 = 0$, and $x_0 = v_0$. Then (8.8) is trivially satisfied.

Now, assume that (8.8) is satisfied for some $k \geq 0$. We want to "increase the rate", by setting $A_{k+1} = A_k + a_{k+1}$, where $a_{k+1} > 0$ is some positive coefficient. Thus, we have

$$\begin{aligned}\tfrac{1}{2}\|x_0 - x\|^2 + A_{k+1} f(x) \;&=\; \tfrac{1}{2}\|x_0 - x\|^2 + a_{k+1} f(x) + A_k f(x) \\ &\overset{(8.8)}{\geq} \tfrac{1}{2}\|v_k - x\|^2 + a_{k+1} f(x) + A_k f(x_k).\end{aligned}$$

Now, when we have a sum of two function values, it is natural to use *convexity*. Denote $\gamma_k = \frac{a_{k+1}}{A_{k+1}} = \frac{a_{k+1}}{A_k + a_{k+1}}$. We have:

$$a_k f(x) + A_k f(x_k) \;=\; A_{k+1}\Big(\gamma_k f(x) + (1-\gamma_k) f(x_k)\Big) \;\geq\; A_{k+1} f(y),$$

where $y := \gamma_k x + (1-\gamma_k) x_k$.

Now, we can continue a lower bound, by using the global linear model, by convexity again:

$$f(y) \;\geq\; f(y_k) + \langle \nabla f(y_k), y - y_k \rangle,$$

where y_k is some point that we have to choose. We have a flexibility in the choice of y_k, but one natural choice is

$$y_k \;=\; \gamma_k v_k + (1-\gamma_k) x_k,$$

as we would then have $y - y_k = \gamma_k (x - v_k)$.

We obtained,

$$\begin{aligned}\tfrac{1}{2}\|x_0 - x\|^2 + A_{k+1} f(x) \;&\geq\; \tfrac{1}{2}\|v_k - x\|^2 + A_{k+1}\Big[f(y_k) + \langle \nabla f(y_k), y - y_k\rangle\Big] \\ &=\; \tfrac{1}{2}\|v_k - x\|^2 + A_{k+1}\Big[f(y_k) + \gamma_k \langle \nabla f(y_k), x - v_k\rangle\Big] \\ &\equiv\; m_k(x).\end{aligned}$$

We minimize the right hand side in x to obtain the next auxiliary point, $v_{k+1} = \arg\min_x m_k(x)$. This leads us to the following update:

$$v_{k+1} \;=\; v_k - a_{k+1}\nabla f(y_k).$$

Hence, by strong convexity of $m_k(\cdot)$, we get

$$\tfrac{1}{2}\|x_0 - x\|^2 + A_{k+1}f(x) \;\geq\; \tfrac{1}{2}\|v_{k+1} - x\|^2 + m_k^\star,$$

where

$$m_k^\star \;=\; m_k(v_{k+1}) \;=\; \tfrac{1}{2}\|v_k - v_{k+1}\|^2 + A_{k+1}\Big[f(y_k) + \gamma_k\langle\nabla f(y_k), v_{k+1} - v_k\rangle\Big].$$

To finish the proof, we need to make it possible that $m_k^\star \geq A_{k+1}f(x_{k+1})$. How we can do that?

We choose $x_{k+1} := \gamma_k v_{k+1} + (1-\gamma_k)x_k$, thus $x_{k+1} - y_k$ is parallel to $v_{k+1} - v_k$. Therefore

$$\begin{aligned} m_k^\star \;&=\; \tfrac{1}{2\gamma_k^2}\|x_{k+1} - y_k\|^2 + A_{k+1}\Big[f(y_k) + \langle\nabla f(y_k), x_{k+1} - y_k\rangle\Big] \\ &=\; A_{k+1}\Big[\tfrac{1}{2\gamma_k^2 A_{k+1}}\|x_{k+1} - y_k\|^2 + \langle\nabla f(y_k), x_{k+1} - y_k\rangle + f(y_k)\Big] \\ &\geq\; A_{k+1}f(x_{k+1}), \end{aligned}$$

as soon as $\frac{1}{\gamma_k^2 A_{k+1}} = \frac{A_{k+1}^2}{a_{k+1}^2 A_{k+1}} = \frac{a_{k+1}+A_k}{a_{k+1}^2} \geq L$. Thus, we have established (8.8) for all $k \geq 0$.

8.3.2 Algorithm

We come to the following algorithmic scheme of the optimization method.

Algorithm 8.1: *Fast Gradient Method.*

Initialization: $x_0 \in \mathbb{R}^n$. Set $v_0 = x_0$ and $A_0 = 0$. Fix $K \geq 1$.

For $k = 0 \ldots K-1$ **iterate:**

1. Choose a new coefficient $a_{k+1} > 0$. Set $A_{k+1} := A_k + a_{k+1}$ and $\gamma_k := \frac{a_{k+1}}{A_{k+1}}$
2. Compute the gradient $\nabla f(y_k)$ at the intermediate point $y_k := \gamma_k v_k + (1-\gamma_k)x_k$
3. Update $v_{k+1} = v_k - a_{k+1}\nabla f(y_k)$
4. Set a new point from the triangle rule: $x_{k+1} := \gamma_k v_{k+1} + (1-\gamma_k)x_k$

Return x_K

In this method, for simplicity we fix the number of iterations K and use it as a stopping condition for the algorithm. A more advanced stopping condition would include to compute an *accuracy certificate* for a solution, that would guarantee a small function residual for the output. We study how to compute accuracy certificates later in the course.

