Lecture 5

5.1 Convex Functions

5.1.1 Motivation

We have studied the following problem classes so far.

1. **Global minimization of smooth functions:**

$$\mathcal{F} \quad = \quad \Big\{ f : \mathbb{R}^n \to \mathbb{R} \ \text{ s.t. } \ f \text{ is continuous / smooth} \Big\}.$$

   The goal: finding a *global solution* $\bar{x}$ s.t. $f(\bar{x}) - f^\star \leq \varepsilon$. We saw that these problems are too hard to be efficiently solvable in general. The optimal algorithm was the simplest grid search.

2. **Finding stationary points of smooth functions.** The same class $\mathcal{F}$, but the goal is much less ambitious: finding $\bar{x}$ s.t. $\|\nabla f(\bar{x})\| \leq \varepsilon$. We have analyzed two algorithms for this class: gradient method and stochastic gradient method, which both possess a dimension-free complexity bounds.

Now, we have two options. *Option 1:* to try finding something in between these classes, which is a difficult (but interesting) path. By using higher-order smoothness, we are able to achieve better complexities to get a stationary point than that one of the gradient method. However, checking whether a point is a local minimum, local maximum, or a saddle point might be NP-hard in general.

Option 2: to find a smaller problem class $\mathcal{F}' \subset \mathcal{F}$ for which the initial goal of finding a global solution is feasible. For example, we want the following property to hold: whenever it holds $\nabla f(\bar{x}) = 0$ then $\bar{x}$ is a global solution, so every stationary point is a global minimum. It appears that such path essentially leads us to *convex functions*.

Convex functions play a central role in optimization theory, as they provide examples of a broad range of globally solvable problem classes. We review the basic facts about convex functions that are most useful for the analysis of optimization algorithms. Convexity is an important concept in mathematics with a rich, well-developed theory, and a wide variety of applications across different domains besides optimization. There are plenty of excellent courses and textbooks on convex analysis, which we recommend for further reading.

5.1.2 Univariate Convex Functions

Let $f : \mathbb{R} \to \mathbb{R}$ be a *continuous* function, as we almost always assume in this course.[1] We say that f is *convex* if for any two points $x, y \in \mathbb{R}$:

$$f\big(\tfrac{x+y}{2}\big) \quad \leq \quad \tfrac{1}{2}\big(f(x) + f(y)\big). \tag{5.1}$$

[1] Otherwise, it would not be called Continuous Optimization.

That is, the value of the function at the midpoint of any interval never exceeds the average of the values at the endpoints.

This simple property, coupled with the fact that we require it *for any two points* $x, y \in \mathbb{R}$, leads to many consequences. We leave the proof of the following facts to the reader.

Proposition 5.1.1. *For any* $x, y \in \mathbb{R}$ *and* $0 \leq \lambda \leq 1$*, it holds*

$$f(\lambda x + (1-\lambda)y) \quad \leq \quad \lambda f(x) + (1-\lambda) f(y). \tag{5.2}$$

Thus, *midpoint* convexity (5.1) leads to a similar inequality along the *entire* segment when f is continuous. Geometrically, this means that the chord connecting any two points $(x, f(x))$ and $(y, f(y))$ always lies on or above the graph of f. The value $\lambda x + (1-\lambda)y$ is called a *convex combination* of x and y for $0 \leq \lambda \leq 1$.

Functions that satisfy (5.1) but not (5.2) are pathologically rare cases; in fact, these definitions are equivalent as soon as the function is measurable (which is true for any continuous function). Moreover, it can be shown that convex functions in the sense of (5.2) are locally *Lipschitz continuous* on the interior of their domain.

Proposition 5.1.2 (Jensen's inequality)**.** *Let* $x_1, \ldots, x_N \in \mathbb{R}$ *be a finite set of points and let* $\lambda \in \Delta_N$ *be from the standard simplex,* $\Delta_N = \{\lambda \in \mathbb{R}^N_+ \,:\, \langle e, \lambda \rangle = 1\}$*, where* $e \in \mathbb{R}^N$ *is the vector of all ones. Then*

$$f\Big(\sum_{i=1}^N \lambda_i x_i\Big) \quad \leq \quad \sum_{i=1}^N \lambda_i f(x_i). \tag{5.3}$$

Inequality (5.3) can be generalized beyond discrete distributions, by taking a limit to an infinite amount of points. The following inequality is very useful when analyzing stochastic algorithms.

