Lecture 3

3.1 Predictable Function Behavior: Smoothness

The key observation that we used to prove the first-order optimality condition is the following one: if at some point x the gradient is non-zero, $\nabla f(x) \neq 0$, then we can move in the direction of anti-gradient to improve the objective function value, for a sufficiently small $\alpha > 0$:

$$f(x - \alpha \nabla f(x)) \;\; = \;\; f(x) - \alpha \|\nabla f(x)\|_2^2 + o(\alpha) \;\; \leq \;\; f(x) - \tfrac{\alpha}{2}\|\nabla f(x)\|_2^2. \tag{3.1}$$

This observation is used in the core of the *gradient descent*, the most popular optimization algorithm. For a new point:

$$x^+ \;\; = \;\; x - \alpha \nabla f(x), \tag{3.2}$$

we can ensure $f(x^+) < f(x)$ when the "step-size" α is sufficiently small. But how small it should be? To implement the method and prove a reasonable rate of convergence, we seek a *quantitative characterization* of α that ensures (3.1). Clearly, it should be related to the behavior of f. In optimization, such a characterization is often called the objective *smoothness*.

3.1.1 Dual Space and Dual Norm

We want to be able to work with arbitrary norms, as the right choice can be crucial in applications.

Assume that we have a fixed norm $\|\cdot\|$ (not necessary Euclidean) in $\mathbb{R}^n$. We define the corresponding *dual norm* $\|\cdot\|_*$ as follows:

$$\|s\|_* \;\; := \;\; \max_{x : \|x\| \leq 1} \langle s, x \rangle \;\; = \;\; \max_{x : \|x\| = 1} \langle s, x \rangle, \qquad s \in \mathbb{R}^n. \tag{3.3}$$

Exercise 3.1.1. Show that all properties of a norm hold for $\|\cdot\|_*$.

Defined this way, the dual norm automatically satisfies the Cauchy-Schwartz inequality:

$$|\langle s, x \rangle| \;\; \leq \;\; \|s\|_* \cdot \|x\|, \qquad x, s \in \mathbb{R}^n. \tag{3.4}$$

Example 3.1.1. Let the primal norm be Euclidean norm: $\|x\| := \|x\|_2 = \langle x, x \rangle^{1/2}$. Then, the dual norm is also Euclidean: $\|s\|_* := \|s\|_2$, which follows from the classical Cauchy-Schwartz inequality.

Example 3.1.2. Let $\|x\| := \langle Bx, x \rangle^{1/2}$, where $B = B^\top \succ 0$ is a fixed positive-definite matrix. Then, the dual norm is given by $\|s\|_* = \langle s, B^{-1}s \rangle^{1/2}$.

Example 3.1.3. Let $\|x\| := \|x\|_p$, for some $p \in [0, \infty]$, where $\|x\|_p := \Bigl(\sum\limits_{i=1}^n |x^{(i)}|^p \Bigr)^{1/p}$ for $p \geq 1$ and $\|x\|_\infty := \max_{i=1}^n |x^{(i)}|$. Then, the dual norm is given by $\|s\|_* = \|s\|_q$ where $q \geq 1$ satisfies $\frac{1}{q} + \frac{1}{p} = 1$. The dual for $\|\cdot\|_\infty$ norm is $\|\cdot\|_1$ and vica versa.

While we use the *primal norm* $\| \cdot \|$ for vectors in our *primal space* $\mathbb{R}^n$, the *dual norm* $\| \cdot \|_*$ is used to measure the size of *linear forms* on $\mathbb{R}^n$, which are the elements of the *dual space*. The main example of a linear form for us is the derivative: $Df(x)[\cdot] \equiv \langle \nabla f(x), \cdot \rangle$.

The definition of the dual norm is very useful as we often have to employ bounds like this:

$$\langle \nabla f(x), h \rangle \quad \overset{(3.4)}{\leq} \quad \|\nabla f(x)\|_* \cdot \|h\|, \qquad x, h \in \mathbb{R}^n.$$

Every matrix $A \in \mathbb{R}^{n \times n}$ can be treated as a bilinear form: $(h, u) \mapsto \langle Ah, u \rangle$ for any $h, u \in \mathbb{R}^n$, and it is convenient to use the following *operator norm*, induced by the primal norm:

$$\|A\| \quad := \quad \max_{h: \|h\| \leq 1} \|Ah\|_* \quad = \quad \max_{\substack{h: \|h\| \leq 1 \\ u: \|u\| \leq 1}} \langle Ah, u \rangle.$$

This definition ensures that we have the following inequality: $\|Ah\|_* \leq \|A\| \cdot \|h\|$.

