Non-Linear Optimization

Francisco Blanco-Silva

CHAPTER 1

Background

Our starting point is, for any positive integer $d \in \mathbb{N}$, the Cartesian products:

$$\mathbb{R}^d = \mathbb{R} \times \overset{(d)}{\cdots} \times \mathbb{R} = \{(x_1, \dots, x_d) : x_k \in \mathbb{R} \text{ for } 1 \le k \le d\}.$$

These sets, endowed with the operations of addition and scalar multiplication, have the structure of a *vector field*:

Addition: For $\boldsymbol{x} = (x_1, \dots, x_d), \boldsymbol{y} = (y_1, \dots, y_d) \in \mathbb{R}^d$,

$$\boldsymbol{x} + \boldsymbol{y} = (x_1 + y_1, \dots, x_d + y_d) \in \mathbb{R}^d.$$

Scalar multiplication: For $\boldsymbol{x} \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$,

$$\lambda \cdot \boldsymbol{x} = \lambda \boldsymbol{x} = (\lambda x_1, \dots, \lambda x_d) \in \mathbb{R}^d.$$

Given $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^d$, $\lambda, \mu \in \mathbb{R}$,

(a) The addition is commutative: $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x}$.
(b) Existence of identity elements for addition: Let $\mathbf{0} = (0, \dots, 0)$. $\boldsymbol{x} + \mathbf{0} = \boldsymbol{x}$.
(c) The addition is associative: $\boldsymbol{x} + (\boldsymbol{y} + \boldsymbol{z}) = (\boldsymbol{x} + \boldsymbol{y}) + \boldsymbol{z}$.
(d) Existence of inverse elements for addition: If $\boldsymbol{x} = (x_1, \dots, x_d)$, the element $-\boldsymbol{x} = (-x_1, \dots, -x_d)$ satisfies $\boldsymbol{x} + (-\boldsymbol{x}) = \mathbf{0}$. We write $\boldsymbol{x} - \boldsymbol{y}$ instead of $\boldsymbol{x} + (-\boldsymbol{y})$.
(e) Scalar multiplication is compatible with field multiplication: $\lambda(\mu \boldsymbol{x}) = (\lambda\mu)\boldsymbol{x}$.
(f) Existence of identity for scalar multiplication: $1 \cdot \boldsymbol{x} = \boldsymbol{x}$.
(g) Scalar multiplication is distributive with respect to addition: $\lambda(\boldsymbol{x} + \boldsymbol{y}) = \lambda\boldsymbol{x} + \lambda\boldsymbol{y}$.
(h) Scalar multiplication is distributive with respect to field addition: $(\lambda + \mu)\boldsymbol{x} = \lambda\boldsymbol{x} + \mu\boldsymbol{x}$.

A *basis* of $\mathbb{R}^d$ is any finite set $\{\boldsymbol{b}_k : 1 \le k \le d\}$ satisfying two properties:

Spanning property: For all $\boldsymbol{x} \in \mathbb{R}^d$ there exist d scalars $\{\lambda_1, \dots, \lambda_d\}$ so that $\boldsymbol{x} = \sum_{k=1}^d \lambda_k \boldsymbol{b}_k$.

Linear independence: If $\{\lambda_1, \dots, \lambda_d\}$ satisfy $\sum_{k=1}^d \lambda_k \boldsymbol{b}_k = \mathbf{0}$, then it must be $\lambda_k = 0$ for all $1 \le k \le d$.

Problem 1.1. Define in $\mathbb{R}^d$, for each $1 \le k \le d$, the element $\boldsymbol{e}_k$ to be the ordered d-tuple with k-th entry equal to one, and zeros on all other entries.

(a) Prove that $\{\boldsymbol{e}_k : 1 \le k \le d\}$ is a basis for $\mathbb{R}^d$.
(b) Set $\boldsymbol{b}_k = \boldsymbol{e}_k - \boldsymbol{e}_{k+1}$ for $1 \le k < d$, $\boldsymbol{b}_d = \boldsymbol{e}_d$. Is $\{\boldsymbol{b}_k : 1 \le k \le d\}$ a basis for $\mathbb{R}^d$?

1. Functions

Given sets X, Y, we define a *function* $f\colon X \to Y$ to be a subset of $X \times Y$ subject to the following condition: for every $\boldsymbol{x} \in X$ there is exactly one element $\boldsymbol{y} \in Y$ such that the ordered pair $(\boldsymbol{x}, \boldsymbol{y})$ is contained in the subset defining f. The sets X and Y are called respectively the *domain* and *codomain* of f.

