Optimization for Data Sciences

Lecture 1

Introduction and Convex Analysis

Samuel Vaiter

---

1 Introduction

Many problems arising in mathematical modeling and machine learning can be casted as an *optimization problem*, *i.e.,*

$$\inf f(x) \quad \text{subject to} \quad x \in \Omega,$$

where $f : \mathcal{X} \to \mathbb{V}$ is a function with adequate "regularity" defined on a space $\mathcal{X}$, $\Omega \subseteq \mathcal{X}$ the *domain* of the problem and $\mathbb{V}$ is a space where the notion of infimum is well-defined, such as $\mathbb{V} = \mathbb{R}$. Depending on the context, we may be interested in proving the existence of a *minimizer*, i.e., a point $x^\star \in \Omega$ that achieves this infimum. In this case, we say that $f(x^\star)$ is the minimum of the function f over Ω.

1.1 Examples of optimization problem

This previous definition is quite vague. It requires some precisions, and to do so, we instantiate it on several important examples.

1.1.1 Toy examples

It is common in optimization to test algorithms on function from $\mathbb{R}^2$ to $\mathbb{R}$ for the single reason that it is easy to visualize their behavior using contour plots, and that their geometry are rich enough in comparaison to the univariate case. Figure 1.1 illustrates two functions:

1. The Rosenbrock function, also known as the "banana function", is a function commonly used in optimization. It was first introduced by Rosenbrock (1960) as a way to test the performance of optimization algorithms.

   $$f(x, y) = (1 - x)^2 + a(y - x^2)^2.$$

   A typical value is $a = 100$. Note that for $a > 0$, the global minimizer $(1, 1)$ is unique. If $a = 0$, every $(1, y)$ is a minimizer; if $a < 0$, the function is unbounded below.

2. A (positive) quadratic form, defined as

   $$f(x, y) = (x - 1)^2 + (y - 1)^2.$$

1.1.2 Ordinary Least Squares

One of the fundamental model in statistics is the *ordinary least squares* (OLS) problem. Given a matrix $A \in \mathbb{R}^{n \times d}$ and a vector $b \in \mathbb{R}^n$, the OLS problem seeks a vector $x \in \mathbb{R}^d$ that minimizes the sum of squared residues:

$$\min_{x \in \mathbb{R}^d} \frac{1}{2} \|Ax - b\|^2,$$

where $\|\cdot\|$ denotes the Euclidean norm. This formulation arises naturally in linear regression, where A represents the data matrix (with each row corresponding to an observation and each

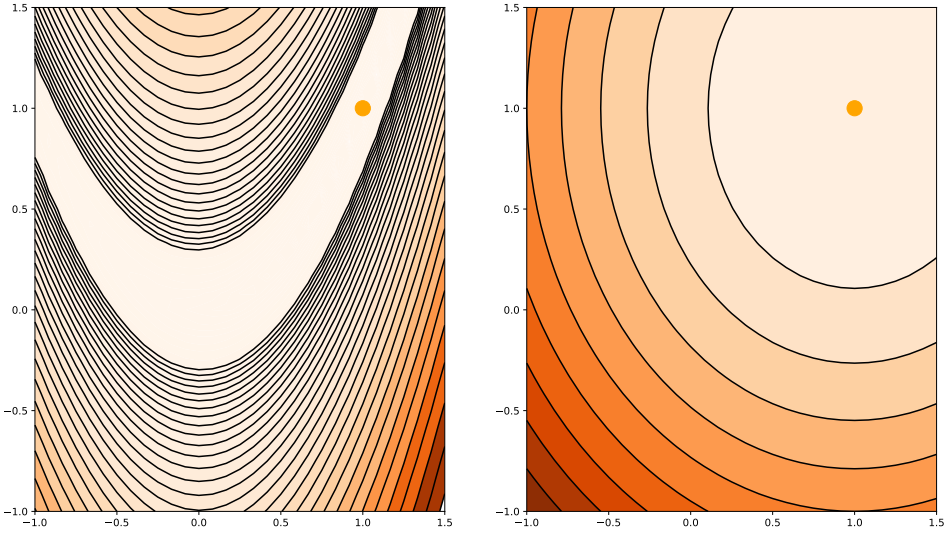


Figure 1.1: Contour plots of functions from $\mathbb{R}^2$ to $\mathbb{R}$. (Left) Rosenbrock function. (Right) Quadratic form.

column to a feature), and b is the vector of observed responses, typically modeled as $b = Ax_0 + \varepsilon$, where x_0 is unknown and ε represents noise or model error. The identity $b = Ax_0$ is the noiseless special case.

The solution to the OLS problem, *when* A has full column rank, is given by the unique minimizer

$$x^\star = (A^\top A)^{-1} A^\top b.$$

If A does not have full rank, the set of minimizers is generally infinite, and the minimum-norm solution can be found using the Moore-Penrose pseudoinverse:

$$x^\star = A^\dagger b.$$

1.1.3 (Empirical) Risk minimization

Given a function $\phi : X \to Y$, a loss function $\ell : Y \times Y \to \mathbb{R}$ and a probability measure $\mathbb{P}$ on $X \times Y$, the *expected risk* of the prediction ϕ is defined as

$$\mathcal{R}(\phi) = \mathbb{E}_{(x,y)\sim\mathbb{P}}[\ell(y, \phi(x))].$$

Minimizing expected risk over a chosen class $\mathcal{F}$ leads to the optimization problem

$$\inf_{\phi \in \mathcal{F}} \mathcal{R}(\phi),$$

Here $\mathcal{F}$ is a class of measurable predictors, and the loss is assumed measurable with well-defined expected risks, for example a nonnegative measurable loss. A minimizer, if it exists, is optimal within $\mathcal{F}$. A Bayes predictor minimizes expected risk over all admissible measurable predictors; a restricted class need not contain one. Under the assumptions stated here, the infimum need not be attained.

