Lecture 1: Introduction and Convexity

In a high-level, optimization packages a decision task into three objects: a variable x, a feasible set Ω, and an objective f. The feasible set records what is allowed; the objective records what is preferred. This language is broad enough to cover resource allocation, profit maximization, training a machine learning model, maximum-likelihood estimation, control, and minimum-energy problems. Once a model is written this way (Definition 1.1), the basic question becomes: among all feasible choices, which one has the smallest cost? In this course, we focus on the continuous optimization problems where the domain E is a finite-dimensional real space.

Definition 1.1 (Optimization problem)**.** Let $f : E \to \mathbb{R}$, and let $\Omega \subseteq E$ be nonempty. The optimization problem associated with (f, Ω) is

$$p^\star := \inf \{f(x) : x \in \Omega\}.$$

A point $x^\star \in \Omega$ is a global minimizer if $f(x^\star) = p^\star$, i.e., $x^\star = \arg\min_{x\in\Omega} f(x)$.

Definition 1.2 (Approximate optimality)**.** Let (f, Ω) be an optimization problem in the sense of Definition 1.1, and let

$$p^\star := \inf \{f(x) : x \in \Omega\}.$$

For every $\varepsilon > 0$, a point $\widehat{x} \in \Omega$ is called ε-optimal if

$$f(\widehat{x}) \le p^\star + \varepsilon.$$

Basic outcomes of an optimization problem. Even at this level, several different things can happen. The feasible set may be empty, the infimum may be finite but not attained, or the problem may be unbounded below. This is one reason a near-optimal solution concept (Definition 1.2) is introduced: exact minimizers may fail to exist, and even when they do exist they may be harder to compute than near-optimal points. Before discussing certificates of optimality, it is therefore natural to ask a more basic question: when does an exact minimizer exist at all? Theorem 1.1 gives the standard topological answer under compactness and continuity.

Theorem 1.1 (Weierstrass)**.** *Let $\Omega \subseteq E$ be nonempty and compact, and let $f : \Omega \to \mathbb{R}$ be continuous. Then there exists $x^\star \in \Omega$ such that*

$$f(x^\star) = \min_{x\in\Omega} f(x).$$

Specification versus computation. An optimization specification tells us what the mathematical problem is: it determines the feasible set, the objective, the optimal value, and the solution notions

we care about, such as exact minimizers or ε-optimal points. But this still does not determine a computational problem. To speak about algorithms and complexity, we must also specify **how the instance is presented, which primitive operations are available, what cost model is being counted, and what kind of output is required**.

For finitely described models such as linear and quadratic programs, the input is a finite list of numbers and one counts arithmetic or bit operations. For large language models (LLMs) pretraining in practice, the input is a huge text corpus and code for model architecture, and one measures the total training wall-clock time on a GPU cluster. The same mathematical specification can therefore lead to very different computational questions under different access models. For example, the computational question will be very different for LLMs pretraining, if the GPU cluster is not owned but rented and the cost is actually the rental cost rather than the wall-clock time. Even for the simple problems which admit closed-form solutions, the computational question will be very different if our unit-cost arithmetic operations are finite-precision or infinite-precision. There are indeed lots of algorithms that are preferred because of their numerical stability.

For the purpose of this course, we will mostly focus on the an abstract and therefore general setting, where we assume the query oracle about local information (such as the value, gradient or Hessian) of the objective at each point, is the relatively cheap primitive. Throughout the course we assume we can work with real numbers with infinite precision.

Definition 1.3 (Differentiability on an open set)**.** Let $U \subseteq E$ be open and let $f : U \to \mathbb{R}$. We say that f is differentiable on U if it is differentiable at every point of U. In that case the first-order object at $x \in U$ is the differential

$$\nabla f(x) := Df(x) \in E^*, \qquad x \in U,$$

characterized by

$$Df(x)[h] = \langle \nabla f(x), h \rangle, \qquad h \in E,$$

where $\langle \xi, h \rangle := \xi(h)$ denotes the natural pairing between E^* and E. Thus $Df(x)[h]$ is the directional derivative of f at x along the direction h.

Remark 1.1 (Ambient convention for Lecture 1). Throughout this lecture, let E be a finite-dimensional real normed space and let E^* be its dual. We write

$$\langle \xi, h \rangle, \qquad \xi \in E^*,\ h \in E,$$

for the dual pairing. If a genuine inner product is being used, we will say so explicitly or decorate the notation, for instance by writing $\langle u, v \rangle_H$ or $\langle u, v \rangle_2$.