If A is any subset of the domain X, then $f(A)$ is the subset of the codomain Y consisting of all images of elements of A. We say that $f(A)$ is the *image* of A under f. The image of f is given by $f(X)$.

If $Y \subset \mathbb{R}$, we say that the function f is real-valued. For a real-valued function $f\colon \mathbb{R}^d \to \mathbb{R}$, we may regard the corresponding ordered pairs $(\boldsymbol{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ as points in a $(d+1)$–dimensional space. We call this set the *graph* of f.

The *inverse image* of a subset B of the codomain Y under a function f is the subset of the domain X defined by $f^{-1}(B) = \{\boldsymbol{x} \in X : f(x) \in B\}$.

For sets X, Y, Z, the *function composition* of $f\colon X \to Y$ with $g\colon Y \to Z$ is the function $g \circ f\colon X \to Z$ defined by $\big(g \circ f\big)(\boldsymbol{x}) = g\big(f(\boldsymbol{x})\big)$.

Unless specifically stated otherwise, all functions in these notes are real-valued functions $f\colon \mathbb{R}^d \to \mathbb{R}$.

Example 1.1 (Linear Functions). We say that a real-valued function is *linear* if it preserves the operations in $\mathbb{R}^d$:

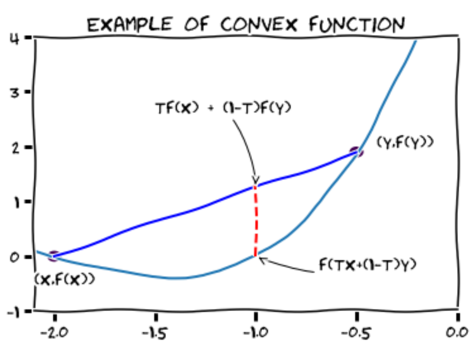
$$f(\boldsymbol{x} + \lambda \boldsymbol{y}) = f(\boldsymbol{x}) + \lambda f(\boldsymbol{y}) \text{ for } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d, \lambda \in \mathbb{R}.$$

With this definition, the function $f(x) = 3x$ is indeed a linear function, but $g(x) = 3x + 5$ is not! It is not hard to see that the only linear constant function is $f(x) = 0$ (since $f(0) = f(x - x) = f(x) - f(x) = 0$). For a non-constant linear function $f(x)$, it is also easy to see that the image is the whole real line.

Example 1.2 (Convex Functions). A subset $C \subset \mathbb{R}^d$ is said to be *convex* if for every $\boldsymbol{x}, \boldsymbol{y} \in C$, and every $\lambda \in [0, 1]$, the point $\lambda \boldsymbol{x} + (1 - \lambda)\boldsymbol{y}$ is also in C. Given such a convex set, we say that a real-valued function $f\colon C \to \mathbb{R}$ is *convex* if

$$f\big(\lambda \boldsymbol{x} + (1 - \lambda)\boldsymbol{y}\big) \le \lambda f(\boldsymbol{x}) + (1 - \lambda) f(\boldsymbol{y})$$

If instead we have $f\big(\lambda \boldsymbol{x} + (1-\lambda) f(\boldsymbol{y})\big) < \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$ for $0 < \lambda < 1$, we say

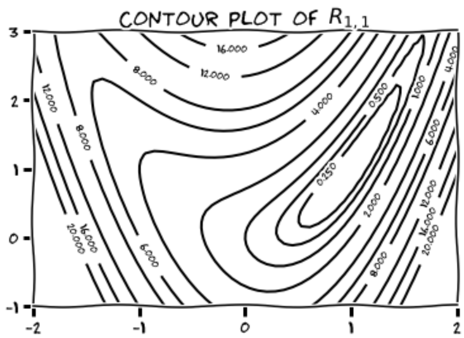


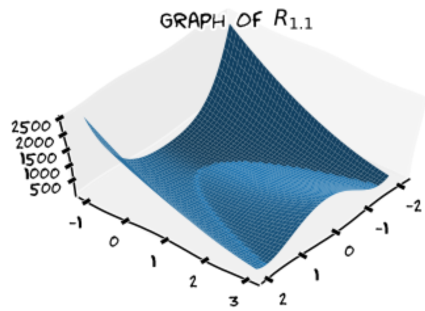
that the function is *strictly convex*. A function f is said to be *concave* (resp. *strictly concave*) if $-f$ is convex (resp. strictly convex).