In practice, in supervised learning, we don't have the knowledge of the distribution $\mathbb{P}$ but only to a training data $(x_i, y_i) \in X \times Y$. The *empirical risk* is defined as

$$\hat{\mathcal{R}}(\phi) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \phi(x_i)).$$

Of particular interest is the structural risk minimization – a term that trace back to Vapnik and Chervonenkis (1974) – where in additional to the loss function, we consider a regularization

function $J : \mathcal{F} \to \mathbb{R}$ that represents an *apriori*, and an hyperparameter $\lambda > 0$ and we seek to solve the minimization problem

$$\inf_{\phi \in \mathcal{F}} \hat{\mathcal{R}}(\phi) + \lambda J(\phi).$$

The topic of supervised learning with an emphasis on optimization is well reviewed in the textbook of Bach (2021) and further theoretical insights are developped for instance in (Shalev-Shwartz and Ben-David, 2014; Mohri, Rostamizadeh, and Talwalkar, 2012).

1.1.4 Portfolio optimization

Consider n economic assets whose returns are denoted $(a_1, \dots, a_n)$. We assume to know the expected returns generated $\mu = (\mu_1, \dots, \mu_n)$ and the risk in/between the assets defined by the covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$ (assumed to be positive definite). The problem of portfolio optimization with minimal variance $\mathrm{Var}(\langle x, a \rangle)$ and expected return r is to find a

$$\inf_{x \in \mathbb{R}^n} f(x) \overset{\text{def.}}{=} \langle \Sigma x, x \rangle \quad \text{subject to} \quad \begin{cases} \mathbb{E}(\langle x, a \rangle) = \langle x, \mu \rangle = r \\ \langle x, \mathbf{1} \rangle = 1. \end{cases}$$

The foundational paper on modern portfolio theory was written by Markowitz (1952).

1.1.5 Optimal control

Given a dynamical system drive by the ordinary differential equation

$$\dot{x}(t) = g(t, x(t), u(t)), \quad x(t_0) = x_0, \tag{1.1}$$

where

- I is an interval of $\mathbb{R}$, $U \subseteq \mathbb{R}^n$ open, $V \subseteq \mathbb{R}^m$ open;
- $(t_0, x_0) \in I \times V$ are the initial conditions;
- $g : I \times V \times U \to \mathbb{R}^m$ is C^1.

For the remainder of this example, take $t_0 = 0$ and let I be an open interval containing 0. Let $\mathcal{U}$ consist of measurable controls $u : I \to U \subseteq \mathbb{R}^n$ such that, on every compact interval $J \subset I$, the values of u lie almost everywhere in some compact set $K_J \subset U$. In particular, these controls belong to $L^\infty_{\mathrm{loc}}(I, \mathbb{R}^n)$. The compact-range condition and $g \in C^1$ give local integrable bounds and local Lipschitz bounds in the state variable. Thus each $u \in \mathcal{U}$ has a unique maximal forward, locally absolutely continuous solution $x_u : [0, t_e(u)) \to V$, satisfying the differential equation almost everywhere and the initial condition at 0.

For $T > 0$ with $[0, T] \subset I$, let $\mathcal{U}_T$ be the controls in $\mathcal{U}$ whose solution exists and remains in V on $[0, T]$. Mere local boundedness in the ambient space, without suitable bounds inside U, does not suffice for the preceding existence claim.

Given a cost function $c : I \times V \times U \to \mathbb{R}$, assumed continuous so the cost integral is finite for each $u \in \mathcal{U}_T$, the *fixed-time optimal control problem* is defined as

$$\inf_{u \in \mathcal{U}} f(u) \overset{\text{def.}}{=} \int_0^T c(t, x_u(t), u(t)) dt \quad \text{subject to} \quad u \in \mathcal{U}_T.$$

Besides its natural applications in engineering (including aerospatial), optimal control is at the heart of reinforcement learning. For a comprehensive treatment of optimal control, see the lecture notes of Evans (1983) and for application to reinforcement learning, see (Recht, 2019).

1.1.6 Optimal transport

Let X, Y two metric spaces, μ a Borel measure on X and $T : X \to Y$ a measurable map. We recall that the *push-forward* of μ by T is the measure $T_\sharp\mu$ defined by the change-of-variable formula

$$\int_Y \phi(y) dT_\sharp\mu(y) = \int_X \phi(T(x)) d\mu(x),$$

for all measurable and bounded $\phi : Y \to R$.