Definition 1.4 (Convex set, convex function, and strict convexity)**.** A set $C \subseteq E$ is convex if

$$\forall x, y \in C,\ \forall \theta \in [0,1], \qquad \theta x + (1-\theta) y \in C.$$

Let $C \subseteq E$ be convex and let $f : C \to \mathbb{R}$. The function f is convex if

$$\forall x, y \in C,\ \forall \theta \in [0,1], \qquad f(\theta x + (1-\theta) y) \le \theta f(x) + (1-\theta) f(y).$$

It is strictly convex if

$$\forall x, y \in C \text{ with } x \neq y,\ \forall \theta \in (0,1), \qquad f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta) f(y).$$

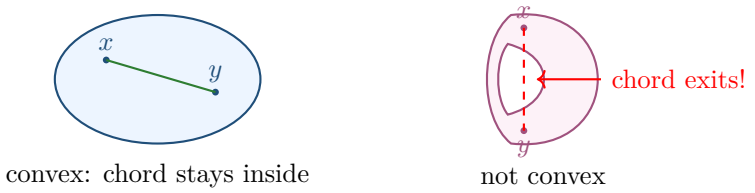

Figure 1: Convex set (left): every chord between two points stays inside. Non-convex set (right): some chord leaves the set.

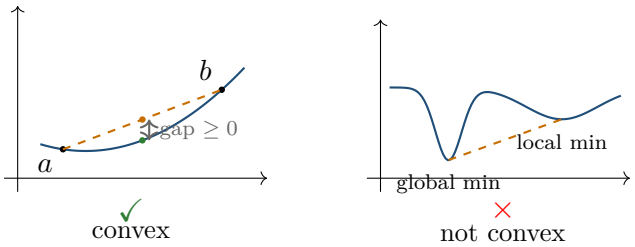

Figure 2: Left: a convex function — the chord between any two points lies above the graph. Right: a non-convex function with two local minima — gradients can be misleading.

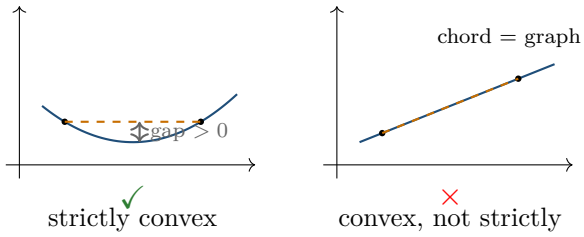

Figure 3: Left: a strictly convex function — the chord lies strictly above the graph between the endpoints. Right: an affine function is convex but not strictly convex, since the chord coincides with the graph.

1.1 Why convexity?

The point of studying convex optimization is not that all realistic optimization problems are convex. Rather, convexity is a good first structural assumption because it is strong enough to yield real global theorems, but not so strong that the subject collapses into a few toy examples.

1. Convexity is strong enough to make local information globally meaningful. Without additional structure, local information is usually only local. A derivative, an affine approximation, or a second-order expansion near one point says very little about what happens elsewhere. Convexity is the first major structural assumption in the course under which this changes in a robust way.

Gradients, subgradients, supporting hyperplanes, and dual certificates stop being merely local descriptions and start becoming global lower certificates.

2. Convexity is broad and stable enough to support a systematic theory. A useful structural assumption should not apply only to a tiny family of specially engineered problems. Convexity still contains linear programs, least squares, logistic regression, constrained quadratic programs, second-order cone programs, semidefinite programs, and many regularized learning models. It is also stable under the operations optimization repeatedly uses, such as nonnegative linear combinations, affine changes of variables, epigraph constructions, partial minimization, and conjugation. Because of that, convex optimization can be developed as a real theory rather than a disconnected collection of tricks.

3. Convex optimization is the right baseline testbed for algorithms and complexity. Even when the eventual application is nonconvex, convex optimization is the cleanest setting in which one can first isolate the role of local primitives, prove global guarantees, and identify genuine complexity barriers. If a method cannot even be explained or stabilized on convex problems, then its behavior on more complicated problems is harder to interpret, not easier.

4. Convex optimization is also a source of ideas that survive beyond the convex setting. Its value is not limited to problems that are themselves convex. Many methods and viewpoints that later matter more broadly were first discovered, justified, or conceptually clarified in convex and online convex optimization. A concrete example is modern LLM training: the objective is highly nonconvex, yet the default optimizer AdamW belongs to a line of adaptive first-order methods whose ancestry runs through AdaGrad, a method that emerged from theoretical work in online convex optimization [DHS11].

1.2 First consequences of convexity

Theorem 1.2 shows that, for convex functions, local optimality is already global. Lemma 1.3 is best read as a companion first-order necessary condition for minimizers on convex feasible sets. Lemma 1.4 gives the complementary global linear lower bound supplied by differentiability and convexity. Theorem 1.5 then shows that, once convexity is added, this same first-order sign condition is not only necessary but also sufficient for global optimality.