Note that in this algorithm, the only unspecified parameter is a sequence $\{a_k\}_{k\geq 1}$ of positive coefficients, that we have to choose. Then, the grows of

$$A_k \;=\; \sum_{i=1}^k a_i$$

defines the rate of convergence.

With our analyses, we have established the following result.

Theorem 8.3.1. *Let $a_{k+1} > 0$ be chosen such that $\frac{a_{k+1}+A_k}{a_{k+1}^2} \geq L$. Then, we have*

$$\frac{1}{2}\|v_k - x\|^2 + A_k(f(x_k) - f(x)) \quad \leq \quad \frac{1}{2}\|x_0 - x\|^2, \qquad x \in \mathbb{R}^n. \tag{8.9}$$

Substituting $x := x^\star$, we obtain

$$f(x_k) - f^\star \quad \leq \quad \frac{\|x_0 - x^\star\|^2}{2A_k}, \qquad k \geq 1. \tag{8.10}$$

8.3.3 The Parameter Choice

We need to establish the following inequality:

$$\frac{A_{k+1}}{a_{k+1}^2} \quad \geq \quad L. \tag{8.11}$$

Note that as larger a_{k+1} than the faster the rate of convergence of the method.

Solving Quadratic Equation. Let us choose $a_{k+1} > 0$ such that inequality (8.11) is satisfied as equation. That is:

$$A_{k+1} \quad = \quad a_{k+1} + A_k \quad = \quad La_{k+1}^2. \tag{8.12}$$

This is quadratic equation in a_{k+1}, which has an explicit formula for a (positive) solution:

$$\boxed{a_{k+1} \quad := \quad \frac{1}{2L} \cdot \left(1 + \sqrt{1 + 4A_k L}\right).} \tag{8.13}$$

Note that in the basic gradient descent we choose $a_{k+1} \approx \frac{1}{L}$. Therefore, formula (8.13) is more aggressive, allowing for larger exploratory steps of the accelerated method.

We need to figure out the rate of convergence. We have

$$\begin{aligned} \sqrt{A_{k+1}} - \sqrt{A_k} \quad &= \quad \frac{A_{k+1} - A_k}{\sqrt{A_{k+1}} + \sqrt{A_k}} \quad = \quad \frac{a_{k+1}}{\sqrt{A_{k+1}} + \sqrt{A_k}} \\ &\overset{(8.12)}{=} \quad \frac{\sqrt{A_{k+1}}}{\sqrt{L}(\sqrt{A_{k+1}} + \sqrt{A_k})} \quad \geq \quad \frac{\sqrt{A_{k+1}}}{2\sqrt{LA_{k+1}}} \quad = \quad \frac{1}{2\sqrt{L}}. \end{aligned}$$

Thus, telescoping this inequality, we obtain

$$\sqrt{A_k} \quad \geq \quad \sqrt{A_0} + \frac{k}{2\sqrt{L}} \quad = \quad \frac{k}{2\sqrt{L}}.$$

Hence, $A_k \geq \frac{k^2}{4L}$ and we obtain the following optimal rate for the fast gradient method.

Corollary 8.3.2. *Let a_{k+1} be chosen according to* (8.13) *in Algorithm 8.1. Then, the rate of convergence is:*

$$f(x_k) - f^\star \quad \leq \quad \frac{2L\|x_0 - x^\star\|^2}{k^2}, \qquad k \geq 1.$$

Remark 8.3.3. This rate matches the lower bound (8.1) up to a numerical constant. Therefore, the fast gradient method is optimal for our problem class.

Remark 8.3.4. The complexity of the fast gradient method to obtain $f(x_K) - f^\star \leq \varepsilon$ is

$$K \quad = \quad \left\lfloor \sqrt{\frac{2L\|x_0 - x^\star\|^2}{\varepsilon}} \right\rfloor + 1$$

first-order oracle calls. This is much better than $O(\frac{1}{\varepsilon})$ of the gradient method.

Predefined Growth. There are other possibilities in choosing sequence a_{k+1} in order to satisfy (8.11). While solving the quadratic equation (8.12) provides us with the best exact formula, in some of more sophisticated situations (such as stochastic accelerated methods, or second-order acceleration), it is more convenient to specify the growth of coefficients explicitely.

For example, let us specify

$$\boxed{a_k \;:=\; \tfrac{1}{2L}k.}$$

Then, we have

$$A_k \;=\; \tfrac{1}{2L}\sum_{i=1}^{k} i \;=\; \tfrac{k(k+1)}{4L},$$

which is the required rate of convergence. It is easy to check that inequality (8.11) is also satisfied for this choice.

Adaptive Search. So far, we discussed accelerated schemes with a fixed Lipscthiz constant $L > 0$. However, by analogy with the gradient method, it is possible to employ a simple adaptive search that estimates parameter L adaptively over iterations. The key inequality comes from the last part of our proof, where we required the following condition to hold:

$$\tfrac{1}{2\gamma_k^2 A_{k+1}}\|x_{k+1} - y_k\|^2 + \langle \nabla f(y_k), x_{k+1} - y_k\rangle + f(y_k) \;\geq\; f(x_{k+1}).$$

Exercise 8.3.1. Develop a version of the fast gradient method with the adaptive search of L. What is the total complexity of the resulting algorithm in terms of the gradient and function value computations?

Literature

The fast gradient method was developed by Yurii E. Nesterov in 1983. The modern versions of this algorithm can be found in [Nes18].

[Nes18] Yurii Nesterov. *Lectures on convex optimization.* Springer, 2018.