Proposition 5.1.3 (Jensen's inequality for expectations)**.** *Let* ξ *be a random variable taking values in* $\mathbb{R}$*. Then*

$$f\big(\mathbb{E}[\xi]\big) \quad \leq \quad \mathbb{E}\big[f(\xi)\big]. \tag{5.4}$$

Of course, it is straightforward to see that (5.4) $\Rightarrow$ (5.3) $\Rightarrow$ (5.2) $\Rightarrow$ (5.1). What is more interesting is that the reverse implications also hold, making them equivalent.

We also say that f is *concave*, if $-f$ is convex.

5.1.3 Maximizing Convex Functions

Convex functions are meant to be minimized. But what if we try to maximize them? It is easy to see that a maximum of a convex function over a set is always at the boundary. Interestingly, this property applied to a function with *all* affine perturbations *defines convexity*. In other words, if we "tilt" a convex function by adding to it a linear slope, we can force the highest point to lie at one of the endpoints of any segment, and this is the characteristic property of convexity.

Theorem 5.1.4. *The following conditions are equivalent:*

- $f : \mathbb{R} \to \mathbb{R}$ *is convex.*
- *For any segment* $[x, y] \in \mathbb{R}$ *and for any affine function* $t \mapsto at + b$ *it holds*

$$\max_{t \in [x,y]} \Big\{ f(t) - at - b \Big\} \quad = \quad \max\Big\{ f(x) - ax - b,\ f(y) - ay - b \Big\}. \tag{5.5}$$

Proof. Let f be convex. Then, since any $t \in [x, y]$ can be represented as the convex combination: $t = \lambda x + (1-\lambda)y$, for some $0 \le \lambda \le 1$, we have

$$\begin{aligned} f(t) + at + b &\overset{(5.2)}{\le} \lambda\Big[f(x) - ax - b\Big] + (1-\lambda)\Big[f(y) - ay - b\Big] \\ &\le \max\Big\{f(x) - ax - b,\ f(y) - ay - b\Big\}, \end{aligned}$$

which is (5.5).

Now assume that (5.5) holds for any $a, b \in \mathbb{R}$. Let us verify that for any two fixed $x, y \in \mathbb{R}$, $x \ne y$, and for all $0 \le \lambda \le 1$ it holds:

$$\begin{aligned} & f(\lambda x + (1-\lambda)y) \ \le\ \lambda f(x) + (1-\lambda) f(y) \\ & \Leftrightarrow \\ & f(y + \lambda(x-y)) - f(y) - \lambda(f(x) - f(y)) \ \le\ 0 \qquad (5.6) \\ & \Leftrightarrow \\ & f(t) - f(y) - (t-y)\cdot \tfrac{f(x)-f(y)}{x-y} \ \le\ 0, \end{aligned}$$

where $t \equiv y + \lambda(x-y) \in [x, y]$. Now, we set $a := \frac{f(x)-f(y)}{x-y}$ and $b := f(y) - y \cdot a$, and consider the following perturbation of f by an affine function:

$$\varphi(t) \ :=\ f(t) - at - b \ =\ f(t) - f(y) - (t-y)\cdot \tfrac{f(x)-f(y)}{x-y},$$

and by assumption (5.5), we have

$$\varphi(t) \ \le\ \max\Big\{\varphi(x),\ \varphi(y)\Big\} \ =\ 0,$$

which proves (5.6). □

Thus, we see that *affine functions* play a fundamental role in the theory of convexity. Note that affine functions are uniquely identified as those and only those that are simultaneously convex and concave.