For a given probability measures μ and ν, and a cost function $c : X \times Y \to \overline{\mathbb{R}}$, the *Monge problem* of optimal transportation is defined as

$$\inf_{T:X\to Y} f(T) = \int_X c(x, T(x)) d\mu(x) \quad \text{subject to} \quad T_\sharp\mu = \nu. \tag{1.2}$$

The optimization is over measurable maps and may be infinite-dimensional, but need not be when the spaces are finite. For maps into a Euclidean space, the push-forward constraint is generally nonconvex, whereas convexity of the objective depends on the cost. For arbitrary metric spaces, convexity of maps requires additional linear structure; these claims are not universal. For more information, see (Villani, 2009) for theoretical and (Peyré and Cuturi, 2018) for computational applications (including machine learning).

1.2 Basic facts on optimization

Suppose we have an constrained minimization problem on $\Omega \subseteq \mathbb{R}^d$ as follows

$$\inf_{x\in\Omega} f(x).$$

Such problem may *not* have a solution. For instance, if $\Omega = \mathbb{R}$, then $f(x) = x$ is not lower-bounded, and the infimum is $-\infty$. But a function may be lower-bounded, for instance $f(x) = 1/x$ on $\Omega = \mathbb{R}_{>0}$, but the infimum is never achieved.

The following theorem shows that if f is continuous, and that we consider a non-empty *compact* – that is bounded and closed – subset of $\mathbb{R}^d$, then the infimum is a minimum.

> **Theorem 1.1: C^0 + compact ⇒ global minimizer**
>
> If f is continuous and Ω is non-empty compact, then there exists at least one global minima of f on Ω.

Proof. We recall that the Bolzano–Weierstrass theorem states that any bounded sequences of $\mathbb{R}^d$ contains a convergent subsequence. Let $f^* = \inf_{x\in\Omega} f(x)$. Consider a minimizing sequence $(x^{(t)})_{t\geq 0}$ of elements of Ω such that $f(x^{(t)})$ converges towards $f^\star$. Using the Bolzano–Weierstrass theorem, there exists a subsequence $(x^{(\varphi(t))})_{t\geq 0}$ converging to some $x^* \in \Omega$. By continuity of f, we have $f(x^{(\varphi(t))})$ converges to $f(x^*)$. Hence, $f(x^*) = f^*$, *i.e.,* x^* is a global minimum of f on Ω. □

Unfortunately, this argument cannot work anymore for *unconstrained* minimization when $\Omega = \mathbb{R}^d$. The trick is to replace the compactness assumption on Ω by the (0-)coercivity of f, *i.e.,*

$$\lim_{\|x\|\to+\infty} f(x) = +\infty.$$

In this case, we have the following result

> **Theorem 1.2: C^0 + coercive ⇒ global minimizer**
>
> If $f : \mathbb{R}^d \to \mathbb{R}$ is continuous and coercive, then there exists at least one global minima of f on $\mathbb{R}^d$.

Proof. Since f is coercive, there exists $M > 0$ such that

$$\forall x \in \mathbb{R}^d, \quad (\|x\| > M) \implies f(x) > f(0).$$

Consider the closed Euclidean ball $\overline{B}(0, M)$. Using theorem 1.1, there exists $x^\star$ a global minima of f on $\overline{B}(0, M)$:

$$\forall y \in \mathbb{R}^d, (\|y\| \leq M) \implies f(y) \geq f(x^\star).$$

In particular, $f(0) \geq f(x^\star)$, and in turns

$$\forall x \in \mathbb{R}^d, f(x) \geq f(x^\star),$$

that is $x^\star$ is a global minima of f on the whole space. □

It is possible to generalize this result by asking merely that f is lower semicontinuous. These two results (theorems 1.1 and 1.2) can be generalized in (infinite dimensional) Hilbert spaces with some additional hypotheses. See the references for more insights in this setting.

Let us finish this section by introducing our usual suspect all through these lecture notes.

> **Example 1.1: Least-square minimization**
>
> Let $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$. The quadratic loss is defined as $f(x) = \frac{1}{2}\|Ax - b\|^2$. It is always lower-bounded by 0 as the norm is a nonnegative function. However, coercivity is verified if, and only if, its nullspace $\mathrm{Ker}(A) = \{0\}$ is reduced to the trivial vector space. To see this, suppose that $\mathrm{Ker}\, A \neq \{0\}$ and consider $x^\star$ a minimum of f. Then, for all $z \in \mathrm{Ker}\, A$, we have $f(x^\star + z) = f(x^\star)$, hence the set of minimizers is unbounded. Reciprocally, suppose that $\mathrm{Ker}\, A = \{0\}$. In particular, A is injective, thus there exists some $c > 0$ such that $\|Ax\| \geq c\|x\|$.
> The reverse triangle inequality gives
>
> $$\|Ax - b\| \geq \|Ax\| - \|b\| \geq c\|x\| - \|b\|.$$
>
> Therefore $\|Ax - b\| \to +\infty$, and hence $f(x) \to +\infty$, as $\|x\| \to +\infty$. We will see later that there is in fact only one minimizer in this case.

2 Basic facts in convex analysis

Definitions.

1. There are in a plane certain terminated bent lines (*καμπύλαι γραμμαὶ πεπερασμέναι*)†, which either lie wholly on the same side of the straight lines joining their extremities, or have no part of them on the other side.