EXAMPLE 1.3 (Rosenbrock Functions). Given strictly positive parameters $a, b > 0$, consider the (a, b)–Rosenbrock function $\mathcal{R}_{a,b} \colon \mathbb{R}^2 \to \mathbb{R}$ defined by:

$$\mathcal{R}_{a,b}(x_1, x_2) = (a - x_1)^2 + b(x_2 - x_1^2)^2.$$

The image of $\mathcal{R}_{a,b}$ is the interval $[0, \infty)$. Indeed, note first that $\mathcal{R}_{a,b}(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^2$. Zero is attained: $\mathcal{R}_{a,b}(a, a^2) = 0$. Note also that $\mathcal{R}_{a,b}(x_1, 0) = (a - x_1)^2 + bx_1^4$ is a polynomial of degree 4, hence unbounded for $x_1 \in \mathbb{R}$. Figure 1.3 illustrates a contour plot with several level lines of $\mathcal{R}_{1,1}$ on the domain $D = [-2, 2] \times [-1, 3]$, as well as its graph





This is a good spot to introduce the goal of these notes. The main purpose of *optimization* is the search for *extrema* of real-valued functions. Given a set $D \subset \mathbb{R}^d$, and a real-valued function $f \colon D \to \mathbb{R}$, we say that a point $\boldsymbol{x}^\star \in D$ is:

(a) A *global minimum* for f on D if $f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in D$.
(b) A *global maximum* for f on D if $f(\boldsymbol{x}^\star) \geq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in D$.
(c) A *strict global minimum* for f on D if $f(\boldsymbol{x}^\star) < f(\boldsymbol{x})$ for all $\boldsymbol{x} \in D \setminus \{\boldsymbol{x}^\star\}$.
(d) A *strict global maximum* for f on D if $f(\boldsymbol{x}^\star) > f(\boldsymbol{x})$ for all $\boldsymbol{x} \in D \setminus \{\boldsymbol{x}^\star\}$.
(e) A *local minimum* for f on D if there exists $\delta > 0$ so that $f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in B_\delta(\boldsymbol{x}^\star) \cap D$.
(f) A *local maximum* for f on D if there exists $\delta > 0$ so that $f(\boldsymbol{x}^\star) \geq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in B_\delta(\boldsymbol{x}^\star) \cap D$.
(g) A *local minimum* for f on D if there exists $\delta > 0$ so that $f(\boldsymbol{x}^\star) < f(\boldsymbol{x})$ for all $\boldsymbol{x} \in B_\delta(\boldsymbol{x}^\star) \cap D$, $\boldsymbol{x} \neq \boldsymbol{x}^\star$.
(h) A *local maximum* for f on D if there exists $\delta > 0$ so that $f(\boldsymbol{x}^\star) > f(\boldsymbol{x})$ for all $\boldsymbol{x} \in B_\delta(\boldsymbol{x}^\star) \cap D$, $\boldsymbol{x} \neq \boldsymbol{x}^\star$.

Let's play around with some more examples of functions, before we proceed to techniques for finding extrema:

EXAMPLE 1.4 (Bilinear Forms). Let $\boldsymbol{A} = \big[a_{jk}\big]_{j,k=1}^d$ be a square matrix with real coefficients. Considering elements in $\mathbb{R}^d$ as horizontal matrices, and by means of matrix products, we construct functions $\mathcal{B}_{\boldsymbol{A}} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ given by

$$\mathcal{B}_{\boldsymbol{A}}(\boldsymbol{x}, \boldsymbol{y}) = \begin{bmatrix} x_1 \cdots x_d \end{bmatrix} \begin{bmatrix} a_{11} & \cdots & a_{1d} \\ \vdots & \ddots & \vdots \\ a_{d1} & \cdots & a_{dd} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_d \end{bmatrix}$$

We call functions constructed in this way *bilinear forms*.

PROBLEM 1.2. Prove that, if the associated matrix is symmetric ($\boldsymbol{A} = \boldsymbol{A}^\intercal$), then $\mathcal{B}_{\boldsymbol{A}}(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{B}_{\boldsymbol{A}}(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$.