Proposition 5.1.5. *The following conditions are equivalent:*

- *f is both convex and concave, i.e. f preserves convex combination, for any $x, y \in \mathbb{R}$:*

$$f(\lambda x + (1-\lambda)y) \ =\ \lambda f(x) + (1-\lambda) f(y), \qquad 0 \le \lambda \le 1.$$

- *f is affine, i.e. $f(t) = at + b$, for some $a, b \in \mathbb{R}$.*

Exercise 5.1.1. Prove Proposition 5.1.5.

5.1.4 Multivariate Convex Functions

A multivariate function $f : \mathbb{R}^n \to \mathbb{R}$ is called *convex* if its restriction to any segment is convex. So, for any two $x, y \in \mathbb{R}^n$ and $0 \leq \lambda \leq 1$ it holds:

$$f(\lambda x + (1-\lambda)y) \quad \leq \quad \lambda f(x) + (1-\lambda) f(y). \tag{5.7}$$

All the properties discussed for the univariate case are naturally inherited by general convex functions. As previously, we say that the function f is *concave* if $-f$ is convex. The only functions that are both convex and concave are *affine functions*:

$$f(x) \quad = \quad \langle a, x \rangle + b, \qquad x \in \mathbb{R}^n,$$

for some $a \in \mathbb{R}^n, b \in \mathbb{R}$.

Thus, we know that the maximum of a convex function over any compact set $Q \subset \mathbb{R}^n$ is achieved on its boundary:

$$\max_{x \in Q} f(x) \quad = \quad \max_{x \in \partial Q} f(x),$$

and the same applies to the minimum of a concave function. In particular, we conclude that the minimum of a *linear function* over any compact set is always achieved on the boundary:

$$\min_{x \in Q} \langle a, x \rangle \quad = \quad \min_{x \in \partial Q} \langle a, x \rangle.$$

When Q is non-compact, we must ensure that the minimum actually exists. This fundamental fact is used in linear programming (where Q is a polyhedron and the minimum is attained at a vertex) and in more general settings such as semidefinite programming (where Q is the intersection of the cone of positive semidefinite matrices with an affine set).

5.1.5 Differentiable Convex Functions

For now, let us assume that $f : \mathbb{R}^n \to \mathbb{R}$ is several times differentiable convex function defined on the whole space (unconstrained minimization). We consider more general non-differentiable convex functions later in the course. We have the following important inequalities, that serve as equivalent definitions of convexity.

Theorem 5.1.6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. Then, the following statements are equivalent:*

- *f is convex* (5.7).
- *The linear approximation of f is its global lower bound:*

$$f(y) \quad \geq \quad f(x) + \langle \nabla f(x), y - x \rangle, \qquad x, y \in \mathbb{R}^n. \tag{5.8}$$

- *The gradient mapping is monotone:*

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \quad \geq \quad 0, \qquad x, y \in \mathbb{R}^n. \tag{5.9}$$

- *The Hessian is positive semidefinite:*

$$\nabla^2 f(x) \quad \succeq \quad 0, \qquad x \in \mathbb{R}^n. \tag{5.10}$$

Proof. Let f be convex. Then, for any $x, y \in \mathbb{R}^n$ and $0 < \alpha < 1$ we have

$$\begin{aligned} f((1-\alpha)x + \alpha y) &\leq (1-\alpha)f(x) + \alpha f(y) \\ \Leftrightarrow & \\ f(y) &\geq f(x) + \tfrac{1}{\alpha}(f(x + \alpha(y-x)) - f(x)). \end{aligned}$$

Taking the limit $\alpha \to 0$ gives (5.8).

At the same time, substituting into (5.8) pairs of points (x, x_α) and (y, x_α), where $x_\alpha = (1-\alpha)x + \alpha y$, we get:

$$\begin{aligned} f(x) &\geq f(x_\alpha) + \langle \nabla f(x_\alpha), x - x_\alpha \rangle &= f(x_\alpha) + \alpha \langle \nabla f(x_\alpha), y - x \rangle, \\ f(y) &\geq f(x_\alpha) + \langle \nabla f(x_\alpha), y - x_\alpha \rangle &= f(x_\alpha) + (1-\alpha) \langle \nabla f(x_\alpha), x - y \rangle. \end{aligned}$$

Multiplying the first by $(1-\alpha)$ and the second by α, gives (5.7) after summation. Thus, we showed that (5.7) and (5.8) are equivalent.