2. I apply the term **concave in the same direction** to a line such that, if any two points on it are taken, either all the straight lines connecting the points fall on the same side of the line, or some fall on one and the same side while others fall on the line itself, but none on the other side.

Figure 2.1: Definition of concavity by Archimedes

Despite the fact that many problems encountered in statistics and machine learning are not convex, convexity is an important to study in order to understand what is the generic behaviour of an optimization procedure.

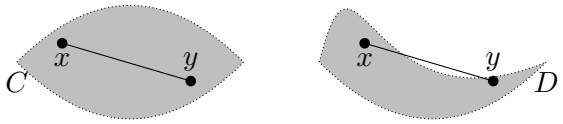


Figure 2.2: Convex and non-convex sets. The set C is convex whereas D is nonconvex.

2.1 Convex set

A set $C \subseteq \mathcal{X}$ is convex if every segment of C is contained in C. It can formalized as the following definition and illustrated in fig. 2.2.

Definition 2.1: Convex set

A set $C \subseteq \mathcal{X}$ is **convex** if

$$\forall (x, y) \in C^2, \forall \lambda \in [0, 1], \quad \lambda x + (1 - \lambda) y \in C.$$

One can observe that it is possible to replace $\lambda \in [0, 1]$ by $\lambda \in (0, 1)$ without any difference (see Exercise 1). In most of these lecture notes, the set $\mathcal{X}$ will the Euclidean space $\mathbb{R}^p$, but remark that this notion only requires that $\mathcal{X}$ is a vector space over the real numbers $\mathbb{R}$, possibly of infinite dimension.

Example 2.1: Basic convex sets

We state several basic examples of convex sets.

1. The convex sets of $\mathbb{R}$ are *exactly* the intervals of $\mathbb{R}$.
2. Any affine hyperplane $H = \{x \in \mathcal{X} \ : \ \varphi(x) = \alpha\}$ where $\mathcal{X}$ is a real vector space, φ a linear functional and $\alpha \in \mathbb{R}$ is convex.
3. Any half-space $H_- = \{x \in \mathcal{X} \ : \ \varphi(x) \leq \alpha\}$ is also convex.
4. The unit simplex of $\mathbb{R}^d$ defined as

$$\Delta^{d-1} = \left\{ x \in \mathbb{R}^d \ : \ x_i \geq 0 \text{ and } \sum_i x_i = 1 \right\}$$

   is convex.
5. The nonnegative orthant $\mathbb{R}^d_{\geq 0} = \left\{ x \in \mathbb{R}^d \ : \ x_i \geq 0 \right\}$ is convex.

The following proposition reviews some important algebraic properties of convex sets.

Proposition 2.1: Operations on convex sets

1. *Arbitrary intersection.* Let $(C_i)_{i \in I}$ convex sets where I is a set. Then $\cap_{i \in I} C_i$ is convex.
2. *Cartesian product.* Let $C_1, \ldots, C_n$ be sets of $\mathbb{R}^{n_1}, \ldots, \mathbb{R}^{n_n}$. If every C_i is convex, their Cartesian product is convex. Conversely, if the product is convex and every factor is nonempty, then every C_i is convex. Without nonemptiness the converse fails, since a product with an empty factor is empty and therefore convex.

3. *Sum.* Let C_1, C_2 two convex sets. Then the (Minkowski) sum $C_1 + C_2$ is convex.
4. *Affine map.* Let $A : \mathbb{R}^n \to \mathbb{R}^m$ be an affine map and $C \subseteq \mathbb{R}^n$ a convex set. Then the image $A(C)$ is convex.

2.2 Convex functions

2.2.1 Definition

The intuitive definition of a convex function is a function such that "the line joining $f(x)$ and $f(y)$ lies above the graph of f between x and y". Formally, this definition makes sense using the definition of the epigraph of f.

Definition 2.2: Epigraph

Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $f \neq +\infty$. The **epigraph** of f is the subset of $\mathbb{R}^n \times \mathbb{R}$ defined by

$$\operatorname{epi} f \stackrel{\text{def.}}{=} \{(x, t) \in \mathbb{R}^n \times \mathbb{R} \ : \ t \geq f(x)\}.$$

(a) Nonconvex function. (b) Convex function.

Figure 2.3: Epigraph of a function.

Equipped with this notion, we can define convex functions using the definition of convex sets (definition 2.1) applied to the epigraph (definition 2.2).

Definition 2.3: Convex function

A function $f : \mathbb{R}^p \to \overline{\mathbb{R}}$ such that $f \neq +\infty$ is **convex** iff its epigraph $\operatorname{epi} f$ is convex.

It is possible to make explicit this definition as follow:

$$\forall x, y \in \operatorname{dom} f, \forall \lambda \in [0, 1], \qquad f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y). \tag{2.1}$$

Following the convention introduced in (Hiriart-Urruty and Lemarechal, 1996), we will denote the set of all convex functions of $\mathbb{R}^p$ as $\operatorname{Conv} \mathbb{R}^p$, and the set of closed – that is, with closed epigraph – convex functions of $\mathbb{R}^p$ as $\overline{\operatorname{Conv}}\mathbb{R}^p$.