3.1.2 Functions with Lipschitz Gradient

We fix a primal norm $\|\cdot\|$ in our space (not necessary Euclidean). We say that a function $f : \mathbb{R}^n \to \mathbb{R}$ has Lipschitz continuous gradient with constant $L > 0$, with respect to this norm, if

$$\|\nabla f(y) - \nabla f(x)\|_* \quad \leq \quad L\|y - x\|, \qquad x, y \in \mathbb{R}^n. \tag{3.5}$$

The functions that satisfy (3.5) are often called *smooth functions* in optimization. Note that in the Euclidean case, we have the same Euclidean norm in the left- and right-hand sides of (3.5).

Intuitively, condition (3.5) says that if the points are close: $x \approx y$, then the gradients should also be uniformly close: $\nabla f(x) \approx \nabla f(y)$.

Note that L is a *global constant* as (3.5) should hold on the entire space $\mathbb{R}^n$. In case of constrained optimization, we can restrict (3.5) onto a given feasible set $Q \subset \mathbb{R}^n$.

For now, we consider the unconstrained optimization:

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

and use definition (3.5).

The following second-order characterization of smoothness is very important.

Theorem 3.1.4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. Then, the following statements are* *equivalent:*

- $\nabla f(\cdot)$ *is Lipschitz condinuous with constant $L > 0$.*
- *For any $x \in \mathbb{R}^n$, we have*

$$\|\nabla^2 f(x)\| \quad \leq \quad L. \tag{3.6}$$

Remark 3.1.5. For the Euclidean norm, condition (3.6) is equivalent to:

$$-LI \quad \preceq \quad \nabla^2 f(x) \quad \preceq \quad LI$$

(all eigenvalues of the Hessian are in $[-L; L]$).

Proof. Assume that the gradient is Lipschitz, and choose an arbitrary direction $h \in \mathbb{R}^n$ of unit length, $\|h\| = 1$, and a small $\varepsilon > 0$. Then, by the definition of the derivative, we have:

$$\nabla^2 f(x)h = \frac{1}{\varepsilon}(\nabla f(x+\varepsilon h) - \nabla f(x)) + o(1).$$

Hence, taking the norm and using triangle inequality, we get

$$\begin{aligned}\|\nabla^2 f(x)h\|_* &\leq \frac{1}{\varepsilon}\|\nabla f(x+\varepsilon h) - \nabla f(x)\|_* + o(1) \\ &\stackrel{(3.5)}{\leq} L + o(1).\end{aligned}$$

Taking the limit $\varepsilon \to 0$ we get $\|\nabla^2 f(x)h\|_* \leq L$. Since h is arbitrary we proved (3.6).

Now assume that (3.6) holds. Using the fundamental theorem of calculus, we have:

$$\begin{aligned}\|\nabla f(y) - \nabla f(x)\|_* &= \|\int_0^1 \nabla^2 f(x+\tau(y-x))(y-x)d\tau\|_* \\ &\leq \int_0^1 \|\nabla^2 f(x+\tau(y-x))\|d\tau \cdot \|y-x\| \stackrel{(3.6)}{\leq} L\|y-x\|,\end{aligned}$$

which finishes the proof. □

Example 3.1.6 (Univariate Functions)**.** The derivative of the following univariate functions is Lipschitz continuous:

- $f(x) = a + bx + cx^2$.
- $f(x) = \sin(x)$.
- $f(x) = \ln(1 + e^x)$.

The derivavtive of the following functions *is not* Lipschitz continuous (globally):

- $f(x) = |x|^3$
- $f(x) = e^x$

Example 3.1.7 (Quadratic Function)**.** Let $f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$ for some $A = A^\top \succeq 0$, $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$. Then, $L = \lambda_{\max}(A)$ (with respect to the Euclidean norm).

Theorem 3.1.8 (Global Model of the Function)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ have Lipschitz continuous gradient with constant $L > 0$. Then,*

$$|f(y) - f(x) - \langle\nabla f(x), y-x\rangle| \leq \frac{L}{2}\|y-x\|^2, \qquad x, y \in \mathbb{R}^n. \tag{3.7}$$

Proof. Using the fundamental theorem of calculus, we have

$$\begin{aligned}|f(y) - f(x) - \langle\nabla f(x), y-x\rangle| &= |\int_0^1 \langle\nabla f(x+\tau(y-x)) - \nabla f(x), y-x\rangle d\tau| \\ &\leq \int_0^1 |\langle\nabla f(x+\tau(y-x)) - \nabla f(x), y-x\rangle| d\tau \stackrel{(3.4)}{\leq} \int_0^1 \|\nabla f(x+\tau(y-x)) - \nabla f(x)\|_* d\tau \cdot \|y-x\| \\ &\stackrel{(3.5)}{\leq} \int_0^1 \tau d\tau \cdot L\|y-x\|^2 = \frac{L}{2}\|y-x\|^2,\end{aligned}$$

which is the required bound. □

3.2 Gradient Method for General Norms

3.2.1 Gradient Step

The main idea in the design and analysis of the gradient method is to use bound (3.7) as the *global upper approximation* of the objective. Staying at a point $x \in \mathbb{R}^n$, we fix a regularization constant $M > 0$ and approximate our objective $f(y)$ by the linear model augmented with quadratic regularizer:

$$f(y) \quad \approx \quad \Omega_M(x; y) \quad := \quad f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{M}{2}\|y - x\|^2, \qquad x, y \in \mathbb{R}^n.$$