To obtain (5.9), we only need to sum up a pair of inequalities (5.8), swapping the roles of x and y.

To show (5.10) we use the definition of the Hessian. For an arbitrary unit direction $h \in \mathbb{R}^n$, $\|h\| = 1$, and a sufficiently small $\varepsilon > 0$, we have:

$$\nabla f(x + \varepsilon h) = \nabla f(x) + \varepsilon \nabla^2 f(x) h + o(\varepsilon).$$

Hence, multiplying this vector equation by h and rearranging the terms, we get:

$$\langle \nabla^2 f(x) h, h \rangle = \tfrac{1}{\varepsilon^2} \langle \nabla f(x + \varepsilon h) - \nabla f(x), \varepsilon h \rangle + o(1) \overset{(5.9)}{\geq} o(1).$$

Taking the limit $\varepsilon \to 0$ proves (5.10).

Finally, using Taylor's formula, we get that (5.10) implies (5.8):

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle = \int_0^1 (1-\tau) \langle \nabla^2 f(x + \tau(y-x))(y-x), y - x \rangle d\tau \overset{(5.10)}{\geq} 0,$$

which completes the proof. □

Second-order condition $\nabla^2 f(x) \succeq 0$ is useful for checking whether a function is convex.

Example 5.1.7. The following univariate functions are convex:

- $f(x) = e^x$
- $f(x) = -\ln x$, for $x > 0$
- $f(x) = x \ln x$, for $x > 0$
- $f(x) = \ln(1 + e^x)$
- $f(x) = |x|^p$, for $p \geq 1$

5.1.6 Global Optimality

We come to the following important implication in convex optimization.

Corollary 5.1.8. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and convex. Then,*

$$\nabla f(x^\star) \;=\; 0 \qquad \Leftrightarrow \qquad x^\star \text{ is a global minimum.} \tag{5.11}$$

Proof. Indeed, substituting $x := x^\star$ into (5.8) gives

$$f(y) \;\geq\; f(x^\star), \qquad y \in \mathbb{R}^n.$$

The other direction is already known as *optimality condition* for local minimum. □

It is interesting that seeking for a functional class that satisfy (5.11), together with some simple natural conditions, we necessarily comes to the class of *convex functions*.

Theorem 5.1.9. *Let $\mathcal{F} \subset \{f : \mathbb{R}^n \to \mathbb{R},$ differentiable$\}$ be a maximal class of functions such that:*

1. *For any $f \in \mathcal{F}$: $\nabla f(\bar{x}) = 0 \Rightarrow \bar{x}$ is a global minimum.*
2. *If $f_1, f_2 \in \mathcal{F}$ then $f_1 + f_2 \in \mathcal{F}$.*
3. *Any affine function belongs to our problem class: $\langle a, x\rangle + b \in \mathcal{F}$, for any $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$.*

Then, $\mathcal{F} \equiv$ convex differentiable functions.

Proof. Let $f \in \mathcal{F}$. Fix $x \in \mathbb{R}^n$. Denote

$$\varphi(y) \;=\; f(y) - \langle \nabla f(x), y\rangle \;\in\; \mathcal{F}.$$

Then,

$$\nabla\varphi(y) \;=\; \nabla f(y) - \nabla f(x).$$

Hence $\nabla\varphi(x) = 0$ and x is a global minimum of φ:

$$\varphi(y) = f(y) - \langle \nabla f(x), y\rangle \;\geq\; \varphi(x) = f(x) - \langle \nabla f(x), x\rangle, \qquad \forall y \in \mathbb{R}^n.$$

This means that $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$ for any $x, y \in \mathbb{R}^n$. Hence f is convex. □

5.2 Convergence of Gradient Method

5.2.1 Smooth Convex Functions

Now, we couple both our assumptions together. We assume that $f : \mathbb{R}^n \to \mathbb{R}$ is both convex and smooth, i.e. its gradient is Lipschitz conitnuous, for any $x, y \in \mathbb{R}^n$: $\|\nabla f(y) - \nabla f(x)\|_* \leq L\|y - x\|$.