EXAMPLE 1.5 (Quadratic Forms). Each symmetric bilinear form has an associated *quadratic form*: A function $\mathcal{Q}_{\boldsymbol{A}} \colon \mathbb{R}^d \to \mathbb{R}$ constructed as follows:

$$\mathcal{Q}_{\boldsymbol{A}}(\boldsymbol{x}) = \mathcal{B}_{\boldsymbol{A}}(\boldsymbol{x}, \boldsymbol{x}) = \begin{bmatrix} x_1 \cdots x_d \end{bmatrix} \begin{bmatrix} a_{11} & \cdots & a_{1d} \\ \vdots & \ddots & \vdots \\ a_{1d} & \cdots & a_{dd} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

We say that the quadratic form (or the associated matrix) is:

positive definite: if $\mathcal{Q}_{\boldsymbol{A}}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$.
positive semidefinite: if $\mathcal{Q}_{\boldsymbol{A}}(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$.
negative definite: if $\mathcal{Q}_{\boldsymbol{A}}(\boldsymbol{x}) < 0$ for all $\boldsymbol{x} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$.
negative semidefinite: if $\mathcal{Q}_{\boldsymbol{A}}(\boldsymbol{x}) \leq 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$.
indefinite: if there exist $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ so that $\mathcal{Q}_{\boldsymbol{A}}(\boldsymbol{x})\mathcal{Q}_{\boldsymbol{A}}(\boldsymbol{y}) < 0$.

EXAMPLE 1.6 (Inner products). We say that a symmetric bilinear form $\mathcal{B}_{\boldsymbol{A}}$ is an *inner product* if its associated quadratic form is positive definite. By extension, we call an inner product any function $\mathcal{F} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ that satisfies the following four properties for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^d, \lambda \in \mathbb{R}$:

(a) $\mathcal{F}(\boldsymbol{x} + \boldsymbol{y}, \boldsymbol{z}) = \mathcal{F}(\boldsymbol{x}, \boldsymbol{z}) + \mathcal{F}(\boldsymbol{y}, \boldsymbol{z})$.
(b) $\mathcal{F}(\lambda\boldsymbol{x}, \boldsymbol{y}) = \lambda\mathcal{F}(\boldsymbol{x}, \boldsymbol{y})$.
(c) $\mathcal{F}(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{F}(\boldsymbol{y}, \boldsymbol{x})$.
(d) $\mathcal{F}(\boldsymbol{x}, \boldsymbol{x}) \geq 0$, $\mathcal{F}(\boldsymbol{x}, \boldsymbol{x}) = 0$ if and only if $\boldsymbol{x} = \boldsymbol{0}$.

PROBLEM 1.3. Prove that $\langle \cdot, \cdot \rangle \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ given by

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{k=1}^{d} x_k y_k$$

is an inner product. What is the matrix associated to its corresponding bilinear form?

PROBLEM 1.4. Prove that, if f is a linear function, then there exist a unique $\boldsymbol{a}_0 \in \mathbb{R}^d$ so that $f(\boldsymbol{x}) = \langle \boldsymbol{a}_0, \boldsymbol{x} \rangle$ for all $\boldsymbol{x} \in \mathbb{R}^d$.

PROBLEM 1.5. We say that $\tau \colon \mathbb{R}^d \to \mathbb{R}^d$ is a translation if there exist a fixed $\boldsymbol{x}_0 \in \mathbb{R}^d$ so that $\tau(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{x}_0$ for all $\boldsymbol{x} \in \mathbb{R}^d$.

An *affine function* $h \colon \mathbb{R}^d \to \mathbb{R}$ is a composition of a linear function $f \colon \mathbb{R}^d \to \mathbb{R}$ with a translation $\tau \colon \mathbb{R} \to \mathbb{R}$.

Prove that for each affine function h there exist a unique $\boldsymbol{a}_0 \in \mathbb{R}^d$ and a unique $\lambda_0 \in \mathbb{R}$ so that $h(\boldsymbol{x}) = \lambda_0 + \langle \boldsymbol{a}_0, \boldsymbol{x} \rangle$ for all $\boldsymbol{x} \in \mathbb{R}^d$. Use this result to prove that the graph of an affine function is a hyperplane in $\mathbb{R}^{d+1}$.