By Theorem 3.1.8, we know that for a sufficiently large regularization parameter (at least, for $M \geq L$), this will be the *global model*: $f(y) \leq \Omega_M(x; y)$ for any $y \in \mathbb{R}^n$. One step of the gradient method consists in minimizing the model $\Omega_M(x; y)$ in y to obtain the next iterate:

$$x^+ \quad = \quad x^+_M(x) \quad = \quad \arg\min_{y \in \mathbb{R}^n} \Big[\, \Omega_M(x; y) \, \Big]. \tag{3.8}$$

Note that a solution to subproblem (3.8) always exists, but may not be unique. If there are many solutions, we can pick any for x^+.

Example 3.2.1 (Euclidean Norm)**.** Let the norm be the standard Euclidean: $\| \cdot \| \equiv \| \cdot \|_2$. To compute x^+ we differentiate $g(y) \equiv \Omega_M(x; y)$ with respect to y:

$$\nabla g(y) \quad = \quad \nabla f(x) + M(y - x),$$

and set the gradient to zero $\nabla g(x^+) = 0$ which gives the unique solution:

$$x^+ \quad = \quad x - \tfrac{1}{M} \nabla f(x),$$

and the minimum of the model is

$$g^\star \quad = \quad \Omega_M(x; x^+) \quad = \quad f(x) - \tfrac{1}{2M} \|\nabla f(x)\|_2^2.$$

Therefore, for the Euclidean norm, computing the minimizer of (3.8) corresponds exactly to the classical gradient descent step (3.2) with step-size $\alpha = 1/M$. □

Example 3.2.2 (General Norm)**.** To solve the subproblem (3.8) for the case of a general norm, let us represent the displacement as follows:

$$y - x \quad = \quad \tau h,$$

where $h \in \mathbb{R}^n : \|h\| = 1$ and $\tau > 0$. Then, the subproblem becomes

$$\begin{aligned} \Omega_M(x; x^+) \quad &= \quad \min_{y \in \mathbb{R}^n} \Big[\Omega_M(x; y) \Big] \quad = \quad \min_{\tau > 0} \min_{h \in \mathbb{R}^n : \|h\| = 1} \Big[f(x) + \tau \langle \nabla f(x), h \rangle + \tfrac{M}{2} \tau^2 \Big] \\ &= \quad \min_{\tau > 0} \Big[f(x) - \tau \|\nabla f(x)\|_* + \tfrac{M}{2} \tau^2 \Big] \quad = \quad f(x) - \tfrac{\|\nabla f(x)\|_*^2}{2M}. \end{aligned} \tag{3.9}$$

The optimum value is achieved for $x^+ - x = \tau^+ h^+$, where $\tau^+ = \frac{\|\nabla f(x)\|_*}{M}$ is the solution to a univariate quadratic minimization, and $h^+ \in \mathbb{R}^n$ is a vector of unit length such that

$$\langle \nabla f(x), h^+ \rangle \quad = \quad -\|\nabla f(x)\|_*.$$

Note that such h^+ always exists, but may not be unique. □

3.2.2 Progress of One Step

Now we have all ingredients to demonstrate the progress of one gradient step (3.8), when regularization parameter $M > 0$ is sufficiently large. We prove the following simple result, which is sometimes called *descent lemma* in the literature.

Proposition 3.2.3. *Let $M \geq L$. Then,*

$$f(x) - f(x^+) \quad \geq \quad \tfrac{1}{2M}\|\nabla f(x)\|_*^2. \tag{3.10}$$

Proof. Indeed, from Theorem 3.1.8 we have that

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

where in the last inequality we used that $M \geq L$. Now, plugging $y := x^+$ where x^+ is any solution to the subproblem (3.8), we get

$$f(x^+) \quad \leq \quad \Omega_M(x; x^+) \quad \overset{(3.9)}{=} \quad f(x) - \tfrac{1}{2M}\|\nabla f(x)\|_*^2,$$

which is the required progress. □

3.2.3 Convergence Rate to a Stationary Point

Let us consider the gradient method in the algorithmic form.