We have the following theorem.

Theorem 5.2.1. *The following conditions are equivalent:*

1. *f is convex and smooth.*
2. *For any $x, y \in \mathbb{R}^n$, it hods: $0 \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle \leq \frac{L}{2}\|y - x\|^2$.*
3. *For any $x, y \in \mathbb{R}^n$, it hods: $0 \leq \langle \nabla f(y) - \nabla f(x), y - x\rangle \leq L\|y - x\|^2$.*

4. *For any $x, h \in \mathbb{R}^n$, it hods: $0 \leq \langle \nabla^2 f(x)h, h\rangle \leq L\|h\|^2$ (for twice continuously differentiable functions).*

5. *For any $x, y \in \mathbb{R}^n$, it holds: $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_*^2$.*

6. *For any $x, y \in \mathbb{R}^n$, it holds: $\langle \nabla f(y) - \nabla f(x), y - x\rangle \geq \frac{1}{L}\|\nabla f(y) - \nabla f(x)\|_*^2$.*

Proof. *1* $\Rightarrow$ *2* $\Rightarrow$ *3* $\Rightarrow$ *4* are trivial. At the same time, *1* follows from *4* by already established second-order characterizations of convexity and smoothness (see Theorem 5.1.6. and Theorem 3.1.4. from Lecture 3).

Let us prove *5*. First consider $x = x^\star$, which is

$$f(y) - f^\star \quad \geq \quad \tfrac{1}{2L}\|\nabla f(y)\|_*^2. \tag{5.12}$$

Note that such progress is satisfied for one gradient step $y \mapsto y^+(L)$ (Proposition 3.2.3 from Lecture 3)! Hence, since $f^\star \leq f(y^+(L))$, (5.12) is established for any $y \in \mathbb{R}^n$.

For an arbitrary x, define the tilted function $\varphi(y) = f(y) - \langle \nabla f(x), y\rangle$ that has a minimum at x. Clearly, $\varphi(\cdot)$ belongs to our problem class, and we already established inequality (5.12):

$$\varphi(y) - \varphi^\star \quad \geq \quad \tfrac{1}{2L}\|\nabla \varphi(y)\|_*^2.$$

Substituting the expression for $\varphi(\cdot)$, we get

$$f(y) - \langle \nabla f(x), y\rangle + f(x) + \langle \nabla f(x), x\rangle \quad \geq \quad \tfrac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_*^2,$$

which proves *5*. To get *6* from 5 we just need to sum up it two times, swapping the role of x and y.

Finally, having *6*, we immediately conclude that f is convex, and by Cauchy-Schwarz inequality,

$$\tfrac{1}{L}\|\nabla f(y) - \nabla f(x)\|_*^2 \quad \leq \quad \langle \nabla f(y) - \nabla f(x), y - x\rangle \quad \leq \quad \|\nabla f(y) - \nabla f(x)\|_* \cdot \|y - x\|,$$

which proves that f is smooth. □

5.2.2 Convergence Rate

Let us study the convergence of the gradient method on our new problem class. By the previous analysis, we have the following progress of one step, using the constant step size (see Proposition 3.2.3 in Lecture 3):

$$f(x_k) - f(x_{k+1}) \quad \geq \quad \tfrac{1}{2L}\|\nabla f(x_k)\|_*^2. \tag{5.13}$$