EXAMPLE 1.7 (Norms). A *norm* in $\mathbb{R}^d$ is a function $\|\cdot\| \colon \mathbb{R}^d \to \mathbb{R}$ that satisfies the following properties: For all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, and for all $\lambda \in \mathbb{R}$,

(a) $\|\boldsymbol{x}\| \geq 0$.
(b) $\|\boldsymbol{x}\| = 0$ if and only if $\boldsymbol{x} = \boldsymbol{0}$
(c) $\|\lambda\boldsymbol{x}\| = |\lambda|\|\boldsymbol{x}\|$.
(d) Triangle inequality: $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$.

Problem 1.6. Consider the function $\|\cdot\| \colon \mathbb{R}^d \to \mathbb{R}$ defined by

$$\|\boldsymbol{x}\| = \langle \boldsymbol{x}, \boldsymbol{x} \rangle^{1/2}.$$

(a) Prove that $\|\cdot\|$ is a norm
(b) Prove the *Cauchy-Schwartz inequality*: For all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$\big| \langle \boldsymbol{x}, \boldsymbol{y} \rangle \big| \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|.$$

2. Topology

The norm introduced in Example 1.7 induces a *metric* $\mathrm{d} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ on the space $\mathbb{R}^d$:

$$\mathrm{d}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\| \text{ for any } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$

Metrics allow us to measure distance between elements. These are the four main properties of these objects: Given $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^d$,

Separation property: $\mathrm{d}(\boldsymbol{x}, \boldsymbol{y}) \geq 0$.
Identity of indiscernibles: $\mathrm{d}(\boldsymbol{x}, \boldsymbol{y}) = 0$ if and only if $\boldsymbol{x} = \boldsymbol{y}$.
Symmetry: $\mathrm{d}(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{d}(\boldsymbol{y}, \boldsymbol{x})$.
Triangle inequality: $\mathrm{d}(\boldsymbol{x}, \boldsymbol{z}) \leq \mathrm{d}(\boldsymbol{x}, \boldsymbol{y}) + \mathrm{d}(\boldsymbol{y}, \boldsymbol{z})$.

Metric spaces like $\big(\mathbb{R}^d, \mathrm{d}(\cdot, \cdot)\big)$ inherit a *topology* in a natural manner, as explained below.

We define the *open ball* of radius $r > 0$ about $\boldsymbol{x}$ as the set $B_d(\boldsymbol{x}, r) = \big\{ \boldsymbol{y} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{y}\| < r \big\}$. We say $\boldsymbol{x}$ is an interior point of $D \subset \mathbb{R}^d$ if $\boldsymbol{x} \in D$ and there exists $r > 0$ so that $B_d(\boldsymbol{x}, r) \subset D$. A subset $G \subset \mathbb{R}^d$ is said to be open if all its points are interior.

A *neighborhood* of the point $\boldsymbol{x}$ is any subset of $\mathbb{R}^d$ that contains an open ball about $\boldsymbol{x}$ as subset.

A *sequence* $(\boldsymbol{x}_n)_{n \in \mathbb{N}}$ in $\mathbb{R}^d$ is an enumerated collection of elements of $\mathbb{R}^d$ in which repetitions are allowed. A sequence is said to *converge* to the limit $\boldsymbol{x} \in \mathbb{R}^d$ if and only if for every $\varepsilon > 0$ there exists $N = N(\varepsilon) \in \mathbb{N}$ so that $\|\boldsymbol{x}_n - \boldsymbol{x}\| < \varepsilon$ for all $n \geq N$. We write then

$$\boldsymbol{x} = \lim_{n \to \infty} \boldsymbol{x}_n = \lim_n \boldsymbol{x}_n, \text{ or } \lim_{n \to \infty} \|\boldsymbol{x}_n - \boldsymbol{x}\| = \lim_n \|\boldsymbol{x}_n - \boldsymbol{x}\| = 0.$$

We say that $(\boldsymbol{x}_n)_{n \in \mathbb{N}}$ is a *Cauchy sequence* if for every $\varepsilon > 0$ there exists $N = N(\varepsilon) \in \mathbb{N}$ so that for any $m, n \geq N$, $\|\boldsymbol{x}_n - \boldsymbol{x}_m\| < \varepsilon$.

Problem 1.7 (Completeness of Euclidean spaces). Prove that all Cauchy sequences converge in $\mathbb{R}^d$ (**Hint**: this is direct consequence of the completeness of $\mathbb{R}$, which you should also prove).

