Lecture 6

6.1 Rates for Gradient Method

6.1.1 On the Choice of the Norm

In the previous lecture, we discussed the class of convex smooth functions $f : \mathbb{R}^n \to \mathbb{R}$, which can be characterized by the following second-order condition:

$$0 \quad \leq \quad \langle \nabla^2 f(x) h, h \rangle \quad \leq \quad L\|h\|^2, \qquad \forall x, h \in \mathbb{R}^n, \tag{6.1}$$

where $\| \cdot \|$ is an arbitrary norm on $\mathbb{R}^n$.

A non-standard choice of norm might be important for the design of optimization algorithms. For example, an important algorithm for training deep learning architectures, called *Adam* [KB15], can be viewed as a stabilized version of the stochastic gradient method under the $\|\cdot\|_\infty$-norm [BN24].

At the same time, the choice of the $\| \cdot \|_1$-norm for the primal space in the gradient method leads to *greedy coordinate descent*.

Therefore, a proper choice of norm can enable desirable features for these algorithms. Other examples include problems with explicitly given geometry (such as the space of probability distributions), when the choice of the corresponding norm is very natural.

However, the Euclidean norm remains the most important choice for the design and analysis of the optimization methods. Unless explicitly specified,

from now on in this course, we will focus on the Euclidean norm $\boxed{\| \cdot \| \equiv \| \cdot \|_2}$.

The condition (6.1) under the Euclidean norm reads as:

$$0 \quad \preceq \quad \nabla^2 f(x) \quad \preceq \quad LI, \qquad x \in \mathbb{R}^n. \tag{6.2}$$

And the gradient method for the Euclidean norm has the following simplest form, starting from some $x_0 \in \mathbb{R}^n$, we iterate:

$$x_{k+1} \quad = \quad x_k - \tfrac{1}{L} \nabla f(x_k), \qquad k \geq 0, \tag{6.3}$$

where we assume the fixed step size $1/L$ for now. Assuming that a minimum $x^\star$ exists, let us look at the distance to the solution. We have

$$\begin{array}{rcl} \tfrac{1}{2}\|x_{k+1} - x^\star\|^2 & = & \tfrac{1}{2}\|x_k - x^\star - \tfrac{1}{L}\nabla f(x_k)\|^2 \\ & = & \tfrac{1}{2}\|x_k - x^\star\|^2 + \tfrac{1}{L}\Bigl(\tfrac{1}{2L}\|\nabla f(x_k)\|^2 + 2\langle x^\star - x_k, \nabla f(x_k)\rangle \Bigr) \\ & \overset{(6.2)}{\leq} & \tfrac{1}{2}\|x_k - x^\star\|^2 + \tfrac{1}{L}\Bigl(f(x^\star) - f(x_k) \Bigr) \quad \leq \quad \tfrac{1}{2}\|x_k - x^\star\|^2. \end{array} \tag{6.4}$$

Hence, in the Euclidean case, we have proved that the iterates of the gradient method remains bounded,

$$\|x_k - x^\star\| \quad \leq \quad \|x_0 - x^\star\|, \qquad \forall k \geq 0. \tag{6.5}$$

Note that $x^\star$ here can be an *arbitrary minimizer.*

6.1.2 Rate for Minimizing Convex Functions

Let us repeat the reasoning from the previous lecture, on the convergence of gradient method on smooth convex functions. We have two main ingredients.

- Progress of each iteration: $f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}\|\nabla f(x_k)\|^2$.
- Convexity: $f(x_k) - f^\star \leq \langle \nabla f(x_k), x_k - x^\star \rangle \leq \|\nabla f(x_k)\| \cdot \|x_k - x^\star\| \overset{(6.5)}{\leq} \|\nabla f(x_k)\| \cdot R$.

where $R := \|x_0 - x^\star\|$ is the explicit distance from the initial point to any fixed solution.