2.2.2 Local and global minima of a convex function

For a (un)constrained minimization problem on a convex set $C \subseteq \operatorname{dom} f$ with a function $f : C \to \mathbb{R}$,

$$\min_{x \in C} f(x), \tag{2.2}$$

we say that $x^\star$ is

- a local solution of (2.2), if there exists a neighborhood $\mathcal{O}$ of $x^\star$ such that for all $x \in C \cap \mathcal{O}$, $f(x) \geq f(x^\star)$;

- a (global) solution if for all $x \in C$, $f(x) \geq f(x^\star)$.

A fundamental result is that every local solution of (2.2) is global solution, and the set of (global) solutions is a convex set.

Theorem 2.1: Local minima of convex functions are global

Assume that f is a convex function and C is a convex **closed** set. Then,

1. Any local solution of (2.2) is a global solution;
2. The set of global solutions of (2.2) is convex.

Proof. 1. By contradiction. Consider a local solution $x^\star$ and assume that $x^\star$ is not a global solution. In particular, there exists $z \in C$ such that $f(z) < f(x^\star)$. By convexity,

$$f((1-t)x^\star + tz) \leq (1-t)f(x^\star) + tf(z) < f(x^\star) \quad \forall t \in (0,1]$$

Let $\mathcal{O}$ be a neighborhood witnessing local minimality. Since $(1-t)x^\star + tz \to x^\star$ as $t \downarrow 0$, this point belongs to $C \cap \mathcal{O}$ for sufficiently small $t > 0$. The strict inequality above then contradicts local minimality.

2. Take two global solutions, and study the segment joining it: let $x^\star$ and z two solutions. We have $f(x^\star) = f(z)$, and in turn,

$$f(x^\star) \leq f((1-t)x^\star + tz) \leq (1-t)f(x^\star) + tf(z) = f(x^\star),$$

where the first inequality comes from global optimality of $x^\star$ and feasibility of $(1-t)x^\star + tz$ for $t \in [0,1]$, and the second by convexity. □

2.2.3 Differentiable properties of a convex function

A differentiable function is convex if, and only if, it is above all its tangents, see fig. 2.4.

Proposition 2.2

Let f be a differentiable function on an open set $\Omega \subseteq \mathbb{R}^p$, and $C \subseteq \Omega$ a convex subset of Ω. Then f is convex on C if, and only if,

$$\forall x, \bar{x} \in C, \quad f(x) \geq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle. \tag{2.3}$$

Proof. Assume f is convex on C. Let $x, \bar{x} \in C$ and $\lambda \in (0,1)$. By definition 2.3, we have

$$f(\lambda x + (1-\lambda)\bar{x}) - f(\bar{x}) \leq \lambda(f(x) - f(\bar{x})).$$

Dividing by λ, and letting λ goes to 0, we obtain (2.3).

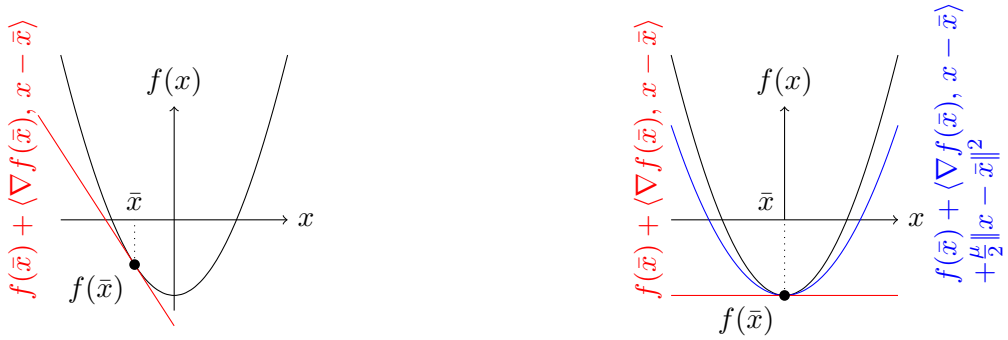
Reciprocally, let $x, y \in C$, $\alpha \in (0,1)$ and define $z = \alpha x + (1-\alpha)y \in C$. Applying (2.3) twice, we have

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

Using a convex combination of these two lines, we get

$$\alpha f(x) + (1-\alpha) f(y) \geq f(z) + \langle \nabla f(z), \alpha x + (1-\alpha)y - z \rangle,$$

which is exactly (2.1). □



(a) Convexity and differentiability. The curve $y = f(x)$ is always above its tangent in red.

(b) Strong convexity and differentiability. There is space for a quadratic form between the curve $y = f(x)$ and the tangent.

Figure 2.4: Convex and differentiable functions. Here $f(x) = x^2 - \cos(x)$.

A similar criterion can be derived for strictly convex function[1].

For an unconstrained local minimizer in an open domain, $\nabla f(x^\star) = 0$, and (2.3) directly gives global optimality. For minimization over a convex set C, local minimality instead implies

$$\langle \nabla f(x^\star),\, x - x^\star \rangle \geq 0 \qquad \forall x \in C,$$

by taking feasible one-sided derivatives along segments. Together with (2.3), this also proves global optimality. The gradient need not vanish at a constrained minimizer: for $f(x) = x$ on $[0, 1]$, the minimizer is 0 and $f'(0) = 1$.