The complement of an open set is called *closed*. In $\mathbb{R}^d$, all subsets F are closed if and only if they are *sequentially closed*: If $\boldsymbol{x}_n \in F$ for all $n \in \mathbb{N}$ and $\lim_n \|\boldsymbol{x}_n - \boldsymbol{x}\| = 0$, then $\boldsymbol{x} \in F$.

We say D is *bounded* if there exists $M > 0$ so that $D \subset B_d(\boldsymbol{0}, M)$. A bounded and closed subset of $\mathbb{R}^d$ is called *compact*.

Theorem 2.1 (Bolzano-Weierstrass). *Every sequence in a compact subset $K \subset \mathbb{R}^d$ contains a convergent subsequence.*

Problem 1.8. Prove Theorem 2.1 for a closed interval $K = [a, b] \subset \mathbb{R}$.

3. Analysis

A real-valued function f is said to be *continuous* at $\boldsymbol{x}_0$ if for any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ so that $|f(\boldsymbol{x}) - f(\boldsymbol{x}_0)| < \varepsilon$ for all $x \in B_d(\boldsymbol{x}_0, \delta)$.

Equivalently, f is continuous at $\boldsymbol{x}_0$ if $\lim_n f(\boldsymbol{x}_n) = f(\boldsymbol{x}_0)$ for any sequence $(\boldsymbol{x}_n)_{n\in\mathbb{N}}$ satisfying $\lim_n \boldsymbol{x}_n = \boldsymbol{x}_0$.

We say that f is continuous in $D \subset \mathbb{R}^d$ if f is continuous at all points $\boldsymbol{x} \in D$.

The image of a continuous functions enjoys nice properties, which are key to the pursue of extrema. Let's start with two basic Theorems.

THEOREM 3.1 (Bounded Value Theorem). *The image $f(K)$ of a continuous real-valued function $f\colon \mathbb{R}^d \to \mathbb{R}$ on a compact set K is bounded: there exists $M > 0$ so that $|f(\boldsymbol{x})| \le M$ for all $\boldsymbol{x} \in K$.*

THEOREM 3.2 (Extreme Value Theorem). *A continuous real-valued function $f\colon K \to \mathbb{R}$ on a compact set $K \subset \mathbb{R}^d$ takes on minimal and maximal values on K.*

Theorem 3.2 guarantees the existence of global *extrema* (maxima/minima) for continuous real-valued functions over compact subsets. What if we do not have compactness?

EXAMPLE 1.8 (Coercive Functions). A continuous real-valued function $f\colon \mathbb{R}^d \to \mathbb{R}$ is said to be *coercive* if the values of $f(\boldsymbol{x})$ cannot remain bounded on any non-bounded set $A \subset \mathbb{R}^d$:

$$\lim_{\|\boldsymbol{x}\|\to\infty} = +\infty.$$

A coercive function always has a global minimum. Indeed: since f is coercive, there exists $r > 0$ so that $f(\boldsymbol{x}) > f(\boldsymbol{0})$ for all $\boldsymbol{x}$ satisfying $\|\boldsymbol{x}\| > r$. On the other hand, the set $K_r = \{\boldsymbol{x} \in \mathbb{R}^d : \|x\| \le r\}$ is compact. The continuity of f guarantees a global minimum $\boldsymbol{x}^\star \in K_r$ with $f(\boldsymbol{x}^\star) \le f(\boldsymbol{0})$. It is then $f(\boldsymbol{x}^\star) \le f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$ trivially.

How about local extrema? Continuity may not be enough:

A real-valued function f is said to be *differentiable* at $\boldsymbol{x}_0$ if there exists a linear function $J\colon \mathbb{R}^d \to \mathbb{R}$ so that

$$\lim_{\boldsymbol{h}\to\boldsymbol{0}} \frac{|f(\boldsymbol{x}_0 + h) - f(\boldsymbol{x}_0) - J(\boldsymbol{h})|}{\|\boldsymbol{h}\|} = 0$$

For any differentiable real-valued function f at a point $\boldsymbol{x}$ of its domain, the corresponding linear function in the definition above guarantees a tangent hyperplane to the graph of f at $\boldsymbol{x}$. It is the behavior of the interaction of the rest of the graph with this hyperplane what will give us clues to the nature of possible extrema.