Denoting the functional residual by $F_k := f(x_k) - f^\star > 0$, we obtain

$$F_k - F_{k+1} \quad = \quad f(x_k) - f(x_{k+1}) \quad \geq \quad \tfrac{1}{2L}\|\nabla f(x_k)\|^2 \quad \geq \quad \tfrac{1}{2LR^2}F_k^2. \tag{6.6}$$

Inequality (6.6) that the sequence $\{F_k\}_{k\geq 0}$ is monotonically decreasing with a certain rate. Notice also that, since $\nabla f(x^\star) = 0$, we have:

$$F_0 \quad = \quad f(x_0) - f^\star \quad \overset{(6.2)}{\leq} \quad \tfrac{LR^2}{2}. \tag{6.7}$$

Now, we observe that

$$\tfrac{1}{F_{k+1}} - \tfrac{1}{F_k} \quad = \quad \tfrac{F_k - F_{k+1}}{F_k F_{k+1}} \quad \overset{(6.6)}{\geq} \quad \tfrac{1}{2LR^2}\tfrac{F_k}{F_{k+1}} \quad \geq \quad \tfrac{1}{2LR^2}.$$

Hence, since we have a constant in the right hand side, this inequality is easy to telescope (or, in the terminology of continuous time, intergrate). We get:

$$\tfrac{1}{F_k} \quad \geq \quad \tfrac{1}{F_0} + \tfrac{k}{2LR^2} \quad \overset{(6.7)}{\geq} \quad \tfrac{4+k}{2LR^2}.$$

We have proved the following result.

Theorem 6.1.1. *Assume that a minimum $x^\star$ exist. On convex smooth functions* (6.2)*, the gradient method* (6.3) *has the following rate of convergence:*

$$f(x_k) - f^\star \quad \leq \quad \tfrac{2LR^2}{k+4}, \qquad k \geq 0. \tag{6.8}$$

Corollary 6.1.2. *In order to find a point x_k such that $f(x_k) - f^\star \leq \varepsilon$, it is enough to perform*

$$k \quad = \quad O\Big(\tfrac{LR^2}{\varepsilon}\Big) \tag{6.9}$$

first-order oracle calls.

6.1.3 Minimizing Gradient Norm

How good this result as compared to what we have seen before for the gradient method?

First of all, notice that the convergence rate (6.8) is in terms of the *last point*, which is the most natural candidate for the output of the algorithm.

In Lecture 3, we have already proved that to reach $\|\nabla f(\bar{x})\| \leq \varepsilon$, it is enough to perform

$$K \;=\; O\Bigl(\tfrac{L(f(x_0) - f^\star)}{\varepsilon^2}\Bigr) \tag{6.10}$$

iterations of the gradient method. The complexity $O(1/\varepsilon^2)$ seems much worse than that one in (6.9). At the same time, technically, complexity bounds (6.9) and (6.10) are not directly *comparable* as they refer to different accuracy measures, and the role of the letter "ε" is different in them!

It appears that using convexity, we can improve the dependence on ε in (6.10) for the gradient norm minimization. Let us consider $2K$ iterations of the gradient method, where $K \geq 1$ is fixed. From the analysis of the gradient norm minimization (Theorem 3.2.4 in Lecture 3), we have:

$$\min_{K \leq i \leq 2K-1} \|\nabla f(x_i)\|^2 \;\leq\; \tfrac{2L(f(x_K) - f^\star)}{K} \;\overset{(6.8)}{\leq}\; \tfrac{4L^2R^2}{K(K+4)} \;=\; O\Bigl(\bigl[\tfrac{LR}{K}\bigr]^2\Bigr). \tag{6.11}$$

Therefore, we have obtained the following result.

Proposition 6.1.3. *In order to find a point $\bar{x}$ such that $\|\nabla f(\bar{x})\| \leq \varepsilon$, the gradient method* (6.3) *on smooth convex functions* (6.2) *needs*

$$K \;=\; O\Bigl(\tfrac{LR}{\varepsilon}\Bigr)$$

iterations (first-order oracle calls).

The convergence in terms of the gradient norm is stronger, as it leads to the convergence in terms of the function value. The gradient norm is also easier to use in practice as the stopping criterion for the algorithms.

