

# Lecture 4: Smooth Functions and Optimality Conditions

Yudong Chen

In this lecture, we use Taylor's Theorem to characterize smooth functions and their local minima. In particular, we derive necessary/sufficient optimality conditions for smooth unconstrained optimization.

## 1 Properties of smooth functions

Recall:  $f$  is called  $L$ -smooth w.r.t.  $\|\cdot\|$  if

$$\forall x, y \in \text{dom}(f) : \|\nabla f(x) - \nabla f(y)\|_* \leq L \|x - y\|.$$

**Lemma 1.** Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be an  $L$ -smooth function w.r.t.  $\|\cdot\|$ . Then,  $\forall x, y \in \text{dom}(f)$ :

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2,$$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle - \frac{L}{2} \|y - x\|^2.$$

*Proof.* We prove the first inequality; second one left as exercise. From Part 1 of Taylor theorem (Theorem 1 in Lecture 3):

$$\begin{aligned} & f(y) - f(x) - \langle \nabla f(x), y - x \rangle \\ &= \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle dt - \int_0^1 \langle \nabla f(x), y - x \rangle dt \\ &= \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt \\ &\leq \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\|_* \|y - x\| dt && \text{Holder} \\ &\leq \int_0^1 Lt \|y - x\|^2 dt && \text{Smoothness} \\ &= \frac{L}{2} \|y - x\|^2. \end{aligned}$$

□

*Remark 1.* In fact, the condition in Lemma 1 is equivalent to  $L$ -smoothness; see Lemma 3.

Recall the Lowner order: For symmetric matrices  $A$  and  $B$ ,

$$A \succcurlyeq B \iff A - B \succcurlyeq 0 \iff A - B \text{ is p.s.d.}$$

In particular,

$$aI \preccurlyeq A \preccurlyeq bI \iff a \leq \lambda_i(A) \leq b, \forall i$$

where  $\lambda_1(A) \leq \dots \leq \lambda_d(A)$  are the eigenvalues of  $A$ .

**Lemma 2.** Suppose that  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is twice continuously differentiable on  $\text{dom}(f)$ . Then  $f$  is  $L$ -smooth w.r.t.  $\|\cdot\|_2$  if and only if

$$-LI \preceq \nabla^2 f(x) \preceq LI, \quad \forall x \in \text{dom}(f).$$

To give the proof, we use the matrix operator norm:

$$\|A\|_2 := \sup_{x: \|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2} \stackrel{\text{for symmetric } A}{=} \max_i |\lambda_i(A)|.$$

Then by definition:

$$\|Ax\|_2 \leq \|A\|_2 \|x\|_2. \quad (1)$$

*Proof.*  $\implies$  direction: Suppose that  $f$  is  $L$  smooth. Want to show:  $\nabla^2 f(x) \preceq LI$ . ( $-LI \preceq \nabla^2 f(x)$  left as exercise.)

Let  $x \in \text{dom}(f)$ ,  $x + \alpha p \in \text{dom}(f)$ ,  $\alpha > 0$ . From Part 4 of Taylor theorem (Theorem 1 in Lecture 3):

$$f(x + \alpha p) = f(x) + \langle \nabla f(x), \alpha p \rangle + \frac{\alpha^2}{2} p^\top \nabla^2 f(x + \gamma \alpha p) p \quad (2)$$

for some  $\gamma \in (0, 1)$ . From Lemma 1:

$$f(x + \alpha p) \leq f(x) + \langle \nabla f(x), \alpha p \rangle + \frac{L}{2} \alpha^2 \|p\|_2^2. \quad (3)$$

Combining (3) and (2):

$$\frac{\alpha^2}{2} p^\top \underbrace{\nabla^2 f(x + \gamma \alpha p)}_{\rightarrow \nabla^2 f(x) \text{ as } \alpha \rightarrow 0} p \leq \frac{L}{2} \alpha^2 \|p\|_2^2.$$

Taking the limit  $\alpha \rightarrow 0$ , we get  $p^\top \nabla^2 f(x) p \leq L \|p\|_2^2$ . Since  $p$  is arbitrary, we have  $\nabla^2 f(x) \preceq LI$ .