Algorithm 3.1: *Gradient Method.*

Initialization: $x_0 \in \mathbb{R}^n$, $\varepsilon > 0$

For $k \geq 0$ **iterate:**

1. If $\|\nabla f(x_k)\|_* \leq \varepsilon$ then
   return x_k
2. Choose $M_k > 0$
3. Perform the gradient step:

$$x_{k+1} \quad = \quad \arg\min_{y \in \mathbb{R}^n}\Bigl[\, \Omega_{M_k}(x_k; y) \quad := \quad f(x_k) + \langle \nabla f(x_k), y - x_k\rangle + \tfrac{M_k}{2}\|y - x_k\|^2 \,\Bigr]$$

In step 2 of this method, we have to choose the regularization parameter $M_k > 0$. A natural choice, which is approved by the condition of Proposition 3.2.3 is the *constant step-size*: $M_k \equiv L$. Of course, for that we have to know the Lipschitz constant.

Another powerful rule is to simply ensure that at each step $k \geq 0$, we have the progress (3.10):

$$\begin{gathered}\text{Choose } M_k > 0 \text{ s.t. for } x_k^+(M_k) := \arg\min_{y \in \mathbb{R}^n} \Omega_{M_k}(x_k; y) \text{ it holds} \\ f(x_k) - f(x_k^+(M_k)) \quad \geq \quad \tfrac{1}{2M_k}\|\nabla f(x_k)\|_*^2.\end{gathered} \tag{3.11}$$

Such condition can be achieved by an *adaptive search* procedure, that we discuss in the next section.

We prove the following convergence result for the gradient method.

Theorem 3.2.4. *Let f be bounded from below: $f^\star := \inf_{y \in \mathbb{R}^n} f(x) > -\infty$. Consider the sequence generated by the gradient method,*

$$x_{k+1} \;=\; x_k^+(M_k), \qquad k \geq 0.$$

for a sequence of regularization parameters $\{M_k\}_{k \geq 0}$.

Assume that all M_k satisfy the progress condition (3.11) and are bounded from above: $M_k \leq M_\star$ for all $k \geq 0$. Then, it holds

$$\frac{2M_\star(f(x_0)-f^\star)}{k} \;\geq\; \frac{1}{k}\sum_{i=0}^{k-1} \|\nabla f(x_i)\|_*^2 \;\geq\; \min_{0 \leq i \leq k-1} \|\nabla f(x_i)\|_*^2. \tag{3.12}$$

Proof. For every iteration, it holds:

$$f(x_i) - f(x_{i+1}) \;\overset{(3.11)}{\geq}\; \frac{1}{2M_k}\|\nabla f(x_i)\|_*^2 \;\geq\; \frac{1}{2M_\star}\|\nabla f(x_i)\|_*^2.$$

Summing up these inequalities for $0 \leq i \leq k-1$, we get

$$f(x_0) - f(x_k) \;\geq\; \frac{1}{2M_\star}\sum_{i=0}^{k-1} \|\nabla f(x_i)\|_*^2.$$

Using the bound: $f(x_k) \geq f^\star$ and multiplying both sides by $\frac{2M_\star}{k}$ completes the proof. □

We see that the gradient method makes the minimal gradient to converge to zero:

$$\min_{0 \leq i \leq k-1} \|\nabla f(x_i)\|_* \;\to\; 0, \qquad \text{with} \qquad k \to +\infty.$$

However, we do not ensure monotonicity of the sequence $\{\|\nabla f(x_k)\|_*\}_{k \geq 0}$, and it does not hold in general.

As a direct consequence of (3.12), we obtain the following complexity bound for our Algorithm 3.1.

Corollary 3.2.5. *To find a point $\bar{x} \in \mathbb{R}^n$ such that $\|\nabla f(\bar{x})\|_* \leq \varepsilon$, the gradient method needs to perform*

$$K \;=\; \left\lfloor \frac{2M_\star(f(x_0)-f^\star)}{\varepsilon^2} \right\rfloor$$

first-order oracle calls, where $M_\star \geq M_k$, $k \geq 0$, is an upper bound on the regularization parameters.

In particular, choosing $M_k \equiv L$, we obtain the complexity:

$$K \;=\; \left\lfloor \frac{2L(f(x_0)-f^\star)}{\varepsilon^2} \right\rfloor. \tag{3.13}$$

In contrast to the complexity bound for global optimization proved in previous lectures: $O((1/\varepsilon)^n)$, we see from (3.13) that

the complexity of the gradient method does not depend on the dimension n,

at least explicitly (it may depend on the dimension indirectly through parameters, such as the Lipschitz constant L). This explains why the gradient method is the most popular approach for solving huge-scale problems, when the dimension is extremely high ($n \to +\infty$).