Finally, we might ask what is the convergence rate in terms of the *distance to the solution* $\|x_k - x^\star\| \to 0$? Unfortunately, in general for convex functions, we cannot guarantee such convergence, even if $\|\nabla f(x_k)\| \to 0$ and $f(x_k) \to f^\star$. However, there is a specific class of functions that enables us to guarantee convergence in terms of the distance between points.

6.1.4 Strongly Convex Smooth Functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. We say that f is *strongly convex* and *smooth*, if the eigenvalues of the Hessian are both uniformly bounded from above and separated from zero:

$$\mu I \;\preceq\; \nabla^2 f(x) \;\preceq\; LI, \qquad x \in \mathbb{R}^n, \tag{6.12}$$

where $0 < \mu \leq L$ are parameters of our problem class. Of course, if $\mu = 0$ in (6.12) than we obtain the class of convex smooth functions, that we discussed before.

The problems that satisfy (6.12) are among the most important in optimization and possesses the fastest rates of convergence. As before, we can formulate the conditions (6.12) in terms of the gradients and in terms of the function values. Indeed, using the fundamental theorems of calculus,

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \;=\; \int\limits_0^1 \langle \nabla^2 f(x + \tau(y-x))(y-x), y - x \rangle d\tau$$

and

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \quad = \quad \int\limits_0^1 (1-\tau)\langle \nabla^2 f(x + \tau(y-x))(y-x), y-x \rangle d\tau,$$

we get the following equivalent characterization of our new problem class.

Theorem 6.1.4. *The following conditions, that hold for any $x, y \in \mathbb{R}^n$, are equivalent to* (6.12)*:*

$$\mu \|y - x\|^2 \quad \leq \quad \langle \nabla f(y) - \nabla f(x), y - x \rangle \quad \leq \quad L\|y - x\|^2 \tag{6.13}$$

and

$$\tfrac{\mu}{2}\|y - x\|^2 \quad \leq \quad f(y) - f(x) - \langle \nabla f(x), y - x \rangle \quad \leq \quad \tfrac{L}{2}\|y - x\|^2. \tag{6.14}$$

Now we obtain the perfect symmetry in our inequalities. Geometrically, inequality (6.14) means that at each point $x \in \mathbb{R}^n$, we have both global upper and lower quadratic models of our function.

Note that if we ignore the upper inequality in (6.12), (6.13), and (6.14), we obtain the class of just *strongly convex functions* (not necessary smooth).

Exercise 6.1.1. Show that every strongly convex function f (with respect to the Euclidean norm), can be represented as $f(x) \equiv \varphi(x) + \frac{\mu}{2}\|x\|^2$, where $\varphi(\cdot)$ is a convex function.

Let us look at some consequences of (6.14). Plugging $x := x^\star$, we get

Proposition 6.1.5. *For a strongly convex smooth function $f : \mathbb{R}^n \to \mathbb{R}$, it holds:*

$$\tfrac{\mu}{2}\|y - x^\star\|^2 \quad \leq \quad f(y) - f^\star \quad \leq \quad \tfrac{L}{2}\|y - x^\star\|^2, \qquad y \in \mathbb{R}^n. \tag{6.15}$$

Therefore, the functional residual and the distance to the solution becomes comparable. If we have a convergence in terms of the functional residual, $f(x_k) - f^\star$, bound (6.15) also leads to a convergence in terms of the distance: $\|x_k - x^\star\| \to 0$ with the same rate.

Note the the lower inequality in (6.15) also proves that the solution $x^\star$ to a strongly convex optimization problem always *exist* and *unique.* This is not the case for convex functions, when $\mu = 0$.

Now, let us rearrange the terms in the lower inequality in (6.14). We obtain:

$$f(y) \quad \geq \quad f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{\mu}{2}\|y - x\|^2,$$

that holds for any $x, y \in \mathbb{R}^n$. Minimizing the left-hand and the right-hand sides independently, we obtain

$$f^\star \quad \geq \quad f(x) - \tfrac{1}{2\mu}\|\nabla f(x)\|^2.$$

Or, rearranging the terms, we obtain the following very important inequality.