$\Leftarrow$  direction: Suppose that  $\forall x : -LI \preceq \nabla^2 f(x) \preceq LI \iff \|\nabla^2 f(x)\|_2 \leq L$ . Want to show:  $\forall x, y \in \text{dom}(f) : \|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2$ .

From Part 3 of Taylor theorem:  $\forall x, y \in \text{dom}(f)$ :

$$\begin{aligned} \|\nabla f(y) - \nabla f(x)\|_2 &= \left\| \int_0^1 \nabla^2 f(x + t(y-x)) (y-x) dt \right\|_2 \\ &\leq \int_0^1 \|\nabla^2 f(x + t(y-x)) (y-x)\|_2 dt && \text{Jensen's} \\ &\leq \int_0^1 \|\nabla^2 f(x + t(y-x))\|_2 \|y-x\|_2 dt && \text{by (1)} \\ &\leq \int_0^1 L \|y-x\|_2 dt \\ &= L \|y-x\|_2. \end{aligned}$$

□

## 2 Characterizing minima of smooth functions

In this part, we consider *unconstrained* optimization, that is,  $\mathcal{X} = \mathbb{R}^d$  in the problem

$$\min_{x \in \mathcal{X}} f(x) \quad (\text{P})$$

## 2.1 Necessary conditions for optimality

### Theorem 1.

1. (First-order necessary condition) Suppose that  $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$  is continuously differentiable. If  $x^*$  is a local minimizer of  $f$ , then  $\nabla f(x^*) = 0$ .
2. (Second-order necessary condition) Suppose that  $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$  is twice continuously differentiable. Then, in addition to 1), we have  $\nabla^2 f(x^*) \succcurlyeq 0$ .

*Remark 2.* A point  $x$  satisfying  $\nabla f(x) = 0$  is called a (first-order) *stationary point* of  $f$ . A point  $x$  satisfying  $\nabla f(x) = 0$  and  $\nabla^2 f(x) \succcurlyeq 0$  is called a *second-order stationary point* (SOSP). Theorem 1 says a local minimizer must be a stationary point if  $f$  is continuously differentiable, and it must be a SOSP if  $f$  is twice continuously differentiable.

*Proof of Theorem 1.* Part 1: Suppose for the purpose of contradiction (f.p.o.c.) that  $\nabla f(x^*) \neq 0$ , but  $x^*$  is a local minimizer. Apply Part 2 of Taylor's Theorem with  $y = x^* - \alpha \nabla f(x^*)$ ,  $x = x^*$ ,  $\alpha > 0$ :

$$f(x^* - \alpha \nabla f(x^*)) = f(x^*) - \alpha \langle \nabla f(x^* - \gamma \alpha \nabla f(x^*)), \nabla f(x^*) \rangle$$

for some  $\gamma \in (0, 1)$ . Note that if  $\alpha$  were equal to 0, then

$$-\langle \nabla f(x^*), \nabla f(x^*) \rangle = -\|\nabla f(x^*)\|_2^2.$$

Since  $\nabla f$  is continuous by assumption, for all sufficiently small  $\alpha > 0$ , it holds that

$$-\langle \nabla f(x^* - \gamma \alpha \nabla f(x^*)), \nabla f(x^*) \rangle \leq -\frac{1}{2} \|\nabla f(x^*)\|_2^2,$$

hence

$$f(x^* - \alpha \nabla f(x^*)) \leq f(x^*) - \underbrace{\frac{\alpha}{2} \|\nabla f(x^*)\|_2^2}_{>0 \text{ by assumption}} < f(x^*).$$

Therefore,  $x^*$  cannot be a local minimizer, a contradiction.