2.3 Strongly convex functions

Throughout this section, $\|\cdot\|$ denotes the Euclidean norm. In particular, the equivalence with subtraction of $\frac{\mu}{2}\|x\|^2$ and the Hessian criterion below use this norm.

A strongly convex function satisfies a stronger statement of the Jensen's inequality.

Definition 2.4: Strongly convex function

A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *strongly convex* of modulus $\mu > 0$ if $\operatorname{dom} f$ is convex and

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) - \frac{1}{2}\mu\lambda(1-\lambda)\|x - y\|^2,$$

for all $x, y \in \operatorname{dom} f$ and $\lambda \in [0, 1]$.

Another way to define strongly convex function is to say that f is μ-strongly convex if the function

$$x \mapsto f(x) - \frac{\mu}{2}\|x\|^2$$

is a convex function.

A finite-valued strongly convex function $f : \mathbb{R}^d \to \mathbb{R}$ has a unique minimizer. More generally, the conclusion holds for a proper (never $-\infty$ and finite at some point) lower semicontinuous strongly convex function. Strong convexity alone does not imply attainment for an arbitrary extended-valued function; for example, $f(x) = x^2$ on $x > 0$ and $f(x) = +\infty$ otherwise has infimum 0 but no minimizer.

For the finite-valued full-domain case, set $h(x) = f(x) - \frac{\mu}{2}\|x\|^2$. This is finite-valued and convex, so it is continuous and has a supporting affine function at 0: for some $p \in \mathbb{R}^d$, $h(x) \geq$

[1] prove it! See exercise 3

$h(0) + \langle p, x\rangle$. Consequently,

$$f(x) \geq f(0) + \langle p, x\rangle + \frac{\mu}{2}\|x\|^2 \longrightarrow +\infty \quad \text{as } \|x\| \longrightarrow +\infty.$$

Thus f is continuous and coercive, and theorem 1.2 ensures existence. Strong convexity implies strict convexity on the effective domain, so two distinct minimizers are impossible. Strict convexity alone gives at most one minimizer, rather than existence.

For a differentiable function on a convex domain, strong convexity is equivalent to a uniform quadratic lower bound above every tangent, with the same modulus $\mu > 0$ for all pairs of points, as in the following proposition. At the point of tangency the bound is an equality. This is a key property to defined lower bounds on F.

Proposition 2.3

Let f be a differentiable and μ-strongly convex function. Then,

$$\forall x, \bar{x} \in \operatorname{dom} f, \quad f(x) \geq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle + \frac{\mu}{2}\|x - \bar{x}\|^2.$$

Note that if f is C^2, f is μ-strongly convex if, and only if, its hessian $\nabla^2 f \geq \mu \operatorname{Id}$.

An easy consequence is that it is possible to establish a lower bound on the values based on a lower bound on the distance between points:

$$f(x) \geq f(x^\star) + \frac{\mu}{2}\|x - x^\star\|^2, \tag{2.4}$$

where $x^\star$ is the minimizer of f.

2.4 Lipschitz continuity of the gradient

Definition 2.5: L-smoothness

A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be L-smooth, with $L > 0$, if ∇f is L-Lipschitz, i.e.

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

Here $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

2.4.1 Quadratic upper-bound

L-smoothness is important in optimization because it ensures a quadratic upper bound of the function.

Proposition 2.4

Let f a L-smooth function. Then, for all $x, y \in \operatorname{dom} f$,

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \leq L\|x - y\|^2. \tag{2.5}$$

Moreover, if $\operatorname{dom} f$ is convex, then (2.5) is equivalent to

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|x - y\|^2, \quad \forall x, y \in \operatorname{dom} f. \tag{2.6}$$

Proof. The fact that L-smoothness implies (2.5) is a direct application of the (generalized) Cauchy-Schwarz inequality.

For the equivalence of (2.5) and (2.6), let's prove the two direction. Given $x, y \in \operatorname{dom} f$, define the real function $g(t) = f(x + t(y - x))$.

Assuming (2.5) holds, then

$$g'(t) - g'(0) = \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle \leq tL\|x - y\|^2.$$

Now,

$$\begin{aligned} f(y) = g(1) = g(0) + \int_0^1 g'(t)\, dt &\leq g(0) + g'(0) + \frac{L}{2}\|x - y\|^2 \\ &= f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2. \end{aligned}$$

Conversely, assuming (2.6), apply it to (x, y) and (y, x) and add the inequalities. The function values cancel, yielding

$$0 \leq \langle \nabla f(x) - \nabla f(y),\, y - x \rangle + L\|x - y\|^2,$$

which is (2.5). □

This result has a strong implication on the minima of f.