Proposition 6.1.6. *For strongly convex functions, it holds*

$$\tfrac{1}{2\mu}\|\nabla f(x)\|^2 \quad \geq \quad f(x) - f^\star. \tag{6.16}$$

Let us apply the new inequality for the analysis of the gradient method. For one gradient step, we have

$$f(x_k) - f(x_{k+1}) \quad \geq \quad \tfrac{1}{2L}\|\nabla f(x_k)\|^2 \quad \overset{(6.16)}{\geq} \quad \tfrac{\mu}{L}(f(x_k) - f^\star), \tag{6.17}$$

or, for the functional residual $F_k := f(x_k) - f^\star$, we have

$$F_{k+1} \quad \overset{(6.17)}{\leq} \quad \Bigl(1 - \tfrac{\mu}{L}\Bigr)F_k \quad \leq \quad \exp\Bigl(-\tfrac{\mu}{L}\Bigr)F_k. \tag{6.18}$$

The ratio $\frac{L}{\mu} \geq 1$ is very important and called the *condition number* of the function. Applying inequality (6.18) for k steps of the method, we prove the following result.

Theorem 6.1.7. *For the iterations of the gradient method on strongly convex smooth functions, we have the linear rate of convergence:*

$$f(x_k) - f^\star \quad \leq \quad \exp(-k\tfrac{\mu}{L})(f(x_0) - f^\star). \tag{6.19}$$

Therefore, to reach $f(x_K) - f^\star \leq \varepsilon$ it is enough to perform

$$K \quad = \quad \tfrac{L}{\mu} \ln \tfrac{f(x_0) - f^\star}{\varepsilon} \quad \overset{(6.14)}{\leq} \quad \tfrac{L}{\mu} \ln \tfrac{LR^2}{2\varepsilon} \tag{6.20}$$

iterations of the algorithm.

6.2 Polyak's Heavy Ball Method

6.2.1 Quadratic Functions

Let us consider the problem of unconstrained minimization of the quadratic function,

$$\min_{x \in \mathbb{R}^n} \Big\{ \ f(x) \quad = \quad \tfrac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle \ \Big\}, \tag{6.21}$$

where $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$ is given data. We assume that $A = A^\top \succeq 0$ (why?). Actually, without loss of generality, we can always assume that $A \succ 0$ (strictly). However, the smallest eigenvalue can be tiny. In this case the problem is called *ill-conditioned*. Unfortunately, ill-conditioned problems are the most frequent in practice.

Computing the gradient, we get

$$\nabla f(x) \quad = \quad Ax - b, \tag{6.22}$$

and at the solution $x^\star$ should solve the linear system:

$$\nabla f(x^\star) \quad = \quad Ax^\star - b \quad = \quad 0 \qquad \Leftrightarrow \qquad b \quad = \quad Ax^\star. \tag{6.23}$$

Therefore, any optimization method for minimizing (6.21) automatically provides us with an algorithm for solving linear systems with symmetric matrices (6.23). In fact, this approach remains the most efficient way to solve large-scale systems, when the dimension n is huge. According to 6.22, to compute the gradient vector at a given point, it requires to perform one *matrix-vector* product. If the matrix A is *sparse*, it can be done efficiently even for a very large dimension n.

From (6.23), we have another representation of the gradient, using the optimal solution:

$$\nabla f(x) \quad = \quad A(x - x^\star). \tag{6.24}$$

Computing the Hessian, we observe that it is constant,

$$\nabla^2 f(x) \quad \equiv \quad A.$$

Therefore, quadratic function (6.21) is strongly convex and smooth with

$$0 \quad < \quad \mu \quad = \quad \lambda_{\min}(A) \quad \leq \quad \lambda_{\max}(A) \quad = \quad L.$$

It is easy to check that the Taylor expansions hold exactly for quadratic functions.