Part 2: Suppose f.p.o.c. that  $\nabla^2 f(x^*)$  has a negative eigenvalue  $-\lambda$ , where  $\lambda > 0$ . Then, there exists  $\theta \in \mathbb{R}^d$ ,  $\|\theta\|_2 = 1$  such that

$$\theta^\top \nabla^2 f(x^*) \theta = -\lambda.$$

Using Part 4 of Taylor's Theorem with  $x = x^*$ ,  $y = x^* + \alpha \theta$ ,  $\alpha > 0$ :

$$f(x^* + \alpha \theta) = f(x^*) + \underbrace{\langle \nabla f(x^*), \alpha \theta \rangle}_{\text{by part 1}} + \frac{\alpha^2}{2} \theta^\top \nabla^2 f(x^* + \gamma \alpha \theta) \theta$$

for some  $\gamma \in (0, 1)$ . As  $\nabla^2 f$  is continuous, for all sufficiently small  $\alpha > 0$ , it holds that

$$\theta^\top \nabla^2 f(x^* + \gamma \alpha \theta) \theta \leq -\frac{\lambda}{2},$$

hence

$$f(x^* + \alpha \theta) \leq f(x^*) - \frac{1}{4} \alpha^2 \lambda < f(x^*).$$

Therefore,  $x^*$  cannot be a local minimizer, a contradiction. □

### 2.1.1 An alternative proof

From calculus, we have the derivative tests for characterizing critical points of 1D functions. Taking these 1D results as given, we can use them to prove the multivariate results in Theorem 1.

Part 1: Define the 1-D function  $\phi(\alpha) = f(x^* - \alpha \nabla f(x^*))$ . If  $x^*$  is a local minimizer of  $f$ , then 0 is a local minimizer of  $\phi$ , then  $\phi'(0) = 0$  by Fermat's Theorem. But

$$\begin{aligned}\phi'(\alpha) &= \langle \nabla f(x^* - \alpha \nabla f(x^*)), -\nabla f(x^*) \rangle, \\ \phi'(0) &= -\|\nabla f(x^*)\|_2^2,\end{aligned}$$

so we must have  $\nabla f(x^*) = 0$ .

Part 2: Fix an arbitrary  $\theta \in \mathbb{R}^d$ , define  $\phi_\theta(\alpha) = f(x^* + \alpha\theta)$ . Use 2nd derivative test on  $\phi_\theta$ .

## 2.2 Sufficient condition for optimality

**Theorem 2** (Second-order sufficient condition). *Let  $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$  be twice continuously differentiable and assume that for some  $x^* \in \text{dom}(f)$ ,*

$$\begin{aligned}\nabla f(x^*) &= 0 \quad \text{and} \\ \nabla^2 f(x^*) &\succ 0.\end{aligned}$$

*Then  $x^*$  is a strict local minimizer of  $f$ .*

*Proof.* Let  $\mathcal{B}$  be a ball centered at  $x^*$  and of radius  $\rho$  that is sufficiently small so that

$$\nabla^2 f(x^* + p) \succ \epsilon I, \quad \forall p : \|p\|_2 \leq \rho$$

for some  $\epsilon > 0$ . (Such a ball must exist because  $\nabla^2 f(x^*) \succ 0$  and  $\nabla^2 f$  is continuous).

Apply Part 4 of Taylor's Theorem with  $x = x^*$ ,  $y = x^* + p$  and arbitrary  $p$  with  $\|p\|_2 \leq \rho$ : for some  $\gamma \in (0, 1)$ ,

$$\begin{aligned}f(x^* + p) &= f(x^*) + \langle \nabla f(x^*), p \rangle + \frac{1}{2} p^\top \nabla^2 f(x^* + \gamma p) p \\ &= f(x^*) + 0 + \frac{1}{2} p^\top \nabla^2 f(x^* + \gamma p) p && \text{by assumption} \\ &\geq f(x^*) + \frac{1}{2} \cdot \epsilon \cdot \|p\|_2^2 \\ &> f(x^*) \quad \text{if } \|p\|_2 \neq 0,\end{aligned}$$

so  $x^*$  is a strict local minimizer. □

*Remark 3.* We notice that there is a gap between the conditions in the last two theorems.