Theorem 1.2 (Local minima are global for convex functions)**.** *Let $\Omega \subseteq E$ be nonempty and convex, let $f : \Omega \to \mathbb{R}$ be convex, and let $x^\star \in \Omega$. If there exists $r > 0$ such that*

$$\forall x \in \Omega \cap \{y \in E : \|y - x^\star\| < r\}, \qquad f(x^\star) \le f(x),$$

then

$$\forall x \in \Omega, \qquad f(x^\star) \le f(x).$$

Lemma 1.3 (First-order necessary condition on a convex feasible set)**.** *Let $\Omega \subseteq E$ be nonempty and convex, let $f : E \to \mathbb{R}$ be differentiable, and let $x^\star \in \Omega$ be a global minimizer of f over Ω. Then*

$$\forall x \in \Omega, \qquad \langle \nabla f(x^\star), x - x^\star \rangle \ge 0.$$

Lemma 1.4 (Differentiable convex functions admit a global linear lower bound)**.** *Let $\Omega \subseteq E$ be nonempty and convex, let $f : E \to \mathbb{R}$ be differentiable, and assume that f is convex on Ω. Then*

$$\forall x, y \in \Omega, \qquad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle .$$

Theorem 1.5 (Differentiable convex first-order characterization on a convex feasible set)**.** *Let $\Omega \subseteq E$ be nonempty and convex, let $f : E \to \mathbb{R}$ be differentiable, assume that f is convex on Ω, and let $x^\star \in \Omega$. Then the following are equivalent:*

1. *$f(x^\star) \leq f(x)$ for every $x \in \Omega$;*
2. *$\langle \nabla f(x^\star), x - x^\star \rangle \geq 0$ for every $x \in \Omega$.*

1.3 Examples of convex functions

Example 1.1 (Least squares)**.** In the Euclidean model $E = \mathbb{R}^d$, given data $(a_i, b_i)_{i=1}^m$ with $a_i \in \mathbb{R}^d$, consider

$$\min_{x \in \mathbb{R}^d} \frac{1}{2m} \sum_{i=1}^m (a_i^\top x - b_i)^2.$$

This problem is explicit and convex. When $A^\top A$ is invertible, it has a closed-form normal-equation solution. Writing $A = (a_1, \ldots, a_m)^\top \in \mathbb{R}^{m \times d}$ and $b = (b_1, \ldots, b_m)^\top$, the minimizer is

$$\hat{x} = (A^\top A)^{-1} A^\top b,$$

which requires $O(d^3)$ time and $O(d^2)$ space to form and invert $A^\top A$. When d is large this is infeasible; when m is large, even forming $A^\top A$ requires a full pass over all data. In the large-data regime, iterative algorithms such as gradient descent or SGD are needed instead. Even a problem with a recognizable formula is therefore not automatically computationally trivial.

Example 1.2 (Logistic regression)**.** Given labels $y_i \in \{\pm 1\}$, consider

$$\min_{x \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^m \log\!\left(1 + e^{-y_i a_i^\top x}\right) + \frac{\lambda}{2} \|x\|_2^2 .$$

This problem is again explicit and convex, but usually has no closed-form minimizer. Its importance is computational rather than symbolic: value and gradient are both natural to evaluate. Convexity here does not mean closed form; it means local information can become globally meaningful.

Example 1.3 (A constrained convex quadratic program)**.** In the Euclidean model $E = \mathbb{R}^n$, let $Q \succeq 0$, $b \in \mathbb{R}^n$, and let

$$\Omega := \{x \in \mathbb{R}^n : Cx = d,\ x \geq 0\} .$$

Then

$$\min\left\{\frac{1}{2}x^\top Qx + b^\top x : x \in \Omega\right\}$$

is a convex optimization problem with both objective geometry and explicit constraints. It previews later themes all at once: constrained optimality, dual variables, and KKT conditions.

Dependency and proof sketch

1. Theorem 1.2 uses only the convexity inequality on the segment joining $x^\star$ to an arbitrary $x \in \Omega$. If $x^\star$ were only locally optimal and some distant x were strictly better, then the convex combination near $x^\star$ would already contradict local optimality.

2. Lemma 1.3 is proved by differentiating the function

   $$\phi(t) := f(x^\star + t(x - x^\star))$$

   at $t = 0^+$, using global minimality of $x^\star$ over a convex feasible set.

3. Lemma 1.4 is proved by applying convexity to

   $$f(x + t(y - x)) \le (1 - t)f(x) + tf(y)$$

   for $t \in (0, 1]$, rearranging, and letting $t \downarrow 0$.