Proposition 2.5

Let f be a L-smooth function with full domain, and assume $x^\star$ is a minimizer of f. Then,

$$\frac{1}{2L}\|\nabla f(z)\|_*^2 \leq f(z) - f(x^\star) \leq \frac{L}{2}\|z - x^\star\|^2 \quad \forall z \in \mathbb{R}^n.$$

Proof. The two side are proved separately.
rhs: Apply (2.6) to $(x^\star, z)$, using $\nabla f(x^\star) = 0$.
lhs: Minimize the quadratic upper bound for $x = z$:

$$\begin{aligned} f(x^\star) = \min_{y \in \mathbb{R}^n} f(y) &\leq \inf_y \left(f(z) + \langle \nabla f(z), y - z \rangle + \frac{L}{2}\|y - z\|^2 \right) && \text{by (2.6)} \\ &= \inf_{\|v\|=1} \inf_t \left(f(z) + t\langle \nabla f(z), v \rangle + \frac{Lt^2}{2} \right), \end{aligned}$$

The last step is an equality: write $w = y - z = tv$ with $\|v\| = 1$ and $t \in \mathbb{R}$. Every $w \neq 0$ admits this representation, and $w = 0$ is represented by $t = 0$ and any unit vector. Allowing signed t is equivalent to allowing both v and $-v$. Factorizing the trinom in t leads to:

$$f(z) + t\langle \nabla f(z), v \rangle + \frac{Lt^2}{2} = \frac{L}{2}\left(t + \frac{1}{L}\langle \nabla f(z), v \rangle \right)^2 - \frac{1}{2L}\langle \nabla f(z), v \rangle^2 + f(z).$$

Hence,

$$\min_{y \in \mathbb{R}^n} f(y) \leq \inf_{\|v\|=1} \left(f(z) - \frac{1}{2L}\langle \nabla f(z), v \rangle^2 \right).$$

By the definition of the dual norm, the maximum of $|\langle \nabla f(z),\, v \rangle|$ over $\|v\| = 1$ is $\|\nabla f(z)\|_*$, attained on the compact unit sphere. In the Euclidean case with a nonzero gradient, one may take $v = \nabla f(z)/\|\nabla f(z)\|_2$. If the gradient is zero, any unit vector suffices. Hence

$$f(x^\star) \leq f(z) - \frac{1}{2L}\|\nabla f(z)\|_*^2,$$

which proves the left-hand inequality. □

2.4.2 Co-coercivity of the gradient of a L-smooth function

Convex L-smooth functions are co-coercive. This result is sometimes coined Baillon–Haddad theorem (Baillon and Haddad, 1977), but one shall note that the original contributionis much more general than its application to the gradient.

Proposition 2.6

Let f be a full-domain convex L-smooth function. Then, ∇f is $1/L$-co-coercive, i.e.,

$$\forall x, y \in \mathbb{R}^n, \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_*^2.$$

Proof. Let $f_x(z) = f(z) - \langle \nabla f(x), z \rangle$ and $f_y(z) = f(z) - \langle \nabla f(y), z \rangle$. The two are convex.

Since f is L-smooth, f_x and f_y are also L-smooth.

Canceling the gradient of f_x shows that $z = x$ minimizes f_x. Hence, using the quadratic upper bound, we have

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x \rangle &= f_x(y) - f_x(x) \\ &\geq \frac{1}{2L} \|\nabla f_x(y)\|_*^2 \\ &= \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|_*^2. \end{aligned}$$

Doing exactly the same thing for f_y leads a similar bound, and combining the two gives the result. □

It turns out that smoothness, co-coercivity and upper quadratic boundness are equivalent properties for full-domain convex function.

Proposition 2.7

Let f be a differentiable convex function with $\operatorname{dom} f = \mathbb{R}^n$. Then the three following properties are equivalent:

1. f is L-smooth.
2. The quadratic upper bound (2.6) holds.
3. ∇f is $1/L$-co-coercive.

Proof. The implication $1 \Rightarrow 2$ is proposition 2.4. For $2 \Rightarrow 3$, define $f_x(z) = f(z) - \langle \nabla f(x),\, z \rangle$ and similarly f_y. Convexity ensures that x minimizes f_x and y minimizes f_y. Both shifted functions inherit the upper bound (2), and the lower-bound argument in proposition 2.5 uses only that upper bound. Applying that argument to f_x at y and f_y at x gives

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x),\, y - x \rangle &\geq \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|_*^2, \\ f(x) - f(y) - \langle \nabla f(y),\, x - y \rangle &\geq \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_*^2. \end{aligned}$$

Adding these yields (3). Finally, for $3 \Rightarrow 1$, let $d = \nabla f(x) - \nabla f(y)$. Co-coercivity and the dual-norm Cauchy–Schwarz inequality give

$$\frac{1}{L} \|d\|_*^2 \leq \langle d,\, x - y \rangle \leq \|d\|_* \|x - y\|.$$

If $d \neq 0$, divide by $\|d\|_*/L$ to obtain (1); if $d = 0$, (1) is immediate. □

For the special case of the ℓ^2 norm, we have:

Proposition 2.8

Let f be a differentiable convex function with $\operatorname{dom} f = \mathbb{R}^n$, and L-smooth with respect to the Euclidean norm. Then,

1. $L\mathrm{Id} - \nabla f$ is monotone, i.e.
$$\forall x, y \in \mathbb{R}^n, \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L\|x - y\|_2^2.$$
2. $\frac{L}{2}\|x\|_2^2 - f(x)$ is convex.
3. If f is twice differentiable, then,
$$\lambda_{\max}(\nabla^2 f(x)) \leq L, \quad \forall x \in \mathbb{R}^n.$$

When f is both L-smooth and μ-strongly convex, we have the following "strenghened" co-coercivity.