EXAMPLE 1.9. Consider a real-valued function $f\colon \mathbb{R} \to \mathbb{R}$ of a real variable. To prove differentiability at a point x_0, we need a linear function: $J(h) = ah$ for some $a \in \mathbb{R}$. Notice how in that case,

$$\frac{|f(x_0 + h) - f(x_0) - J(h)|}{|h|} = \left| \frac{f(x_0) - f(x_0)}{h} - a \right|;$$

therefore, we could pick $a = \lim_{h\to 0} h^{-1}\big(f(x_0 + h) - f(x_0)\big)$—this is the definition of derivative we learned in Calculus.

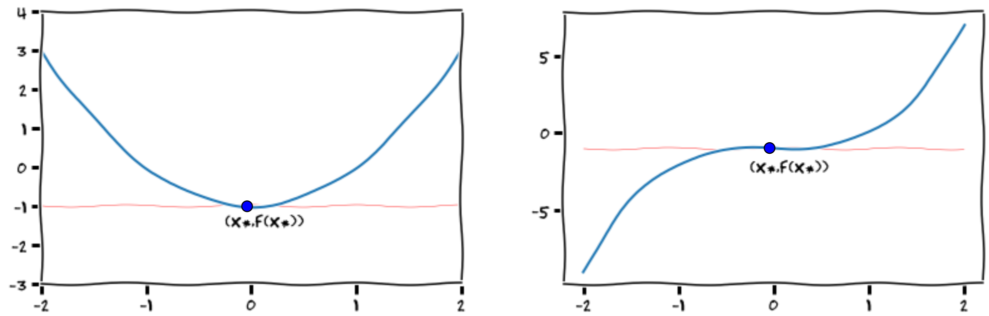


FIGURE 1. On the left: the function is locally *above* the tangent hyperplane at $\big(\boldsymbol{x}^\star, f(\boldsymbol{x}^\star)\big)$. We have a local minimum at that location. On the right, the graph crosses the tangent hyperplane. We have no extrema at that location.

PROBLEM 1.9. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a real-valued function. To prove that f is differentiable at a point $\boldsymbol{x}_0 \in \mathbb{R}^d$ we need a linear function $J(h) = \langle \boldsymbol{a}, h\rangle$ for some $\boldsymbol{a} \in \mathbb{R}^d$. Prove that in this case, we can use

$$\boldsymbol{a} = \nabla f(\boldsymbol{x}_0) = \left(\frac{\partial f}{\partial x_1}(\boldsymbol{x}_0), \dots, \frac{\partial f}{\partial x_d}(\boldsymbol{x}_0)\right).$$

EXAMPLE 1.10 (Weierstrass Function). For any positive real numbers a, b satisfying $0 < a < 1 < b$ and $ab \geq 1$, consider the Weierstrass function $\mathcal{W}_{a,b}\colon \mathbb{R} \to \mathbb{R}$ given by

$$\mathcal{W}_{a,b}(x) = \sum_{n=0}^{\infty} a^n \cos(b^n \pi x)$$

This function is continuous everywhere, yet *nowehere* differentiable! For a proof, see e.g. [**1**]

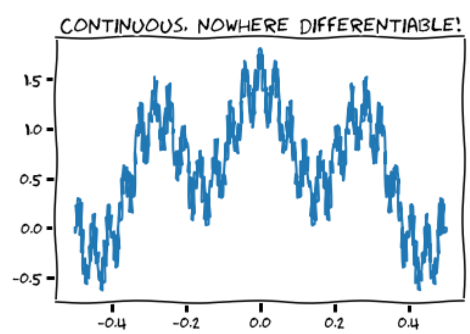


FIGURE 2. Detail of the graph of $\mathcal{W}_{0.5,7}$

It is possible to extend the notion to higher derivatives. We would say, for instance, that a function is *twice differentiable* if the derivative is differentiable. For the case of such a real-valued function $f\colon \mathbb{R}^d \to \mathbb{R}$, this would mean in particular that all second partial derivatives exist, and are continuous over the domain of f.