4. Theorem 1.5 is the combination of Lemmas 1.3 and 1.4,

   $$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle .$$

Proofs

Proof of Theorem 1.1. Because Ω is compact and f is continuous on Ω, the image set $f(\Omega) \subseteq \mathbb{R}$ is compact. In particular, $f(\Omega)$ is nonempty and closed and bounded below, so it contains its infimum. Choose $m \in f(\Omega)$ such that

$$m = \inf f(\Omega).$$

Then there exists $x^\star \in \Omega$ with $f(x^\star) = m$. For every $x \in \Omega$ one has $f(x) \in f(\Omega)$ and therefore $m \le f(x)$. Hence

$$f(x^\star) = m = \min_{x\in\Omega} f(x).$$

□

Proof of Theorem 1.2. Assume for contradiction that there exists $x \in \Omega$ such that

$$f(x) < f(x^\star).$$

Because Ω is convex, for every $\theta \in (0, 1)$ the point

$$x_\theta := \theta x + (1 - \theta)x^\star$$

belongs to Ω. By convexity of f,

$$f(x_\theta) \le \theta f(x) + (1 - \theta)f(x^\star) < f(x^\star).$$

Moreover,

$$\|x_\theta - x^\star\| = \theta \|x - x^\star\|.$$

Choosing $\theta > 0$ sufficiently small makes $\|x_\theta - x^\star\| < r$. Then $x_\theta \in \Omega \cap \{y \in E : \|y - x^\star\| < r\}$ and

$$f(x_\theta) < f(x^\star),$$

contradicting the assumed local minimality of $x^\star$. Therefore no such x exists, and so

$$\forall x \in \Omega, \qquad f(x^\star) \le f(x).$$

□

Proof of Lemma 1.3. Fix any $x \in \Omega$ and define

$$\phi(t) := f\big(x^\star + t(x - x^\star)\big), \qquad t \in [0, 1].$$

Because Ω is convex, one has $x^\star + t(x - x^\star) \in \Omega$ for every $t \in [0, 1]$. Since $x^\star$ is a global minimizer of f over Ω,

$$\phi(t) \ge \phi(0) \qquad \forall t \in [0, 1].$$

Hence the right derivative of ϕ at 0 is nonnegative:

$$\phi'_+(0) \ge 0.$$

Because f is differentiable,

$$\phi'_+(0) = \langle \nabla f(x^\star), x - x^\star \rangle.$$

Therefore

$$\langle \nabla f(x^\star), x - x^\star \rangle \ge 0.$$

Since $x \in \Omega$ was arbitrary, the claim follows. □

Proof of Lemma 1.4. Fix $x, y \in \Omega$. For every $t \in (0, 1]$, convexity of f on Ω gives

$$f\big(x + t(y - x)\big) \le (1 - t) f(x) + t f(y).$$

Rearranging, we obtain

$$\frac{f\big(x + t(y - x)\big) - f(x)}{t} \le f(y) - f(x).$$

Because f is differentiable at x, letting $t \downarrow 0$ yields

$$\langle \nabla f(x), y - x \rangle \le f(y) - f(x).$$

Equivalently,

$$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle.$$

Since $x, y \in \Omega$ were arbitrary, the claim follows. □

Proof of Theorem 1.5. We first prove that item (1) implies item (2). If $f(x^\star) \le f(x)$ for every $x \in \Omega$, then $x^\star$ is a global minimizer of f on Ω. Applying Lemma 1.3 yields

$$\forall x \in \Omega, \qquad \langle \nabla f(x^\star), x - x^\star \rangle \ge 0.$$

We next prove that item (2) implies item (1). Fix any $x \in \Omega$. Applying Lemma 1.4 with $x^\star$ and x, we obtain

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

By item (2),

$$\langle \nabla f(x^\star), x - x^\star \rangle \geq 0.$$

Hence

$$f(x) \geq f(x^\star).$$

Because $x \in \Omega$ was arbitrary, item (1) follows. □

Exercises

1. Give an example of a nonconvex differentiable function on $\mathbb{R}^2$ that has a strict local minimizer that is not global. Then explain precisely where Theorem 1.2 fails.

2. Prove that if $C \subseteq \mathbb{R}^n$ is convex and $f : C \to \mathbb{R}$ is strictly convex, then f has at most one global minimizer on C.

3. Let $f(x) = \max\{x_1, x_2, 0\}$ on $\mathbb{R}^2$. Determine all global minimizers and explain why Theorem 1.5 does not apply.

4. A set $C \subseteq \mathbb{R}^n$ is *midpoint-convex* if

$$\forall x, y \in C, \qquad \frac{x+y}{2} \in C.$$

   Prove that every convex set is midpoint-convex. Then prove that every *closed* midpoint-convex set is convex. Finally, give a counterexample showing that midpoint-convexity alone does not imply convexity.

References

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.