At the same time, by convexity, we have

$$f^\star \quad \geq \quad f(x_k) + \langle \nabla f(x_k), x^\star - x_k\rangle,$$

or, rearranging the terms,

$$\begin{aligned} F_k \quad &:= \quad f(x_k) - f^\star \quad \leq \quad \langle \nabla f(x_k), x_k - x^\star\rangle \\ &\leq \quad \|\nabla f(x_k)\|_* \cdot \|x_k - x^\star\| \quad \leq \quad \|\nabla f(x_k)\|_* \cdot D, \end{aligned} \tag{5.14}$$

where we denote by D a global bound for all the iterates:

$$\|x_k - x^\star\| \quad \leq \quad D, \qquad k \geq 0. \tag{5.15}$$

For example, noticing the the method is *monotone* in function value (see (5.13)), we conclude that all iterates belong to the initial sublevel set:

$$x_k \;\in\; \mathcal{S}_0 \;:=\; \Big\{x \in \mathbb{R}^n \;:\; f(x) \leq f(x_0)\Big\}.$$

Hence, denoting

$$D \;:=\; \sup_{x \in \mathcal{S}_0} \|x - x^\star\|$$

and assuming that $D < +\infty$, we ensure (5.15). In some other cases, as for example, in the gradient method for the Euclidean norm, we can explicitly show that the distance to the solution is non-increasing and bounded.

Combining (5.13) with (5.14) we get the recursion:

$$F_k - F_{k+1} \;\geq\; \tfrac{1}{2LD^2} F_k^2. \tag{5.16}$$

Notice that in the imaginary case of *continuous* time, the corresponding dynamic $t \mapsto F_t$ can be analyzed in a differential form:

$$-\dot{F}_t \;\geq\; cF_t^2, \tag{5.17}$$

where $c > 0$ is a constant, $\dot{F}_t$ is the time derivative, which becomes a finite difference in the discrete-time case. Hence, we obtain

$$\tfrac{d}{dt}\big[\tfrac{1}{F_t}\big] \;=\; -\tfrac{\dot{F}_t}{F_t^2} \overset{(5.17)}{\geq} c,$$

and, after integrating, $F_t^{-1} \geq F_0^{-1} + ct$. Thus, $F_t = O(1/t)$, and we use these observations to analyze (5.16). We have, for every $k \geq 0$:

$$\tfrac{1}{F_{k+1}} - \tfrac{1}{F_k} \;=\; \tfrac{F_k - F_{k+1}}{F_{k+1} \cdot F_k} \;\geq\; \tfrac{1}{2LD^2} \cdot \tfrac{F_k^2}{F_k F_{k+1}} \;\geq\; \tfrac{1}{2LD^2}.$$

Telescoping, we get

$$\tfrac{1}{F_k} \;\geq\; \tfrac{1}{F_0} + \tfrac{k}{2LD^2} \geq \tfrac{k+4}{2LD^2},$$

where we used that $F_0 = f(x_0) - f^\star = f(x_0) - f(x^\star) - \langle \nabla f(x^\star), x_0 - x^\star \rangle \leq \tfrac{L}{2}\|x_0 - x^\star\|^2 \leq \tfrac{LD^2}{2}$. We have proved the following convergence result.

Theorem 5.2.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and smooth. For the iterates $\{x_k\}_{k \geq 0}$ generated by the gradient method, it holds:*

$$f(x_k) - f^\star \;\leq\; \tfrac{2LD^2}{k+4}, \qquad k \geq 0.$$

As a consequence, we see that the gradient method converges to the global solution, and to find a point x_k such that $f(x_k) - f^\star \leq \varepsilon$ it is enough to perform

$$k \;=\; \Big\lfloor \tfrac{2LD^2}{\varepsilon} \Big\rfloor + 1$$

first-order oracle calls. In the following lectures we will discuss the optimality of this complexity bound and whether it can be improved.

Literature

For an additional reading on convexity, we refer to [Hör94, BL06, NP06, Roc15]. Theorem 5.1.4 is from [Hör94], and Theorem 5.1.9 is from [Nes18]. See Section 2.1.5 in [Nes18] for the analysis of the gradient method on smooth convex functions.

[BL06] Jonathan Borwein and Adrian Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, 2006.

[Hör94] Lars Hörmander. *Notions of convexity*. Springer, 1994.

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

[NP06] Constantin Niculescu and Lars-Erik Persson. *Convex functions and their applications*, volume 23. Springer, 2006.

[Roc15] Ralph Tyrell Rockafellar. Convex analysis. 2015.