We define for these functions the *Hessian* of f at $\boldsymbol{x} \in D$ to be the following matrix of second partial derivatives:

$$\operatorname{Hess} f(\boldsymbol{x}) = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2}(\boldsymbol{x}) & \dfrac{\partial^2 f}{\partial x_1 \partial x_2}(\boldsymbol{x}) & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_d}(\boldsymbol{x}) \\ \dfrac{\partial^2 f}{\partial x_2 \partial x_1}(\boldsymbol{x}) & \dfrac{\partial^2 f}{\partial x_2^2}(\boldsymbol{x}) & \cdots & \dfrac{\partial^2 f}{\partial x_2 \partial x_d}(\boldsymbol{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_d \partial x_1}(\boldsymbol{x}) & \dfrac{\partial^2 f}{\partial x_d \partial x_2}(\boldsymbol{x}) & \cdots & \dfrac{\partial^2 f}{\partial x_d^2}(\boldsymbol{x}) \end{bmatrix}$$

The following three results aid in our search for local extrema for twice-differentiable real-valued functions of one variable.

THEOREM 3.3 (Rolle's Theorem). *If $f\colon [a,b] \to \mathbb{R}$ is a continuous function on a closed interval $[a,b]$, differentiable on (a,b), and $f(a) = f(b)$, then there exists $c \in (a,b)$ so that $f'(c) = 0$.*

THEOREM 3.4 (Mean Value Theorem). *If $f\colon [a,b] \to \mathbb{R}$ is a continuous function on the closed interval $[a,b]$ and differentiable on (a,b), then there exists $c \in (a,b)$ so that*

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

THEOREM 3.5 (Extended Law of the Mean). *If $f\colon D \to \mathbb{R}$ is a twice differentiable function on a domain $D \subset \mathbb{R}$ containing the closed interval $[a,b]$, then there exists $c \in (a,b)$ so that*

$$f(b) = f(a) + f'(a)(b-a) + \tfrac{1}{2} f''(c)(b-a)^2$$

This last result can be extended to a real-valued function $f\colon \mathbb{R}^{\to}\mathbb{R}$ as follows:

THEOREM 3.6 (Taylor). *Given two points $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$, let $f\colon G \to \mathbb{R}$ be a twice-differentiable real-valued function on an open set $G \subset \mathbb{R}^d$ containing the segment $[\boldsymbol{a}, \boldsymbol{b}] = \{\boldsymbol{a} + t(\boldsymbol{b} - \boldsymbol{a}) : t \in [0,1]\}$. There exists $\boldsymbol{c} \in [\boldsymbol{a}, \boldsymbol{b}]$ so that*

$$f(\boldsymbol{x}) = f(\boldsymbol{a}) + \langle \nabla f(\boldsymbol{a}), \boldsymbol{x} - \boldsymbol{a} \rangle + \frac{1}{2} \mathcal{Q}_{\mathbf{Hess} \boldsymbol{f}(\mathbf{c})}(\boldsymbol{x} - \boldsymbol{a})$$

EXAMPLE 1.11 (Rosenbrock functions, continued). In Example 1.3 we showed that the image of $\mathcal{R}_{a,b}$ is the interval $[0, \infty)$. We also found (by inspection) that the point (a, a^2) is a global minimum for this function. A straightforward computation shows that it is actually a strict global minimum. A different approach to obtain this result can be obtained using the previous technique:

- Notice $\mathcal{R}_{a,b}$ is twice differentiable. Its gradient and Hessian are given respectively by

$$\nabla \mathcal{R}_{a,b}(\boldsymbol{x}) = \big(2(x_1 - a) + 4bx(x_1^2 - x_2), b(x_2 - x_1^2)\big)$$

$$\operatorname{Hess}\mathcal{R}_{a,b}(\boldsymbol{x}) = \begin{bmatrix} 12bx_1^2 - 4bx_2 + 2 & -4bx_1 \\ -4bx_1 & 2b \end{bmatrix}$$

- The search for critical points $\nabla \mathcal{R}_{a,b} = \mathbf{0}$ gives only the point (a, a^2).
- The Hessian at that point is positive definite:

$$\mathrm{Hess}\mathcal{R}_{a,b}(a, a^2) = \begin{bmatrix} 8ba^2 + 2 & -4ab \\ -4ab & 2b \end{bmatrix}$$

4. Optimization

Notes

CHAPTER 2

Unconstrained Optimization via Calculus

Bibliography

[1] Godefroy Harold Hardy. Weierstrasss non-differentiable function. *Trans. Amer. Math. Soc*, 17(3):301–325, 1916.

[2] Anthony L Peressini, Francis E Sullivan, and J Jerry Uhl. *The mathematics of nonlinear programming*. Springer-Verlag New York, 1988.