- The condition  $\nabla f(x^*) = 0, \nabla^2 f(x^*) \succ 0$  in Theorem 1 is necessary but not sufficient: it is possible that a point  $x$  satisfies this condition but is not a local min (e.g.,  $f(x) = x^3$  and  $x^* = 0$ ).
- The condition  $\nabla f(x^*) = 0, \nabla^2 f(x^*) \succ 0$  in Theorem 2 is sufficient but not necessary: it is possible that a local minimizer  $x^*$  has  $\nabla^2 f(x^*) = 0$  (e.g.,  $f(x) = x^4$  and  $x^* = 0$ ).

In general, it is computationally hard to check whether a point  $x$  is a local min, even for smooth unconstrained problems. For example, consider a function of the form

$$f(x) = (x_1^2, x_2^2, \dots, x_d^2) D(x_1^2, x_2^2, \dots, x_d^2)^\top = \sum_{i=1}^d \sum_{j=1}^d D_{ij} x_i^2 x_j^2,$$

which is a degree-4 polynomial in  $x$ . It is NP hard to decide whether  $x = 0$  is a local min (by reduction from Subset Sum; see Murty-Kabadi 1987),

*Remark 4.* Theorem 2 only guarantees *local* optimality, not global optimality.

## Appendices

**Lemma 3.** Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be a continuously differentiable function. If it holds that

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2} \|y - x\|_2^2, \quad \text{for all } x, y \in \mathbb{R}^d, \quad (4)$$

then  $f$  is an  $L$ -smooth function w.r.t.  $\|\cdot\|_2$ .

*Proof.* Let  $x, y \in \mathbb{R}^d$  be arbitrary and  $p \in \mathbb{R}^d$  be chosen later. Under the assumption we have the upper bound

$$\begin{aligned} \rho &:= f(y + p) - f(x) + f(x - p) - f(y) \\ &\leq \langle \nabla f(x), y + p - x \rangle + \frac{L}{2} \|y + p - x\|_2^2 + \langle \nabla f(y), x - p - y \rangle + \frac{L}{2} \|x - p - y\|_2^2 \\ &= -\langle \nabla f(x) - \nabla f(y), x - y - p \rangle + L \|x - y - p\|_2^2 \end{aligned}$$

and the lower bound

$$\begin{aligned} \rho &= f(y + p) - f(y) + f(x - p) - f(x) \\ &\geq \langle \nabla f(y), p \rangle - \frac{L}{2} \|p\|_2^2 + \langle \nabla f(x), -p \rangle - \frac{L}{2} \|p\|_2^2 \\ &= -\langle \nabla f(x) - \nabla f(y), p \rangle - L \|p\|_2^2. \end{aligned}$$

Combining the two bounds and rearranging, we get

$$\langle \nabla f(x) - \nabla f(y), x - y - 2p \rangle \leq L \|x - y - p\|_2^2 + L \|p\|_2^2.$$

Taking  $p = \frac{1}{2} [x - y - \frac{1}{L} (\nabla f(x) - \nabla f(y))]$  gives

$$\begin{aligned} \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_2^2 &\leq \frac{L}{4} \left\| x - y + \frac{1}{L} (\nabla f(x) - \nabla f(y)) \right\|_2^2 + \frac{L}{4} \left\| x - y - \frac{1}{L} (\nabla f(x) - \nabla f(y)) \right\|_2^2 \\ &= \frac{L}{2} \|x - y\|_2^2 + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2, \end{aligned}$$

Rearranging terms gives

$$\|\nabla f(x) - \nabla f(y)\|_2^2 \leq L^2 \|x - y\|_2^2,$$

which is the definition of  $L$ -smoothness. □

*Remark 5.* The condition (4) is equivalent to

$$|\langle \nabla f(x) - \nabla f(y), x - y \rangle| \leq L \|x - y\|_2^2 \quad \text{for all } x, y \in \mathbb{R}^d. \quad (5)$$

Proof left as exercise.

*Remark 6.* Suppose that  $f$  is a convex function satisfying the upper bound

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2} \|y - x\|_2^2 \quad \text{for all } x, y \in \mathbb{R}^d$$

or equivalently

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L \|x - y\|_2^2 \quad \text{for all } x, y \in \mathbb{R}^d.$$

Then  $f$  satisfies (4) and hence  $f$  is  $L$ -smooth.