Proposition 2.9

Let $f : \mathbb{R}^n \to \mathbb{R}$ have full effective domain and be L-smooth and μ-strongly convex with respect to the Euclidean norm, where $0 < \mu \leq L$. All norms in this proposition and its proof are Euclidean. For any $x, y \in \mathbb{R}^n$, we have

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L}\|x - y\|^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|^2.$$

Proof. Consider the convex function $\phi(x) = f(x) - \frac{\mu}{2}\|x\|^2$. Its gradient reads $\nabla\phi(x) = \nabla f(x) - \mu x$. Using the L-smoothness of f, we have that

$$\begin{aligned}\langle \nabla\phi(x) - \nabla\phi(y), x - y \rangle &= \langle \nabla f(x) - \nabla f(y), x - y \rangle - \mu\langle x - y, x - y \rangle \\ &\leq (L - \mu)\|x - y\|^2.\end{aligned}$$

Since ϕ is convex on the full space, when $L > \mu$ its displayed upper inner-product bound implies the quadratic upper bound with parameter $L - \mu$ by proposition 2.4. The equivalence in proposition 2.7 then shows that ϕ is $(L - \mu)$-smooth. The upper inner-product bound alone would not imply smoothness without convexity.

If $L = \mu$, the inner-product bound for $\nabla\phi$ is both nonnegative and at most zero. Its zero quadratic upper bound and convexity give $\phi(y) = \phi(x) + \langle \nabla\phi(x), y - x \rangle$ for all x, y, so ϕ is affine. Consequently $\nabla f(x) - \nabla f(y) = \mu(x - y)$, and the claimed strengthened inequality holds with equality. Otherwise, the $(L - \mu)$-co-coercivity of ϕ (proposition 2.6) gives us

$$\langle \nabla\phi(x) - \nabla\phi(y), x - y \rangle \geq \frac{1}{L - \mu}\|\nabla\phi(x) - \nabla\phi(y)\|^2.$$

Explicing the expression of $\nabla\phi(x)$ allows us to conclude. □

2.5 Exercises

Exercise 1 Prove that C is convex if, and only if,

$$\forall (x, y) \in C^2, \forall \lambda \in (0, 1), \quad \lambda x + (1 - \lambda)y \in C.$$

Exercise 2 Show that the following sets are convex.

1. The positive orthant: $\mathbb{R}^d_{\geq 0} = \{x \in \mathbb{R}^d \ : \ x_i \geq 0, \forall i\}$.
2. An hyperplane directed by $a \in \mathbb{R}^d$: $\{x \in \mathbb{R}^d \ : \ \langle a, x \rangle = b\}$, for $b \in \mathbb{R}$.

3. An half-space directed by $a \in \mathbb{R}^d$: $\big\{x \in \mathbb{R}^d \ : \ \langle a, x\rangle \leq b\big\}$, for $b \in \mathbb{R}$.
4. The unit-ball of a norm $\|\cdot\|$ on $\mathbb{R}^d$: $\big\{x \in \mathbb{R}^d \ : \ \|x\| \leq 1\big\}$.

Exercise 3 Let f be a differentiable function on an open set $\Omega \subseteq \mathbb{R}^p$, and $C \subseteq \Omega$ a convex subset of Ω. Prove that f is strictly convex on C if, and only if,

$$\forall x, \bar{x} \in C, \quad x \neq \bar{x} \implies f(x) > f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle.$$

References

Bach, Francis (Apr. 4, 2021). *Learning Theory from First Principles*.

Baillon, Jean-Bernard and Georges Haddad (June 1, 1977). "Quelques propriétés des opérateurs angle-bornés etn-cycliquement monotones". In: *Israel J. Math.* 26.2, pp. 137–150.

Evans, Lawrence C (1983). *An Introduction to Mathematical Optimal Control Theory*. Department of Mathematics University of California, Berkeley, p. 126.

Hiriart-Urruty, Jean-Baptiste and Claude Lemarechal (Oct. 30, 1996). *Convex Analysis and Minimization Algorithms I: Fundamentals*. Corrected edition. Berlin: Springer. 418 pp.

Markowitz, Harry (1952). "Portfolio Selection". In: *The Journal of Finance* 7.1, pp. 77–91. JSTOR: 2975974.

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar (Aug. 17, 2012). *Foundations of Machine Learning*. Cambridge, Mass.: The MIT Press. 432 pp.

Peyré, Gabriel and Marco Cuturi (Mar. 1, 2018). "Computational Optimal Transport". arXiv: 1803.00567 [stat].

Recht, Benjamin (2019). "A Tour of Reinforcement Learning: The View from Continuous Control". In: *Annual Review of Control, Robotics, and Autonomous Systems* 2, pp. 253–279.

Rosenbrock, H. H. (Jan. 1, 1960). "An Automatic Method for Finding the Greatest or Least Value of a Function". In: *The Computer Journal* 3.3, pp. 175–184.

Shalev-Shwartz, Shai and Shai Ben-David (May 19, 2014). *Understanding Machine Learning: From Theory to Algorithms*. 1 edition. Cambridge New York, NY; Port Melbourne Delhi Singapore: Cambridge University Press. 410 pp.

Vapnik, Vladimir and Alexey Chervonenkis (1974). *Theory of Pattern Recognition*. Nauka, Moscow.

Villani, Cédric (2009). *Optimal Transport*. Grundlehren Der Mathematischen Wissenschaften. XXII, 976.