Proposition 6.2.1. *For quadratic functions, it holds:*

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x \rangle &\equiv \tfrac{1}{2}\langle \nabla f(y) - \nabla f(x), y - x \rangle \\ &\equiv \tfrac{1}{2}\langle \nabla^2 f(x)(y-x), y - x \rangle \\ &\equiv \tfrac{1}{2}\langle A(y-x), y-x \rangle \;\equiv\; \tfrac{1}{2}\|y-x\|_A^2, \end{aligned} \tag{6.25}$$

where $\|y - x\|_A$ stands for the generalized Euclidean norm with matrix $A = A^\top \succ 0$.

Exercise 6.2.1. Check (6.25).

From (6.25), we obtain the following interesting formula,

$$f(y) - f(x) \;=\; \tfrac{1}{2}\langle \nabla f(y) + \nabla f(x), y - x \rangle, \qquad x, y \in \mathbb{R}^n. \tag{6.26}$$

6.2.2 Heavy Ball Method

We discuss a faster method for minimizing a quadratic function, developed by B.T. Polyak [Pol87]. Instead of performing a simple gradient step, we add some *inertia* to the method. Each iteration reads as follows:

$$x_{k+1} \;=\; x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}), \qquad k \geq 1. \tag{6.27}$$

where $\alpha > 0$ is a step-size and $0 \leq \beta \leq 1$ is an extra parameter. This algorithm is called the heavy ball method. The dynamical system (6.27) corresponds to a discretized version of the motion of a body ("the heavy ball") in a potential field $\nabla f(\cdot)$ under the force of *friction*, where β is a parameter of this force ($\beta = 0$ correspond to no friction).

Another common interpretation of this algorithm is the gradient method with *momentum*. Indeed, iterations (6.27) can be rewritten as follows, starting from some initialization $x_0 \in \mathbb{R}^n$ and $s_0 = 0$, we update, for $k \geq 0$:

$$\begin{aligned} s_{k+1} &= \beta s_k + \nabla f(x_k), \\ x_{k+1} &= x_k - \tfrac{1}{L} s_{k+1}, \end{aligned} \tag{6.28}$$

and $0 \leq \beta \leq 1$ has an interpretation of *momentum parameter* (how fast we forget the history), and we use the constant step-size $1/L$ in front of s_{k+1}.

Exercise 6.2.2. Check that iterations (6.28) and (6.27) are equivalent.

This is a very popular technique in machine learning, where it helps the method to behave more stable, especially when the gradients are stochastic and noisy, and the objective landscape is non-convex.

Let us analyze algorithm (6.28) for the quadratic case. We consider the simplest choice of momentum parameter, $\boxed{\beta := 1}$. Hence, we have

$$s_{k+1} \;=\; \sum_{i=0}^{k} \nabla f(x_i). \tag{6.29}$$

The consequence of the fact that the gradient mapping $\nabla f(\cdot)$ is affine is that

$$\tfrac{1}{k} s_k \;=\; \tfrac{1}{k}\sum_{i=0}^{k-1} \nabla f(x_i) \;=\; \nabla f(\bar{x}_k), \qquad \text{where} \qquad \bar{x}_k \;:=\; \tfrac{1}{k}\sum_{i=0}^{k-1} x_i. \tag{6.30}$$

Let us substitute direction from (6.28) into the global quadratic upper bound:

$$\begin{array}{rcl} f(x_{k+1}) & \leq & f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ & \overset{(6.28)}{=} & f(x_k) - \frac{1}{L}\langle \nabla f(x_k), s_{k+1}\rangle + \frac{1}{2L}\|s_{k+1}\|^2 \\ & = & f(x_k) - \frac{1}{L}\langle s_{k+1} - s_k, s_{k+1}\rangle + \frac{1}{2L}\|s_{k+1}\|^2 \\ & = & f(x_k) - \frac{1}{2L}\|s_{k+1}\|^2 + \frac{1}{L}\langle s_k, s_{k+1}\rangle. \end{array}$$

Therefore, we get the following "progress" of each step.

Proposition 6.2.2. *For every $k \geq 0$:*

$$f(x_k) - f(x_{k+1}) \quad \geq \quad \frac{1}{2L}\|s_{k+1}\|^2 - \frac{1}{L}\langle s_k, s_{k+1}\rangle. \tag{6.31}$$

The last term has an interpretation of the "correlation" between partial sums of the gradients.

Let us substitute two consequitive points into equation (6.26), that is valid only for quadratic functions:

$$\begin{array}{rcl} f(x_k) - f(x_{k+1}) & = & \frac{1}{2}\langle \nabla f(x_k) + \nabla f(x_{k+1}), x_k - x_{k+1}\rangle \\ & = & \frac{1}{2L}\langle s_{k+2} - s_k, s_{k+1}\rangle \quad = \quad \frac{1}{2L}\bigl(\langle s_{k+2}, s_{k+1}\rangle - \langle s_k, s_{k+1}\rangle\bigr). \end{array}$$

Rearranging the terms, we express next correlation using the previous one and the function difference,

$$f(x_k) - f(x_{k+1}) + \frac{1}{2L}\langle s_k, s_{k+1}\rangle \quad = \quad \frac{1}{2L}\langle s_{k+1}, s_{k+2}\rangle. \tag{6.32}$$

Telescoping this inequality, and using that $s_0 = 0$, we get the following bound on the correlations.

Proposition 6.2.3. *For every $k \geq 0$:*

$$f(x_0) - f^\star \quad \geq \quad f(x_0) - f(x_k) \quad \overset{(6.32)}{=} \quad \frac{1}{2L}\langle s_k, s_{k+1}\rangle. \tag{6.33}$$

Now, we can substitute bound (6.33) into (6.31). We obtain:

$$f(x_k) - f(x_{k+1}) \quad \geq \quad \frac{1}{2L}\|s_{k+1}\|^2 - 2(f(x_0) - f^\star).$$

Telescoping this inequality, we get

$$f(x_0) - f^\star \quad \geq \quad f(x_0) - f(x_k) \quad \geq \quad \frac{1}{2L}\sum_{i=1}^{k}\|s_i\|^2 - 2k(f(x_0) - f^\star).$$

Theorem 6.2.4. *For the iterations of the heavy ball method (6.28), with $\beta = 1$, it holds:*

$$2L(1+2k)(f(x_0) - f^\star) \quad \geq \quad \sum_{i=1}^{k}\|s_i\|^2 \quad \overset{(6.30)}{=} \quad \sum_{i=1}^{k} i^2 \Bigl\|\nabla f(\bar{x}_i)\Bigr\|^2. \tag{6.34}$$

Therefore, denoting the smallest gradient among averaged points, $g_k := \min\Bigl\{\|\nabla f(\bar{x}_1)\|, \ldots, \|\nabla f(\bar{x}_k)\|\Bigr\}$, we have

$$g_k^2 \quad \leq \quad \frac{2L(1+2k)(f(x_0)-f^\star)}{\sum_{i=1}^{k} i^2} \quad = \quad \frac{12L(f(x_0)-f^\star)}{k(k+1)} \quad = \quad O\Bigl(\Bigl[\frac{L(f(x_0)-f^\star)}{k^2}\Bigr]\Bigr). \tag{6.35}$$

Exercise 6.2.3. Compare the convergence rate (6.35) with that of the gradient method in (6.11).

Literature

See Section 3.2.1 in [Pol87] for the direct convergence analysis of the heavy ball method on strongly convex quadratic functions and the tuning of the momentum parameter $0 \leq \beta \leq 1$.

Our analysis in Section 6.2.2 is inspired by [MT25], in which the authors establish the convergence of the heavy ball method with restarts for non-convex optimization.

[BN24] Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

[KB15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

[MT25] Naoki Marumo and Akiko Takeda. Universal heavy-ball method for nonconvex optimization under Hölder continuous Hessians. *Mathematical Programming*, 212(1):147–175, 2025.

[Pol87] Boris T Polyak. Introduction to optimization. 1987.